Approximations for Heavily-Loaded $G/GI/n + GI$ Queues

Yunan Liu*[1], Ward Whitt†[2], and Yao Yu‡[1]

[1]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695-7906
[2]Department of Industrial Engineering and Operations Research, Columbia University, New York City, NY 10027-6699

February 16, 2016

Abstract

Motivated by applications to service systems, we develop simple engineering approximation formulas for the steady-state performance of heavily-loaded $G/GI/n + GI$ multi-server queues, which can have non-Poisson and non-renewal arrivals and non-exponential service-time and patience-time distributions. The formulas are based on recently established Gaussian many-server heavy-traffic limits in the efficiency-driven (ED) regime, where the traffic intensity is fixed at $\rho > 1$, but the approximations also apply to systems in the quality-and-efficiency-driven (QED) regime, where $\rho > 1$ but ρ is close to 1. Good performance across a wide range of parameters is obtained by making a heuristic refinements, the main one being truncation of the queue length and waiting time approximations to nonnegative values. Simulation experiments show that the proposed approximations are effective for large-scale queueing systems for a significant range of the traffic intensity ρ and the abandonment rate θ, roughly for $\rho > 1.02$ and $\theta < 2.0$.

keywords: many-server queues, queues with customer abandonment, queueing performance approximations, steady-state performance, queues with non-exponential distributions.

*yunan_liu@ncsu.edu
†ww2040@columbia.edu
‡yyu15@ncsu.edu

1 Introduction

In this paper we develop and evaluate new engineering approximation formulas for heavily-loaded non-Markovian queueing systems with customer abandonment. Models that account for customer abandonment from queue due to customer impatience have generated substantial interest in recent years because of their application to call centers and other service systems. This has led to renewed interest in the Markovian $M/M/n+M$ Erlang-A model [11, 30, 40]. However, data analysis from service systems has also shown that the distributions of the service and patience times are often not nearly exponential [6]. Data analysis also has suggested that in some cases the arrival process might not be well modeled by a Poisson process; see [1, 17, 19, 21, 45].

Thus, we focus on the stationary $G/GI/n+GI$ queueing system, allowing a non-Poisson (and even non-renewal) stationary arrival process (the G), independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), multiple (n) servers working in parallel, unlimited waiting space, customer abandonment according to i.i.d. patience times with a general distribution (the $+GI$) and the first-come first-served service discipline. We provide simple formulas to approximate the mean, variance and distribution of important steady-state performance measures, including the number of customers in the system, the number in queue and the waiting time (which we take to be the potential waiting time, i.e., the time a potential arrival at time t that is infinitely patient would have to wait before starting service). We also give formulas to approximate important steady-state probabilities, including the probability of delay (PoD) and probability of abandonment (PoA).

The stationary $G/GI/n+GI$ queueing model has several model elements. Even the basic $M/M/n+M$ Erlang-A model has four parameters: the arrival rate λ, the individual service rate μ, the number of servers n and the abandonment rate θ. Without loss of generality (by choosing units to measure time), we can let $\mu = 1$ so that there is the parameter triple (λ, n, θ) or equivalently (ρ, n, θ), with $\rho \equiv \lambda/n\mu \equiv \lambda/n$ being the traffic intensity. In addition, the stationary $G/GI/n+GI$ queueing model has a general service-time cdf G with mean μ^{-1}, a general patience-time cdf F with mean θ^{-1} and a general stationary arrival process with rate λ, allowing complex dependence among interarrival times. We develop performance approximations for that general $G/GI/n+GI$ model and we conduct extensive experiments to study how all its model elements affect (i) the performance of the system and (ii) the accuracy of the proposed approximations. We summarize our conclusions in §10.

1.1 A Basis in Many-Server Heavy-Traffic Limits

We primarily base these new engineering approximations on the many-server heavy-traffic (MSHT) functional central limit theorem (FCLT) for the time-varying non-Markovian $G_t/M/n_t+GI$ queueing model in [26], with Gaussian process limits, but we also draw on the MSHT functional weak law of large numbers (FWLLN) for the more general time-varying $G_t/GI/n_t+GI$ queueing model in [24, 25, 42], yielding deterministic fluid limits. Scaled fluid limits appear as centering terms in the FCLT and so provide the mean values of the resulting Gaussian approximations. These MSHT limits apply to the stationary $G/GI/n+GI$ model as a special case. In this paper we investigate the quality of those approximations and develop heuristic refinements that are more effective.

For service systems, there is strong motivation for the time-varying feature, because service systems typically have arrival rates that vary strongly by the time of day. Nevertheless, what we do here still has significant relevance for service systems because stationary models often can be applied when the service times are relatively short, as in most call centers. With short service times,

stationary models can be used in a nonstationary way via the pointwise stationary approximation [12, 35], as reviewed in [13]. Indeed, that is the common approach used to staff service systems in practice.

The MSHT limits indicate when the approximations should be effective. Because MSHT limits involve letting the arrival rate and number of servers grow without bound, we expect the approximations to be more effective with large scale. Hence, our base case has $n = 100$ servers, but we also find that the approximations are quite effective for smaller scale, e.g., $n = 20$, 10 and 5.

The MSHT limits for the time-varying non-Markovian $G_t/M/n_t + GI$ model in [26] are for systems with a continuous time-varying arrival-rate function that makes the system alternate between overloaded (OL) intervals, where we locally have $\rho > 1$, and underloaded (UL) intervals, where we locally have $\rho < 1$, without being critically loaded (CL), where we locally have $\rho \approx 1$, over a positive interval. That limit yields one-sided approximations in each of the OL and UL intervls. Nevertheless, simulation experiments showed that the approximations are remarkably effective for difficult time-varying systems that are mostly not nearly critically loaded. However, for stationary models, these MSHT limits hold only for OL models with $\rho > 1$ or UL models with $\rho < 1$. Moreover, only the OL approximations are effective for the queue length and waiting time processes. For that reason, the approximations in this paper are only for heavily loaded models. We do provide information about the corresponding (less useful) approximations for UL systems in §8 and in the appendix [27].

We investigate the direct application of the MSHT Gaussian process limit in [26], but we find that the direct approximations are ineffective when the traffic intensity ρ is near 1. Thus, a significant contribution here is to develop effective heuristic refinements. We conduct extensive simulation experiments investigating when these approximations are effective. We find that the effectiveness of the refined Gaussian approximations for heavily loaded models primarily depends on two parameters: the traffic intensity ρ and the abandonment rate θ. Assuming that the scale (which is characterized by the relevant number of servers n) is not too small, we find that the refined approximations are effective roughly for $\rho > 1.02$ and $\theta < 2.0$. Since our approximations tend to work better when the system is heavily loaded, the quality of the approximations tends to improve as ρ increases and as θ decreases. We also find that our approximations are relatively robust to the variability in the arrival process, service times and patience times, in a reasonable range.

Even though we focus on heavily loaded models with $\rho > 1$, as observed previously, e.g. [41, 42], these are practical cases that often occur in practice, because the abandonment always keeps the system stable. (For example, see the $M/M/n + M$ base case with $\rho = 1.05$ in Table 2. The steady-state mean number in system and abandonment probabilities are representative of what is often seen in practice.) For these cases, the proposed approximations not only cover general non-Markovian models, but the accuracy of the approximations and simplicity of the formulas makes the proposed approximations attractive alternatives even for the Markovian $M/M/n+M$ Erlang-A model, as in [11, 30]; e.g., for quick approximations of OL models it may not be necessary to solve any birth-and-death equations.

From the range of abandonment rates θ, it should be evident that we are only considering models with abandonment, and only at a typical level; we are not considering the $G/G/n/0$ loss model or the $G/G/n/\infty$ delay model, which already have been quite extensively studied, e.g., see [22, 39, 37] and references therein. While the proposed approximations should be useful, they are far from universal approximations; we are *not* claiming that the approximations apply to all multi-server queueing models. To appreciate the limitations, recall that the waiting time distribution in $M/M/n$ queue and $G/GI/n$ generalizations has an exponential tail [14], unlike the much more rapidly decaying tails of our Gaussian approximations, which arise because of the abandonment.

The nice Gaussian approximations here can be understood as a generalization of exact results for the $M/M/n+M$ model that hold when $\theta = \mu$, where μ is the individual service rate. As discussed in §6 of [9], for that special parameter choice, the number in system has the same structure as in the associated $M/M/\infty$ infinite-server model, so that the steady-state distribution is exactly Poisson, which is approximately Gaussian provided that the arrival rate is not too small. Limit theorems supporting such Gaussian approximations for the $M/M/\infty$ model go back to [18] and there has been much work since then. The limits in [26] can be viewed as generalizations.

1.2 Overview of the Proposed Approach

We now give a brief overview of our approach. We start by applying the MSHT limits to generate a Gaussian approximation for the number of customers in the system. Unfolding the limit in the usual way, we obtain a direct approximation for the total number of customers in the system at time t, $X_n(t)$, as

$$X_n(t) \approx nX(t) + \sqrt{n}\hat{X}(t) \stackrel{\mathrm{d}}{=} \mathcal{N}(nX(t), n\sigma^2_{\hat{X}}(t)), \tag{1}$$

where X is the fluid approximation analyzed directly in [24] and obtained as a limit in the MSHT FWLLN in [25], while $\hat{X}$ is a zero-mean Gaussian process with variance $\sigma^2_{\hat{X}}(t) \equiv Var(\hat{X}(t))$ obtained from the MSHT FCLT in [26] (which assumes M service), and $\mathcal{N}(m, \sigma^2)$ denotes a Gaussian random variable with mean m and variance σ^2. To obtain associated approximations for the steady-state variable, denoted by $X_n(\infty)$, we take the direct approach and simply let $t \to \infty$ in (1) and obtain

$$X_n(\infty) \approx nX(\infty) + \sqrt{n}\hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(nX(\infty), n\sigma^2_{\hat{X}}(\infty)). \tag{2}$$

Of course, the associated number in queue satisfies $Q_n(t) \equiv (X_n(t) - n)^+$, where $(a)^+ \equiv \max\{a, 0\}$ and $\equiv$ denotes equality by definition, but if $\rho > 1$, then its MSHT limit holds without that truncation. If $\rho > 1$, then $X(\infty) > 1$ in (2), so that the direct MSHT limit for the number in queue yields

$$Q_n(\infty) \approx n(X(\infty) - 1) + \sqrt{n}\hat{X}(\infty), \tag{3}$$

without any truncation. That makes the approximation have $P(Q_n(\infty) < 0) > 0$, even though that is not possible. Similarly, the direct approximation allows $P(B_n(\infty) > n) > 0$, where $B_n(\infty)$ is the steady-state number in service, even though that is not possible. As a consequence, these direct MSHT approximations produce probability distributions in regions that cannot occur. Moreover, this defect can seriously degrade the approximations; e.g., see Table 2.

Because these direct Gaussian approximations are ineffective except when the system is significantly OL or UL, e.g., when $\rho > 1.2$, we investigate a simple refinement based on truncation that was proposed in (1.2) of [26], but never tested; indeed that is the purpose of the present paper. As in (1.2) of [26], the first step is to make the obvious refinement for the queue length, letting

$$Q_n(\infty) \equiv (X_n(\infty) - n)^+ \approx (nX(\infty) + \sqrt{n}\hat{X}(\infty) - n)^+. \tag{4}$$

Similarly, we let the number in service be

$$B_n(\infty) \equiv X_n(\infty) \wedge n \approx (nX(\infty) + \sqrt{n}\hat{X}(\infty) \wedge n,$$

where $a \wedge b \equiv \min\{a, b\}$. We call these approximations based on the FCLT for the $G/M/n+GI$ model in [26] the *truncated Gaussian approximations* (TGAs); see §3 for the details. In this paper we show that the approximations based on the MSHT limit plus this simple truncation refinement are remarkably effective for OL models.

1.3 Related literature

There are two streams of important related work. The first stream is the literature on MSHT limits, starting with models without abandonment in [18, 14] and then continuing with stationary models with abandonment [7, 8, 11, 16, 20, 33], and then time-varying Markov models in [28] and non-Markovian models in [24, 25, 26]; we refer to those papers for further discussion of the MSHT literature.

The second stream is the literature on exact results and approximations for the $M/GI/n+GI$ model [2, 4, 5, 44]. The exact results for the $M/M/n+GI$ model in [44] led to further studies that provided better understanding, such as [29]. However, exact results for the $M/GI/n+GI$ model with non-exponential service remains an important open problem. Approximations for the $M/GI/n+GI$ model were developed in [41]. As we have confirmed in our simulation experiments, these approximations from [41] are remarkably effective, but they are substantially more complicated, requiring approximation by a state-dependent $M/M/s+M(n)$ model and then a numerical algorithm.

A main conclusion in [41, 42] was that the steady state performance of the $M/GI/n+GI$ model tends to be nearly insensitive to the service-time distribution beyond its mean. Our experiments confirm that conclusion for the performance measures considered, but not for all performance measures. In particular, our simulations show that the mean values of the steady-state queue length and waiting time have this near-insensitivity property, but the variance and distribution do not; they depend significantly on the service-time distribution beyond its mean. Moreover, the simulations show that the new approximations of these quantities for the $M/GI/n+GI$ model with a non-exponential service times are effective; e.g., see Tables 7 and 8. Since the limit in [26] is only for the $G_t/M/s_t+GI$ model, with exponential service times, our approximation in this step is only heuristic. It does follow from a natural modification to account for GI service, but it remains to be better justified theoretically.

Refinements to the direct MSHT approximations for the time-varying Markovian $M_t/M/s_t+M$ model in [28] have also been developed by [31]. These Gaussian and skewness closure approximations are remarkably effective, even for critically loaded models, but so far they are restricted to Markovian models.

1.4 Organization of the Paper

In §2 we give a brief review of the heavy-traffic limits which are the building blocks for our approximations. In §3 we develop the TGA formulas for the $GI/GI/n+GI$ queueing systems. In §§4-7 we present results of numerical examples to test the performance of the Gaussian approximations: In §4 we consider the Markovian $M/M/n+M$ model; in §5 we consider the extensions to non-M arrival processes and non-M patience distributions; in §6 we consider the extension to non-M service; and in §7 we consider examples with fewer servers and a corresponding lower arrival rate.

We provide additional supporting material in §§8-9. In §8 we give examples exposing the limitations of the approximations. In §8.1 and §8.2 we compare our approximations to previous approximations developed by Whitt [41] and Reed and Tezcan [33]; these tend to perform better in the QED regime. In §8.3 we expose the limitations of the approximations for UL models. In §9 we elaborate on the simulation methodology. We draw conclusions in §10. We present additional material in an appendix [27], which is available as supporting information on the journal and author web pages.

2 The Many-Server Heavy-Traffic Limits for Stationary Models

The MSHT limits are obtained by considering a sequence of stationary queueing models indexed by the integers n. In the n^{th} model, there are $s_n = \lceil ns \rceil$ servers and the arrival rate is $\lambda_n = n\lambda$, where λ is the base arrival rate. (The scaling factor becomes the number of servers when we let $s \equiv 1$.) The service times and patience times are unscaled; they are assumed to come from independent sequences of i.i.d. random variables with cumulative distribution functions (cdfs) G and F with means $E[S] = 1/\mu = 1$ and $E[A] = 1/\theta$ and finite second moments. Let f and $\bar{F}$ be the probability density function (pdf) and complementary cumulative distribution function (ccdf) of F; and let $\bar{G}$ be the ccdf of G. Thus the traffic intensity is $\rho_n = \lambda_n/s_n\mu = \lambda/s\mu \equiv \rho$ for all n.

