ESTIMATING WAITING TIMES WITH THE TIME-VARYING LITTLE'S LAW

Song-Hee Kim and Ward Whitt

Industrial Engineering and Operations Research
Columbia University
New York, NY, 10027
{sk3116, ww2040}@columbia.edu

December 20, 2012

Abstract

When waiting times cannot be observed directly, Little's law can be applied to estimate the average waiting time by the average number in system divided by the average arrival rate, but that simple indirect estimator tends to be biased significantly when the arrival rates are time-varying and the service times are relatively long. Here it is shown that the bias in that indirect estimator can be estimated and reduced by applying the time-varying Little's law (TVLL). If there is appropriate time-varying staffing, then the waiting time distribution may not be time-varying even though the arrival rate is time varying. Given a fixed waiting time distribution with unknown mean, there is a unique mean consistent with the TVLL for each time t. Thus, under that condition, the TVLL provides an estimator for the unknown mean wait, given estimates of the average number in system over a subinterval and the arrival rate function. Useful variants of the TVLL estimator are obtained by fitting a linear or quadratic function to arrival data. When the arrival rate function is approximately linear (quadratic), the mean waiting time satisfies a quadratic (cubic) equation. The new estimator based on the TVLL is a positive real root of that equation. The new methods are shown to be effective in estimating the bias in the indirect estimator and reducing it, using simulations of multi-server queues and data from a call center.

Keywords: Little's law; time-varying Little's law; $L = \lambda W$; estimation; estimation bias; estimating the average wait

Short Title: Time-Varying Little's Law

Contact Author: Song-Hee Kim, sk3116@columbia.edu

1 Introduction

Little's law (LL, $L = \lambda W$) is a useful tool for analyzing operations; e.g., if the average number of patients in a hospital is 400 and the arrival rate is 100 per day, then the average length of stay should be about $W = L/\lambda = 400/100 = 4$ days. Many applications of LL are quick, like this example, but others are more elaborate and require more care, e.g, see [16, 17, 18, 19].

Little [15] and Stidham [24] first showed that LL can be put on a sound theoretical foundation. There is now a well developed theory supporting LL and related conservation laws, as reviewed recently by [16, 29], and earlier by [6, 26, 28] and others. This supporting theory concerns expected values of steady state distributions in stochastic models and limits of sample path averages. In contrast, as emphasized by Little in his recent review [16], most applications involve measurements over finite time intervals.

Applications with measurements over finite time intervals motivated us in [12] to consider ways to perform statistical analysis with LL. We focused on the scenario in which we start with an observation of $L(s)$, the number of items (which we call customers) in a system at time s, for $0 \le s \le t$, for some finite time interval $[0, t]$. From that sample path, we can directly observe the number $R(0)$ of arrivals before time 0 remaining in the system at time 0, and the number $A(s)$ of new arrivals (jumps up) in the interval $[0, s]$, $0 \le s \le t$, but based only on the available information, we typically cannot determine the waiting time W_k, the time arrival k spends in the system, for each k, because the customers need not depart in the same order that they arrived.

Within that framework, LL suggests considering the three finite averages

$$\bar{\lambda}(t) \equiv \frac{A(t)}{t}, \quad \bar{L}(t) \equiv \frac{1}{t}\int_0^t L(s)\,ds, \quad \bar{W}(t) \equiv \frac{1}{A(t)}\sum_{k=R(0)+1}^{R(0)+A(t)} W_k, \tag{1.1}$$

where the waiting times are ordered according to the arrival times, including those before time 0. Given that the finite averages $\bar{L}(t)$ and $\bar{\lambda}(t)$ in (1.1) have been observed, but the waiting times cannot be directly observed, it is natural to use the *indirect estimator*

$$\bar{W}_{L,\lambda}(t) \equiv \frac{\bar{L}(t)}{\bar{\lambda}(t)}. \tag{1.2}$$

as a substitute for $\bar{W}(t)$. It is well known that $\bar{L}(t) = \bar{\lambda}(t)\bar{W}(t)$, so that $\bar{W}_{L,\lambda}(t) = \bar{W}(t)$, if the system starts and ends empty (see Theorem 1 of [12]), but not otherwise. Theorem 2 of [12] gives the exact relation more generally:

$$\bar{W}_{L,\lambda}(t) - \bar{W}(t) = \frac{T_W^{(r)}(0) - T_W^{(r)}(t)}{A(t)}, \tag{1.3}$$

where $T_W^{(r)}(t)$ is the total residual waiting time of all customers in the system at time t, which typically is not known if the waiting times are not directly observed.

In [12] we considered estimation in two cases: when the system is stationary and when it is not. When the system can be assumed to be stationary, $E[T_W^{(r)}(0)] = E[T_W^{(r)}(t)]$ in (1.3), so that it is reasonable to use the indirect estimator in (1.2). Moreover, the very general conditions required for LL to hold should be satisfied, so that $L = \lambda W$ for underlying parameters L, λ and W. In that setting, we regarded the finite averages in (1.1) and (1.2) as estimates of these parameters and showed how to estimate confidence intervals for $\bar{W}_{L,\lambda}(t)$ using a single observation of $L(s)$, $0 \le s \le t$, by applying the method of batch means. We showed, first, that statistical analysis with LL can show how well finite averages determine the underlying parameters L, λ and W and, second, that the indirect estimation using (1.2) can be effective, illustrating with data from the call center data repository of Mandelbaum [20]; also see [13].

In the second case, the system was assumed to be nonstationary, as commonly occurs when the arrival rate is time-varying. That typically occurs in service systems, with the arrival rate increasing at the beginning of each day and decreasing at the end of each day, as illustrated by the call center example in [12]. Given that there is some underlying stochastic model for which the mean $E[\bar{W}(t)]$ is well defined, $\bar{W}(t)$ can be regarded as an estimate of $E[\bar{W}(t)]$. In this nonstationary setting, we considered ways to refine the estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) in order to reduce its bias, i.e., to reduce $|E[\Delta_W(t)]|$, where $\Delta_W(t) \equiv \bar{W}_{L,\lambda}(t) - \bar{W}(t)$. Assuming that the waiting time (total time in system) distribution is approximately a fixed exponential distribution over the measurement interval, as is often approximately appropriate in large-scale service systems, in equation (28) of [12] we proposed the following refined estimator based on (1.3):

$$\bar{W}_{L,\lambda,r}(t) \equiv \bar{W}_{L,\lambda}(t) - \frac{(R(0) - L(t))\bar{W}_{L,\lambda}(t)}{A(t)} = \bar{W}_{L,\lambda}(t)\left(1 - \frac{R(0) - L(t)}{A(t)}\right). \tag{1.4}$$

The exponential distribution assumption was used to justify approximating the residual waiting time distribution for each customer in the system at time t by the ordinary waiting time distribution, which in turn is estimated by $\bar{W}_{L,\lambda}(t)$. In [12] we showed that the proposed bias-reduction scheme in (1.4) is effective by making comparisons with call center data. Since we also had waiting time data, we could evaluate the actual bias by comparing the estimates to direct estimates of $E[\bar{W}(t)]$.

The purpose of the present paper is to investigate a different way to reduce the estimation bias in the nonstationary case. We apply the *time-varying Little's law* (TVLL), as developed by Bertsimas and Mourtzinou [1], extended and elaborated upon by Fralix and Riano [8], and reviewed here in §2, instead of using the classical LL. We show that indeed the TVLL can be used to reduce estimation bias. We also show that both the new estimators based on the TVLL and the previous refined estimator in (1.4) have advantages.

Moreover, we establish a connection between the two. Hence, the present paper together with [12] provides improved understanding.

When the staffing can decrease, we need to specify how the system operates when the staffing level is scheduled to decrease but all servers are busy. We assume that service assignments can be switched, so that a server becomes available to release at the next service completion by any server. Thus, we assume that the server scheduled to depart when the servers are all busy remains in the system until the next service completion by any server. At that time, the server scheduled to depart leaves and the server completing service starts serving the customer that was being served by the departing server. In our simulations with decreasing staffing, for simplicity, we keep track of the remaining service times of all customers and assume instead that the customer with the least remaining service time completes service immediately to allow the staffing level to decrease. In the actual system, that server would remain in the system and only leave after that minimum remaining service time is complete. We separately account for this effect by keeping track of the number of customers leaving early and the total time that there would be an additional customer in the system. In that way we show that this effect is negligible in our examples.

There appears to be only limited related literature about statistical analysis with LL. There have been studies about exploiting LL with a known arrival rate to estimate L or W more efficiently using an estimate of the other; see [9] and references therein. More generally, there have been many papers on the statistical analysis of queueing models, including inference with limited data, as illustrated by [14]. Evidently this is the first paper to exploit TVLL for estimation.

Here is how the rest of this paper is organized: In §2 we review the TVLL, including the close connection between the TVLL and infinite-server queueing models. In §3 we show how the TVLL can be applied by assuming that the waiting time distribution is fixed and specified except for its mean. Theorem 3.1 there shows that the TVLL uniquely characterizes the fixed mean $E[W]$. In §4 we develop a refined estimator based on a linear approximation for the arrival rate function. Formulas (4.6) and (4.9) there also provide an estimate of the amount of bias in the indirect estimator $\bar{W}_{L,\lambda}(t)$, showing when bias reduction is important. In §5 we apply a perturbation argument to develop an alternative estimator to the estimator in §4 to use when the estimated derivative of the arrival rate function is small. In §6 we establish Theorem 6.1 showing the connection between the TVLL approach and the estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4) when $R(0)$ and $L(t)$ cannot be observed and so are estimated using the TVLL. In §6 we also extend the estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4) to non-exponential waiting time distributions. In §7 we develop a new refined estimator based on approximating the arrival rate function by a quadratic function. In §8 we conduct simulation experiments of multi-server queueing models to compare the estimators. In §9 we compare the bias in the alternative estimators using

the call center data from [12, 13]. Finally, in §10 we draw conclusions. Additional material appears in an online appendix.

2 The Time-Varying Little's Law (TVLL)

The TVLL is a time-varying generalization of LL developed by Bertsimas and Mourtzinou [1], extended and elaborated upon by Fralix and Riano [8]. For the basic TVLL as in [1], we assume that arrivals occur one at a time to a system that was empty in the distant past and that the arrival process has a well-defined arrival-rate function $\lambda(t)$. (We can specify starting empty at any time t_0 by letting $\lambda(t) = 0$ for $t < t_0$.) For an interval I of the real line, let $A(I)$ be the number of arrivals in I. The arrival rate over the interval $[0, t]$ is specified by requiring that

$$E[A([t_1, t_t])] \equiv \Lambda(t_1, t_2) = \int_{t_1}^{t_2} \lambda(s)\, ds, \quad -\infty < t_1 < t_2 < +\infty \tag{2.1}$$

for some function λ integrable over $[t_1, t_2]$ for $-\infty < t_1 < t_2 < +\infty$, which is the arrival rate function.

As in §2 of [8], let $W(t)$ be the waiting time of the last customer to arrive at or before time t, with $W(t) \equiv 0$ if no customers have arrived by time t. We assume that the conditional *cumulative distribution function* (cdf) $G_t(x) \equiv P(W(t) \le x|\mathcal{A}_t)$, $x \ge 0$, of the waiting time (time in system) for a new arrival at time t, given that an arrival occurs at time t (the event $\mathcal{A}_t$) is well defined for all t and a measurable function on $[0, \infty)$. (The precise meaning of the cdf G_t is somewhat complicated; see [1] and [8]. In the most general form, the cdf $G_t(x)$ corresponds to a Palm measure P_t from a collection of Palm measures $\{P_t : t \ge 0\}$; see §2 of [8], but that framework supports the interpretation above. The precise meaning is not too important for this paper, because in §3 below we will make the stronger assumption that the cdf $G_t(x)$ is independent of t.)

For any time t, let $T_{-k}(t)$ be the time of the k^{th} arrival before time t (less than or equal to t), so that $T_{-(k+1)}(t) < T_{-k}(t) \le t$ for all $k \ge 1$. Let $W_{-k}(t) \equiv W(T_{-k}(t))$ be the waiting time by the arrival at time $T_{-k}(t)$. Then the number in system can be expressed as an infinite sum of random variables or, equivalently, as an elementary stochastic integral via

$$L(t) \equiv \sum_{k=1}^{\infty} 1_{\{W_{-k}(t) \ge t - T_{-k}(t)\}} = \sum_{k=1}^{\infty} 1_{\{W(T_{-k}(t)) > t - W(T_{-k}(t))\}} = \int_{-\infty}^{t} 1_{\{W(s) > t-s\}}\, dA(s). \tag{2.2}$$

Taking expectations in (2.2) and letting $G_s^c(x) \equiv P(W(s) > x|\mathcal{A}_s)$ (or, more rigorously, by applying the Campbell-Mecke formula as in the proof in [8]), we get the TVLL:

Theorem 2.1 (*the TVLL, from* [1, 8]) *Under the conditions above,*

$$E[L(t)] = \int_{-\infty}^{t} G_s^c(t-s)\lambda(s)\, ds. \tag{2.3}$$

Just like LL, the TVLL in Theorem 2.1 has important connections to infinite-server (IS) queueing models, and thus has some history prior to [1]. The connection between LL and the IS queueing model was discussed and emphasized by the sentence in italics on p. 238 of [26]; a corresponding representation holds for the TVLL. The TVLL can be regarded as part of the theory for IS models, because the abstract system can be regarded as a general IS model if we simply call the waiting time the service time in the IS model. This observation is supported by observing that the TVLL formula (2.3) coincides with the expected number of busy servers in the $M_t/GI_t/\infty$ IS model in (6) of [11], where the waiting times coincide with the service times. The M_t means that the arrival process in the IS model is a nonhomogeneous Poisson process, while the GI_t means that the service times are mutually independent and independent of the arrival process, with a general time-varying service-time cdf G_t. Thus, if we made the stronger assumption that our system can be approximated by the $M_t/GI_t/\infty$ IS model, where the waiting times play the role of the service times, then we would obtain the same formula in (2.3). Indeed, that is a setting in which the meaning of the cdf $G_t(x)$ is straightforward. The remaining content of Theorem 2.1 is the conclusion that the formula remains valid if the stochastic assumptions of the $M_t/GI_t/\infty$ IS model are relaxed.

The remaining issue is: Does the formula (2.3) for the mean $E[L(t)]$ in the $M_t/GI_t/\infty$ IS model remain valid if the stochastic assumptions are replaced by much weaker conditions? Such general conditions are provided by Theorem 1 of [1] and Theorems 2.1 and 3.1 of [8], but without discussing IS models. The fact that the stochastic assumptions in the $M_t/GI_t/\infty$ IS model can be relaxed was observed previously in §5 of [11] and Remark 2.3 of [22]. The martingale arguments in §2 of [22] are in the spirit of the earlier martingale argument of [23].

The greater validity of (2.3) occurs primarily because the expectation is a linear operator; i.e., the expected value of a sum of random variables is the sum of the expectations, without any stochastic assumptions. (Recall that an integral is essentially a sum.) The biggest drawback of (2.3) is that it is hard to specify the time-varying conditional waiting-time distribution $G_t(x)$ for a general model, beyond the $M_t/GI_t/\infty$ IS model. In general, the cdf G_t should be a derived quantity. Nevertheless, [1] and [8] show that there are important applications of the TVLL.

Remark 2.1 (*mean values versus sample path relations*) The TVLL in (2.3) is obtained by taking expected values in the sample path relation (2.2). Thus, the finite-interval relation in (1.3), which is the basis for the previously refined estimator in (1.4), parallels (2.2) rather than the TVLL (2.3). In that respect, (1.3) is stronger than the TVLL in (2.3). When we focus on the bias, which involves the expected value of the estimator, there is no difference. Thus, we anticipate that refinements of the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) based on the TVLL should compare favorably with $\bar{W}_{L,\lambda,r}(t)$ in (1.4) when viewed as an estimator of

$E[\bar{W}(t)]$. In contrast, the estimator $\bar{W}_{L,\lambda,r}(t)$ has an advantage when viewed as an estimator of $\bar{W}(t)$ for one sample path. This insight is substantiated in our experiments.

3 The TVLL with Fixed Waiting Time Distribution

It is not immediately apparent how to apply the TVLL in Theorem 2.1 to estimate waiting times, but we show that the TVLL can be used to reduce estimation bias under two additional assumptions. First, as in [12], we assume that the waiting time distribution remains fixed throughout the measurement interval.

Assumption 3.1 *The distribution of $W(t)$ is distributed as W, independent of t.*

Second, we make the statistical estimation parametric by assuming that the fixed waiting time W has a cdf that is known except for its mean.

Assumption 3.2 *There is a specified cdf G with mean 1 such that $P(W \le xE[W]) = G(x)$, $x \ge 0$.*

Given Assumptions 3.1 and 3.2, we will be concerned with estimating the mean $E[W]$ for given cdf G.

Fortunately, in many applications, Assumptions 3.1 and 3.2 are reasonable. For example, in well-managed call centers, the waiting times often remain approximately stationary, even though the arrival rate may be time varying. That is primarily achieved by using appropriate time-varying staffing. With appropriate staffing, the time spent in queue usually is relatively short compared to the service time, so that the waiting times tend to not greatly exceed the service times. Even if the service times and waiting times do vary over the day, they often change relatively slowly compared to the rate of change of the arrival rate, so that it is often reasonable to regard the waiting times as stationary over subintervals, and to have approximately the form of the service time distribution.

Furthermore, as in [12], it is often reasonable to assume that the service time and waiting time distributions are exponential, in which case $G(x) = 1 - e^{-x}$. In fact, there is now considerable evidence that service times are better fit to lognormal distributions than exponential distributions, e.g., see [3], but those lognormal distributions often can be regarded as approximately exponential, because the squared coefficient of variation (SCV, c^2, variance divided by the square of the mean) is often very close to 1. That was the case for the call center data studied in [12]. However, Assumption 3.2 also holds more generally, e.g., when the cdf G is the two-parameter lognormal distribution with specified SCV and unknown mean.

Under Assumption 3.1, the TVLL in (2.3) reduces to the corresponding $M_t/GI/\infty$ IS formula in Theorem 1 of [5], i.e.,

$$E[L(t)] = \int_{-\infty}^{t} P(W > s)\lambda(t-s)\,ds = E[\lambda(t - W_e)]E[W], \tag{3.1}$$

where the W and W_e are random variables with the fixed waiting-time cdf and the associated stationary-excess cdf, i.e.,

$$P(W_e \le x) \equiv \frac{1}{E[W]} \int_0^x P(W > u)\, du, \quad E[W_e^k] = \frac{E[W^{k+1}]}{(k+1)E[W]}. \tag{3.2}$$

Equivalently, in terms of G, the cdf of $W/E[W]$ defined in Assumption 3.2, we have

$$E[L(t)] = \int_{-\infty}^{t} G^c(s/E[W])\lambda(t-s)\, ds. \tag{3.3}$$

We now show under minor regularity conditions that this version of the TVLL uniquely determines the mean $E[W]$, both for a single t and for an average over $[0,t]$. Paralleling the definition of $\bar{L}(t)$ as the average over the interval $[0,t]$, let

$$\bar{\lambda}_t(s) \equiv \frac{1}{t} \int_0^t \lambda(u-s)\, du \tag{3.4}$$

for each s under consideration.

Theorem 3.1 (*characterization of $E[W]$*) *If the complementary cdf $G^c \equiv 1 - G(x)$ defined in Assumption 3.2 is positive, continuous and strictly decreasing for all x with $G^c(x) \to 0$ as $x \to \infty$, and if*

$$E[L(t)] < G^c(0) \int_{-\infty}^{t} \lambda(s)\, ds, \tag{3.5}$$

then the mean $E[W]$ is characterized as the unique solution to equation (3.3). If

$$E[\bar{L}(t)] < G^c(0) \int_{-\infty}^{t} \bar{\lambda}_t(s)\, ds, \tag{3.6}$$

where $\bar{\lambda}_t(s)$ is defined in (3.4), then the mean $E[W]$ is the unique solution to the equation

$$E[\bar{L}(t)] = \int_{-\infty}^{t} G^c(s/E[W])\bar{\lambda}_t(s)\, ds. \tag{3.7}$$

Proof. Under the conditions, $G^c(s/E[W])$ in the integrand is strictly increasing in $E[W]$ for each s, converging to 0 as $E[W] \downarrow 0$ and converging to $G^c(0)$ as $E[W] \uparrow \infty$. Hence, the right side of equation (3.3) is continuous and strictly increasing in $E[W]$. Condition (3.5) then implies that there is a unique solution. The same reasoning applies to (3.7) under condition (3.6). ∎

Given estimates of $\bar{L}(t)$ and the arrival rate function $\lambda(s)$ for $s \le t$, Theorem 3.1 provides an estimator for $E[W]$. In particular, let the estimator $\bar{W}_{tvll}(t)$ be the unique solution to equation (3.7) after replacing $E[\bar{L}(t)]$ by its estimate $\bar{L}(t)$ and after replacing $\lambda(s)$ in (3.4) and (3.7) for $s \le t$ by its estimate based on arrival data. An algorithm can be based on bisection search, exploiting the monotonicity used in the proof of Theorem 3.1.

Given that the values of $E[\bar{L}(t)]$ and $\bar{\lambda}_t(s)$, $s \le t$, are being estimated, it may be useful to employ the following elementary corollary.

Corollary 3.1 (*bounds for* $E[W]$) *Suppose that the assumptions of Theorem* 3.1 *hold for* $i = 1, 2$ *with*

$$E[\bar{L}_i(t)] = \int_{-\infty}^{t} G^c(s/E[W_i])\bar{\lambda}_{i,t}(s)\, ds \quad \textit{for} \quad i = 1, 2$$

for some t. *If* $E[\bar{L}_1(t)] \le E[\bar{L}(t)] \le E[\bar{L}_2(t)]$ *and* $\bar{\lambda}_{1,t}(s) \ge \bar{\lambda}_t(s) \ge \bar{\lambda}_{2,t}(s)$, $s \le t$, *then the assumptions of Theorem* 3.1 *hold for the unsubscripted system and* $E[W_1] \le E[W] \le E[W_2]$.

We now proceed to develop simple alternatives to the estimator provided by Theorem 3.1. Since these new estimators involve roots of equations, these estimators could fail to be unique, but it is always possible to confirm these estimators by applying Theorem 3.1.

4 An Approximating Linear Arrival Rate Function

Even though the arrival rate function is typically highly time-varying over a day, it is often approximately linear over subintervals, such as an hour or two. We thus can apply a linear approximation,

$$\lambda(s) \approx \lambda_l(s) \equiv a + bs, \quad 0 \le s \le t, \tag{4.1}$$

where a and b are constants such that $\lambda_l(s) \ge 0$, $0 \le s \le t$, with $[0, t]$ denoting the designated time interval. Since the number in system at any time depends on the arrival rate prior to that time, it is important that this approximation also be reasonable prior to time 0 as well as over the interval $[0, t]$. It usually suffices to go back a few (e.g., 4) mean waiting times. (That is supported by the rate of convergence to steady-state in IS models, as given in (20) of [5]. A mean waiting time can be roughly estimated by $\bar{W}_{L,\lambda}(t)$ in (1.2).) Obviously no non-constant linear approximation can be valid on the entire real line, because it would necessarily be negative in one semi-infinite interval. Nevertheless, judiciously applied, the linear approximation can be very useful, as we will show.

For a specified smooth arrival rate function, the approximation (4.1) can be obtained by a Taylor series approximation, as discussed in [5]. Ways to fit a linear arrival rate function to data from a nonhomogeneous Poisson process were studied in [21]. Experiments conducted there show that it suffices to use an ordinary least square fit unless the arrival rate is nearly 0 at one endpoint. Then an iterated least squares fit is better; it produces the maximum likelihood estimator.

Given the linear approximation (4.1), Theorem 3.1 is immediately applicable with $\lambda_t(s) = (a + b(t - 2s)/2)^+$, where $(x)^+ \equiv \max\{x, 0\}$. However, if the linear approximation is really appropriate, then it is not necessary to apply Theorem 3.1. As noted in (7) of [5], the mean number of busy servers in the $M_t/GI/\infty$ model, and thus the TVLL with non-time-varying waiting time W in (3.1), simplifies if we assume that the

arrival rate is approximately linear as in (4.1), i.e., if we can ignore the nonnegativity constraint. The simple expression involves the parameter

$$\gamma_W^2 \equiv (c_W^2 + 1)/2, \quad \text{where} \quad c_W^2 \equiv Var(W)/E[W]^2 \tag{4.2}$$

is the SCV of the cdf G. (The parameters c_W^2 and γ_W^2 provide partial characterizations of the cdf G and are independent of the mean $E[W]$. For the common case in which G is exponential, $\gamma_W^2 = c_W^2 = 1$.) Using equations (3.1) and (3.2), we we get the associated *linear approximation* for $E[L(t)]$:

$$E[L(t)] \approx \lambda_l(t - E[W_e])]E[W] = \lambda_l(t - \gamma_W^2 E[W])]E[W] = (a + bt)E[W] - b\gamma_W^2 E[W]^2. \tag{4.3}$$

By integrating over $[0, t]$ and dividing by t in (4.3), we obtain

$$E[\bar{L}(t)] \equiv t^{-1} \int_0^t E[L(s)]\, ds \approx (a + b(t/2))E[W] - b\gamma_W^2 E[W]^2. \tag{4.4}$$

Observing that the linear equations as a function of t in (4.3) and (4.4) can also be viewed as *quadratic equations* as a function of $E[W]$, we immediately obtain the following result. (The second formula in (4.6) follows from (4.5) below by rearranging terms.) Let $\bar{\lambda}_l(t)$ be the average arrival rate over $[0, t]$.

Theorem 4.1 (*exact expression for* $E[W]$) *If the linear approximations in* (4.1) *and* (4.3) *can be taken to be exact, then* $E[W]$ *is a solution to the quadratic equation*

$$\gamma_W^2 \lambda_l' x^2 - \bar{\lambda}_l(t)x + E[L(t)] = 0, \tag{4.5}$$

and the bias in $E[\bar{L}(t)]/\bar{\lambda}_l(t)$ *can be expressed as*

$$\frac{(E[\bar{L}(t)]/\bar{\lambda}_l(t)) - E[W]}{E[W]} = -\frac{\gamma_W^2 \lambda_l' E[W]}{\bar{\lambda}_l(t)}. \tag{4.6}$$

Now continuing to the estimation, we use equation (4.4) to estimate $E[W]$ based on the estimate $\bar{L}(t)$ of $E[\bar{L}(t)]$ and estimates $\bar{a}$ and $\bar{b}$ of the parameters a and b. We use the average arrival rate over $[0, t]$ of the linear approximation, $\bar{\lambda}_l(t) \equiv \bar{a} + (\bar{b}t/2)$ and $\bar{\lambda}_l' \equiv \bar{b}$, where the prime denotes a derivative. (Typically, we will have $\bar{\lambda}_l(t) = \bar{\lambda}(t)$.) Then, paralleling (4.5), we obtain the new refined estimator based on a linear approximation of the arrival rate function by solving the quadratic equation

$$\gamma_W^2 \bar{\lambda}_l' x^2 - \bar{\lambda}_l(t)x + \bar{L}(t) = 0, \tag{4.7}$$

from which we get

$$\bar{W}_{L,\lambda,l}(t) \equiv x \equiv \frac{B \pm \sqrt{B^2 - 4C}}{2}, \tag{4.8}$$

for $B \equiv \bar{\lambda}_l(t)/\gamma_W^2 \bar{\lambda}_l'$ and $C \equiv \bar{L}(t)/\gamma_W^2 \bar{\lambda}_l'$.

Formula (4.6) is very important because it provides an a priori estimate of the bias in the indirect estimator, assuming that the linear arrival rate function is a suitable approximation. In particular, we can estimate the relative bias in the indirect estimator $\bar{W}_{L,\lambda}(t)$, given $\bar{W}_{L,\lambda}(t)$ and estimates of the linear arrival rate function by

$$\frac{\bar{W}_{L,\lambda}(t) - E[W]}{E[W]} \approx \frac{(E[\bar{L}(t)]/\bar{\lambda}_l(t)) - E[W]}{E[W]} = -\frac{\gamma_W^2 \lambda_l' E[W]}{\bar{\lambda}_l(t)} \approx -\frac{\gamma_W^2 \bar{\lambda}_l' \bar{W}_{L,\lambda}(t)}{\bar{\lambda}_l(t)}, \tag{4.9}$$

where only available estimates appear on the right hand side. Formula (4.9) provides an estimate of the bias in $\bar{W}_{L,\lambda}(t)$ once $\bar{W}_{L,\lambda}(t)$ has been determined. Formulas (4.6) and (4.9) show that the estimated bias reduction is directly proportional to three separate factors: (i) the variability of the waiting time distribution, as quantified by the scale-free parameter $\gamma_W^2 \equiv (c_W^2 + 1)/2$, (ii) the relative slope of the arrival rate function, as quantified by the ratio $\bar{\lambda}_l'/\bar{\lambda}(t)$ and (iii) the mean waiting time itself, $E[W]$, as estimated by $\bar{W}_{L,\lambda}(t)$. We can thus anticipate the change in bias reduction when one or all of these factors change. For example, for given arrival rate function and given waiting time variability, the bias reduction (as quantified by the relative error) is directly increasing in the expected waiting time $E[W]$. This shows that the bias reduction is more important when the mean waiting time is larger and quantifies the impact.

Since the coefficients of the quadratic equation in (4.7) are estimated, they should be regarded as random variables. Hence the estimator $\bar{W}_{L,\lambda,l}(t)$ in (4.8) is the root of a random polynomials, as in [2] and [10]. From Theorem 3.1, we know that the multiple roots in (4.8) is a consequence of the linear approximation for the arrival rate function. If $\bar{\lambda}_l' < 0$, then $C < 0$, so that $\sqrt{B^2 - 4C} > B$ and both roots are real, one positive and one negative; then x is the one positive root. If $\bar{\lambda}_l' > 0$, then we require as a condition that $B^2 - 4C > 0$ to obtain a real root. We can then check the roots in equation (3.7).

There can be numerical instability if $|\gamma_W^2 \bar{\lambda}_l'|$ is too small, because we divide by $\gamma_W^2 \bar{\lambda}_l'$ when calculating B and C above. In that case, we provide an alternative estimator in the next section.

5 Perturbation Analysis with a Linear Arrival-Rate Function

We have observed that the estimation can be unstable if $\bar{\lambda}_l'$ in (4.7) is small, because we divide by it in the solution (4.8). An alternative estimator to the estimator in (4.8) when $\bar{\lambda}_l'$ is small, and additional insight, can be gained by performing perturbation analysis.