The arrival process $N_n(t)$ is assumed to satisfy a FCLT

$$(N_n(t) - n\lambda t)/\sqrt{n} \Rightarrow c_\lambda \mathcal{B}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty, \tag{5}$$

where $\mathcal{B}$ is a standard Brownian motion (BM), $c_\lambda > 0$ is a variability parameter, $\Rightarrow$ denotes weak convergence and $\mathbb{D}$ denotes the space of right-continuous functions that have left limits; see [3, 26, 38] for details.

Let $Q_n(t)$ and $B_n(t)$ to be the number of customers waiting in queue and in service at time t. Let $X_n(t) = Q_n(t) + B_n(t)$ be the total number in the system. Let $W_n(t)$ and $V_n(t)$ to be the head-of-line waiting time (the elapsed waiting time of the head-of-line customer if there is any) and the potential waiting time at time t (the waiting time of an infinitely patient arrival at time t if there were an arrival at that time). We next review the MSHT FWLLN and FCLT limits for the stationary model, which are the building blocks for our Gaussian approximations.

2.1 The FWLLN Yielding Fluid Limits for the $G/GI/n + GI$ Queue

Let the LLN-scaled processes be

$$\bar{X}_n(t) = \frac{X_n(t)}{n}, \quad \bar{Q}_n(t) = \frac{Q_n(t)}{n}, \quad \bar{B}_n(t) = \frac{B_n(t)}{n}, \quad \bar{W}_n(t) = W_n(t) \quad \text{and} \quad \bar{V}_n(t) = V_n(t).$$

The waiting times need no spatial scaling because the service-time and patience-time cdf's are not scaled. By Theorem 1 in [25], we have the joint convergence for the LLN-scaled functions

$$\left(\bar{X}_n, \bar{Q}_n, \bar{B}_n, \bar{W}_n, \bar{V}_n\right) \Rightarrow (X, Q, B, w, v), \quad \text{in } \mathbb{D}^5, \tag{6}$$

where the limit is a vector of deterministic functions, specified in [42, 24]. We next summarize the steady state performance.

Theorem 2.1 (Theorem 3.1 in [42] and Theorem 4.1 in [23])**.** *The stationary fluid model with capacity s arising as the MSHT FWLLN limit of stationary $G/GI/n+GI$ model with s_n servers has a steady state characterized by the deterministic vector $(b(\infty), q(\infty), B(\infty), Q(\infty), X(\infty), w(\infty), v(\infty))$ in $\mathbb{R}^7$, where $X(\infty) = B(\infty) + Q(\infty)$ and the other variables depend on the value of the traffic intensity $\rho \equiv \lambda/s\mu = \lambda/s$.*

(a) If $\rho \le 1$, then for $x \ge 0$,

$$B(\infty) = \int_0^\infty b(x)\,dx = s\rho, \quad b(x) = \lambda\bar{G}(x), \quad Q(\infty) = \int_0^\infty q(x)\,dx = w(\infty) = v(\infty) = q(x) = 0.$$

(b) If $\rho > 1$, then

$$B(\infty) = \int_0^\infty b(x)\,dx = s, \quad b(x) = s\mu\bar{G}(x) \quad \textit{for} \quad x \geq 0,$$
$$v(\infty) = w(\infty) = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad q(x) = \lambda\bar{F}(x) \quad \textit{for} \quad 0 \leq x \leq w(\infty),$$
$$Q(\infty) = \int_0^\infty q(x)\,dx = \lambda\int_0^{w(\infty)} \bar{F}(x)dx.$$

2.2 The FCLT Yielding Gaussian limits for the $G/M/n + GI$ Queue

As in [26], we now restrict to the $G/M/n + GI$ model having M service. To provide the FCLT limits, let the CLT-scaled processes be

$$\hat{X}_n(t) = \frac{X_n(t) - nX(t)}{\sqrt{n}}, \quad \hat{Q}_n(t) = \frac{Q_n(t) - nQ(t)}{\sqrt{n}}, \quad \hat{B}_n(t) = \frac{B_n(t) - nB(t)}{\sqrt{n}},$$
$$\hat{W}_n(t) = \sqrt{n}\,(W_n(t) - w(t))\,, \quad \hat{V}_n(t) = \sqrt{n}\,(V_n(t) - v(t))\,, \tag{7}$$

where n is the number of servers, while $X(t), Q(t), B(t), w(t)$ and $v(t)$ are the deterministic limit functions in (6), i.e., the deterministic fluid functions in [24]. The FCLT follows directly from the FCLT for the time-varying $G_t/M/s_t + GI$ model (Theorems 4.2 and 5.1 in [26]) by simply letting $\lambda(t) = \lambda$ and $s(t) = s$.

Theorem 2.2 (MSHT FCLT limits for the $G/M/n + GI$ queues)**.** *Consider the sequence of $G/M/n + GI$ queueing models having s_n servers. Under regularity conditions in [26], including appropriate initial convergence at time 0,*

$$\left(\hat{X}_n, \hat{B}_n, \hat{Q}_n, \hat{W}_n, \hat{V}_n\right) \Rightarrow \left(\hat{X}, \hat{B}, \hat{Q}, \hat{W}, \hat{V}\right) \quad \textit{in } \mathbb{D}^5 \quad \textit{as } n \to \infty.$$

(a) When $\rho < 1$, i.e. in an underloaded (UL) interval, $\hat{Q} = \hat{W} = \hat{V} = 0$, and $\hat{B} \stackrel{\mathrm{d}}{=} \hat{X}$ satisfies the stochastic differential equation (SDE)

$$d\hat{X}(t) = -\mu\hat{X}(t)dt + \sqrt{c_\lambda^2\lambda + \mu X(t)}\,d\mathcal{B}(t),$$

where $\mathcal{B}$ is a standard Brownian motion, so that $\hat{X}(t)$ is a Gaussian process with

$$\sigma_X^2(t) \;=\; (c_\lambda^2 - 1)\int_0^t \bar{G}^2(t-u)\lambda du + \int_0^t \bar{G}(t-u)\lambda du.$$

(b) When $\rho > 1$, i.e. in an overloaded (OL) interval, $\hat{B}(t) = 0$, $\hat{Q}(t) = \hat{X}(t)$ where

$$\hat{X}(t) \;=\; \hat{X}(0)\bar{F}_w(t) + \sum_{i=1}^3 \int_0^t K_i(t,u)\,d\mathcal{B}_i(u),$$
$$\hat{W}(t) \;=\; \hat{W}(0)H(t,0) + \sum_{i=1}^3 \int_0^t H(t,u)I_i(u)\,d\mathcal{B}_i(u),$$
$$\hat{V}(t) \;=\; \frac{\hat{W}(t + v(t))}{1 - \dot{w}(t + v(t))},$$

where $\mathcal{B}_1 \equiv \mathcal{B}_\lambda$, $\mathcal{B}_2 \equiv \mathcal{B}_s$ and $\mathcal{B}_3 \equiv \mathcal{B}_a$ are independent standard BMs, that are the FCLT limits of the scaled arrival process (the subscript "λ"), service process (the subscript "s") and abandonment process (the subscript "a"). Both $\hat{X}(t)$ and $\hat{W}(t)$ are zero mean Gaussian processes with variance process

$$\begin{aligned}\sigma_W^2(t) &= \int_0^t H^2(t,u) I^2(u)\, du + H^2(t,0)\, Var\left(\hat{W}(0)\right), \qquad (8)\\ \sigma_X^2(t) &= \int_{t-w(t)}^t \lambda\, \bar{F}(t-u)\left((c_\lambda^2-1)\bar{F}(t-u)+1\right) du \\ &\quad +\lambda^2 \bar{F}^2(w(t))\sigma_W^2(t) + Var\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2. \qquad (9)\end{aligned}$$

Finally, $H(t,u)$, $I_i(t)$, $K_i(t)$ are deterministic analytic functions given by

$$\begin{aligned}H(t,u) &= \exp\left\{\int_u^t h(v)dv\right\} = \exp\left\{\int_u^t -\frac{f(w(\infty))}{\bar{F}(w(\infty))}dv\right\} = e^{-\frac{f(w(\infty))}{\bar{F}(w(\infty))}(t-u)},\\ I_1^2(t) &= \frac{c_\lambda^2 \bar{F}(w(u))b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{c_\lambda^2 \bar{F}(w(\infty))s\mu}{\lambda^2\bar{F}^2(w(\infty))} = c_\lambda^2\lambda^{-1},\\ I_2^2(t) &= \frac{b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{s\mu}{\lambda^2\bar{F}^2(w(\infty))} = \rho/\lambda,\\ I_3^2(t) &= \frac{F(w(u))b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{F(w(\infty))s\mu}{\lambda^2\bar{F}^2(w(\infty))} = (\rho-1)/\lambda,\\ I^2(t) &= I_1^2(t)+I_2^2(t)+I_3^2(t) = \frac{(c_\lambda^2-1)\rho^{-1}+2}{\mu s},\\ K_1(t,u) &= c_\lambda \bar{F}(t-u)\sqrt{\lambda(u)}\mathbb{1}_{\{t-w(t)u<t\}} + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_1(L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},\\ K_2(t,u) &= -\sqrt{b(t,0)-\dot{s}(t)}H(t,u),\\ K_3(t,u) &= -\sqrt{\lambda(u)F(t-u)\bar{F}(t-u)}\mathbb{1}_{\{t-w(t)\le u\le t\}}\\ &\quad + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_3(L^{-1}(u))H(t,L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},\end{aligned}$$

where

$$\bar{I}_1(t) = \frac{c_\lambda \bar{F}(w(u))b(u,0)}{\tilde{q}(u,w(u))} = \frac{c_\lambda \bar{F}(w(\infty))s\mu}{\lambda\bar{F}(w(\infty))}, \quad \bar{I}_3(t) = -\frac{\sqrt{\bar{F}(w(u))b(u,0)}}{\tilde{q}(u,w(u))} = -\frac{\bar{F}(w(\infty))s\mu}{\lambda\bar{F}(w(\infty))},$$

$L(t) = t - w(\infty)$, $\lambda(t) = \lambda t$ and $\mathbb{1}_A$ is an indicator random variable of A.

We next provide steady-state distribution for the FCLT limits of the $G/M/n+GI$ model.

Theorem 2.3 (Steady-state of the MSHT FCLT limit of $G/M/n+GI$ queues)**.** *The steady-state of the Gaussian process arising in the MSHT FCLT limit for the sequence of $G/M/n+GI$ models with s_n servers in model n is given by the vector $(\hat{Q}(\infty), \hat{B}(\infty), \hat{X}(\infty), \hat{W}(\infty), \hat{V}(\infty))$ specified below:*

(a) If $\rho < 1$, then

$$\hat{Q}(\infty) = \hat{W}(\infty) = \hat{V}(\infty) = 0 \quad \textit{and} \quad \hat{B}(\infty) = \hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0, \sigma_X^2\right),$$

where the variance is

$$\sigma_X^2 \equiv \lambda(c_\lambda^2-1)\int_0^\infty \bar{G}^2(u)\,du + \lambda\int_0^\infty \bar{G}(u)\,du. \qquad (10)$$

(b) If $\rho > 1$, then

$$\hat{Q}(\infty) \stackrel{\mathrm{d}}{=} \hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0, \sigma_X^2\right), \quad \hat{B}(\infty) = 0, \quad \hat{W}(\infty) \stackrel{\mathrm{d}}{=} \hat{V}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(0, \sigma_W^2),$$

where the variances are

$$\sigma_W^2 \equiv \frac{(c_\lambda^2 - 1) + 2\rho}{2\mu s \rho^2 f(w(\infty))} \quad \textit{and} \quad \sigma_X^2 \equiv (\mu s)^2 \sigma_W^2 + \lambda \int_0^{w(\infty)} \bar{F}(u) \left(1 + (c_\lambda^2 - 1)\bar{F}(u)\right) du, \quad (11)$$

with $w(\infty)$ being the steady-state fluid waiting time in the OL case, given in Theorem 2.1.

Proof. Because the limit process is a zero-mean Gaussian process, it suffices to show that the variances converge as $t \to \infty$. In particular, it suffices to show that $\sigma_W^2(t) \to \sigma_W^2$ and $\sigma_X^2(t) \to \sigma_X^2$ as $\to \infty$. That is easy to check in underloaded interval, thus we focus on the limit in the overloaded interval.

$$\begin{aligned}
\sigma_W^2 &= \lim_{t\to\infty} \sigma_W^2(t) = \lim_{t\to\infty} \left(\int_0^t I^2(u) \cdot H^2(t,u) \mathrm{d}u + H^2(t,0) \mathrm{Var}(\hat{W}(0)) \right) \\
&= \lim_{t\to\infty} \left(\frac{(c_\lambda^2 - 1)\rho^{-1} + 2}{\mu s} \int_0^t e^{-\frac{2f(w(\infty))}{\bar{F}(w(\infty))}(t-u)} \mathrm{d}u + e^{-\frac{2f(w(\infty))}{\bar{F}(w(\infty))}t} \mathrm{Var}(\hat{W}(0)) \right) \\
&= \frac{\bar{F}(w(\infty))}{2f(w(\infty))} \frac{(c_\lambda^2 - 1)\rho^{-1} + 2}{\mu s} = \frac{(c_\lambda^2 - 1) + 2\rho}{2\mu s \rho^2 f(w(\infty))},
\end{aligned}$$

where the last equality holds because $\bar{F}(w(\infty)) = 1/\rho$. Inserting it to (9), we get the remaining limit

$$\begin{aligned}
\sigma_X^2 \equiv \lim_{t\to\infty} \sigma_X^2(t) &= \lim_{t\to\infty} \mathrm{Var}\left(\hat{X}(0)\right) \cdot \left(\bar{F}_w(t)\right)^2 + \lim_{t\to\infty} \lambda^2 \bar{F}^2(w(t)) \sigma_W^2(t) \\
&\quad + \lim_{t\to\infty} \int_{t-w(t)}^t \lambda \bar{F}(t-s) \left((c_\lambda^2 - 1)\bar{F}(t-s) + 1\right) \mathrm{d}s \\
&= (\lambda/\rho)^2 \sigma_W^2 + \lim_{t\to\infty} \lambda \int_0^{w(t)} \bar{F}(u) \left((c_\lambda^2 - 1)\bar{F}(u) + 1\right) \mathrm{d}u \\
&= (\mu s)^2 \sigma_W^2 + \lambda \int_0^{w(\infty)} \bar{F}(u) \left((c_\lambda^2 - 1)\bar{F}(u) + 1\right) \mathrm{d}u,
\end{aligned}$$

where $\lim_{t\to\infty} \bar{F}(w(t)) = \bar{F}(w(\infty)) = 1/\rho$, $\lim_{t\to\infty} \mathrm{Var}\left(\hat{X}(0)\right) \cdot \left(\bar{F}_w(t)\right)^2 = 0$ since $\bar{F}(t) \to 0$ as $t \to \infty$ but $\mathrm{Var}\left(\hat{X}(0)\right)$ is bounded.