Lemma 5.1 *Consider the quadratic equation $c_2 x^2 - c_1 x + c_0 = 0$ with $c_1 > 0$ and $c_0 > 0$, and let $\epsilon(c_2) \equiv c_2 c_0 / c_1^2$. If $4\epsilon(c_2) < 1$, then the equation has two positive real roots and the minimum positive root*

can be expressed as

$$x = \frac{c_0}{c_1}\left(1 + \epsilon(c_2) + o(c_2)\right) \quad \textit{as} \quad c_2 \to 0. \tag{5.1}$$

Proof. Apply the Taylor series expansion

$$\sqrt{x+\epsilon} = \sqrt{x} + \frac{\epsilon}{2\sqrt{x}} - \frac{\epsilon^2}{8x^{3/2}} + o(\epsilon^2) \quad \text{as} \quad \epsilon \to 0. \quad \blacksquare$$

Based on Lemma 5.1, and assuming that $\bar{\lambda}_l(t) = \bar{\lambda}(t)$, we can approximate the minimum positive root of the quadratic equation in (4.7) by the *perturbation approximation*

$$\bar{W}_{L,\lambda,l,p}(t) \equiv \bar{W}_{L,\lambda}(t)\left(1 + \bar{W}_{L,\lambda}(t)\left(\frac{\gamma_W^2 \bar{\lambda}_l'}{\bar{\lambda}(t)}\right)\right). \tag{5.2}$$

The estimator $\bar{W}_{L,\lambda,l,p}(t)$ is to be preferred to the estimator $\bar{W}_{L,\lambda,l}(t)$ when $\gamma_W^2 \bar{\lambda}_l'$ is small. The advantage may be apparent by much smaller confidence intervals when confidence intervals are estimated for both, which is desirable.

Like formula (4.9), formula (5.2) quantifies the importance of the bias reduction, because it too provides a simple estimate of the approximate relative change in going from $\bar{W}_{L,\lambda}(t)$ to the refinement $\bar{W}_{L,\lambda,l}(t)$, yielding essentially the same result as (4.9); i.e.,

$$\frac{\bar{W}_{L,\lambda,l}(t) - \bar{W}_{L,\lambda}(t)}{\bar{W}_{L,\lambda}(t)} \approx \frac{\bar{W}_{L,\lambda,l,p}(t) - \bar{W}_{L,\lambda}(t)}{\bar{W}_{L,\lambda}(t)} \equiv \frac{\gamma_W^2 \bar{\lambda}_l' \bar{W}_{L,\lambda}(t)}{\bar{\lambda}(t)}. \tag{5.3}$$

6 Combining the Two Approaches: Estimating $R(0) - L(t)$ in $\bar{W}_{L,\lambda,r}(t)$

We may want to apply the previous refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4), but we may be unable to observe $R(0)$ and $L(t)$, because we only have available $\bar{L}(t)$ and arrival process data, and do not have a full observation of $L(s)$, $0 \le s \le t$. If that is the case, then we might elect to use an estimate of $E[L(0)] - E[L(t)]$ instead (assuming that $L(0) = R(0)$, as is the case w.p.1 with an arrival rate function). We now show that we can apply the TVLL to obtain such an estimate.

We first fit the arrival rate function to a linear function. Then we can use (4.3) to estimate $E[L(0)] - E[L(t)]$. From (4.3), we get

$$E[L(0)] - E[L(t)] \approx -btE[W] = -\lambda' tE[W]. \tag{6.1}$$

If we estimate $E[W]$ by $\bar{W}_{L,\lambda}(t)$ and λ' by $\bar{\lambda}' = \bar{b}$, then we can estimate $R(0) - L(t)$ by $-\bar{\lambda}' t \bar{W}_{L,\lambda}(t)$. Let $\bar{W}_{L,\lambda,r,e}(t)$ be the resulting estimator:

$$\bar{W}_{L,\lambda,r,e}(t) \equiv \bar{W}_{L,\lambda}(t)\left(1 + \frac{\bar{\lambda}' t \bar{W}_{L,\lambda}(t)}{A(t)}\right). \tag{6.2}$$

Before connecting this new estimator to what we have already done, we extend the estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4) to non-exponential distributions. As a natural approximation, we approximate the expected residual waiting time of each customer in service at time t by $E[W_e] = E[W]\gamma_W^2$, where W_e has the stationary-excess waiting-time distribution in (3.2). We obtain the new approximation

$$\bar{W}_{L,\lambda,r,\gamma}(t) \equiv \bar{W}_{L,\lambda}(t) - \frac{(R(0) - L(t))\bar{W}_{L,\lambda}(t)\gamma_W^2}{A(t)} = \bar{W}_{L,\lambda}(t)\left(1 - \frac{\gamma_W^2(R(0) - L(t))}{A(t)}\right), \tag{6.3}$$

which of course reduces to the previous approximation in (1.4) for exponential waiting times. Now let $\bar{W}_{L,\lambda,r,\gamma,e}(t)$ be the resulting estimator based on (6.3) and estimating $R(0) - L(t)$ as described above:

$$\bar{W}_{L,\lambda,r,\gamma,e}(t) \equiv \bar{W}_{L,\lambda}(t)\left(1 + \frac{\gamma_W^2\bar{\lambda}'t\bar{W}_{L,\lambda}(t)}{A(t)}\right). \tag{6.4}$$

Theorem 6.1 (*connection between the estimators*) *If we* (i) *fit the arrival rate function to a linear function,* (ii) *use* (4.3) *to estimate* $E[L(0)] - E[L(t)]$ *as described above and* (iii) *use that to estimate* $R(0) - L(t)$ *in* (6.3) *to obtain the estimator* $\bar{W}_{L,\lambda,r,\gamma,e}(t)$*, then*

$$\bar{W}_{L,\lambda,r,\gamma,e}(t) = \bar{W}_{L,\lambda,l,p}(t)$$

for $\bar{W}_{L,\lambda,l,p}(t)$ *in* (5.2).

Proof. The conclusion follows directly from the expressions above, using $\bar{\lambda}(t) = A(t)/t$. ∎

7 An Approximating Quadratic Arrival Rate Function

If the arrival rate function is neither approximately constant nor approximately linear, then we can consider the quadratic approximation

$$\lambda(s) \approx \lambda_q(s) \equiv a + bs + cs^2, \quad 0 \le s \le t. \tag{7.1}$$

Let $\gamma_W^2 \equiv (c_W^2 + 1)/2$ as before and

$$\theta_W^3 \equiv E[W^3]/6E[W]^3, \tag{7.2}$$

noting that $\gamma_W^2 = \theta_W^3 = 1$ if W has an exponential distribution. Given (7.1) with $\lambda_q'' \equiv 2c$, we can apply (14) of [5] and the moment formula in (3.2) to obtain the formula:

$$\begin{aligned} E[L(t)] &\approx E[L_q(t)] \equiv \lambda_q(t - E[W_e])E[W] + (\lambda_q''/2)Var(W_e)E[W] \\ &= (a + bt + ct^2)E[W] - (b + 2ct)\gamma_W^2 E[W]^2 + 2c\theta_W^3 E[W]^3 \end{aligned}$$

$$= \lambda_q(t)E[W] - \gamma_W^2\lambda_q'(t)E[W]^2 + \theta_W^3\lambda_q''E[W]^3, \tag{7.3}$$

which gives us

$$E[\bar{L}(t)] \equiv t^{-1}\int_0^t E[L(s)]\,ds \approx \bar{\lambda}_q(t)E[W] - \gamma_W^2\bar{\lambda}_q'(t)E[W]^2 + \theta_W^3\lambda_q''E[W]^3, \tag{7.4}$$

where

$$\bar{\lambda}_q(t) \equiv \frac{1}{t}\int_0^t \lambda_q(s)\,ds = a + (bt/2) + (ct^2)/3, \quad \bar{\lambda}_q'(t) \equiv \frac{1}{t}\int_0^t \lambda_q'(s)\,ds = b + ct.$$

Given the estimator $\bar{L}(t)$ for $E[\bar{L}(t)]$ and the estimators $\bar{\lambda}_q(t)$, $\bar{\lambda}_q'(t)$ and λ_q'' associated with the quadratic equation in (7.1) fit to the arrival rate data, we obtain a new refined estimator of $E[W]$, denoted by $\bar{W}_{L,\lambda,q}(t)$, by solving the following *cubic equation*

$$\theta_W^3\lambda_q''x^3 - \gamma_W^2\bar{\lambda}_q'(t)x^2 + \bar{\lambda}_q(t)x - \bar{L}(t) = 0, \tag{7.5}$$

Paralleling §5, we can do a perturbation analysis assuming that $\lambda_q'' << \bar{\lambda}_q'(t) << \bar{\lambda}_q(t)$ in (7.5) to get the approximation

$$x \equiv \bar{W}_{L,\lambda,q}(t) \approx \bar{W}_{L,\lambda,q,p}(t) \equiv w\left(1 + w\delta - w^2\epsilon\left(\frac{1}{1-2w\delta}\right)\right), \tag{7.6}$$

for $w \equiv \bar{W}_{L,\lambda}(t)$ in (1.2), $\delta \equiv \gamma_W^2\bar{\lambda}_q'(t)/\bar{\lambda}_q(t)$ and $\epsilon \equiv \theta_W^3\lambda_q''/\bar{\lambda}_q(t)$ with $\epsilon << \delta << 1$. (We assumed that $x = x_0 + \epsilon x_1 + o(\epsilon)$ and $\lambda_q'' = O(\epsilon)$ as $\epsilon \to 0$ and then used (5.2) for the $O(1)$ terms.)

8 Simulation Experiments

We now report the results of simulation experiments to evaluate the new waiting time estimators. We consider the $M_t/GI/s_t + M$ multi-server queueing model, having a nonhomogeneous Poisson arrival process (the M_t), i.i.d. service times distributed according to a random variable S with a general distribution (the GI), a time-varying staffing level (number of servers, the s_t) and customer abandonment with i.i.d. exponentially random patience times (the $+M$). The arrival process, service times and patience times are mutually independent. We let the mean service time be $E[S] = 1$, so that we are measuring time in units of mean service times. Consistent with many call centers, we let the mean patience time be 2.

In §8.1 we describe the experimental design for the main experiment. The base case has an increasing linear arrival rate function, but we also consider quadratic and constant arrival rate functions. In §8.2 we discuss two important theoretical reference points, for which we can do exact mathematical analysis for comparison: (i) the corresponding $M_t/GI/\infty$ infinite-server (IS) queueing models and (ii) the corresponding stationary $M/M/s + M$ models. In §8.3 we present the results of our main simulation experiments for

exponential (M) and Erlang (E_4) service times. In §8.4 we present corresponding results for hyperexponential (H_2) service times. In order to provide an example with greater bias, in §8.5 we consider an example with longer service times, specifically, $E[S] = 4$. In order to consider the impact of decreasing staffing, in §8.6 we consider an example with decreasing linear arrival rate and thus decreasing staffing. In order to illustrate how the procedures work for more general arrival rate functions, in §8.7 we consider the case of a sinusoidal arrival rate function.

8.1 Experimental Design

We consider three service time distributions: exponential (M, having parameters $\gamma_S^2 \equiv (c_S^2 + 1)/2 = \theta_S^3 \equiv E[S^3]/6E[S]^3 = 1$), Erlang E_4 (less variable, a sum of four i.i.d. exponentials, having parameters $\gamma_S^2 = 0.6125$ and $\theta_S^3 = 0.3125$) and hyperexponential H_2 (more variable, a mixture of two exponentials, having parameters $\gamma_S^2 = 3.0$ and $\theta_S^3 = 15.0$). (The parameters γ_W^2 and θ_W^3 are defined in (4.2) and (7.2), respectively. The same definition applies to the service time S.) The third H_2 parameter is chosen to produce balanced means as in (3.7) of [25]; the cdf is $P(S \le x) \equiv 1 - p_1 e^{-\lambda_1 x} - p_2 e^{-\lambda_2 x}$, where $p_1 = 0.0918$, $p_2 = 0.9082$, $\lambda_i = 2p_i$, yielding $p_i/\lambda_i = 1/2$, $c^2 = 5$ and $E[S^3] = 90$. The H_2 case is included to illustrate more difficult cases caused by high variability.

Since we consider multi-server queues with reasonable staffing (specified below), the waiting times (time spent in system) do not differ greatly from the service times. For customers that are served, the waiting times are somewhat longer because of the time spent in queue, but that usually is relatively short compared to the service times. Longer waiting times in queue are reduced by customer abandonment. Thus, in our estimation procedures, we approximate the unknown (γ_W^2, θ_W^3) by the specified (γ_S^2, θ_S^3).

Our base case for the time-varying arrival rate is the linear arrival rate function $\lambda(t) = 36 + 3t$ over the interval $[0, t]$ for $t = 4$ and $t = 8$, assuming the system starts empty at time $t = -12$, with $\lambda(-12) = 0$. To avoid a startup effect, i.e., serious deviations from the approximating linear and quadratic arrival rate functions, we start empty in the past. The time lag in the linear approximation in (4.3) is $E[W_e] \approx E[S_e] = \gamma_S^2 E[S] = \gamma_S^2$. Since $\gamma_S^2 = 3$ for our H_2 distribution, the system starts empty 4 time lags in the past for H_2. For the other distributions, the system starts empty more time lags in the past.

The average arrival rate $\bar{\lambda}(t)$ is 42 over $[0, 4]$ and 48 over $[0, 8]$. Thus, the expected total number of arrivals in $[0, 4]$ is 168, while it is 384 in $[0, 8]$. For comparison, we also consider quadratic and constant arrival rate functions, also starting empty at time -12. Figure 1 shows the three different arrival rate functions. The quadratic arrival rate function has the form $\lambda(t) = 53.333 + 2.222t - 0.185t^2$ based on $\lambda(-12) = \lambda(24) = 0$ and peak of 60.0 at $t = 6$. Thus, $\lambda(0) = 53.333$ and $\lambda(4) = \lambda(8) = 59.262$. The case of constant arrival

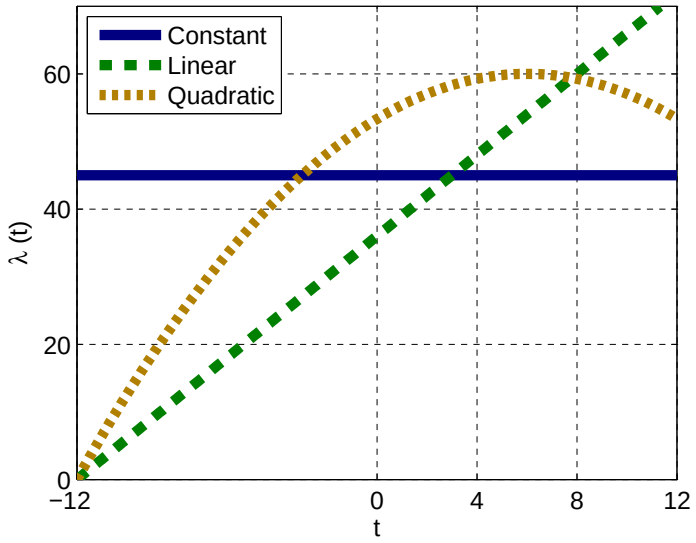

Figure 1: The three arrival rate functions used in simulation experiments: constant ($\lambda(t) = 45$), linear ($\lambda(t) = 36 + 3t$) and quadratic ($\lambda(t) = 53.333 + 2.222t - 0.185t^2$). More details in the fourth paragraph of §8.1 and §2.1 of the appendix.

rate has $\lambda = 45$. For these alternative arrival rate functions, we also estimate average waiting times over the intervals $[0, 4]$ and $[0, 8]$ via the direct estimator in (1.1).

The time-varying staffing is chosen to stabilize the performance at typical performance levels, following the method of [11] and [7]. In particular, the staffing is set using the square root staffing formula

$$s(t) \equiv \lceil m(t) + \beta\sqrt{m(t)} \rceil, \tag{8.1}$$

where $m(t)$ is the offered load and $\lceil x \rceil$ is the least integer greater than or equal to x. The offered load is $m(t) \equiv E[L(t)]$ in the associated IS model, which has formula (3.1) with the service time S playing the role of the waiting time W there. We consider three cases for the quality-of-service (QoS) parameter β in (8.1): 0, 1 and 2. With abandonment in the model, the first two cases produce typical performance, while $\beta = 2$ corresponding to high QoS, producing performance close to the IS model.

For the three arrival rate functions, we obtain explicit expressions for the offered load $m(t)$ using the linear and quadratic approximations in (4.3) and (7.3), and then the staffing via (8.1). For the linear arrival rate function $36 + 3t$, the offered load has approximately the form in (4.3), yielding $m(t) = 36 + 3t - 3\gamma_W^2$. For the quadratic arrival rate, the offered load has approximately the form in (7.3), yielding $m(t) = 53.333 - 2.222\gamma_W^2 - 0.370\theta_W^3 + (2.222 + 0.370\gamma_W^2)t - 0.185t^2$. (All these offered loads are nondecreasing, so that the staffing is nondecreasing. We consider cases with decreasing staffing in §8.6 and §8.7.) For the constant

arrival rate, $m(t) = \lambda(t) = 45$. We simulated these models using matlab, performing 100 replications in each case. We report the halfwidths of 95% confidence intervals for all estimates.

8.2 Theoretical Reference Points

There are two special cases for which we can analyze the performance analytically. These two useful theoretical reference points are: (i) the corresponding IS models and (ii) the corresponding models with constant arrival rate and exponential service times.

Infinite-Server Model. For the IS model, the waiting times coincide with the service times, so that we can calculate everything analytically. For the IS model, $E[L(t)] = m(t)$ in (3.1). We first consider the case of the linear arrival rate function. Since the approximation (4.3) is accurate for our example, there is essentially no bias at all in the refined estimator $\bar{W}_{L,\lambda,l}(t)$ in §4. For the linear arrival rate function $\lambda(t) = 36 + 3t$, the relative slope is $\lambda'/\bar{\lambda}(t) = 3/42 = 0.071$ for the interval $[0, 4]$ and $3/48 = 0.0625$ for $[0, 8]$. Thus, by formula (4.6), we anticipate that the estimation bias in $\bar{W}_{L,\lambda}(t)$ is $\gamma_W^2(\lambda'/\bar{\lambda}(t))E[W] = 7.1\gamma_W^2\%$ for $[0, 4]$ and $6.25\gamma_W^2\%$ for $[0, 8]$.

For the M, H_2 and E_4 service time distributions, and the linear arrival rate function, we have respectively, $E[\bar{L}(t)] = 39.0$, 33.0 and 40.125 over $[0, 4]$ and 45, 39.0 and 46.125 over $[0, 8]$. Since the average arrival rate $\bar{\lambda}(t)$ is 42.0 over $[0, 4]$ and 48.0 over $[0, 8]$, for the three M, H_2 and E_4 service time distributions, the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) takes the values $39/42 = 0.929$, $33/42 = 0.786$ and $40.125/42 = 0.955$ over $[0, 4]$, and $45/48 = 0.938$, $39.0/48 = 0.813$ and $46.125/48 = 0.961$ over $[0, 8]$. That means that the estimation bias in $\bar{W}_{L,\lambda}(t)$ in (1.2) is, respectively, 7.1%, 21.4% and 4.5% over $[0, 4]$, and 6.2%, 18.7% and 3.9% over $[0, 8]$, as predicted above (by (4.6)).

For the quadratic arrival rate function, we can do a corresponding analysis. For the M, H_2 and E_4 service time distributions, and the quadratic arrival rate function, we have respectively, $E[\bar{L}(t)] = 56.173$, 55.432 and 58.140 over $[0, 4]$ and 62.840, 60.617 and 65.363 over $[0, 8]$. Since the average arrival rate $\bar{\lambda}(t)$ is 58.765 over $[0, 4]$ and 66.173 over $[0, 8]$, for the three M, H_2 and E_4 service time distributions, the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) takes the values $56.173/58.765 = 0.956$, $55.432/58.765 = 0.943$ and $58.140/58.765 = 0.989$ over $[0, 4]$, $62.840/66.173 = 0.950$, $60.617/66.173 = 0.916$ and $65.363/66.173 = 0.988$ over $[0, 8]$. That means that the estimation bias in $\bar{W}_{L,\lambda}(t)$ in (1.2) is, respectively, 4.4%, 5.7% and 1.1% over $[0, 4]$, and 5.0%, 9.2% and 1.2% over $[0, 8]$. The bias is approximately $\delta - \epsilon/(1 - 2\delta) \times 100\%$, consistent with (7.6), especially for the M and E_4 service time distributions. We see for the IS model that the bias is less in the quadratic case than in the linear case. We will see that tends to be true for the simulations below as well.

Constant Arrival Rate and Exponential Service Times. For a constant arrival rate function and exponential service times, we have the stationary $M/M/s + M$ Erlang-A model, so that we can calculate the steady-state performance measures. (We used the algorithm described in [27], which also applies as an approximation to non-exponential patience times.) Since the constant arrival rate is $\lambda = 45$ and $E[S] = 1$, the stationary offered load is $m = \lambda E[S] = 45$, so that the staffing level with QoS parameter $\beta = 0$, 1 and 2 is $s = 45$, 52 and 59. In these three cases of β, the mean waiting time (in system) is 1.043, 1.0077 and 1.0008; the variance of the waiting time is 0.923, 0.938 and 0.986; the probability of delay is 0.602, 0.185 and 0.028; the probability of abandonment is 0.049, 0.0084 and 0.0008 and the mean number in system L is 47.21, 45.38 and 45.04. First, we see that the mean waiting time differs little from the mean service time $E[S] = 1$, but the variance is reduced (less than $Var(S) = 1$), evidently because the abandonments produces some short waiting times. The probability of abandonment is less than 0.01 for $\beta \geq 1$, but significant for $\beta = 0$. From [11] and [7], we anticipate that the performance in the $M_t/M/s_t + M$ model with the same service and abandonment distributions should be similar for these same QoS parameters β. Thus, we can predict the performance in advance.

8.3 Simulation Results for M and E_4 Service

In each case, we first fit constant, linear and quadratic arrival rate functions to the arrival data using least squares methods, as in [21]. (The exact arrival rates are as in §8.1 and are treated as unknown.) For the target intervals $[0, 4]$ and $[0, 8]$, we used data from the intervals $[-4, 4]$ and $[-8, 8]$, respectively. The estimates with 95% confidence intervals are shown in Table 1.

		Constant	Linear		Quadratic		
Int.	*Arrival*	$\bar{\lambda}(t)$	a	b	a	b	c
$[-4, 4]$	*Constant*	45.2 ± 0.7	44.9 ± 0.5	0.099 ± 0.197	44.9 ± 0.7	0.099 ± 0.197	-0.013 ± 0.084
	Linear	41.7 ± 0.6	35.8 ± 0.5	2.907 ± 0.162	35.4 ± 0.6	2.907 ± 0.162	0.069 ± 0.083
	Quadratic	56.6 ± 0.7	52.1 ± 0.5	2.167 ± 0.212	52.8 ± 0.8	2.167 ± 0.212	-0.120 ± 0.112
$[-8, 8]$	*Constant*	45.1 ± 0.5	45.0 ± 0.4	0.017 ± 0.058	44.7 ± 0.6	0.017 ± 0.058	0.011 ± 0.018
	Linear	48.0 ± 0.5	35.9 ± 0.3	3.025 ± 0.064	35.6 ± 0.5	3.025 ± 0.064	0.016 ± 0.016
	Quadratic	58.3 ± 0.6	49.5 ± 0.4	2.185 ± 0.071	53.1 ± 0.5	2.185 ± 0.071	-0.167 ± 0.015

Table 1: Fitting constant, linear and quadratic arrival rate functions over the intervals $[-4, 4]$ and $[-8, 8]$ to the arrival data for each arrival process; estimates with associated 95% confidence intervals based on 100 replications.

In Table 1, the actual arrival process model is shown in the rows of the second column, while the results for the different fitting methods are shown in the subsequent columns of that row. Table 1 shows that fitting

a more complex model to the arrival rate function (e.g., linear for constant) still produces pretty good results.

We next verified that the performance was indeed stabilized over the intervals $[0, 4]$ and $[0, 8]$ in all cases. For each case, we plotted the time-dependent average waiting time, the percentage of arrivals delayed, and the percentage of arrivals abandoning over the interval $[-12, 12]$, each estimated over intervals of length 0.5. We illustrate here with Figures 2, 3 and 4 for the case of linear arrival rate and M service. The story

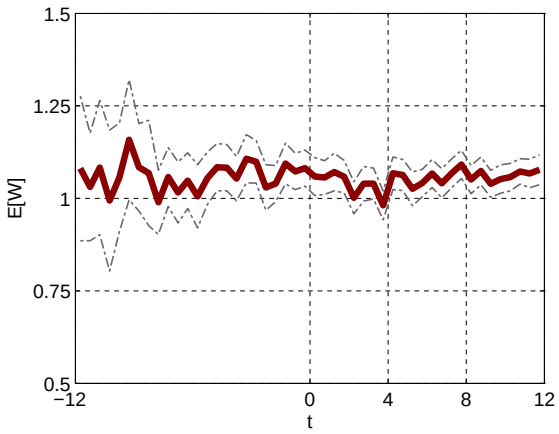

Figure 2: Linear arrival rate and M service with QoS parameter $\beta = 0$ - Average waiting time over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

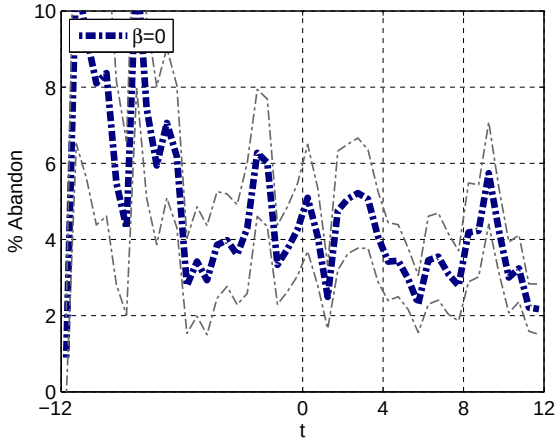

Figure 3: Linear arrival rate and M service with QoS parameter $\beta = 0$ - Average percent of arrivals abandoning over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

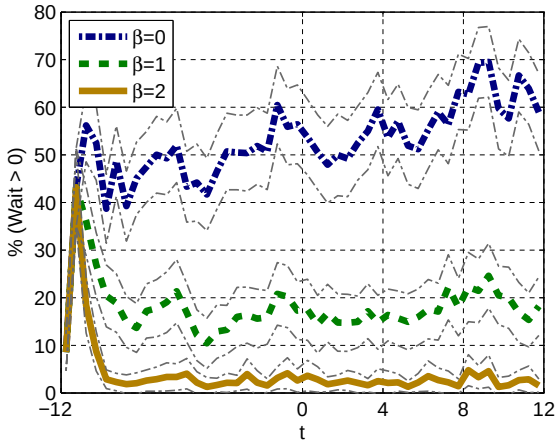

Figure 4: Linear arrival rate and M service - Average percent of arrivals delayed over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

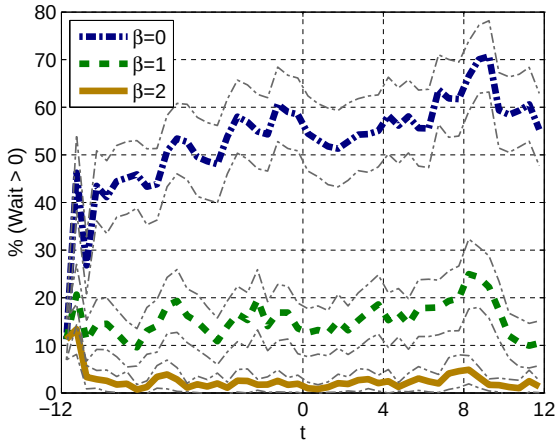

Figure 5: Linear arrival rate and E_4 service - Average percent of arrivals delayed over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

is clearest for the average waiting times in Figure 2. Since the waiting times tend to differ little from the service times, the average waiting time tend to be stabilized approximately from the beginning at time -12.

However, estimates of the time-dependent abandonment probability and probability of delay have more statistical error and require some time to stabilize. It is reasonable to conclude that they are stabilized by time $t = -8$, after $4E[W_e] \approx 4E[S] = 4$. Figure 5 shows the estimated time-dependent probability of delay for E_4 service, in which the less variability makes stabilization faster. Overall, we conclude that, as expected, the performance is stabilized over [0,4] and [0,8] by the staffing method we have used. Similar figures for all cases are given in Figures 2-55 of the appendix.

Table 2 shows the estimated waiting times by ten different methods for the linear arrival rate function. The first estimator is the direct average $\bar{W}(t)$ in (1.1), which we could not use if the waiting times were not actually observed. The second is the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) based on LL, whose bias we want to reduce. Then we give the estimators $\bar{W}_{L,\lambda,r}(t)$ in (1.4) from [12] and its extension $\bar{W}_{L,\lambda,r,\gamma}(t)$ in (6.3), which are based on the sample path relation in (1.3). Next is the estimator $\bar{W}_{L,\lambda,l}(t)$ from §4 based on the fitted linear arrival rate function, its perturbation refinement $\bar{W}_{L,\lambda,l,p}(t)$ from §5 and the estimated best of these two, $\bar{W}_{L,\lambda,l,b}(t)$, chosen as the one with the smaller confidence interval. Finally there are the corresponding three estimators from §7 based on the fitted quadratic arrival rate function.