Full convergence of the multivariate Gaussian distribution in $\mathbb{R}^5$ requires convergence of the pairwise covariances too. That also follows from the representation in Theorem 2.2, but we omit the details, because we will not be applying that. □

Remark 2.1 (Interchange of the two limits)**.** Our MSHT approximation is based on an iterated limit in which first $n \to \infty$ and then $t \to \infty$, but we need the iterated limit in the other order. As often is done in MSHT approximations, we are assuming without proof that a limit interchange is valid, in particular,

$$\lim_{t\to\infty} \hat{X}(t) = \lim_{t\to\infty} \lim_{n\to\infty} \hat{X}_n(t) = \lim_{n\to\infty} \lim_{t\to\infty} \hat{X}_n(t) = \lim_{n\to\infty} \hat{X}_n(\infty). \quad (12)$$

We conjecture that (12) holds for the $G/GI/n+GI$ model under mild regularity conditions, if any. For the more elementary $M/M/n$, $M/M/n+M$ and $M/M/n+GI$ models, that limit interchange was proved in [14, 11, 44].

Remark 2.2 (The underloaded case)**.** The MSHT limits above are relatively elementary in the underloaded case with $\rho < 1$. In that case the results reduce to corresponding infinite-server results in [32] and references there, as shown for the $M/M/n$ model in [18].

3 The Steady-State Gaussian Approximations

In this section, we develop the Gaussian approximation formulas for the steady-state distribution of the stationary $G/GI/n+GI$ model. Henceforth, we focus on steady-state random variables and no longer discuss stochastic processes. Thus, for simplicity, we omit the time argument t, which would become ∞ in steady state; i.e., we let Q_n denote the steady-state queue length in model n with s_n servers and we let Q and $\hat{Q}$ denote the steady-state of the fluid and diffusion limits, respectively, and similarly for the other variables. (Previously, these were denoted by $Q_n(\infty)$, $Q(\infty)$ and $\hat{Q}(\infty)$.) Since the steady-state head-of-line waiting time W_n and potential waiting time V_n should coincide, as do the limits in Theorem 2.1 (b) and Theorem 2.3 (b), we henceforth use only the notion W_n, w and $\hat{W}$ for the waiting time (instead of $W_n(\infty)$, $V_n(\infty)$, $w(\infty)$, $v(\infty)$, $\hat{W}(\infty)$ and $\hat{V}(\infty)$).

We start in §3.1 with direct Gaussian approximations (DGA) and then turn to the basic truncation refinement TGA for the $G/M/n+GI$ model in §3.2. Afterwards, in §3.3 we generalize TGA from M service to GI service, which we refer to as TGA-G.

3.1 Direct Gaussian Approximations

The most straightforward performance approximation is a direct application of the steady-state of the deterministic fluid and the zero-mean Gaussian limits. We we use superscript "DGA" to denote these approximations, which are

$$X_n \approx X_n^{DGA} \equiv nX + \sqrt{n}\hat{X}, \tag{13}$$

$$B_n \approx B_n^{DGA} \equiv nB + \sqrt{n}\hat{B}, \quad Q_n^{DGA} \equiv nQ + \sqrt{n}\hat{Q}, \tag{14}$$

$$W_n \approx W_n^{DGA} \equiv w + \frac{1}{\sqrt{n}}\hat{W}, \tag{15}$$

where X, B, Q and w are given in Theorem 2.1 and $\hat{X}$, $\hat{B}$, $\hat{Q}$ and $\hat{W}$ are given in Theorem 2.3. (Recall that we have eliminated the infinite time argument from the steady-state quantities given in Theorems 2.1 and 2.3.)

In addition to approximating means and variances, defined in the obvious direct way, we have associated approximations for the PoD and PoA, namely,

$$PoD \approx P_D^{DGA}(n) \equiv \mathbb{P}\left(W_n^{DGA} > 0\right) \quad \text{and} \tag{16}$$

$$PoA \approx P_A^{DGA}(n) \equiv \mathbb{P}\left(W_n^{DGA} > A\right), \tag{17}$$

where A is a generic independent patience time.

3.2 Truncated Gaussian Approximations

It is natural to refine the DGA approximations by truncation because $Q_n = (X_n - s_n)^+$ and $B_n = (X_n \wedge s_n)^+$, where a^+ is $\max\{0, a\}$ and $a \wedge b = \min\{a, b\}$. Thus, our TGA approximations

for Q_n and B_n are

$$Q_n^{TGA} \equiv \left(X_n^{DGA} - ns\right)^+ = \sqrt{n}\sigma_X \left(\frac{\hat{X}}{\sigma_X} + \frac{\sqrt{n}(X-s)}{\sigma_X}\right)^+ \equiv \sqrt{n}\sigma_X \left(\mathcal{Z} \vee -a_X(n)\right) + n(X-s),$$
$$B_n^{TGA} \equiv \left(X_n^{DGA} \wedge ns\right)^+ = \sqrt{n}\sigma_X \left(\left(\mathcal{Z} \wedge -a_X(n)\right) + \frac{\sqrt{n}}{\sigma_X} X\right)^+, \tag{18}$$

where $\mathcal{Z}$ is a standard Gaussian random variable, σ_X is given in (11), and

$$a_X(n) = \sqrt{n}(X-s)/\sigma_X. \tag{19}$$

In the same spirit, we truncate the DGAs for waiting times (15) to obtain their TGAs.

$$W_n^{TGA} \quad = \quad V_n^{TGA} = \left(W_n^{DGA}\right)^+ = \left(w + \frac{\hat{W}}{\sqrt{n}}\right)^+ = w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right), \tag{20}$$

where σ_W is given in (11) and

$$a_W(n) = \sqrt{n}w/\sigma_W. \tag{21}$$

The means, variances and distributions are then approximated in the obvious way. The formulas can be conveniently expressed as functions of the standard normal ccdf $\bar{\Phi}$ and pdf ϕ via

$$\mathrm{E}[V_n^{TGA}] = \mathrm{E}[w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right)] = w\left(\Phi(a_W(n)) + \frac{\phi(a_W(n))}{a_W(n)}\right),$$
$$\begin{aligned}\mathrm{Var}(V_n^{TGA}) &= \mathrm{Var}(w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right)), \\ &= \frac{(a_W(n)\sigma_W)^2}{n}\left[\left(1 + \frac{1}{a_W^2(n)}\right)\bar{\Phi}(a_W(n)) - \frac{\phi(a_W(n))}{a_W(n)} - \left(\frac{\phi(a_W(n))}{a_W(n)} - \bar{\Phi}(a_W(n))\right)^2\right],\end{aligned}$$
$$\mathrm{E}[Q_n^{TGA}] = \mathrm{E}[n(X-s) + \frac{\sigma_X}{\sqrt{n}}\left(\mathcal{Z} \vee -a_X(n)\right)] = n(X-s)\left(\Phi(a_X(n)) + \frac{\phi(a_X(n))}{a_X(n)}\right),$$
$$\begin{aligned}\mathrm{Var}(Q_n^{TGA}) &= \mathrm{Var}(n(X-s) + \frac{\sigma_X}{\sqrt{n}}\left(\mathcal{Z} \vee -a_X(n)\right)) \\ &= n(a_X(n)\sigma_X)^2\left[\left(1 + \frac{1}{a_X^2(n)}\right)\bar{\Phi}(a_X(n)) - \frac{\phi(a_X(n))}{a_X(n)} - \left(\frac{\phi(a_X(n))}{a_X(n)} - \bar{\Phi}(a_X(n))\right)^2\right],\end{aligned}$$

where $a_W(n)$ and $a_X(n)$ are given in (21) and (19).

The PoD and PoA are natural analogs of (16) and (17), i.e.,

$$\begin{aligned} PoD \quad &\approx \quad P_D^{TGA}(n) \equiv \mathbb{P}\left(W_n^{TGA} > 0\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > 0\right) \\ &= \quad \bar{\Phi}\left(-a_W(n)\right), \end{aligned} \tag{22}$$

$$\begin{aligned} PoA \quad &\approx \quad P_A^{TGA}(n) \equiv \mathbb{P}\left(W_n^{TGA} > A\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > A\right) \\ &= \quad \int_0^\infty \bar{\Phi}\left(a_W(n)\left(\frac{x}{w} - 1\right)\right) f(x)\mathrm{d}x, \end{aligned} \tag{23}$$

where f is the pdf of the patience time, as before.

Notice that several of the TGA approximations coincide with their DGA counterparts, i.e.,

$$(X_n^{TGA}, \mathrm{E}[X_n^{TGA}], \mathrm{Var}(X_n^{TGA}), P_A^{TGA}(n), P_D^{TGA}(n)) = (X_n^{DGA}, \mathrm{E}[X_n^{DGA}], \mathrm{Var}(X_n^{DGA}), P_A^{DGA}(n), P_D^{DGA}(n))$$

3.3 The Refinement for Non-Exponential Service: TGA-G

Recall that both DGAs and TGAs are developed based on Theorems 2.2 and 2.3 for the $G/M/n+GI$ queue. However, the form of the limit in 2.3 allows us to identify the impact of the service process. In particular, we see that the three sources of variability – the arrival process, service times and patience times – produce identifiable impacts on the limiting Gaussian processes through three independent Brownian motions. Thus, even though the impact of the service process is complicated, we can exploit this separability to see how to introduce a good candidate heuristic approximation.

First, for the OL model, where asymptotically all servers are busy all the time, we can act as if the servers are continuously busy. Because of the assumed exponential service time distribution, the approximating service-completion process is a Poisson process with rate $\mu s(t)$, which yields a Brownian FCLT limit, corresponding to the BM $\mathcal{B}_2 = \mathcal{B}_s$ in Theorem 2.2 (See Theorem 4.2 in [26] for details; also see (4.5) and (6.64) there). When service becomes GI, and when the servers are all busy, the service process is the superposition of i.i.d. renewal processes. Because the number of servers n grows in the limit, that superposition departure process is complicated, as discussed in §9.8 of [38]. However, one candidate approximation is to act as if n is fixed. Then the departure process again satisfies a FCLT with a Brownian motion limit, but with the prefactor c_s, just as if the service process were a single renewal process with c_s^2 being the squared coefficient of variation (scv, variance divided by the square of the mean) of the time between renewals; see §9.4 of [38].

Consequently, in our heuristic approximation we capture the nonexponential service distribution by scaling the BM $\mathcal{B}_2$ by adding the prefactor c_s. That leads to replacing the "1" by "c_s" in the numerator of (11), yielding

$$\sigma^2_{W_G} \quad = \quad \frac{(c_\lambda^2 - 1) + (c_s + 1)\rho}{2\mu s \rho^2 f(w)}. \tag{24}$$

By replacing σ_W by σ_{W_G} in (11), (18)–(20), we obtain TGA-G. It is significant that TGA-G reduces to TGA when service is M, for which $c_s^2 = 1$.

4 Evaluating the Gaussian Approximations for Markov Models

Since the Markovian $M/M/n+M$ model is relatively tractable, we primarily want to develop effective approximations for other non-Markov $G/GI/n+GI$ models. Nevertheless, it is convenient to start examining the proposed Gaussian approximations by making comparisons with exact numerical results for the $M/M/n+M$ model because numerical algorithms are readily available. A minimum requirement for our proposed approximations is that they perform well for this basic model.

Hence, we start evaluating the proposed approximations in this section by comparing with exact numerical results for $M/M/n+M$ model. For that purpose, we use the numerical algorithm from [41]. That algorithm was developed to treat the more general $M/GI/n+GI$ model by approximating it by the Markovian $M/M/n+M(n)$ model with state-dependent abandonment rate, but it applies to the $M/M/n+M$ model as a special case. That model includes a finite waiting room; we let it be so large that it does not affect the formulas.

Our base case has $n = 100$ servers, but we also examine smaller systems later. For the overloaded systems of primary interest, we considered a range of traffic intensities from $\rho = 1.5$ down to 1.001. The case $\rho = 1.5$ is so overloaded that there is little need for truncation, so that the TGA and DGA approximations nearly coincide, and the performance is very good; see Table 18 in the appendix. We thus start by showing the experimental results for $\rho = 1.2$ with abandonment rates $0.1 \le \theta \le 4.0$

in Table 1, where $M(m)$ refers to an exponential distribution with mean m. For the lower three abandonments rates ($\theta \leq 0.5$), the system is still highly heavily loaded. For these two cases in Table 1, TGA and DGA are quite close with all errors less than 1%. However, otherwise we see that TGA provides significant improvement.

Table 1: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.2$ for six values of θ, $0.10 \leq \theta \leq 4.00$.

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	3.00E+2	3.00E+2	same	1.80E+2	1.80E+2	same	1.40E+2	1.40E+2	same
rel. err.		0%			0%			0%	
Var(X)	1.20E+3	1.20E+3	same	4.80E+2	4.80E+2	same	2.39E+2	2.40E+2	same
rel. err.		0%			0%			0%	
E[Q]	2.00E+2	2.00E+2	2.00E+2	8.00E+1	7.99E+1	7.99E+1	4.00E+1	3.99E+1	3.99E+1
rel. err.		0%	0%		0%	0%		0%	0%
Var(Q)	1.20E+3	1.20E+3	1.20E+3	4.80E+2	4.80E+2	4.80E+2	2.38E+2	2.40E+2	2.38E+2
rel. err.		0%	0%		0%	0%		1%	0%
E[W]	1.83E+0	1.82E+0	1.82E+0	7.34E-1	7.29E-1	7.29E-1	3.70E-1	3.65E-1	3.65E-1
rel. err.		0%	0%		1%	1%		1%	1%
Var(W)	1.00E-1	1.00E-1	1.00E-1	4.00E-2	4.00E-2	4.00E-2	1.99E-2	2.00E-2	1.98E-2
rel. err.		0%	0%		0%	0%		1%	0%
PoD	1.00E+0	1.00E+0	same	1.00E+0	1.00E+0	same	9.97E-1	9.95E-1	same
rel. err.		0%			0%			0%	
PoA	1.67E-1	1.66E-1	same	1.67E-1	1.66E-1	same	1.67E-1	1.65E-1	same
rel. err.		0%			1%			1%	

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.20E+2	1.20E+2	same	1.10E+2	1.10E+2	same	1.04E+2	1.05E+2	same
rel. err.		0%			0%			1%	
Var(X)	1.20E+2	1.20E+2	same	6.35E+1	5.99E+1	same	3.76E+1	2.99E+1	same
rel. err.		0%			6%			20%	
E[Q]	2.01E+1	1.98E+1	2.00E+1	1.02E+1	9.86E+0	1.02E+1	5.22E+0	4.92E+0	5.47E+0
rel. err.		1%	1%		3%	0%		6%	5%
Var(Q)	1.14E+2	1.20E+2	1.13E+2	5.22E+1	5.99E+1	5.00E+1	2.29E+1	2.99E+1	2.14E+1
rel. err.		5%	1%		15%	4%		31%	7%
E[W]	1.88E-1	1.82E-1	1.83E-1	9.76E-2	9.10E-2	9.43E-2	5.17E-2	4.60E-2	5.09E-2
rel. err.		3%	3%		7%	3%		11%	2%
Var(W)	9.61E-3	1.00E-2	9.40E-3	4.48E-3	5.00E-3	4.19E-3	2.04E-3	2.50E-3	1.81E-3
rel. err.		4%	2%		11%	7%		23%	11%
PoD	9.72E-1	9.66E-1	same	9.06E-1	9.01E-1	same	8.01E-1	8.21E-1	same
rel. err.		1%			1%			2%	
PoA	1.68E-1	1.64E-1	same	1.70E-1	1.65E-1	same	1.74E-1	1.73E-1	same
rel. err.		2%			3%			1%	

We regard the case $\rho = 1.2$ as quite heavily loaded, much more in the ED heavy-traffic regime than the QED regime. Hence, we primarily focus on models with lower traffic intensities. For the traffic intensity, we regard $\rho = 1.05$ as our base case; it corresponds to levels often encountered in practice. Moreover, for $n = 100$ and $\rho = 1.05$, the system is operating in the more practical QED regime, which can be characterized by the quality-of-service (QoS) parameter $\beta \equiv (1 - \rho)\sqrt{n} = -0.5$; see [14, 11]. Table 2 shows the main performance measures for six abandonment rates with $0.1 \leq \theta \leq 4.0$. Table 2 shows that DGA performs poorly in this case, but TGA provides dramatic improvement, having all errors in the means $E[Q]$ and $E[W]$ and variances $Var(Q)$ and $Var(W)$ less than 10% for $\theta \leq 2.0$.