GI	Int	β	$\bar{W}(t)$	$\bar{W}_{L,\lambda}(t)$	$\bar{W}_{L,\lambda,r}(t)$	$\bar{W}_{L,\lambda,r,\gamma}(t)$	$\bar{W}_{L,\lambda,l}(t)$	$\bar{W}_{L,\lambda,l,p}(t)$	$\bar{W}_{L,\lambda,l,b}(t)$	$\bar{W}_{L,\lambda,q}(t)$	$\bar{W}_{L,\lambda,q,p}(t)$	$\bar{W}_{L,\lambda,q,b}(t)$
M	[0, 4]	0	1.038 ± 0.019	0.980 ± 0.020	1.047 ± 0.020	1.047 ± 0.020	1.058 ± 0.023	1.046 ± 0.022	1.046 ± 0.022	1.062 ± 0.021	1.046 ± 0.021	1.046 ± 0.021
		1	1.002 ± 0.016	0.939 ± 0.015	1.005 ± 0.015	1.005 ± 0.015	1.011 ± 0.016	1.000 ± 0.016	1.011 ± 0.016	1.014 ± 0.015	1.000 ± 0.015	1.014 ± 0.015
		2	0.996 ± 0.016	0.933 ± 0.014	1.000 ± 0.014	1.000 ± 0.014	1.003 ± 0.015	0.993 ± 0.015	1.003 ± 0.015	1.007 ± 0.013	0.993 ± 0.014	1.007 ± 0.013
	[0, 8]	0	1.051 ± 0.013	0.983 ± 0.015	1.050 ± 0.015	1.050 ± 0.015	1.052 ± 0.017	1.044 ± 0.016	1.044 ± 0.016	1.054 ± 0.016	1.045 ± 0.016	1.045 ± 0.016
		1	1.010 ± 0.010	0.944 ± 0.011	1.006 ± 0.011	1.006 ± 0.011	1.008 ± 0.012	1.001 ± 0.012	1.001 ± 0.012	1.009 ± 0.011	1.002 ± 0.011	1.009 ± 0.011
		2	1.003 ± 0.009	0.938 ± 0.010	0.998 ± 0.010	0.998 ± 0.010	1.000 ± 0.011	0.993 ± 0.011	0.993 ± 0.011	1.002 ± 0.010	0.994 ± 0.010	1.002 ± 0.010
	Avg		1.017	0.953	1.018	1.018	1.022	1.013	1.016	1.025	1.013	1.021
E_4	[0, 4]	0	1.039 ± 0.009	0.997 ± 0.012	1.069 ± 0.010	1.042 ± 0.011	1.045 ± 0.013	1.040 ± 0.013	1.040 ± 0.013	1.048 ± 0.012	1.041 ± 0.012	1.048 ± 0.012
		1	1.010 ± 0.008	0.963 ± 0.010	1.039 ± 0.008	1.011 ± 0.008	1.008 ± 0.010	1.004 ± 0.010	1.008 ± 0.010	1.012 ± 0.009	1.005 ± 0.009	1.012 ± 0.009
		2	1.005 ± 0.007	0.959 ± 0.009	1.033 ± 0.008	1.005 ± 0.008	1.003 ± 0.010	0.998 ± 0.010	1.003 ± 0.010	1.006 ± 0.008	1.000 ± 0.009	1.006 ± 0.008
	[0, 8]	0	1.048 ± 0.008	1.001 ± 0.008	1.070 ± 0.008	1.044 ± 0.008	1.043 ± 0.010	1.040 ± 0.009	1.040 ± 0.009	1.044 ± 0.009	1.041 ± 0.008	1.041 ± 0.008
		1	1.011 ± 0.005	0.967 ± 0.005	1.033 ± 0.006	1.008 ± 0.005	1.006 ± 0.006	1.003 ± 0.005	1.003 ± 0.005	1.007 ± 0.006	1.005 ± 0.005	1.005 ± 0.005
		2	1.004 ± 0.005	0.960 ± 0.005	1.026 ± 0.005	1.001 ± 0.005	0.999 ± 0.005	0.997 ± 0.005	0.997 ± 0.005	1.000 ± 0.005	0.998 ± 0.005	0.998 ± 0.005
	Avg		1.020	0.974	1.045	1.018	1.017	1.014	1.015	1.020	1.015	1.018

Table 2: Waiting time estimates by ten different methods (described in the fourth paragraph of §8.3) with associated 95% confidence intervals in the $M_t/GI/s_t$ model with linear arrival rate function and the staffing set using the square root staffing formula in (8.1) with the QoS parameter β. Results are based on 100 replications of the model over the intervals $[0, 4]$ and $[0, 8]$.

First, consistent with the very low probabilities of abandonment and delay for $\beta = 1$ and 2, we see that $E[W]$ is very close to $E[S] = 1$ in those cases, but is $3 - 5\%$ higher for the QoS parameter $\beta = 0$. Second, the bias is evident in the indirect estimator $\bar{W}_{L,\lambda}(t)$; the confidence intervals are approximately the same as the others, but the correct values are not inside these confidence intervals. The same is true for

the estimator $\bar{W}_{L,\lambda,r}(t)$ for the non-exponential E_4 distribution, as expected from §6. All other estimators produce estimates and confidence intervals much like the direct estimator $\bar{W}(t)$ itself. Very roughly, the confidence interval halfwidth is approximately $\sqrt{\gamma_W^2}\%$ for $[0, 8]$ and $\sqrt{2\gamma_W^2}\%$ for $[0, 4]$.

We estimate the bias reduction achieved by our estimators by computing the absolute difference between (i) the average of the estimate of interest over the 100 replications and (ii) the average of the direct estimate $\bar{W}(t)$ over the same 100 replications. Table 3 summarizes these results for all cases; see Tables 9-11 in the appendix for details. Since the mean waiting time is approximately 1 in each case, these also are approximately percentage errors.

Arrival	*GI*	$\bar{W}_{L,\lambda}(t)$	(r)	(r,γ)	(l)	(l,p)	(l,b)	(q)	(q,p)	(q,b)
Constant	M	0.4	0.2	0.2	0.4	0.4	0.4	0.3	0.3	0.2
	E_4	0.4	0.1	0.1	0.4	0.4	0.4	0.4	0.2	0.2
Linear	M	6.4	0.4	0.4	0.7	0.7	0.8	0.8	0.6	0.6
	E_4	4.5	2.5	0.2	0.4	0.6	0.5	0.4	0.5	0.5
Quadratic	M	3.5	0.4	0.4	0.8	0.7	0.7	0.5	0.9	0.9
	E_4	0.8	0.9	0.1	0.8	0.6	0.6	0.2	0.3	0.4

Table 3: Absolute difference of the waiting time estimates by ten different methods from the direct estimate $\bar{W}(t)$ averaged over varying QoS parameter $\beta = 0$, 1 and 2 and two estimate intervals $[0, 4]$ and $[0, 8]$. $(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$. Results are in units of 10^{-2}. Details in Tables 9-11 of the appendix.

As expected, Table 3 shows that there is very little bias in $\bar{W}_{L,\lambda}(t)$ for the constant arrival rate case. In contrast, Table 3 shows that there is substantial bias for the linear and quadratic cases, agreeing closely with the values predicted by the IS results in §8.2. Moreover, the refined estimators do succeed in significantly reducing that bias. Indeed, the estimated bias is less than 1%, the confidence interval halfwidth for the direct estimator $\bar{W}(t)$, in all cases by all methods.

Overall, as illustrated by Table 3, we find that the quadratic methods are less reliable and contribute relatively little improvement over the linear approximation even to nonlinear arrival rate functions. Also, we find that the best estimator is $\bar{W}_{L,\lambda,r,\gamma}(t)$ in (6.3). However, the three estimators based on the linear approximation are also consistently good. By Theorem 6.1, we know that the advantage of $\bar{W}_{L,\lambda,r,\gamma}(t)$ over $\bar{W}_{L,\lambda,l,p}(t)$ is due to knowing $R(0) - L(t)$ rather than estimating it.

8.4 Simulation Results for H_2 Service

In this section, we consider H_2 service times, in order to illustrate difficult cases caused by higher variability. Figures 6 and 7 show that the performance is approximately stabilized over the intervals $[0, 4]$ and $[0, 8]$ for the H_2 case as well. However, the greater variability makes stabilization slower and statistical estimation

less precise for the same amount of data; e.g., compare to Figures 2 and 4 for M service times.

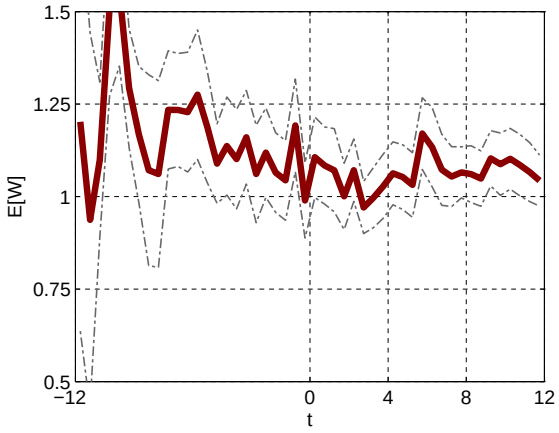

Figure 6: Linear arrival rate and H_2 service with QoS parameter $\beta = 0$ - Average waiting time over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

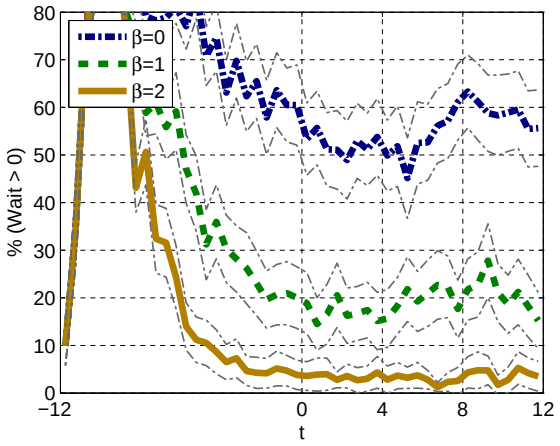

Figure 7: Linear arrival rate and H_2 service - Average percent of arrivals delayed over periods of length 0.5 with associated 95% confidence interval based on 100 replications.

Table 4 shows the estimated waiting times by ten different methods for the linear arrival rate function with H_2 service times. We first did 100 replications, just as in Section 8.3. As expected from §8.2, the bias in the indirect estimator $\bar{W}_{L,\lambda}(t)$ is much greater now than in Table 2. Second, we see that the halfwidths of the confidence intervals are larger, as expected because of the greater variability. As before, the estimates and the confidence intervals for the estimators $\bar{W}_{L,\lambda}(t)$ and $\bar{W}_{L,\lambda,r}(t)$ show bias. The other estimates and confidence intervals are close to those for the direct estimator $\bar{W}(t)$, except now poor performance is seen for the quadratic approximation. Fortunately, the confidence intervals remain quite accurate. When there is poor performance, it is revealed by the large confidence intervals. We see that some of the problem is caused by dividing by small numbers; there is clear improvement in going from $\bar{W}_{L,\lambda,q}(t)$ to $\bar{W}_{L,\lambda,q,p}(t)$ in all cases except for $\beta = 1$ over $[0, 4]$. However, for the interval $[0, 4]$, the performance of both methods remains weak.

We anticipated that the poor performance was caused by the greater variability. To better understand, we investigated further. As an initial step, we performed 1000 replications instead of 100. However, this does not help. (More results for H_2 service with $1,000$ replications can be found in §2.5 of the appendix). The results in Table 4 for $1,000$ replications show that the poor performance is *not* due to the small sample size. The poor cases remain bad.

Since we know that much of the problem with $\bar{W}_{L,\lambda,q}(t)$ is caused by dividing by small numbers, we focus on understanding the result of $\bar{W}_{L,\lambda,q,p}(t)$ on $[0, 4]$ better. Figure 8 shows the histogram of $W_{L,\lambda,q,p}(t)$

from the 100 replications for $\beta = 1$ and $[0, 4]$. It turns out that it has five outlier estimate values ($W_{L,\lambda,q,p}(t)$ less than 0 or greater than 2) that badly influence the average and confidence interval; the estimates for these outlier cases were: -1234.700, $-.838$, $-.122$, 2.320 and 18.218. We observe that in all these cases, either $L(0)$ is too low or $L(t)$ is too high, making corrections to the estimators invalid. For instance, Figure 9 shows the sample path of the case with $W_{L,\lambda,q,p}(t) = -1234.700$. In this sample path, the H_2 random service times have unusually large values: we observe 191 arrivals in the interval $[0, 4]$, and their average service times is 1.8 with maximum value of 35.5, and 10 arrivals whose service time is greater than 10.

We observe similar patterns in other cases. For the cases with $\beta = 0$ and $\beta = 2$, we used the same arrival process and service times for different values of β and observe that the six outlier cases when $\beta = 0$ and four outlier cases when $\beta = 2$ are caused by the same set of sample paths that caused problems for the $\beta = 1$ case. By getting rid of these (at most six) outlier values, we get $\bar{W}_{L,\lambda,q,p'}(t) = 0.902 \pm 0.066$ for $\beta = 0$, $\bar{W}_{L,\lambda,q,p'}(t) = 0.863 \pm 0.049$ for $\beta = 1$ and $\bar{W}_{L,\lambda,q,p'}(t) = 0.866 \pm 0.050$ for $\beta = 2$, whose confidence intervals are now much smaller. For more detailed discussion on this, see §2.4 of the appendix.

N	Int	β	$\bar{W}(t)$	$\bar{W}_{L,\lambda}(t)$	$\bar{W}_{L,\lambda,r}(t)$	$\bar{W}_{L,\lambda,r,\gamma}(t)$	$\bar{W}_{L,\lambda,l}(t)$	$\bar{W}_{L,\lambda,l,p}(t)$	$\bar{W}_{L,\lambda,l,b}(t)$	$\bar{W}_{L,\lambda,q}(t)$	$\bar{W}_{L,\lambda,q,p}(t)$	$\bar{W}_{L,\lambda,q,b}(t)$
100	[0, 4]	0	1.041 ± 0.035	0.854 ± 0.026	0.909 ± 0.029	1.017 ± 0.043	1.157 ± 0.061	1.007 ± 0.036	1.007 ± 0.036	0.348 ± 1.511	1.230 ± 0.568	1.230 ± 0.568
		1	1.006 ± 0.035	0.811 ± 0.020	0.868 ± 0.021	0.981 ± 0.033	1.058 ± 0.041	0.948 ± 0.027	0.948 ± 0.027	0.567 ± 1.488	-11.3 ± 24.2	0.567 ± 1.488
		2	0.998 ± 0.035	0.802 ± 0.017	0.858 ± 0.018	0.971 ± 0.028	1.043 ± 0.038	0.935 ± 0.021	0.935 ± 0.021	0.520 ± 1.490	1.444 ± 1.305	1.444 ± 1.305
	[0, 8]	0	1.063 ± 0.027	0.873 ± 0.019	0.931 ± 0.022	1.048 ± 0.029	1.116 ± 0.040	1.018 ± 0.026	1.018 ± 0.026	0.852 ± 0.244	1.009 ± 0.026	1.009 ± 0.026
		1	1.021 ± 0.026	0.831 ± 0.014	0.884 ± 0.015	0.991 ± 0.021	1.038 ± 0.027	0.962 ± 0.019	0.962 ± 0.019	0.853 ± 0.202	0.954 ± 0.019	0.954 ± 0.019
		2	1.013 ± 0.025	0.822 ± 0.012	0.874 ± 0.013	0.980 ± 0.018	1.020 ± 0.021	0.949 ± 0.015	0.949 ± 0.015	0.970 ± 0.095	0.942 ± 0.015	0.942 ± 0.015
	Avg		1.024	0.832	0.887	0.998	1.072	0.970	0.970	0.685	-0.959	1.024
1000	[0, 4]	0	1.052 ± 0.012	0.863 ± 0.008	0.915 ± 0.009	1.018 ± 0.013	1.177 ± 0.018	1.019 ± 0.011	1.019 ± 0.011	-0.65 ± 0.23	0.969 ± 0.219	0.969 ± 0.219
		1	1.008 ± 0.011	0.815 ± 0.006	0.868 ± 0.006	0.974 ± 0.010	1.070 ± 0.013	0.952 ± 0.007	0.952 ± 0.007	-0.47 ± 0.18	0.366 ± 0.888	-0.47 ± 0.18
		2	0.999 ± 0.011	0.806 ± 0.005	0.859 ± 0.006	0.965 ± 0.009	1.051 ± 0.011	0.941 ± 0.006	0.941 ± 0.006	-0.40 ± 0.17	0.852 ± 0.048	0.852 ± 0.048
	[0, 8]	0	1.058 ± 0.009	0.885 ± 0.007	0.940 ± 0.008	1.051 ± 0.011	1.135 ± 0.014	1.033 ± 0.010	1.033 ± 0.010	0.695 ± 0.106	1.029 ± 0.010	1.029 ± 0.010
		1	1.009 ± 0.008	0.835 ± 0.005	0.885 ± 0.005	0.986 ± 0.007	1.039 ± 0.008	0.966 ± 0.006	0.966 ± 0.006	0.942 ± 0.049	0.962 ± 0.007	0.962 ± 0.007
		2	0.999 ± 0.007	0.825 ± 0.004	0.875 ± 0.005	0.974 ± 0.006	1.022 ± 0.007	0.953 ± 0.005	0.953 ± 0.005	1.005 ± 0.022	0.949 ± 0.006	0.949 ± 0.006
	Avg		1.021	0.838	0.890	0.995	1.082	0.977	0.977	0.188	0.855	0.715

Table 4: Waiting time estimates by ten different methods (described in the fourth paragraph of §8.3) with associated 95% confidence intervals in the $M_t/H_2/s_t$ model with linear arrival rate function and the staffing set using the square root staffing formula in (8.1) with the QoS parameter β. Results are based on $N = 100$ and $1,000$ replications of the model over the intervals $[0, 4]$ and $[0, 8]$.

Table 5 shows the bias reduction achieved by our estimators. Similar to Table 3, since the mean waiting time is approximately 1 in each case, these also are approximately percentage errors. The results are less spectacular compared to the M and E_4 service in Table 3, but still quite good considering that the halfwidth of confidence intervals for the direct estimator $\bar{W}(t)$ are $2 - 4\%$ for H_2. From that perspective, all but the quadratic methods are consistently good, yielding bias estimates less than 6%. Also, note that if we remove

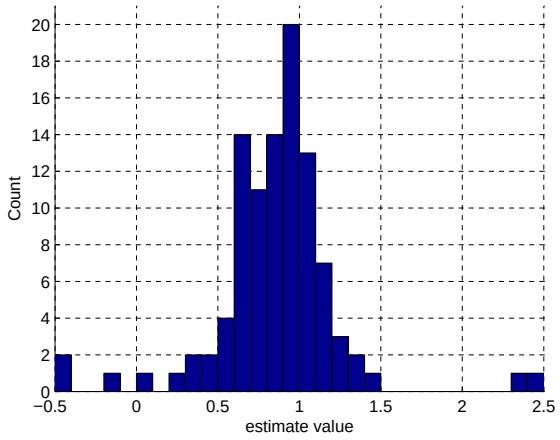

Figure 8: Linear arrival rate and H_2 service with QoS parameter $\beta = 1$ - Histogram of $W_{L,\lambda,q,p}(t)$ on $[0, 4]$ from 100 replications; counts at -0.5 (2.5) indicate the number of estimates that are less (greater) than -0.5 (-2.5).

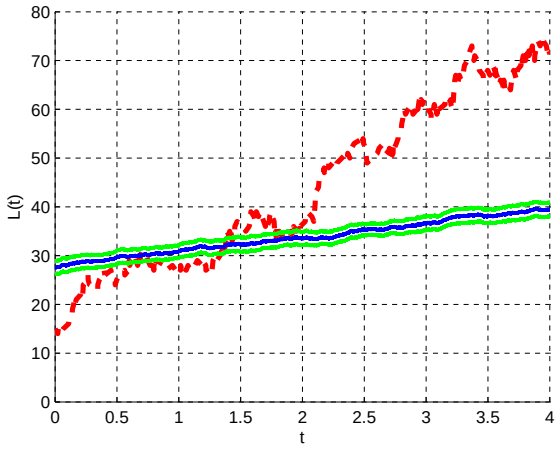

Figure 9: Linear arrival rate and H_2 service with QoS parameter $\beta = 1$ - $L(t)$ of a sample path (dashed line) with an extreme value of $W_{L,\lambda,q,p}(t)$ on $[0, 4]$. Average $L(t)$ with associated 95% confidence interval based on 100 replications are also shown.

the (at most six) outlier cases for the linear arrival rate case, the absolute difference of the estimates reduces dramatically. For instance, the value 219.4 for linear arrival under (q, p) in Table 5 becomes 10.1, as shown under $(q, p)'$. These results suggest that in order to improve the performance of quadratic estimators for H_2 service time, one can increase the number of sample size, but it is more effective if one can detect and remove the outlier values of $W_{L,\lambda,q,p}(t)$.

N	*Arrival*	$\bar{W}_{L,\lambda}(t)$	(r)	(r,γ)	(l)	(l,p)	(l,b)	(q)	(q,p)	(q,b)	(q,p')
100	*Constant*	4.3	3.0	1.0	3.6	4.2	4.2	40.7	7.1	7.1	7.1
	Linear	19.1	13.6	2.6	4.8	5.4	5.4	33.9	219.4	21.1	10.1
	Quadratic	15.0	11.8	5.5	3.0	6.0	6.0	73.4	6.5	6.5	5.7
$1,000$	*Constant*	2.8	2.3	1.4	2.6	3.1	3.0	40.4	6.4	20.5	6.2
	Linear	18.3	13.0	2.6	6.2	4.3	4.3	83.5	16.6	30.5	7.3
	Quadratic	13.0	10.0	3.9	1.8	3.8	3.8	86.2	3.4	3.4	3.1

Table 5: Absolute difference of the waiting time estimates by ten different methods from the direct estimate $\bar{W}(t)$, averaged over varying QoS parameter $\beta = 0$, 1 and 2 and two estimate intervals $[0, 4]$ and $[0, 8]$]. Estimates are based on $N = 100$ and $1,000$ replications of the model over the intervals $[0, 4]$ and $[0, 8]$. $(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$ and results under (q, p') are computed using the new waiting time estimate after removing extreme values of $W_{L,\lambda,q,p}(t)$. Results are in units of 10^{-2}. More details in §2.3 - §2.5 of the appendix.

8.5 Longer Service Times

Formulas (4.6) and (4.9) show that the bias in $\bar{W}_{L,\lambda}(t)$ should be proportional to $E[W]$. Thus there should be more bias in $\bar{W}_{L,\lambda}(t)$ and we should achieve more bias reduction with longer service times. We illustrate that now by assuming that $E[S] = 4$ instead of 1. However, for these longer service times, the linear and quadratic approximations become less appropriate. Hence, we now use Theorem 3.1 with the exact arrival rate function, which is 0 before $t = -12$, as well as the other methods to do the estimation. We consider the previous case of the linear arrival rate function with exponential service. Since the system starts empty at time -12, the linear approximation is valid three mean service times in the past, and so should still be reasonable.

Table 6 presents a summary of the results; see §2.6 of the appendix for more. First, for the previous case $E[S] = 1$, the results using Theorem 3.1 coincide with the results for the linear approximation in the precision we use, so in that case it does indeed suffice to consider the linear approximation, as claimed before. However, when $E[S] = 4$, Table 6 shows that the bias in $\bar{W}_{L,\lambda}(t)$ is $106.1/4.00 \approx 26.5\%$, consistent with formulas (4.6) and (7.6). Moreover, that bias is reduced to just over 1%, and thus essentially removed, by an application of the estimator $\bar{W}_{tvll}(t)$ based on Theorem 3.1. The larger bias in $\bar{W}_{L,\lambda}(t)$ make all approximation methods that use it less accurate, including $\bar{W}_{L\lambda,r}(t)$. The estimators $\bar{W}_{tvll}(t)$ and $\bar{W}_{L\lambda,l}(t)$ based on the TVLL are clearly superior to all other methods in this case.

$E[S]$	$\bar{W}_{L,\lambda}(t)$	(r)	$(tvll)$	(l)	(l,p)	(l,b)	(q)	(q,p)	(q,b)
1	6.4	0.4	0.7	0.7	0.7	0.8	0.8	0.6	0.6
4	106.1	29.1	5.3	4.3	50.0	50.0	151.4	69.4	69.4

Table 6: Absolute difference of the waiting time estimates from the direct estimate $\bar{W}(t)$, averaged over varying QoS parameter $\beta = 0$, 1 and 2 and two estimate intervals $[0, 4]$ and $[0, 8]$. Estimates are based on 100 replications of the model. $(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$. Results are in units of 10^{-2}. More details in §2.6 of the appendix.

8.6 Decreasing Staffing

In this section, we consider a minor modification of the previous linear arrival rate function, with time reversed. Specifically, the arrival rate function is $\lambda(t) = 48 - 3t$ over [0,4] and [0,8]. Otherwise the experimental design is just as in §8.1. Table 7 gives the estimated parameters for the linear decreasing arrival rate function over 100 replications.

As indicated in §1, if a server is scheduled to depart when all servers are busy, then in our simulations we let that server depart immediately and force the customer with the least remaining service time to complete

	Constant	Linear		Quadratic		
$Int.$	$\bar{\lambda}(t)$	a	b	a	b	c
$[-4,4]$	42.6 ± 0.6	48.4 ± 0.5	-2.877 ± 0.187	48.6 ± 0.8	-2.877 ± 0.187	-0.026 ± 0.100
$[-8,8]$	36.1 ± 0.4	48.2 ± 0.4	-3.018 ± 0.068	48.4 ± 0.5	-3.018 ± 0.068	-0.009 ± 0.018

Table 7: Fitting constant, linear and quadratic arrival rate functions over the intervals [4, 4] and [8, 8] to the linear decreasing arrival rate function; estimates with associated 95% confidence intervals over 100 replications.

service at that time. In fact, we assume that the server scheduled to leave would actually depart only after that minimum remaining service time has elapsed. At that time, the server completing service can take over the service of the departing server's customer, because service switching is allowed.

To study this effect, Table 8 shows the number of staffing decreases ($\#dec$), the number of departures ($\#dep$), the number of violation ($\#v$) and the percentage of departures that are violations ($\%v$) in each case. From Table 8 we see that we could estimate the number of violations in advance, before doing the simulation, by $\#dec \times P(W > 0)$; see §2.7 of the appendix for more discussion. Table 9 shows other key performance estimates, including the total early termination time ($TETT$), which can be divided by the number of arrivals to estimate the addition to the mean waiting time. We show that, for M and E_4 service, the average waiting time is consistently increased by about 0.1% for $\beta = 0$ and much less for $\beta = 1$ and 2. For H_2 service, the average waiting time is consistently increased by about 1.0% or less, which is still negligible. The H_2 case is relatively more problematic, because the remaining waiting times tend to be much longer than $E[W] \approx E[S] = 1$, usually having mean close to the larger of the two exponential means. Nevertheless, this effect is still relatively small.

Int.		[0, 4]						[0, 8]					
GI	β	$\#dec$	$Pr(Delay)$	$E[\#v]$	$\#dep$	$\#v$	$\%v$	$\#dec$	$Pr(Delay)$	$E[\#v]$	$\#dep$	$\#v$	$\%v$
M	0	12	0.68 ± 0.06	8.18 ± 0.69	180.7 ± 2.7	7.97 ± 0.68	4.36 ± 0.36	24	0.67 ± 0.04	16.09 ± 1.06	314.3 ± 3.7	15.61 ± 1.02	4.93 ± 0.30
	1	13	0.25 ± 0.05	3.19 ± 0.63	182.0 ± 2.8	3.00 ± 0.64	1.59 ± 0.34	26	0.23 ± 0.04	5.89 ± 0.93	313.4 ± 3.6	5.25 ± 0.90	1.63 ± 0.27
	2	14	0.05 ± 0.02	0.69 ± 0.24	182.5 ± 2.9	0.64 ± 0.26	0.33 ± 0.14	28	0.04 ± 0.01	1.11 ± 0.34	313.3 ± 3.6	0.96 ± 0.32	0.29 ± 0.10
H_2	0	12	0.37 ± 0.06	4.43 ± 0.74	177.6 ± 2.5	4.14 ± 0.73	2.30 ± 0.40	24	0.41 ± 0.05	9.83 ± 1.30	307.8 ± 3.5	9.42 ± 1.30	3.04 ± 0.41
	1	13	0.07 ± 0.02	0.88 ± 0.31	179.2 ± 2.6	0.83 ± 0.31	0.46 ± 0.17	26	0.09 ± 0.03	2.29 ± 0.71	308.6 ± 3.4	2.23 ± 0.77	0.72 ± 0.25
	2	14	0.01 ± 0.00	0.08 ± 0.06	179.5 ± 2.6	0.07 ± 0.06	0.04 ± 0.03	28	0.01 ± 0.01	0.27 ± 0.19	308.8 ± 3.4	0.29 ± 0.25	0.09 ± 0.08
E_4	0	12	0.67 ± 0.06	8.04 ± 0.67	181.0 ± 2.2	7.95 ± 0.67	4.33 ± 0.34	24	0.66 ± 0.04	15.86 ± 1.06	314.2 ± 3.4	15.24 ± 1.08	4.78 ± 0.31
	1	12	0.20 ± 0.04	2.45 ± 0.50	182.2 ± 2.5	2.35 ± 0.52	1.23 ± 0.26	26	0.19 ± 0.03	4.89 ± 0.81	313.2 ± 3.5	4.33 ± 0.74	1.34 ± 0.22
	2	13	0.04 ± 0.02	0.56 ± 0.21	182.5 ± 2.6	0.53 ± 0.22	0.28 ± 0.11	27	0.03 ± 0.01	0.93 ± 0.30	313.1 ± 3.5	0.71 ± 0.26	0.22 ± 0.08

Table 8: Early service termination in the 9 different models $M_t/GI/s_t$ with linear decreasing arrival rate and the staffing set using the square root staffing formula in (8.1) with the QoS parameter β. Results are based on 100 replications of the model over the intervals $[0, 4]$ and $[0, 8]$; $\#dec$ indicates the number of staffing decreases, $\#dep$ indicates the number of departures and v means violations.