As indicated earlier, we find good performance for $\theta \leq 2.0$. For $\theta = 0.1$ and 0.25 with $(n, \rho) = (100, 1.05)$, the maximum percentage error among the means E[X], E[Q], E[W] and the probabilities

Table 2: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.05$ for six values of θ, $0.10 \leq \theta \leq 4.00$.

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.52E+2	1.50E+2	same	1.22E+2	1.20E+2	same	1.11E+2	1.10E+2	same
rel. err.		1%			2%			1%	
Var(X)	9.25E+2	1.05E+3	same	3.47E+2	4.20E+2	same	1.81E+2	2.10E+2	same
rel. err.		16%			21%			16%	
E[Q]	5.22E+1	4.99E+1	5.08E+1	2.30E+1	1.99E+1	2.17E+1	1.27E+1	9.94E+0	1.21E+1
rel. err.		4%	2%		14%	6%		22%	5%
Var(Q)	8.99E+2	1.05E+3	9.41E+2	3.05E+2	4.20E+2	3.11E+2	1.35E+2	2.10E+2	1.33E+2
rel. err.		19%	7%		37%	2%		56%	1%
E[W]	5.14E-1	4.88E-1	4.96E-1	2.29E-1	1.95E-1	2.12E-1	1.28E-1	9.80E-2	1.18E-1
rel. err.		5%	3%		15%	8%		23%	7%
Var(W)	8.59E-2	1.00E-1	8.97E-2	2.93E-2	4.00E-2	2.97E-2	1.31E-2	2.00E-2	1.27E-2
rel. err.		19%	7%		36%	1%		53%	3%
PoD	9.67E-1	9.39E-1	same	8.90E-1	8.35E-1	same	8.03E-1	7.56E-1	same
rel. err.		3%			6%			6%	
PoA	4.97E-2	4.80E-2	same	5.47E-2	5.09E-2	same	6.04E-2	5.75E-2	same
rel. err.		4%			7%			5%	

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.05E+2	1.05E+2	same	1.01E+2	1.02E+2	same	9.86E+1	1.01E+2	same
rel. err.		0%			1%			3%	
Var(X)	1.05E+2	1.05E+2	same	6.85E+1	5.24E+1	same	5.00E+1	2.61E+1	same
rel. err.		0%			23%			48%	
E[Q]	7.03E+0	4.92E+0	7.01E+0	3.88E+0	2.36E+0	4.22E+0	2.12E+0	1.13E+0	2.65E+0
rel. err.		30%	0%		40%	8%		47%	25%
Var(Q)	5.92E+1	1.05E+2	5.71E+1	2.57E+1	5.24E+1	2.50E+1	1.10E+1	2.61E+1	1.13E+1
rel. err.		77%	3%		103%	3%		139%	3%
E[W]	7.22E-2	4.90E-2	6.91E-2	4.09E-2	2.40E-2	4.18E-2	2.32E-2	1.20E-2	2.65E-2
rel. err.		32%	4%		42%	2%		48%	14%
Var(W)	5.88E-3	1.00E-2	5.49E-3	2.64E-3	5.00E-3	2.42E-3	1.18E-3	2.50E-3	1.10E-3
rel. err.		70%	7%		89%	9%		112%	6%
PoD	7.00E-1	6.88E-1	same	5.92E-1	6.33E-1	same	4.87E-1	5.95E-1	same
rel. err.		2%			6%			22%	
PoA	6.70E-2	6.43E-2	same	7.40E-2	7.59E-2	same	8.07E-2	9.31E-2	same
rel. err.		3%			3%			16%	

PoD and PoA for TGA was 7%. On the other hand, Table 21 shows that the performance of TGA degrades for $\theta = 4$ and 10, but then the high abandonment rate makes the system far from being overloaded; e.g., $\mathrm{E}[Q] = 1.47$ and $\mathrm{E}[W] = 0.017$ for $\theta = 4$.

Most of the rest of this paper is devoted to showing that the good results in Table 2 extend to a wide class of models and parameters. We illustrate in Figures 1-3, which graphically show the performance for the $M/M/n + M$ model as a function of ρ and θ. Figures 1 and 2 show the performance as a function of ρ for two values of θ: 0.5 to 2.0. Figure 3 shows the performance as a function of θ for $\rho = 1.05$.

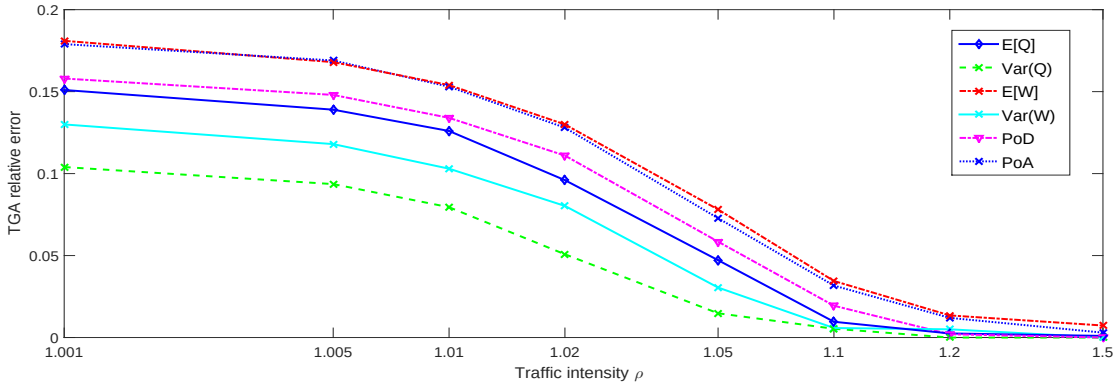

Figure 1: The relative error in the approximations of six performance measures as a function of the traffic intensity ρ for $1.001 \leq \rho \leq 1.500$ (with $\rho - 1$ in log scale) in the $M(1/100\rho)/M(1)/100 + M(2.0)$ model with abandonment rate $\theta = 0.5$.

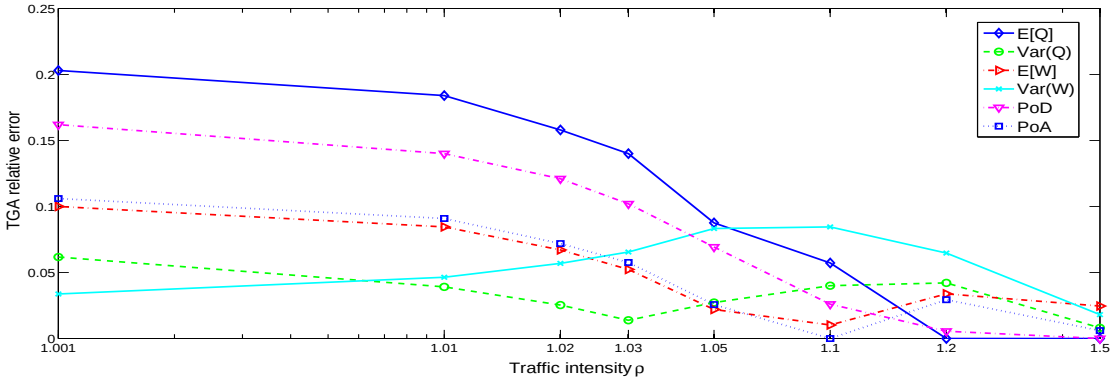

Figure 2: The relative error in the approximations of six performance measures as a function of the traffic intensity ρ for $1.001 \leq \rho \leq 1.500$ (with (with $\rho - 1$ in log scale) in the $M(1/100\rho)/M(1)/100 + M(0.5)$ model with abandonment rate $\theta = 2.0$.

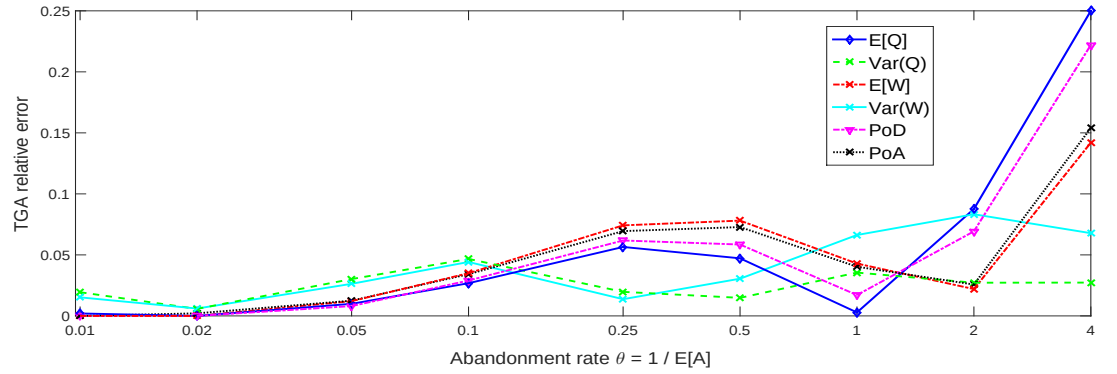

Figure 3: The relative error in the approximations of six performance measures as a function of the abandonment rate θ for $0.01 \leq \theta \leq 4.00$ (with θ in log scale) in the $M(1/105)/M(1)/100 + M(1/\theta)$ model having traffic intensity $\rho = 1.05$.

5 Evaluating the Approximations for $G/M/n + GI$ Models

We now start investigating the approximations for non-Markov models, first considering the $G/M/n + GI$ models for which the MSHT limits have been established in [26]. For these and all other non-Markov models, we will use simulation to estimate the exact values of the performance measures. We estimated all the performance measures using 2000 independent replications over the time interval $[0, 100]$, starting empty in each case. To ensure the system is nearly in steady state, we use the data after time $T' = 40$. To construct confidence intervals, on each sample path, we sample values for the performance measures (e.g., queue length and waiting times) at evenly-spaced time points in the interval $[40, 100]$. We describe the detailed simulation methodology in §9.

5.1 Non-Poisson arrivals

Theorems 2.1-2.3 show that the MSHT limits depend on the non-Poisson arrival process only through the asymptotic variance c_λ^2 in the FCLT assumed for the arrival process in (5). For a renewal arrival process the parameter c_λ^2 coincides with the scv of an interarrival time, but not more generally.

Renewal processes. To illustrate a non-Markovian arrival process, we consider a non-Poisson renewal process. We let the interarrival-time distribution be lognormal, denoted by $LN(\lambda^{-1}, c_\lambda^2)$, where λ^{-1} is its mean, which is the reciprocal of the fixed arrival rate $\lambda = n\rho = 105$, and c_λ^2 denotes its scv. Recall that an $LN(\lambda^{-1}, c_\lambda^2)$ random variable is distributed as $e^{\hat{\mu}+\hat{\sigma}\mathcal{Z}}$, where $\mathcal{Z}$ is a standard Gaussian random variable, $\hat{\mu} = \log\left(\lambda^{-1}/\sqrt{1+c_\lambda^2}\right)$ and $\hat{\sigma} = \sqrt{\log\left(1+c_\lambda^2\right)}$.

Table 3 and Figure 4 show the experimental results for five values of c_λ^2 with $0.25 \le c_\lambda^2 \le 4.0$. Table 3 also shows the experimental results for the case of a Poisson (M) arrival process, which is interesting, because the $LN(\lambda^{-1}, 1)$ distribution is different from the corresponding $M(\lambda^{-1})$ exponential distribution, even though they have the same mean and variance.

Table 3 compares the Gaussian approximations to simulations for the $LN/M/n + M$ queueing model with $(n, \rho, \theta) = (100, 1.05, 0.5)$. First, Table 3 shows that the interarrival time distribution has a significant impact upon performance. For example, $E[Q]$ increases from 11.5 to 16.1 as c_λ^2 increases from 0.25 to 4.0. Moreover, by comparing the results for $M(\lambda^{-1})$ and $LN(\lambda^{-1}, 1)$, where the approximations coincide, we see that the interarrival-time distribution matters little beyond its mean and variance, just as predicted.

Second, Table 3 shows that, just like Table 2, TGA performs consistently well, whereas DGA does not. Figure 5 adds to the story by showing that the full distributions of the queue length Q_n and potential waiting time W_n are well approximated by TGA as well.

Markov-Modulated Poisson processes (MMPP's) We next consider an alternative non-Poisson arrival process: Markov-modulated Poisson process (MMPP), which is a Poisson process having a random rate modulated by a continuous-time Markov chain (CTMC) $\{\Gamma(t), t \ge\}$, e.g., see [10]. Specifically, we can construct an MMPP by composition:

$$N_n(t) \equiv M\left(n\rho \int_0^t \alpha(\Gamma(u))\, du\right),$$

where M is a rate-1 Poisson process, and the random rate $\alpha(\Gamma(t)) = \alpha_i$ when the CTMC $\Gamma(t) = i$. In particular, we now consider an MMPP with an underlying CTMC $\{\Gamma(t), t \ge 0\}$ that is a

Table 3: A comparison of the TGA and DGA approximations to simulation estimates in the $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ model with i.i.d. lognormal interarrival times and $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ for six values of the interarrival time scv c_λ^2, $0.25 \leq c_\lambda^2 \leq 4.00$.

	$c_\lambda^2 = 0.25$			$c_\lambda^2 = 0.5$			$c_\lambda^2 = 1$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.11E+2	1.10E+2	same	1.11E+2	1.10E+2	same	1.11E+2	1.10E+2	same
rel. err.	±1.16E-1	1%		±1.25E-1	1%		±1.44E-1	1%	
Var(X)	1.16E+2	1.31E+2	same	1.38E+2	1.57E+2	same	1.80E+2	2.10E+2	same
rel. err.	±2.62E+1	13%		±2.83E+1	14%		±3.27E+1	17%	
E[Q]	1.15E+1	9.94E+0	1.12E+1	1.20E+1	9.94E+0	1.15E+1	1.27E+1	9.94E+0	1.21E+1
rel. err.	±1.04E-1	14%	3%	±1.10E-1	17%	5%	±1.24E-1	22%	5%
Var(Q)	9.22E+1	1.31E+2	9.23E+1	1.07E+2	1.57E+2	1.06E+2	1.33E+2	2.10E+2	1.33E+2
rel. err.	±3.18E+0	42%	0%	±3.62E+0	48%	0%	±4.46E+0	58%	0%
E[W]	1.18E-1	9.80E-2	1.10E-1	1.23E-1	9.80E-2	1.13E-1	1.28E-1	9.80E-2	1.18E-1
rel. err.	±1.05E-3	17%	7%	±1.10E-3	20%	8%	±1.22E-3	23%	7%
Var(W)	9.46E-3	1.29E-2	9.03E-3	1.07E-2	1.52E-2	1.03E-2	1.30E-2	2.00E-2	1.27E-2
rel. err.	±1.05E-3	36%	5%	±1.10E-3	42%	4%	±1.22E-3	54%	2%
PoD	8.22E-1	8.06E-1	same	8.08E-1	7.86E-1	same	7.80E-1	7.56E-1	same
rel. err.	±3.02E-3	2%		±3.15E-3	3%		±3.31E-3	3%	
PoA	5.60E-2	5.36E-2	same	5.83E-2	5.49E-2	same	6.01E-2	5.75E-2	same
rel. err.	±6.45E-4	4%		±6.70E-4	6%		±6.99E-4	4%	
	$M, c_\lambda^2 = 1$			$c_\lambda^2 = 2$			$c_\lambda^2 = 4$		
Perf.	Exact	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.11E+2	1.10E+2	same	1.12E+2	1.10E+2	same	1.13E+2	1.10E+2	same
rel. err.		1%		±1.69E-1	2%		±2.17E-1	3%	
Var(X)	1.81E+2	2.10E+2	same	2.60E+2	3.15E+2	same	4.08E+2	5.25E+2	same
rel. err.		16%		±3.88E+1	21%		±5.02E+1	29%	
E[Q]	1.27E+1	9.94E+0	1.21E+1	1.39E+1	9.94E+0	1.31E+1	1.61E+1	9.94E+0	1.50E+1
rel. err.		22%	5%	±1.40E-1	29%	6%	±1.73E-1	38%	7%
Var(Q)	1.35E+2	2.10E+2	1.33E+2	1.79E+2	3.15E+2	1.82E+2	2.61E+2	5.25E+2	2.75E+2
rel. err.		56%	1%	±5.81E+0	76%	2%	±8.52E+0	101%	6%
E[W]	1.28E-1	9.80E-2	1.18E-1	1.38E-1	9.80E-2	1.28E-1	1.56E-1	9.80E-2	1.45E-1
rel. err.		23%	7%	±1.36E-3	29%	7%	±1.63E-3	37%	7%
Var(W)	1.31E-2	2.00E-2	1.27E-2	1.68E-2	2.95E-2	1.72E-2	2.32E-2	4.86E-2	2.57E-2
rel. err.		53%	3%	±1.36E-3	76%	3%	±1.63E-3	110%	11%
PoD	8.03E-1	7.56E-1	same	7.51E-1	7.16E-1	same	7.26E-1	6.72E-1	same
rel. err.		6%		±3.46E-3	5%		±3.76E-3	8%	
PoA	6.04E-2	5.75E-2	same	6.46E-2	6.22E-2	same	7.29E-2	7.02E-2	same
rel. err.		5%		±7.69E-4	4%		±8.58E-4	4%	

birth-and-death process having three states 0, 1 and 2. Let CTMC-dependent arrival rate be $(\alpha_0, \alpha_1, \alpha_2) = (3, 1, 1/3)$. The long-run rate of the MMPP arrival process is

$$\lambda_n = n\rho\lambda^*, \quad \lambda^* \equiv \lim_{t\to\infty} t^{-1} \int_0^t \alpha(\Gamma(u))\, du = \sum_{j=0}^{2} \pi_j\, \alpha_j,$$

where $\pi \equiv (\pi_0, \pi_1, \pi_2)$ is the steady state distribution for the CTMC. We consider four sets of birth and deaths rates: (i) $\hat{\lambda}_0 = 20/81$, $\hat{\lambda}_1 = 5/27$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/81$, (ii) $\hat{\lambda}_0 = 20/27$, $\hat{\lambda}_1 = 5/9$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/27$, (iii) $\hat{\lambda}_0 = 20/9$, $\hat{\lambda}_1 = 5/3$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/9$, and (iv) $\hat{\lambda}_0 = 40/9$, $\hat{\lambda}_1 = 10/3$, $\hat{\mu}_1 = \hat{\mu}_2 = 20/9$, which yield the same steady state $\pi = (1/6, 1/3, 1/2)$ and asymptotic rate $\lambda^* = 1$, but different asymptotic variability parameter: (i) $c_\lambda^2 = 10$, (ii) $c_\lambda^2 = 4$, (iii) $c_\lambda^2 = 2$ and (iv) $c_\lambda^2 = 1.5$,

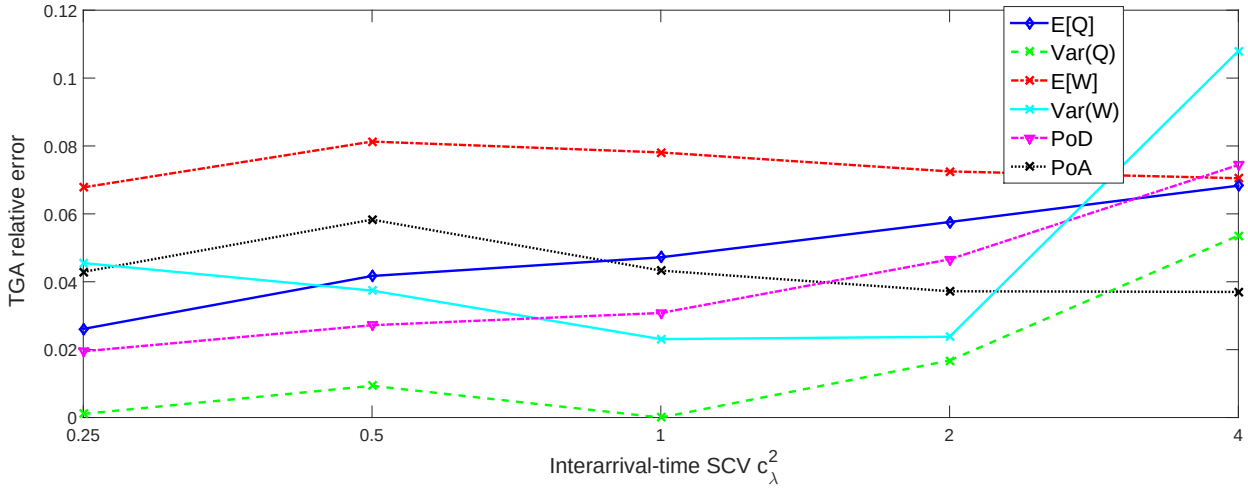

Figure 4: The relative error in the approximations of six performance measures as a function of the interarrival-time scv c_λ^2 for $0.25 \le c_\lambda^2 \le 4.0$ (with c_λ^2 in log scale) in the $LN(1/105, c_\lambda^2)/M(1)/100 + M(2)$ model.

where

$$c_\lambda^2 = 1 + c_C^2, \quad \text{where} \quad c_C^2 \equiv 2 \sum_{j=1}^{2} \frac{1}{\hat{\lambda}_j \pi_j} \left(\sum_{i=1}^{j} \pi_i \left(\alpha_i - \lambda_C \right) \right)^2 .$$

See Proposition 1 in [36] and also see [15].

We denote by $MMPP(\lambda^{-1}, c_\lambda^2)$ our MMPP arrival process having rate λ and variability parameter c_λ^2. Table 4 compares TGA to the simulation results for $MMPP/M/n + M$ models with $\lambda = 105$, $n = 100$, $\mu = 1$, $\theta = 0.5$, and different variability parameters $c_\lambda^2 = 1.5, 2, 4$ and 10.

5.2 Non-exponential patience

It has been shown that the full abandonment distribution has a significant impact on the performance; e.g., see [41, 42]. We confirm that here when we study how our TGAs works in $M/M/n+GI$ models using different abandonment distributions, again using the lognormal distribution. Figure 6 and Tables 5 and 6 compare the DGAs and TGAs to the simulations of the $M/M/n + LN(2, c_{ab}^2)$ model with the same parameter triple $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ we have been using, where scv of abandonment distribution c_{ab}^2 ranges from 0.25 to 4.0. Paralleling Table 3, we add a column for the results of Erlang-A models.

Figure 6 and Tables 5 and 6 show that TGA is again consistently quite accurate for all the first-moment measures (mean and probability) in all cases, but the accuracy degrades to $20 - 30\%$ for the variances when c_{ab}^2 is low. Moreover, by comparing the results for the $M/M/n + LN(2, 1)$ and $M/M/n+M(2)$ models in Table 6, we can see that the patience distribution has a significant impact on the queueing system beyond its mean and variance, unlike the interarrival-time distribution. Approximations of distributions of Q_n and W_n are given in Figure 7.

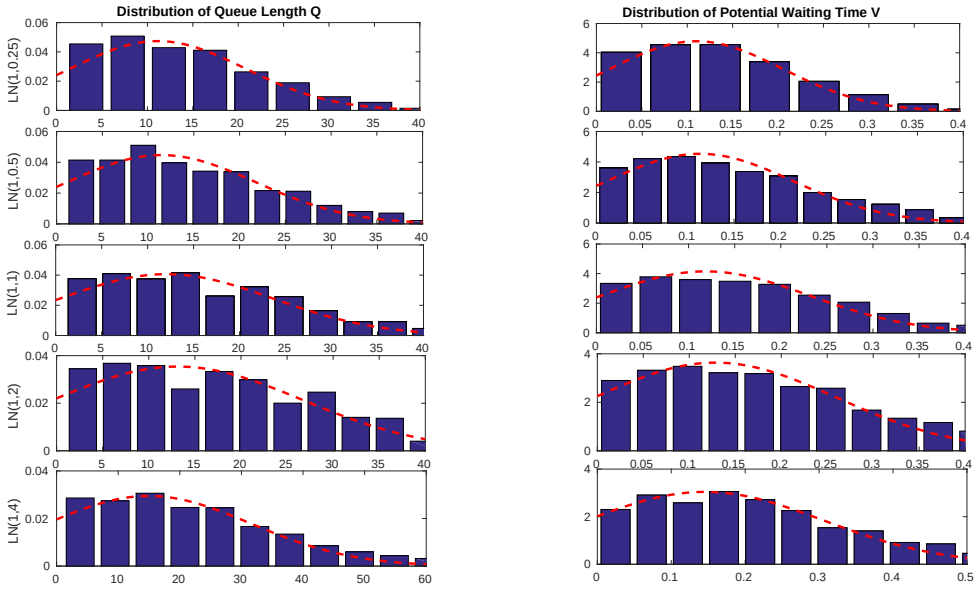

Figure 5: Simulation estimates (histograms) of the TGA approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) for the $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

Table 4: A comparison of the TGA approximations to simulation estimates in the $MMPP(\lambda^{-1}, c_\lambda^2)/M/100 + M(\theta^{-1})$ model with $(\lambda, \rho, \theta) = (105, 1.05, 0.5)$ for four values of the arrival process variability parameter c_λ^2 in (5), $1.5 \le c_\lambda^2 \le 10.0$.

	$c_\lambda^2 = 1.5$		$c_\lambda^2 = 2$		$c_\lambda^2 = 4$		$c_\lambda^2 = 10$	
Perf.	Sim	TGA	Sim	TGA	Sim	TGA	Sim	TGA
E[X]	1.12E+2	1.10E+2	1.12E+2	1.10E+2	1.13E+2	1.09E+2	1.15E+2	1.10E+2
rel. err.	±1.62E-1	2%	±1.77E-1	2%	±2.16E-1	3%	±3.20E-1	5%
Var(X)	2.22E+2	2.62E+2	2.65E+2	3.15E+2	4.30E+2	5.25E+2	9.19E+2	1.15E+3
rel. err.	±3.71E+1	18%	±4.08E+1	19%	±5.15E+1	22%	±8.05E+1	26%
E[Q]	1.34E+1	1.26E+1	1.38E+1	1.31E+1	1.60E+1	1.50E+1	2.06E+1	1.91E+1
rel. err.	±1.37E-1	6%	±1.48E-1	5%	±1.78E-1	7%	±2.55E-1	7%
Var(Q)	1.62E+2	1.58E+2	1.88E+2	1.82E+2	2.93E+2	2.75E+2	6.02E+2	5.35E+2
rel. err.	±5.46E+0	2%	±6.21E+0	3%	±9.55E+0	6%	±1.95E+1	11%
E[W]	1.34E-1	1.24E-1	1.37E-1	1.28E-1	1.55E-1	1.45E-1	1.91E-1	1.85E-1
rel. err.	±1.33E-3	8%	±1.43E-3	6%	±1.67E-3	6%	±2.26E-3	3%
Var(W)	1.53E-2	1.50E-2	1.74E-2	1.72E-2	2.56E-2	2.57E-2	4.68E-2	4.94E-2
rel. err.	±5.16E-4	2%	±5.76E-4	1%	±8.21E-4	0%	±1.50E-3	6%
PoD	7.83E-1	7.33E-1	7.61E-1	7.16E-1	7.22E-1	6.72E-1	6.61E-1	6.18E-1
rel. err.	±3.33E-3	6%	±3.50E-3	6%	±3.55E-3	7%	±4.00E-3	7%
PoA	6.32E-2	5.82E-2	6.44E-2	6.02E-2	7.15E-2	6.72E-2	8.61E-2	8.28E-2
rel. err.	±7.49E-4	8%	±7.86E-4	7%	±8.71E-4	6%	±1.10E-3	4%

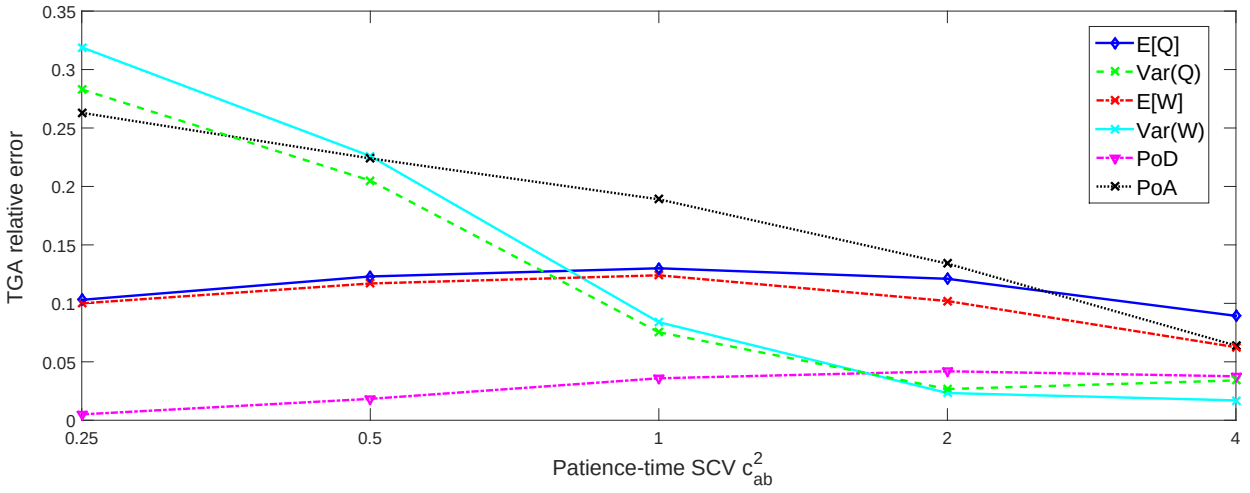

Figure 6: The relative errors in the approximations of six performance measures as a function of the patience-time scv c_{ab}^2 for $0.25 \leq c_{ab}^2 \leq 4.0$ (with c_{ab}^2 in log scale) in the $M(1/105)/M(1)/100 + LN(2, c_{ab}^2)$ model.