Int.		[0, 4]					[0, 8]				
GI	β	$E[W]$	$\%Delayed$	$\%Aban.$	$\%EarlyTer.$	$TETT$	$E[W]$	$\%Delayed$	$\%Aban.$	$\%EarlyTer.$	$TETT$
M	0	1.12 ± 0.02	67.7 ± 5.7	4.15 ± 0.78	4.85 ± 0.42	0.20 ± 0.02	1.11 ± 0.02	65.4 ± 4.4	4.97 ± 0.74	5.35 ± 0.34	0.45 ± 0.04
	1	1.04 ± 0.02	23.7 ± 4.7	0.61 ± 0.21	1.61 ± 0.32	0.06 ± 0.02	1.03 ± 0.01	21.0 ± 3.4	0.66 ± 0.19	1.68 ± 0.27	0.12 ± 0.02
	2	1.02 ± 0.02	4.6 ± 1.6	0.04 ± 0.03	0.35 ± 0.14	0.01 ± 0.01	1.02 ± 0.01	3.4 ± 1.1	0.04 ± 0.03	0.32 ± 0.10	0.02 ± 0.01
H_2	0	1.05 ± 0.04	36.3 ± 6.2	1.93 ± 0.66	3.44 ± 0.49	2.09 ± 0.76	1.04 ± 0.03	41.0 ± 5.5	3.62 ± 1.02	4.13 ± 0.50	4.38 ± 1.25
	1	1.01 ± 0.04	6.5 ± 2.3	0.13 ± 0.09	0.91 ± 0.26	1.70 ± 0.66	1.00 ± 0.03	8.8 ± 2.9	0.52 ± 0.35	1.25 ± 0.29	3.80 ± 1.18
	2	1.01 ± 0.04	0.5 ± 0.4	0.01 ± 0.01	0.37 ± 0.12	1.51 ± 0.62	0.99 ± 0.03	1.0 ± 0.8	0.04 ± 0.06	0.54 ± 0.14	3.37 ± 1.12
E_4	0	1.10 ± 0.02	66.6 ± 5.6	3.05 ± 0.63	4.76 ± 0.38	0.20 ± 0.02	1.08 ± 0.01	64.3 ± 4.4	3.84 ± 0.62	5.17 ± 0.32	0.43 ± 0.04
	1	1.02 ± 0.01	19.5 ± 4.1	0.37 ± 0.14	1.39 ± 0.30	0.05 ± 0.01	1.01 ± 0.01	17.1 ± 2.8	0.39 ± 0.11	1.42 ± 0.22	0.10 ± 0.02
	2	1.01 ± 0.01	4.0 ± 1.5	0.04 ± 0.03	0.28 ± 0.12	0.01 ± 0.00	1.00 ± 0.01	2.9 ± 0.9	0.04 ± 0.03	0.23 ± 0.08	0.01 ± 0.00

Table 9: Performance of the 9 different models $M_t/GI/s_t$ with linear decreasing arrival rate and the staffing set using the square root staffing formula in (8.1) with the QoS parameter β, averaged over periods of length 0.5. Results are based on 100 replications of the model over the intervals $[0, 4]$ and $[0, 8]$; $TETT$ is the total early termination time.

8.7 A Sinusoidal Arrival Rate Function

In order to illustrate how the estimation procedures should apply for a realistic arrival rate function arising in applications, which will not be exactly linear or quadratic, we now consider a sinusoidal arrival rate function, as is often done when studying staffing with time-varying arrival rates; e.g., see [7]. Specifically, we now consider the arrival rate function $\lambda(t) = 40 + 25\sin(t/2)$ over the intervals $[0, 4]$ and $[0, 8]$, starting empty at time -36. As before, let the mean service time be $E[S] = 1$. We consider the cases of M and H_2 service, using the same distributions as before.

Assuming that the system starts empty in the infinite past, as in (3.3), exact expressions for the offered load with M and H_2 service, respectively, are $m(t) = 40 + 20(\sin(t/2) - (1/2)\cos(t/2))$ and $m(t) = 40 + 25(0.5242\sin(t/2) - 0.2897\cos(t/2))$ by (15) and (29) of [4], after correcting an error in (29); see the short appendix here. We consider the same three levels of staffing according to (8.1) with QoS parameter $\beta = 0$, 1 and 2. For M and H_2 service, the arrival rate, offered load and staffing with $\beta = 1$ are shown in Figures 10 and 11. These figures show that there is no staffing decrease in the interval $[0, 4]$, but there is in the interval $[4, 8]$, so that we also study the effect of server release over $[0, 8]$ when all servers are busy. As shown in §2.8 of the appendix, the impact is negligible.

Figures 10 and 11 show that a linear quadratic approximation to the arrival rate function should be appropriate over the interval $[0, 4]$, but not over $[0, 8]$. However, a quadratic approximation to the arrival rate function should be appropriate over both intervals $[0, 4]$ and $[0, 8]$. The experiment involves the same methods as before, after fitting linear and quadratic functions to simulation data for the arrival process. For the target intervals $[0, 4]$ and $[0, 8]$, we base the estimation on data from the intervals $[-2, 4]$ and $[-2, 8]$. We then simulate the systems over the interval $[-36, 12]$, starting empty at time -36. Table 10 shows the performance of the alternative estimators.

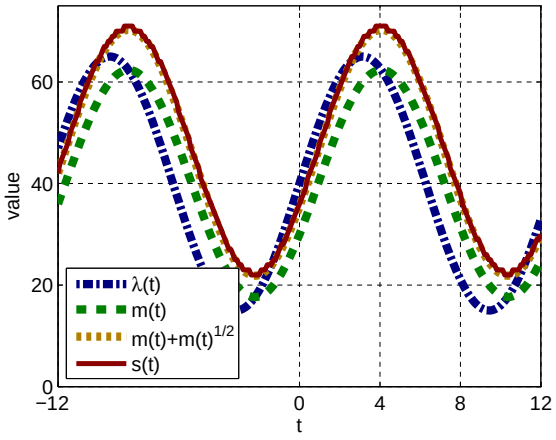

Figure 10: The sinusoidal arrival rate, offered load and staffing for M service according to (8.1) with QoS parameter $\beta = 1$.

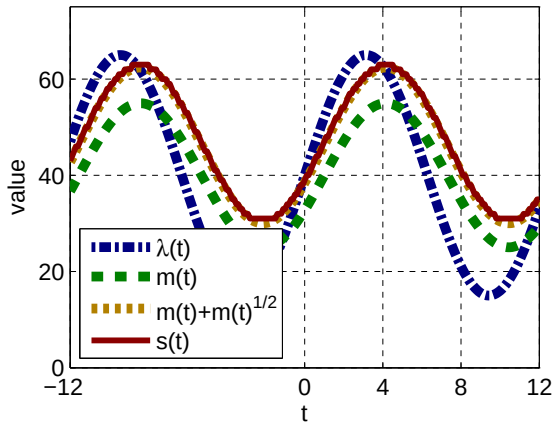

Figure 11: The sinusoidal arrival rate, offered load and staffing for H_2 service according to (8.1) with QoS parameter $\beta = 1$.

GI	Int	β	$\bar{W}_{L,\lambda}(t)$	$\bar{W}_{L,\lambda,r}(t)$	$\bar{W}_{L,\lambda,r,\gamma}(t)$	$\bar{W}_{L,\lambda,l}(t)$	$\bar{W}_{L,\lambda,l,p}(t)$	$\bar{W}_{L,\lambda,l,b}(t)$	$\bar{W}_{L,\lambda,q}(t)$	$\bar{W}_{L,\lambda,q,p}(t)$	$\bar{W}_{L,\lambda,q,b}(t)$	$\bar{W}_{L,\lambda,q,p'}(t)$
M	$[0,4]$	0	14.0	1.8	1.8	5.8	2.1	2.1	4.1	2.1	2.1	2.1
		1	14.1	2.1	2.1	4.9	2.6	2.6	3.0	2.7	2.7	2.7
		2	14.1	2.1	2.1	4.6	2.7	2.7	2.6	2.8	2.8	2.8
	Avg		14.1	2.0	2.0	5.1	2.4	2.4	3.2	2.5	2.5	2.5
	$[0,8]$	0	0.3	0.1	0.1	9.6	0.7	0.7	1.1	0.2	0.2	0.2
		1	0.5	0.1	0.1	9.3	0.5	0.5	0.7	0.5	0.5	0.5
		2	0.5	0.1	0.1	9.2	0.5	0.5	0.7	0.5	0.5	0.5
	Avg		0.4	0.1	0.1	9.3	0.6	0.6	0.8	0.4	0.4	0.4
H_2	$[0,4]$	0	20.4	12.2	4.0	6.9	11.7	6.9	263.2	100.2	100.2	44.7
		1	20.5	12.9	2.2	10.1	8.8	8.8	260.2	91.0	91.0	47.8
		2	20.6	13.0	2.0	10.5	8.4	8.4	259.7	93.1	93.1	50.6
	Avg		20.5	12.7	2.8	9.2	9.6	8.1	261.0	94.7	94.7	47.7
	$[0,8]$	0	6.8	6.5	6.0	5.1	4.1	4.1	219.8	49.2	219.8	44.1
		1	7.1	6.6	5.6	4.3	4.6	4.6	216.5	43.3	216.5	43.3
		2	7.1	6.5	5.4	4.3	4.5	4.5	215.9	42.6	215.9	42.6
	Avg		7.0	6.6	5.7	4.6	4.4	4.4	217.4	45.0	217.4	43.3

Table 10: Absolute difference of the waiting time estimates from the direct estimate $\bar{W}(t)$ for varying QoS parameter $\beta = 0$, 1 and 2 and averages over them in the $M_t/H_2/s_t$ model with sinusoidal arrival rate function and the staffing set using the square root staffing formula in (8.1) with the QoS parameter β. Two estimate intervals are $[0,4]$ and $[0,8]$. Estimates are based on 100 replications of the model. $(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$. $\bar{W}_{L,\lambda,q,p'}(t)$ is the new waiting time estimate after removing the outliers ($W_{L,\lambda,q,p}(t) < 0$ or $W_{L,\lambda,q,p}(t) > 2$). Results are in units of 10^{-2}. More details in §2.8 of the appendix.

As expected, all the refined estimators perform very well for M service over $[0,4]$, while all but the linear estimators do over $[0,8]$. Evidently, the extra time lag for H_2 service prevents the linear and quadratic approximations for the arrival rate perform well, so that the refined estimators do not perform nearly as well for H_2. The estimators $\bar{W}_{L,\lambda,r,\gamma}(t)$ and $\bar{W}_{L,\lambda,l}(t)$ clearly help, especially over $[0,4]$, but the quadratic

estimators fail completely.

9 Comparisons of Estimators Using Call Center Data

We now compare the performance of the different estimators of the mean waiting time using the same call center data as in [12, 13]. The data are for one class of customers from an American bank on 18 weekdays in May 2001. As in §8, we have data for waiting times as well as arrivals and the number in the system, so that we can compare all the estimators for $E[W]$ based on $\bar{L}(t)$ and the estimated arrival rate to the direct sample mean $\bar{W}(t)$ in (1.1).

It is natural to start by computing and plotting the finite averages $\bar{\lambda}(t)$ and $\bar{W}_{L,\lambda}(t)$ in (1.1) and (1.2), as shown in Figures 12 and 13 below for the day May 7. From such plots we can evaluate when Assumption 3.1 is approximately valid and when the arrival rate is approximately constant, linear or quadratic.

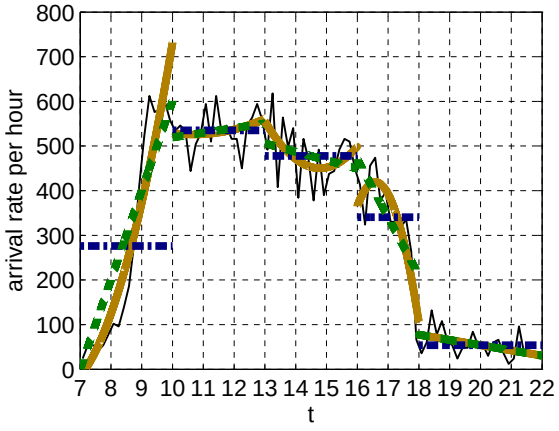

Figure 12: Arrival rate and its approximations by constant, linear and quadratic functions fitted to 5 intervals, $[7, 10]$, $[10, 13]$, $[13, 16]$, $[16, 18]$ and $[18, 22]$, on May 7.

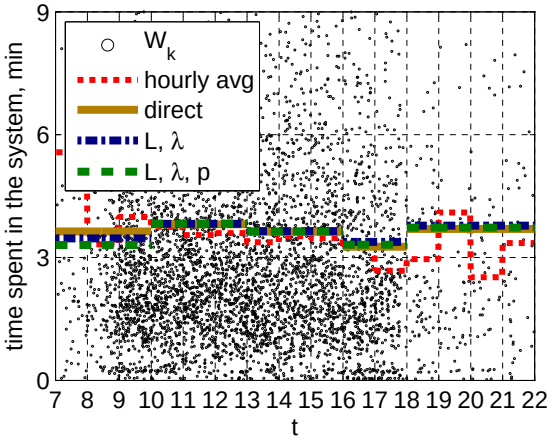

Figure 13: Scatter plot of the waiting times and its hourly averages, the direct estimator $\bar{W}(t)$ in (1.1), the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) and the refined estimator $\bar{W}_{L,\lambda,p}(t)$ from §5 of each hour in $[7, 22]$ on May 7.

First, Figure 12 shows that the arrival rate is increasing in $[7, 10]$, approximately stationary in $[10, 13]$ and $[13, 16]$, decreasing in $[16, 18]$ and again decreasing in $[18, 22]$ but with less steep slope. Hence, we divide each day into 5 intervals, $[7, 10]$, $[10, 13]$, $[13, 16]$, $[16, 18]$ and $[18, 22]$. Figure 12 also shows the fit to constant, linear and quadratic functions in each interval. Second, Figure 13 shows, in addition to a scatter plot of the waiting times, (i) the direct estimator $\bar{W}(t)$ in (1.1), (ii) the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) and (iii) the refined estimator $\bar{W}_{L,\lambda,p}(t)$ from §5. Thus, we see that Assumption 3.1 tends to be good for the actual waiting times, that it is revealed by our indirect estimator $\bar{W}_{L,\lambda}(t)$, and that the bias and bias

reduction are not great, as predicted by formulas (4.6) and (4.9).

Figures 3, 5 and 6 of [12] show that the waiting times are relatively stationary over the day, unlike the arrival rate and the number in the system. Nevertheless, the waiting times do fluctuate over time substantially for some days, especially outside of normal business hours ($[9, 17]$, i.e., nine to five). Possible reasons are inappropriate time-varying staffing and the lower call volumes outside of normal business hours. Hence, among the 18 days, we picked three days for which Assumption 3.1 holds approximately holds up to 6 pm. The three days are May 7, 18 and 21. Figure 13 shows the results for May 7; see §3 of the appendix for the others.

The estimated values of c_W^2 and $E[W^3]/E[W]^3$ for the three days were compared to the exponential values 1 and 6. With rare exceptions, the estimates of c_W^2 consistently fell in the interval $[0.90, 1.10]$, while the estimates of $E[W^3]/E[W]^3$ tended to fall in the interval $[5.5, 8.5]$, with average about 7.0, which is somewhat higher that 6.0. Hence, Assumption 3.2 is approximately valid too for G exponential.

Tables 11 and 12 show the average absolute errors of the different estimators. for the selected 3 days and all 18 days. As expected, there is more bias and bias reduction at the ends of the day when the system is nonstationary, and the bias is reduced by the refined estimators.

Int.	$\bar{W}_{L,\lambda}(t)$	(r)	(l)	(l,p)	(q)	(q,p)
$[7,10]$	3.77	0.76	4.69	0.81	110.96	5.54
$[10,13]$	0.25	0.19	0.56	0.56	58.88	0.70
$[13,16]$	0.50	0.55	0.54	0.54	108.87	0.87
$[16,18]$	3.30	0.44	0.64	0.66	101.88	2.42
$[18,23]$	1.44	0.78	1.26	1.07	91.09	1.22

Table 11: Comparison of the different estimators using call center data: Average absolute error of the estimates for each time interval over 3 days ($(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$), in units of 10^{-2}. More details in §3.2 of the appendix.