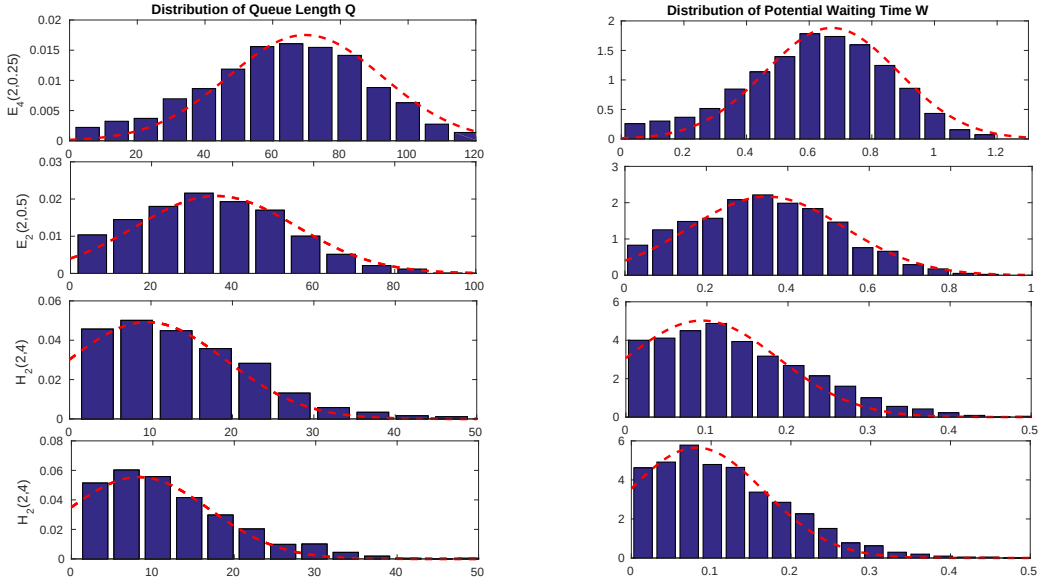

Figure 7: Simulation estimates (histograms) of the TGA approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) in the $M/M/100 + LN(\theta^{-1}, c_{ab}^2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

Table 5: A comparison of the TGA and DGA approximations to simulation estimates in the $M/M/100 + LN(\theta^{-1}, c_{ab}^2)$ model with with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ for four values of the patience scv c_{ab}^2, $0.25 \leq c_{ab}^2 \leq 4.00$.

	$LN(2, 0.25)$			$LN(2, 0.5)$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.77E+2	1.85E+2	same	1.52E+2	1.59E+2	same
rel. err.	±4.92E-1	4%		±3.94E-1	4%	
Var(X)	6.39E+2	4.54E+2	same	5.22E+2	4.04E+2	same
rel. err.	±1.63E+2	29%		±1.15E+2	23%	
E[Q]	7.67E+1	8.46E+1	8.46E+1	5.21E+1	5.85E+1	5.85E+1
rel. err.	±4.89E-1	10%	10%	±3.88E-1	12%	12%
Var(Q)	6.33E+2	4.54E+2	4.54E+2	5.07E+2	4.04E+2	4.03E+2
rel. err.	±6.55E+1	28%	28%	±3.72E+1	20%	21%
E[W]	7.39E-1	8.13E-1	8.13E-1	5.05E-1	5.64E-1	5.64E-1
rel. err.	±4.54E-3	10%	10%	±3.64E-3	12%	12%
Var(W)	5.42E-2	3.69E-2	3.69E-2	4.46E-2	3.45E-2	3.45E-2
rel. err.	±5.80E-3	32%	32%	±3.30E-3	22%	23%
PoD	9.95E-1	1.00E+0	same	9.81E-1	9.99E-1	same
rel. err.	±8.22E-4	1%		±1.47E-3	2%	
PoA	4.83E-2	6.10E-2	same	4.90E-2	6.00E-2	same
rel. err.	±8.35E-4	26%		±8.02E-4	22%	

	$LN(2, 2)$			$LN(2, 4)$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.18E+2	1.21E+2	same	1.10E+2	1.11E+2	same
rel. err.	±1.75E-1	2%		±1.13E-1	1%	
Var(X)	2.25E+2	2.23E+2	same	1.43E+2	1.42E+2	same
rel. err.	±4.14E+1	1%		±2.49E+1	0%	
E[Q]	1.90E+1	2.07E+1	2.13E+1	1.12E+1	1.10E+1	1.22E+1
rel. err.	±1.58E-1	9%	12%	±9.19E-2	1%	9%
Var(Q)	1.87E+2	2.23E+2	1.92E+2	9.96E+1	1.42E+2	1.03E+2
rel. err.	±6.78E+0	19%	3%	±2.74E+0	43%	3%
E[W]	1.87E-1	2.01E-1	2.06E-1	1.12E-1	1.08E-1	1.19E-1
rel. err.	±1.52E-3	7%	10%	±8.96E-4	3%	6%
Var(W)	1.72E-2	2.02E-2	1.76E-2	9.43E-3	1.31E-2	9.59E-3
rel. err.	±6.30E-4	17%	2%	±2.62E-4	39%	2%
PoD	8.84E-1	9.21E-1	same	7.97E-1	8.27E-1	same
rel. err.	±2.85E-3	4%		±3.07E-3	4%	
PoA	5.45E-2	6.18E-2	same	6.10E-2	6.49E-2	same
rel. err.	±7.34E-4	13%		±7.19E-4	6%	

Table 6: The impact of the abandonment distribution beyond its mean and variance: a performance comparison between $M(\lambda^{-1})/M/n + M(2)$ and $M(\lambda^{-1})/M/n + LN(2,1)$ models, where $(\lambda, \rho, n) = (105, 1.05, 100)$.

	$M(2)$			$LN(2,1)$		
Perf.	Exact	DGA	TGA	Sim	DGA	TGA
E[X]	1.05E+2	1.05E+2	same	1.32E+2	1.36E+2	same
rel. err.		0%		±2.79E-1	3%	
Var(X)	1.05E+2	1.05E+2	same	3.58E+2	3.18E+2	same
rel. err.		0%		±7.27E+1	11%	
E[Q]	7.03E+0	4.92E+0	7.01E+0	3.24E+1	3.65E+1	3.66E+1
rel. err.		30%	0%	±2.67E-1	13%	13%
Var(Q)	5.92E+1	1.05E+2	5.71E+1	3.32E+2	3.18E+2	3.07E+2
rel. err.		77%	3%	±1.74E+1	4%	8%
E[W]	7.22E-2	4.90E-2	6.91E-2	3.15E-1	3.53E-1	3.54E-1
rel. err.		32%	4%	±2.54E-3	12%	12%
Var(W)	5.88E-3	1.00E-2	5.49E-3	2.98E-2	2.82E-2	2.73E-2
rel. err.		70%	7%	±2.54E-3	6%	9%
PoD	7.00E-1	6.88E-1	same	9.48E-1	9.82E-1	same
rel. err.		2%		±2.19E-3	4%	
PoA	6.70E-2	6.43E-2	same	5.08E-2	6.04E-2	same
rel. err.		3%		±7.74E-4	19%	

6 Non-Exponential Service

6.1 Refined Gaussian Approximations for the $M/GI/n+M$ Model

We now evaluate the heuristic approximation TGA-G developed in §3.3. We let the service-time distribution be phase-type, denoted by PH with fixed mean $1/\mu = 1$ and scv c_s^2 ranging in $[0.25, 4]$. To be specific, for cases with $c_s^2 = 0.25, 0.5 < 1$, we used Erlang 4 (E_4) and Erlang 2 (E_2) distribution and for cases with $c_s^2 = 2, 4 > 1$, we used the two-phase hyperexponential distribution (H_2) with balanced means, see [34] for more details.

Table 7 compares the approximations for all three Gaussian approximations for the $M/PH/n+M$ model. Table 7 shows that TGA-G outperforms TGA, while both are far better than DGA. The new approximation TGA-G is especially important for the variances, which depend quite strongly on the service-time distribution, unlike the means. Figure 8 shows that the TGA-G approximation accuracy tends to be independent of the service-time distribution scv c_s^2, consistent with the observations in [41, 42]. Figure 9 shows that corresponding TGA-G approximations of the distributions remain good.

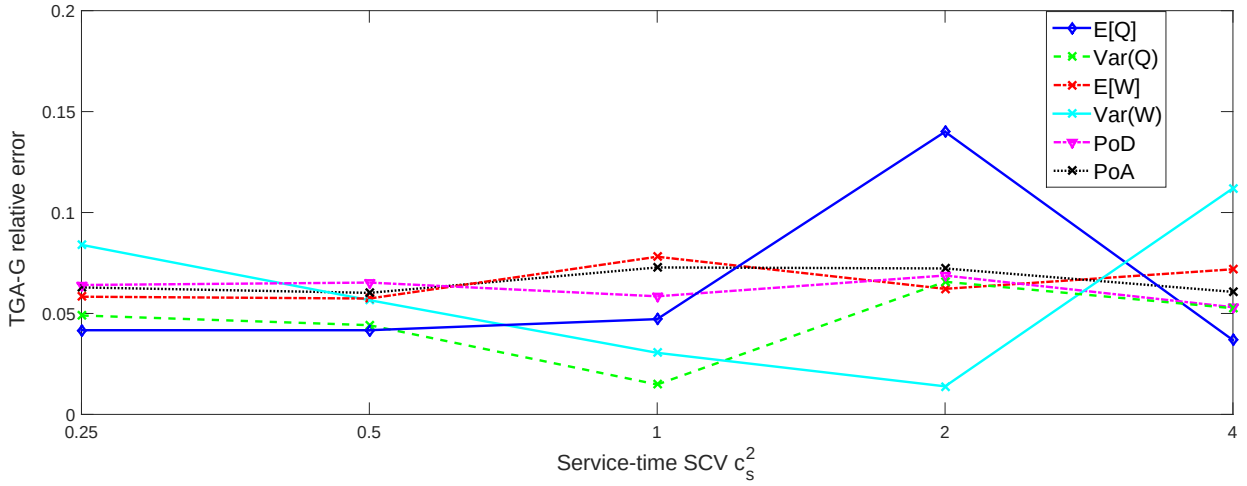

Figure 8: The relative error in the TGA-G approximations of six performance measures as a function of the service-time scv c_s^2 for $0.25 \le c_s^2 \le 4.0$ (with c_s^2 in log scale) in the $M(1/105)/PH(1, c_s^2)/100+M(2)$ model

In Table 8 we consider the $M/H_2/n+M$ model, with hyperexponential service according to $H_2(1/\mu, c_s^2)$, which denotes mean $1/\mu = 1$ and SCV $c_s^2 = 2$, for a range of abandonment rates with $0.25 \le \theta \le 2.0$. We observe that the performance of TGA-G is acceptable. However, just as for the $M/M/n+M$ model, the approximation accuracy degrades when θ increases; see the appendix for more examples.

6.2 The General $GI/GI/n+GI$ Model

We have also considered examples in which all three stochastic components of the $G/GI/n+GI$ model are non-exponential. We illustrate some of these now. Table 9 shows comparisons of TGA and TGA-G to simulation estimates for the $H_2/PH/n+H_2$ model, having a renewal arrival process with H_2 inter-arrival times, PH service times (in the settings of Table 7), $n = 100$ servers, and H_2 patience times. We fix the scv's for arrival and patience times at $c_\lambda^2 = c_{ab}^2 = 2$ and consider a range of service scv: $0.25 \le c_s^2 \le 4.0$.

Table 7: A comparison of the TGA-G, TGA and DGA approximations to simulation estimates in the $M(\lambda^{-1})/PH(1,c_s^2)/100 + M(\theta^{-1})$ model with $(\lambda,\rho,\theta) = (100, 1.05, 0.5)$ for four different phase-type (Ph) service distributions characterized by their scv c_s^2, $0.25 \le c_s^2 \le 4.00$.

	$c_s^2 = 0.25$				$c_s^2 = 0.5$			
Perf. Meas.	Sim.	DGA	TGA	TGA-G	Sim.	DGA	TGA	TGA-G
E[X]	1.11E+2	1.10E+2	same	same	1.11E+2	1.10E+2	same	same
rel. err.	±1.21E-1	1%			±1.30E-1	1%		
Var(X)	1.34E+2	1.60E+2	same	same	1.50E+2	1.81E+2	same	same
rel. err.	±2.72E+1	19%			±2.94E+1	20%		
E[Q]	1.20E+1	9.94E+0	1.21E+1	1.15E+1	1.22E+1	9.94E+0	1.21E+1	1.17E+1
rel. err.	±1.06E-1	17%	1%	4%	±1.13E-1	19%	1%	4%
Var(Q)	1.02E+2	1.60E+2	1.33E+2	1.07E+2	1.13E+2	1.81E+2	1.33E+2	1.18E+2
rel. err.	±3.45E+0	57%	30%	6%	±3.83E+0	60%	18%	5%
E[W]	1.20E-1	9.80E-2	1.18E-1	1.13E-1	1.22E-1	9.80E-2	1.18E-1	1.15E-1
rel. err.	±1.01E-3	18%	1%	6%	±1.10E-3	20%	3%	6%
Var(W)	9.41E-3	1.50E-2	1.27E-2	1.02E-2	1.06E-2	1.71E-2	1.27E-2	1.12E-2
rel. err.	±3.20E-4	59%	35%	8%	±3.64E-4	60%	19%	6%
PoD	8.42E-1	7.88E-1	same	same	8.27E-1	7.73E-1	same	same
rel. err.	±2.67E-3	6%			±2.81E-3	6%		
PoA	5.72E-2	5.36E-2	same	same	5.81E-2	5.46E-2	same	same
rel. err.	±7.60E-4	6%			±7.81E-4	6%		

	$c_s^2 = 2$				$c_s^2 = 4$			
Perf. Meas.	Sim.	DGA	TGA	TGA-G	Sim.	DGA	TGA	TGA-G
E[X]	1.01E+2	1.02E+2	same	same	1.12E+2	1.10E+2	same	same
rel. err.	±6.90E-2	1%			±2.17E-1	2%		
Var(X)	7.32E+1	6.28E+1	same	same	2.57E+2	3.10E+2	same	same
rel. err.	±1.38E+1	14%			±5.03E+1	20%		
E[Q]	3.93E+0	2.36E+0	4.22E+0	4.48E+0	1.36E+1	9.94E+0	1.21E+1	1.31E+1
rel. err.	±3.68E-2	40%	7%	14%	±1.84E-1	27%	11%	4%
Var(Q)	2.74E+1	6.28E+1	2.50E+1	2.92E+1	1.90E+2	3.10E+2	1.33E+2	1.80E+2
rel. err.	±5.49E-1	129%	9%	7%	±7.74E+0	63%	30%	5%
E[W]	4.18E-2	2.40E-2	4.18E-2	4.44E-2	1.39E-1	9.80E-2	1.18E-1	1.29E-1
rel. err.	±3.85E-4	43%	0%	6%	±1.90E-3	29%	15%	7%
Var(W)	2.88E-3	6.03E-3	2.42E-3	2.84E-3	1.97E-2	3.00E-2	1.27E-2	1.75E-2
rel. err.	±5.93E-5	109%	16%	1%	±8.32E-4	53%	35%	11%
PoD	5.81E-1	6.21E-1	same	same	7.54E-1	7.14E-1	same	same
rel. err.	±3.36E-3	7%			±4.50E-3	5%		
PoA	7.47E-2	8.01E-2	same	same	6.43E-2	6.04E-2	same	same
rel. err.	±9.21E-4	7%			±1.06E-3	6%		

Just as in Table 7, Table 9 shows that the mean values such as $E[Q]$ and the probabilities such as PoD are relatively insensitive to the service-time distribution beyond its mean; these entries differ little in the three cases. However, as before, we see differences in the variances. Table 9 shows that both TGA and TGA-G are effective for the mean values such as $E[Q]$ and the probabilities such as PoD, but TGA-G provides significant improvement for the variances. Additional results on other combinations of arrival processes, patience times and service times are given in Table 20 in the appendix.