Int.	$\bar{W}_{L,\lambda}(t)$	(r)	(l)	(l,p)	(q)	(q,p)
$[7,10]$	3.28	0.57	2.01	0.89	165.03	4.99
$[10,13]$	0.87	0.34	0.82	0.82	749.41	2.19
$[13,16]$	0.58	0.51	0.57	0.57	96.30	1.10
$[16,18]$	3.14	0.78	1.50	1.68	756.52	4.85
$[18,23]$	1.23	1.05	1.56	1.59	93.62	1.08

Table 12: Comparison of the different estimators using call center data: Average absolute error of the estimates for each time interval over 18 days ($(\cdot)$ refers to $\bar{W}_{L,\lambda,(\cdot)}(t)$), in units of 10^{-2}. More details in §3.3 of the appendix.

10 Conclusions

When waiting times cannot be observed directly, Little's law can be applied to estimate the average waiting time by the average number in system divided by the average arrival rate. However, for estimation based on data over a finite time interval $[0, t]$, that simple indirect estimator $\bar{W}_{L,\lambda}(t) \equiv \bar{L}(t)/\bar{\lambda}(t)$ in (1.1) and (1.2) tends to be biased significantly when the arrival rates are time-varying and the service times are relatively long, as we have shown in examples here, see especially §8.5.

In this paper we have shown how the time-varying LL (TVLL, in Theorem 2.1) can be used to estimate the bias in $\bar{W}_{L,\lambda}(t)$ and produce refined estimators that reduce that bias under Assumptions 3.1 and 3.2, stipulating that the waiting time distribution is not time-varying and is specified except for its mean. Theorem 3.1 shows that the TVLL uniquely characterizes the mean waiting time under those assumptions and thus produces a well defined estimator for the expected wait $E[W]$ given an estimate of $\bar{L}(t)$ and the arrival rate function over the subinterval. When $E[W]$ is relatively large, the estimator $\bar{W}_{tvll}(t)$ based on Theorem 3.1 can perform much better than all other methods, as shown for the case $E[S] = 4$ in Table 6 in §8.5.

The TVLL estimator $\bar{W}_{tvll}(t)$ based on Theorem 3.1 in §3 is somewhat complicated, requiring numerical integration, search and an estimation of the arrival rate function. However, we show that convenient modifications of the TVLL estimator can be developed if we fit a linear or quadratic function to the arrival rate data. We have shown that the arrival rate function can be fit to linear and quadratic functions using least squares methods, as shown in [21]. We developed the estimators $\bar{W}_{L,\lambda,l}(t)$ in §4 and $\bar{W}_{L,\lambda,q}(t)$ in §7 based on approximating the arrival rate function by, respectively, linear and quadratic functions over a subinterval. When the arrival rate function can be regarded as approximately linear (quadratic) over the intended interval and some time into the past (a few mean waiting times), then it suffices to specify only the second moment or SCV (second and third moments) of the cdf G instead of the full cdf G. For multi-server queueing models, the waiting times do not differ greatly from the service times, so we may use the service time distribution as an approximation for the shape of the waiting time distribution, i.e., to specify the parameters γ_W^2 in (4.2) and θ_W^3 in (7.2), representing the scaled second and third moments. When the arrival rate function is approximately linear (quadratic), the mean waiting time satisfies a quadratic (cubic) equation. The new estimator based on the TVLL is a positive real root of that equation.

Solving the quadratic and cubic equations can lead to dividing by small quantities. To address that problem, we developed alternative perturbation estimators $\bar{W}_{L,\lambda,l,p}(t)$ and $\bar{W}_{L,\lambda,q,p}(t)$ in (5.2) and (7.6), respectively. These are appropriate when the first derivative of the approximating linear arrival rate function or the second derivative of the approximating quadratic arrival rate function are too small. When confidence

intervals are estimated, the perturbation estimator should be used if its confidence intervals are much smaller.

For the common case of an approximating linear arrival rate function, formulas (4.6) and (4.9) based on Theorem 4.1 provides valuable insight, giving a simple expression for the bias in $\bar{W}_{L,\lambda}(t)$ in (1.2), all of which could be removed if there were no noise in the estimation. Moreover, our experience indicates that this bias estimate is also good for nonlinear arrival rate functions. Formulas (4.6) and (4.9) show that the estimated bias is directly proportional to three separate factors: (i) the variability of the waiting time distribution, as quantified by the scale-free parameter $\gamma_W^2 \equiv (c_W^2 + 1)/2$, (ii) the relative slope of the arrival rate function, as quantified by the ratio $\bar{\lambda}_l'/\bar{\lambda}(t)$ and (iii) the mean waiting time itself, $E[W]$, as estimated by $\bar{W}_{L,\lambda}(t)$. We can obtain a rough estimate of the bias in $\bar{W}_{L,\lambda}(t)$ before considering any refined estimators. We can also see what happens when one or all of these factors change. We clearly see that the bias reduction is more important when the mean waiting time is large, with the relative error removed being directly proportional to $E[W]$. As predicted, the bias for the $M_t/GI/s_t+M$ queueing models in §8 is about four times greater when the mean service time is increased from $E[S] = 1$ to $E[S] = 4$. As predicted, the bias was also roughly proportional to the variability parameter $\gamma_W^2 \equiv (c_W^2 + 1)/2$ for the three service time distributions M, H_2 and E_4 considered in §8. We could predict in advance that the bias is relatively low (about 3%) in the call center example in §9.

From the results of the simulation experiment for the $M_t/GI/s_t + M$ model in §8 with $E[S] = 1$, as summarized by Table 3, we can draw several conclusions. First, the new refined estimator $\bar{W}_{L,\lambda,r,\gamma}(t)$ in (6.3), which is an extension of the previous refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4) for exponential waiting-time distributions studied in [12], was found to consistently provide the most bias reduction. As explained in Remark 2.1, that is not too surprising, because $\bar{W}_{L,\lambda,r,\gamma}(t)$ is based on the sample path relation in (1.3), whereas the TVLL in Theorem 2.1 is an expression for the mean. However, the estimator $\bar{W}_{L,\lambda,r,\gamma}(t)$ in (6.3) does require knowledge of $R(0)$ and $L(t)$, the number in system at the interval endpoints.

In §6 we showed that, if we use a linear approximation for the arrival rate function, then the TVLL can be applied to estimate the expected value $E[R(0) - L(t)]$ when $R(0)$ and $L(t)$ are not known. Theorem 6.1 shows that the resulting estimator $\bar{W}_{L,\lambda,r,\gamma,e}(t)$ reduces to the estimator $\bar{W}_{L,\lambda,l,p}(t)$ in §5 based directly on the TVLL. Hence, the advantage of the previous refined estimator $\bar{W}_{L,\lambda,r}(t)$ in (1.4) and its refinement to non-exponential service times $\bar{W}_{L,\lambda,r,\gamma}(t)$ in (6.3) based on the sample path relation (1.3) over the estimator $\bar{W}_{L,\lambda,l,p}(t)$ based on TVLL is due to exploiting knowledge of $R(0)$ and $L(t)$.

In §8.5 we also considered examples of the $M_t/M/s_t+M$ model with longer service times, in particular, with $E[S] = 4$ instead of $E[S] = 1$. For these examples, the bias in $\bar{W}_{L,\lambda}(t)$ was approximately 25%. With such a large bias, the estimator $\bar{W}_{L,\lambda,r,\gamma}(t) = \bar{W}_{L,\lambda,r}(t)$ performs poorly. For this example with very long

service times, the estimators $\bar{W}_{tvll}(t)$ and $\bar{W}_{L,\lambda,l}(t)$ were far superior. As a consequence, we conclude that the estimator $\bar{W}_{L,\lambda,r,\gamma}(t)$ based on the sample path relation (1.3) and the estimators $\bar{W}_{tvll}(t)$, $\bar{W}_{L,\lambda,l}(t)$ and $\bar{W}_{L,\lambda,l,p}(t)$ based on the TVLL all can be useful.

We found that these refined estimators were also effective in the call center example in §9. However, because the waiting times there were relatively short ($3-4$ minutes), the bias in the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2) was relatively small, less than 4%. In the call center example, the estimator $\bar{W}_{L,\lambda,l,p}(t)$ in §5 tended to perform best, roughly equivalent to the refined estimator in §5.2.2 of [12] based on Theorem 2 of [12]. However, overall, the estimators $\bar{W}_{L,\lambda,q}(t)$ and $\bar{W}_{L,\lambda,q,p}(t)$ based on a quadratic approximation for the arrival rate function were less useful.

It is also noteworthy that the confidence intervals for all the refined estimators (shown in the appendix) were found to be roughly the same as for the indirect estimator $\bar{W}_{L,\lambda}(t)$ in (1.2), provided that division by small values did not require using the perturbation estimators. Thus, we deduce that no additional variance must be incurred in order to reduce the bias.

In summary, formulas (4.6) and (4.9) based on Theorem 4.1 provides valuable insight, giving a simple approximate expression for the bias in $\bar{W}_{L,\lambda}(t)$ in (1.2), most of which can be removed by the methods here if there are ample data. If there is significant variation in the arrival rate (as measured by $\lambda'/\bar{\lambda}(t)$ in a linear approximation of the arrival rate function) and the waiting times are relatively long (as measured by $E[W]$ and estimated by $\bar{W}_{L,\lambda}(t)$), then there can be significant bias, which can be estimated and reduced by the methods here.

Acknowledgement

The authors thank Avishai Mandelbaum, Galit Yom-Tov, Ella Nadjharov and the Center for Service Enterprise Engineering (SEE) at the Technion for access to the SEE call center data and advice about its use. The authors thank the Samsung Foundation and NSF for support (NSF grant CMMI 1066372).

References

[1] Bertsimas, D. and Mourtzinou, G. (1997). Transient laws of nonstationary queueing systems and their applications. *Queueing Systems* 25:315–359.

[2] Bharucha-Reid, A. T. and Sambandham, M. (1986). *Random Polynomials*. New York: Academic Press.

[3] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* 100:36–50.

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

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

[6] El-Taha, M. and Jr., S. S. (1999). *Sample-Path Analysis of Queueing Systems*. Boston: Kluwer.

[7] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54(2):324–338.

[8] Fralix, B. H. and Riano, G. (2010). A new look at transient versions of Little's Law. *Journal of Applied Probability* 47:459–473.

[9] Glynn, P. W. and Whitt, W. (1989). Indirect estimation via $L = \lambda W$. *Operations Research* 37:82–103.

[10] Hamblen, J. W. (1956). Distribution of roots of quadratic equations with random coefficients. *The Annals of Mathematical Statistics* 27:1136–1143.

[11] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science* 42:1383–1394.

[12] Kim, S.-H. and Whitt, W. (2012). Statistical analysis with Little's Law. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[13] Kim, S.-H. and Whitt, W. (2012). Statistical analysis with Little's Law, supplementary material: Technical report. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[14] Larson, R. C. (1990). The queue inference engine: Deducing queue statistics from transactional data. *Management Science* 36:586–601.

[15] Little, J. D. C. (1961). A proof of the queueing formula: $L = \lambda W$. *Operations Research* 9:383–387.

[16] Little, J. D. C. (2011). Little's Law as viewed on its 50^{th} anniversary. *Operations Research* 59:536–539.

[17] Little, J. D. C. and Graves, S. C. (2008). *Building Intuition: Insights from Basic Operations Management Models and Principles*, chapter 5. Little's Law. New York: Springer, pp. 81–100.

[18] Lovejoy, W. S. and Desmond, J. S. (2011). Little's Law flow analysis of observation unit impact and sizing. *Academic Emergency Medicine* 18:183–189.

[19] Mandelbaum, A. (2010). Lecture notes on Little's Law, course on service engineering. The Technion, Israel, http://iew3.technion.ac.il/serveng/Lectures/lectures.html.

[20] Mandelbaum, A. (2012). Service Engineering of Stochastic Networks web page: http://iew3.technion.ac.il/serveng/.

[21] Massey, W. A., Parker, G. A. and Whitt, W. (1996). Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems* 5:361–388.

[22] Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1):183–250.

[23] Mazumdar, R., Kannurpatti, R. and Rosenberg, C. (1991). On rate conservation for non-stationary processes. *Journal of Applied Probability* 28:762–770.

[24] Stidham, S., Jr. (1974). A last word on $L = \lambda W$. *Operations Research* 22:417–421.

[25] Whitt, W. (1982). Approximating a point process by a renewal process: two basic methods. *Operations Research* 30:125–147.

[26] Whitt, W. (1991). A review of $L = \lambda W$. *Queueing Systems* 9:235–268.

[27] Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science* 51:221–235.

[28] Wolfe, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall.

[29] Wolfe, R. W. (2011). *Wiley Encyclopedia of Operations Research and Management Science: Little's Law and related results*. New York: Wiley.

A More on the Offered Load

In this appendix we correct formula (29) of [4] for $m(t)$, the time-varying mean number of busy servers, in the $M_t/H_k/\infty$ model with sinusoidal arrival rate $\lambda(t) \equiv \bar{\lambda} + \beta \sin(\gamma t)$ as in (6) of [4], service-time cdf

$$G(x) \equiv P(S \le x) \equiv 1 - \sum_{i=1}^{k} p_i e^{-\mu_i x}, \quad x \ge 0, \tag{A.1}$$

where $E[S] = \sum_{i=1}^{k} (p_i/\mu_i) = 1$, as in (28) of [4] and starting empty in the infinite past. The following replaces formula (29) in [4], correcting errors in the constants A_i and B_i in Proposition A.1 below.

Proposition A.1 *For the $M_t/H_k/\infty$ model above,*

$$m(t) = \bar{\lambda} + \beta \sum_{i=1}^{k} \left(A_i \sin(\gamma t) - B_i \cos(\gamma t) \right), \tag{A.2}$$

where

$$A_i \equiv \frac{p_i \mu_i}{\mu_i^2 + \gamma^2} \quad \textit{and} \quad B_i \equiv \frac{p_i \gamma}{\mu_i^2 + \gamma^2}. \tag{A.3}$$

Proof. Formula (A.2) can be derived from the general formula for $m(t)$ in the $M_t/GI/\infty$ model with the sinusoidal arrival rate function above given in Theorem 4.1 of [4] in two different ways. One way is to directly derive the distribution of S_e given the distribution of S in (A.1) above, which turns out also to be H_k with the same parameters μ_i but new parameters p_i. We will use another way, which is to represent the system as the sum of k independent $M_t/M/\infty$ models, with model i having arrival rate $\lambda(t)p_i$ and exponential service times having mean $1/\mu_i$. Then we can write $m(t) = m_1(t) + \cdots + m_k(t)$; i.e., we consider the different exponential phases of service as types of customers and thin the original nonhomogeneous Poisson arrival process into k independent Poisson processes with rates $\lambda(t)p_i$. From this representation, we immediately obtain (A.2) above with

$$A_i \equiv \frac{p_i E[\cos(\gamma X/\mu_i)]}{\mu_i} \quad \text{and} \quad B_i \equiv \frac{p_i E[\sin(\gamma X/\mu_i)]}{\mu_i}, \tag{A.4}$$

where X is an exponential random variable with mean 1. We then can apply the formulas $E[\cos(cX)] = 1/(1+c^2)$ and $E[\sin(cX)] = c/(1+c^2)$ given in the beginning of §5 of [4] to (A.4) to obtain (A.3). ∎