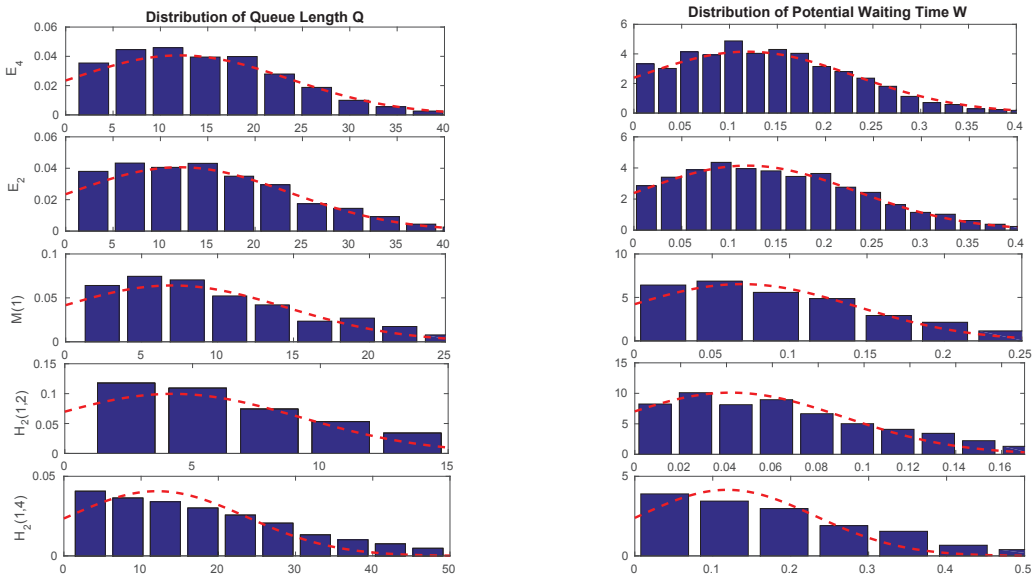

Figure 9: Simulation estimates (histograms) of the TGA-G approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) in the $M(105^{-1})/PH/100 + M(2)$ model for five service distributions with $0.25 \leq c_s^2 \leq 4.0$.

Table 8: A comparison of the TGA-G approximations to simulation estimates in the $M(\lambda^{-1})/H_2(1,2)/100 + M(\theta^{-1})$ model with $\lambda = 100$, $\rho = 1.05$ and five different abandonment rates θ, $0.1 \leq \theta \leq 2.0$.

	$\theta = 0.1$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
Perf.	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G
E[X]	1.53E+2	1.50E+2	1.23E+2	1.20E+2	1.12E+2	1.10E+2	1.05E+2	1.05E+2	1.01E+2	1.02E+2
rel. err.	±8.30E-1	2%	±3.20E-1	2%	±1.75E-1	2%	±1.07E-1	0%	±6.90E-2	1%
Var(X)	1.24E+3	1.26E+3	4.31E+2	5.03E+2	2.15E+2	2.51E+2	1.18E+2	1.26E+2	7.32E+1	6.28E+1
rel. err.	±2.69E+2	2%	±8.23E+1	17%	±4.00E+1	17%	±2.27E+1	6%	±1.38E+1	14%
E[Q]	5.38E+1	5.12E+1	2.37E+1	2.22E+1	1.32E+1	1.25E+1	7.19E+0	7.35E+0	3.93E+0	4.48E+0
rel. err.	±8.16E-1	5%	±3.00E-1	6%	±1.49E-1	5%	±7.62E-2	2%	±3.68E-2	14%
Var(Q)	1.19E+3	1.09E+3	3.76E+2	3.57E+2	1.59E+2	1.53E+2	6.67E+1	6.62E+1	2.74E+1	2.92E+1
rel. err.	±1.05E+2	9%	±1.91E+1	5%	±5.78E+0	4%	±1.84E+0	1%	±5.49E-1	7%
E[W]	5.32E-1	5.01E-1	2.37E-1	2.18E-1	1.34E-1	1.23E-1	7.42E-2	7.26E-2	4.18E-2	4.44E-2
rel. err.	±8.06E-3	6%	±3.00E-3	8%	±1.51E-3	8%	±7.83E-4	2%	±3.85E-4	6%
Var(W)	1.16E-1	1.05E-1	3.71E-2	3.43E-2	1.60E-2	1.47E-2	6.81E-3	6.40E-3	2.88E-3	2.84E-3
rel. err.	±1.03E-2	10%	±1.90E-3	8%	±5.93E-4	8%	±1.94E-4	6%	±5.93E-5	1%
PoD	9.48E-1	9.20E-1	8.64E-1	8.13E-1	7.82E-1	7.36E-1	6.80E-1	6.72E-1	5.81E-1	6.21E-1
rel. err.	±2.84E-3	3%	±3.47E-3	6%	±3.72E-3	6%	±3.74E-3	1%	±3.36E-3	7%
PoA	5.12E-2	4.83E-2	5.60E-2	5.19E-2	6.29E-2	5.79E-2	6.87E-2	6.72E-2	7.47E-2	8.01E-2
rel. err.	±9.66E-4	6%	±9.03E-4	7%	±9.25E-4	8%	±9.48E-4	2%	±9.21E-4	7%

Table 9: A comparison of the TGA-G and TGA approximations to simulation estimates in the $H_2(\lambda^{-1}, 2)/PH/100 + H_2(1/\theta, 2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ and four different phase-type service distributions characterized by the scv c_s^2, $0.25 \leq c_s^2 \leq 4.00$

	$c_s^2 = 0.25$			$c_s^2 = 0.5$		
Perf.	Sim	TGA	TGA-G	Sim	TGA	TGA-G
E[X]	1.09E+2	1.07E+2	1.07E+2	1.09E+2	1.07E+2	1.07E+2
rel. err.	±1.21E-1	1%	1%	±1.31E-1	1%	1%
Var(X)	1.84E+2	2.37E+2	2.00E+2	1.95E+2	2.37E+2	2.15E+2
rel. err.	±2.64E+1	29%	8%	±2.89E+1	22%	11%
E[Q]	1.07E+1	1.05E+1	1.01E+1	1.09E+1	1.05E+1	1.03E+1
rel. err.	±9.32E-2	2%	6%	±1.02E-1	3%	5%
Var(Q)	1.12E+2	1.29E+2	1.12E+2	1.21E+2	1.29E+2	1.19E+2
rel. err.	±3.01E+0	15%	0%	±3.44E+0	7%	1%
E[W]	1.06E-1	1.03E-1	9.84E-2	1.07E-1	1.03E-1	1.00E-1
rel. err.	±8.77E-4	2%	7%	±9.71E-4	4%	7%
Var(W)	1.00E-2	1.22E-2	1.05E-2	1.10E-2	1.22E-2	1.12E-2
rel. err.	±2.65E-4	22%	5%	±3.09E-4	12%	3%
PoD	7.52E-1	6.88E-1	7.04E-1	7.42E-1	6.88E-1	6.97E-1
rel. err.	±3.05E-3	9%	6%	±3.18E-3	7%	6%
PoA	6.54E-2	6.34E-2	6.09E-2	6.66E-2	6.34E-2	6.19E-2
rel. err.	±8.09E-4	3%	7%	±8.67E-4	5%	7%

	$c_s^2 = 2$			$c_s^2 = 4$		
Perf.	Sim	TGA	TGA-G	Sim	TGA	TGA-G
E[X]	1.09E+2	1.07E+2	1.07E+2	1.09E+2	1.07E+2	1.07E+2
rel. err.	±1.64E-1	1%	1%	±1.95E-1	1%	1%
Var(X)	2.30E+2	2.37E+2	2.69E+2	2.50E+2	2.37E+2	3.13E+2
rel. err.	±3.66E+1	3%	17%	±4.36E+1	5%	25%
E[Q]	1.13E+1	1.05E+1	1.09E+1	1.13E+1	1.05E+1	1.14E+1
rel. err.	±1.29E-1	7%	3%	±1.52E-1	7%	0%
Var(Q)	1.47E+2	1.29E+2	1.43E+2	1.61E+2	1.29E+2	1.62E+2
rel. err.	±4.69E+0	12%	3%	±5.75E+0	20%	0%
E[W]	1.13E-1	1.03E-1	1.07E-1	1.14E-1	1.03E-1	1.11E-1
rel. err.	±1.28E-3	9%	5%	±1.54E-3	10%	2%
Var(W)	1.42E-2	1.22E-2	1.36E-2	1.60E-2	1.22E-2	1.55E-2
rel. err.	±4.58E-4	14%	4%	±5.93E-4	23%	3%
PoD	7.11E-1	6.88E-1	6.77E-1	6.94E-1	6.88E-1	6.64E-1
rel. err.	±3.88E-3	3%	5%	±4.51E-3	1%	4%
PoA	6.80E-2	6.34E-2	6.53E-2	6.92E-2	6.34E-2	6.78E-2
rel. err.	±9.75E-4	7%	4%	±1.10E-3	8%	2%

7 Smaller Scale: Lower Arrival Rates and Fewer Servers

Since the MSHT limits involve a sequence of queueing systems with increasing scale, the MSHT approximations DGA and TGA-G should perform better as the scale increases. Thus, we considered the base case with $n = 100$, because it is large but also small enough to be of practical value. However, we also want to apply the approximations to even smaller scale systems. Thus, in this section, we examine the effectiveness of DGA, TGA and TGA-G for smaller systems.

In order to set the parameters for these smaller systems, it is good to exploit the MSHT limits. When the system is in the QED regime, we know that the scaling factor n (number of servers) and the traffic intensity ρ_n should roughly satisfies the relation

$$\sqrt{n}\,(1-\rho_n) \approx \beta, \quad -\infty < \beta < \infty, \tag{25}$$

where the β is the QoS factor. Since $\beta = \sqrt{n}(1-\rho_n) = -0.5$ when $\rho_n = 1.05$ and $n = 100$ as in previous tables, we now fix β at -0.5 as we change n. Note that this increases the traffic intensity as ρ decreases.

Table 10 shows the results for an $H_2/H_2/n + H_2$ model for three values of n: $20, 10, 5$, which correspond to traffic intensities $\rho_n = 1.11, 1.16, 1.22$. Table 10 shows that TGA-G remains effective

Table 10: Smaller scale: A comparison of the TGA-G and DGA approximations to simulation estimates in the $H_2(\lambda^{-1})/H_2(1,2)/n + H_2(\theta^{-1},2)$ model with three pairs (n,ρ) with $\rho = 1-\beta/\sqrt{n}$, $\lambda = n\rho$, $\beta = -0.5$ and $n = 20$, 10 and 5

	$n=20$			$n=10$			$n=5$		
Perf.	Sim	DGA	TGA-G	Sim	DGA	TGA-G	Sim	DGA	TGA-G
E[X]	2.40E+1	2.34E+1	2.34E+1	1.29E+1	1.24E+1	1.24E+1	7.08E+0	6.69E+0	6.69E+0
rel. err.	±7.46E-2	3%	3%	±5.38E-2	4%	4%	±3.96E-2	5%	5%
Var(X)	4.86E+1	5.06E+1	5.69E+1	2.52E+1	2.65E+1	2.97E+1	1.32E+1	1.41E+1	1.57E+1
rel. err.	±3.84E+0	4%	17%	±1.55E+0	5%	18%	±6.61E-1	6%	19%
E[Q]	5.16E+0	4.82E+0	4.98E+0	3.68E+0	3.46E+0	3.57E+0	2.65E+0	2.49E+0	2.57E+0
rel. err.	±5.93E-2	7%	4%	±4.31E-2	6%	3%	±3.21E-2	6%	3%
Var(Q)	3.19E+1	2.74E+1	3.01E+1	1.68E+1	1.42E+1	1.56E+1	9.04E+0	7.50E+0	8.19E+0
rel. err.	±1.03E+0	14%	6%	±5.46E-1	15%	7%	±3.01E-1	17%	9%
E[W]	2.61E-1	2.28E-1	2.36E-1	3.77E-1	3.18E-1	3.30E-1	5.60E-1	4.44E-1	4.61E-1
rel. err.	±2.96E-3	13%	10%	±4.38E-3	16%	13%	±6.90E-3	21%	18%
Var(W)	7.54E-2	6.01E-2	6.70E-2	1.61E-1	1.19E-1	1.33E-1	3.67E-1	2.34E-1	2.62E-1
rel. err.	±2.55E-3	20%	11%	±5.72E-3	26%	18%	±1.45E-2	36%	29%
PoD	7.22E-1	6.85E-1	6.74E-1	7.26E-1	6.82E-1	6.71E-1	7.34E-1	6.79E-1	6.68E-1
rel. err.	±3.74E-3	5%	7%	±3.75E-3	6%	8%	±3.74E-3	7%	9%
PoA	1.43E-1	1.28E-1	1.31E-1	1.92E-1	1.67E-1	1.71E-1	2.54E-1	2.15E-1	2.19E-1
rel. err.	1.72E-3	11%	9%	2.15E-3	13%	11%	±2.70E-3	16%	14%

for systems with fewer servers. Figure 10 shows that accuracy consistently degrades as the scale decreases, but remains reasonable. Figure 11 shows that TGA-G can be used to approximate the distributions of key performance measures. Figure 11 shows that the simulated histograms for the $H_2/M/n + H_2$ model are well approximated by the pdfs of the corresponding Gaussian approximations, for $n = 100, 50, 20$ and 5. Additional simulation results appear in the appendix.

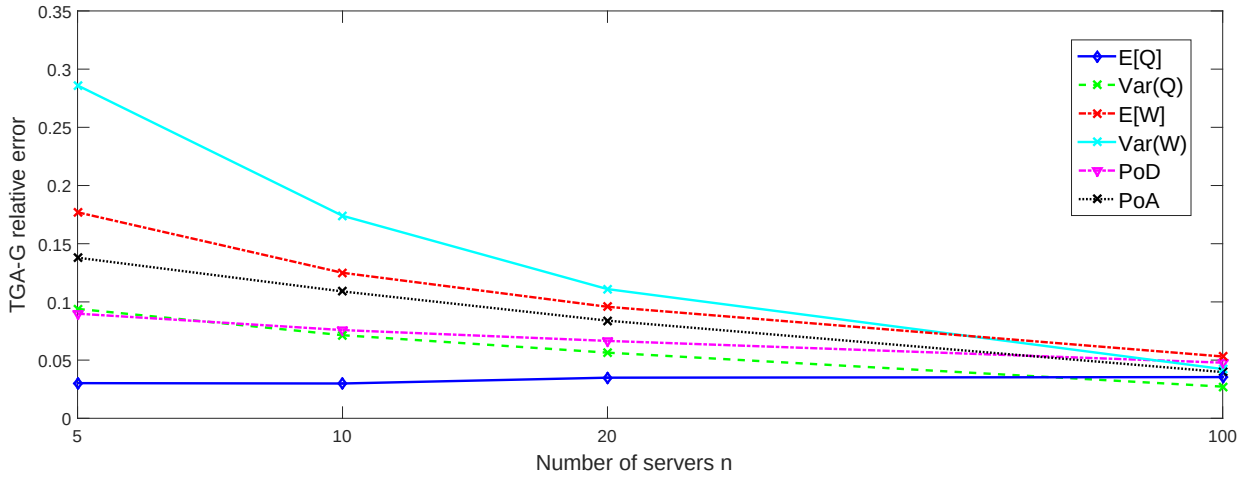

Figure 10: Simulation estimates of the relative errors in the approximations for six performance measures as a function of the number of servers, n, for $n = 5, 10, 20, 100$ (with n in log scale) in the $H_2(1/n\rho, 2)/H_2(1, 2)/n + H_2(2, 2)$ model.

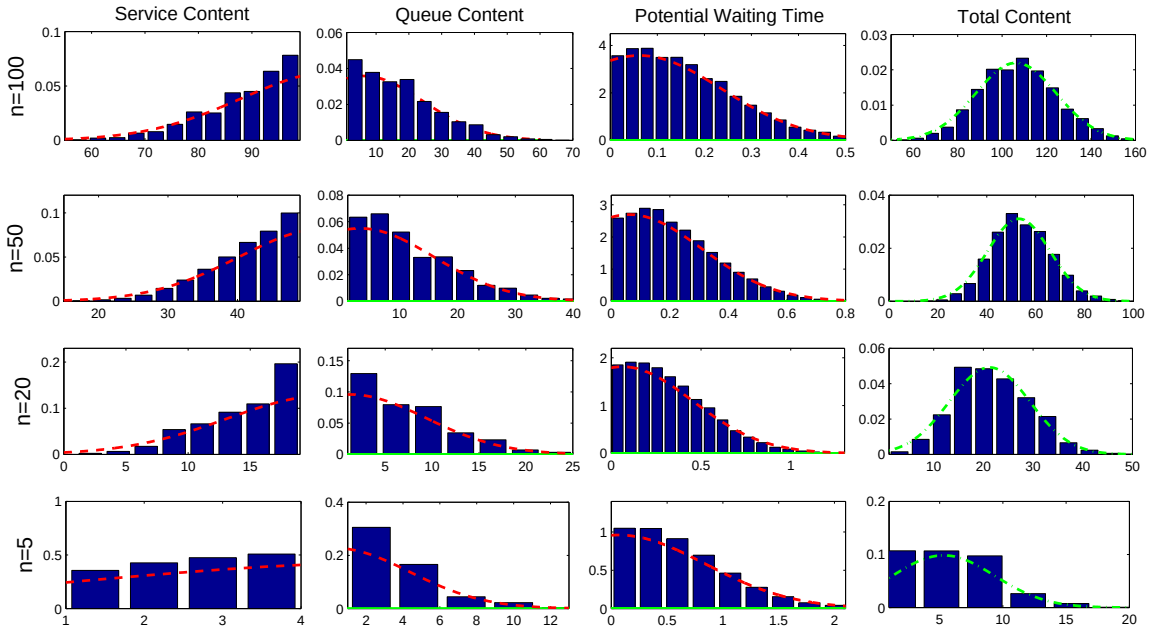

Figure 11: Simulation estimates (histograms) of the TGA-G approximating distributions for B_n, Q_n, V_n and X_n in the $H_2(\lambda^{-1}, 2)/M/n + H_2(\theta^{-1}, 2)$ model with $\beta = -0.5$ and four values of n, $5 \le n \le 100$.

8 Limitations of the Proposed Approximations

We first compare our TGA-G approximation to previous approximations in [41] and [33]. Then we show that the performance is not good for UL models.

8.1 Comparison with Approximations in [41]

A numerical approximation algorithm for the $M/GI/n+GI$ model was developed and evaluated in [41]. It was based on an application of an exact analysis of an associated state-dependent Markov $M/M/n+M(n)$ queue, after approximating the GI abandonment by an appropriate state-dependent $M(n)$ abandonment mechanism. (The GI service was simply approximated by M.) That numerical procedure has the advantage that it applies to all loadings (underloaded, critically loaded and overloaded), but it is much more computationally intensive. That approximation was shown to be quite effective. A shortcoming of [41] that we address here is that it does not describe the impact of non-M arrival processes and service times.

We now compare our new TGA-G approximation to the approximation developed in [41] by comparing to the displayed results in Tables 6 and 7 of [41], which are also for $n = 100$, but for the relatively light loading $\rho = 1.02$, which is at the edge of the range of effectiveness for TGA-G.

As expected for this relatively light loading, Tables 11 and 12 show that the engineering approximations in [41], labeled as *Eng. Approx. (W05)*, are more accurate overall. However, TGA-G is better for D service.

First, Table 11 shows that TGA-G performs reasonably well for the Erlang E_2 patience distribution, except for the variance $\mathrm{Var}(Q)$, even if not as accurate as Table 7 of [41]. In fact, TGA-G performs better for D service. On the other hand, Table 12 shows that the performance of TGA-G degrades significantly for the $LN(1,1)$ patience distribution. Nevertheless, the TGA-G (=DGA) approximation for $\mathrm{E}[X]$ remains good.

In summary, we have seen in previous sections that our proposed TGA-G approximation for heavily-loaded $G/GI/n+GI$ model is remarkably effective for a wide class of models. Nevertheless, there are limitations, as exposed by Table 12. Lack of accuracy is most likely as the loading decreases toward critical loading. That breakdown is likely to occur sooner (for higher ρ) if the component model elements deviate more from M. The examples show difficulties for low loading ($\rho = 1.02$) and non-M patience distributions.

Table 11: Comparison of the TGA-G and DGA approximations with the $M/M/n+M(n)$ approximation in [41] and simulation estimates for the $M(102^{-1})/GI(1,c_s^2)/100/200+E_2$ model with four service-time cdf's with a range of scv's: $0.0 \leq c_s^2 \leq 4.0$.

Serv. Dist.	$D, c_s^2=0$				$E_2, c_s^2=0.5$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGAG	Sim	Eng. Approx. (W05)	DGA	TGAG
$\mathbb{P}(W=0)$	1.80E-1	2.50E-1	1.05E-1	same	2.17E-1	2.50E-1	2.08E-1	same
rel. err.	±1.30E-3	28%	42%		±2.10E-3	13%	4%	
PoA	3.09E-2	3.81E-2	2.86E-2	same	3.51E-2	3.81E-2	3.83E-2	same
rel. err.	±1.70E-4	19%	7%		±2.90E-4	8%	9%	
E[Q]	1.11E+1	1.14E+1	1.08E+1	1.13E+1	1.15E+1	1.14E+1	1.08E+1	1.24E+1
rel. err.	±4.20E-2	3%	3%	2%	±7.50E-2	1%	6%	8%
Var(Q)	8.93E+1	1.22E+2	7.80E+1	6.42E+1	1.12E+2	1.22E+2	1.79E+2	1.22E+2
rel. err.	±4.00E-1	27%	12%	28%	±7.10E-1	8%	60%	9%
E[X]	1.10E+2	1.10E+2	1.11E+2	same	1.10E+2	1.10E+2	1.11E+2	same
rel. err.	±4.90E-2	0%	1%		±9.20E-2	0%	1%	

Serv. Dist.	$M, c_s^2=1$				$LN(1,1)$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGAG	Sim	Eng. Approx. (W05)	DGA	TGAG
$\mathbb{P}(W=0)$	2.46E-1	2.50E-1	2.33E-1	same	2.33E-1	2.50E-1	2.33E-1	same
rel. err.	±2.00E-3	2%	5%		±2.10E-3	7%	0%	
PoA	3.78E-2	3.81E-2	4.18E-2	same	3.70E-2	3.81E-2	4.18E-2	same
rel. err.	±3.20E-4	1%	11%		±2.70E-4	3%	13%	
E[Q]	1.18E+1	1.14E+1	1.08E+1	1.29E+1	1.17E+1	1.14E+1	1.08E+1	1.29E+1
rel. err.	±7.50E-2	3%	8%	9%	±6.30E-2	3%	8%	10%
Var(Q)	1.29E+2	2.20E+2	1.43E+2	1.80E+2	1.23E+2	1.22E+2	2.20E+2	1.43E+2
rel. err.	±9.40E-1	70%	11%	39%	±7.20E-1	1%	79%	16%
E[X]	1.10E+2	1.10E+2	1.11E+2	same	1.10E+2	1.10E+2	1.11E+2	same
rel. err.	±9.10E-2	0%	1%		±7.20E-1	0%	1%	

Table 12: Comparison of the TGA-G and DGA approximations with the $M/M/n+M(n)$ approximation in [41] and simulation estimates for the $M(102^{-1})/GI(1,c_s^2)/100/200+LN(1,1)$ model with four service-time cdf's with a range of scv's: $0.0 \leq c_s^2 \leq 4.0$.

Serv. Dist.	$E_2, c_s^2=0.5$				$M, c_s^2=1$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGA-G	Sim	Eng. Approx. (W05)	DGA	TGA-G
$\mathbb{P}(W=0)$	2.11E-1	2.47E-1	1.76E-1	same	2.42E-1	2.47E-1	1.95E-1	same
rel. err.	±1.30E-3	15%	17%		±2.60E-3	2%	20%	
PoA	3.48E-2	3.79E-2	5.13E-2	same	3.76E-2	3.79E-2	5.51E-2	same
rel. err.	±2.10E-4	8%	47%		±3.20E-4	1%	47%	
E[Q]	1.14E+1	1.10E+1	1.29E+1	1.43E+1	1.14E+1	1.10E+1	1.29E+1	1.46E+1
rel. err.	±3.90E-2	3%	13%	25%	±7.10E-2	4%	13%	27%
Var(Q)	1.03E+2	1.07E+2	1.99E+2	1.43E+2	1.16E+2	1.07E+2	2.31E+2	1.61E+2
rel. err.	±3.90E-1	4%	94%	40%	±4.60E-1	8%	100%	39%
E[X]	1.10E+2	1.09E+2	1.13E+2	1.13E+2	1.10E+2	1.09E+2	1.13E+2	same
rel. err.	±5.30E-2	1%	3%		±9.20E-2	0%	3%	

Serv. Dist.	$LN(1,1)$				$LN(1,4)$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGA-G	Sim	Eng. Approx. (W05)	DGA	TGA-G
$\mathbb{P}(W=0)$	2.29E-1	2.47E-1	1.95E-1	same	2.11E-1	2.47E-1	2.41E-1	same
rel. err.	±1.50E-3	7%	15%		±1.30E-3	15%	14%	
PoA	3.66E-2	3.79E-2	5.51E-2	same	3.48E-2	3.79E-2	6.66E-2	same
rel. err.	±2.40E-4	3%	51%		±2.10E-4	8%	91%	
E[Q]	1.14E+1	1.10E+1	1.29E+1	1.46E+1	1.14E+1	1.10E+1	1.29E+1	1.55E+1
rel. err.	±5.10E-2	4%	13%	27%	±3.90E-2	3%	13%	36%
Var(Q)	1.11E+2	1.07E+2	2.31E+2	1.61E+2	1.03E+2	1.07E+2	3.40E+2	2.16E+2
rel. err.	±4.30E-1	3%	109%	45%	±3.90E-1	4%	231%	111%
E[X]	1.10E+2	1.09E+2	1.13E+2	same	1.10E+2	1.09E+2	1.13E+2	same
rel. err.	±6.20E-2	1%	3%		±5.30E-2	1%	3%	

8.2 Comparison with Approximations in [33]

A MSHT limit was established for the $GI/M/n + GI$ model in [33] and its corresponding numerical approximation was developed and evaluated. The MSHT limit in [33] was in the QED regime, whereas ours in Theorem 2.3 from [26] is in the ED regime. Nevertheless, it is natural to compare the two candidate approximations for systems with $\rho > 1$, because any such system can be regarded as one in a sequence of systems satisfying a QED limit or an ED limit.

Before we present numerical comparison results, we first summarize the differences between TGA and [33]. First, just like [26], the MSHT limit in [33] is for the model with exponential service times, and the resulting approximation in [33] is limited to that case. Second, the formulas in [33] require the knowledge of the entire patience-time cdf F or, equivalently, of its hazard-rate function $h(t) \equiv f(t)/\bar{F}(t)$. In contrast, the ED fluid limit used in TGA depends on the patience-time cdf F only on the value of its inverse F^{-1} at the limiting fluid abandonment rate $(\rho - 1)/\rho$. The TGA fluid head-of-line waiting time $w(\infty)$ is determined by $w(\infty) = F^{-1}((\rho - 1)/\rho)$. Third, the approximation formulas in [33] are rather complicated, requiring computation of the triple integrals in (11), (12) and (13) in [33]. In contrast, the TGA approximations are easier to compute. Finally, the hazard-rate scaling in the MSHT limit in [33] makes that approximation especially effective for models in which the patience-time hazard-rate function changes rapidly near the origin.

The examples in [33] considered three different patience distributions. The first two are hyperexponential (H_2) distributions with parameter triples (p, θ_1, θ_2). The first is a regular H_2 distribution with $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. The second is a more extreme H_2 distribution with $(p, \theta_1, \theta_2) = (0.9, 1, 200)$, which changes rapidly near the origin. The third is another relatively extreme hazard-rate function that increases rapidly near the origin. In [33] the new approximations were compared to exact values and approximations "Z&M" from [44].

We now compare our new TGA approximation to the approximation developed in [33] by comparing our results to the displayed results in [33] for the cases with $\rho > 1$. These appear in Tables 3, 6 and 9 of [33]. As in [33], we consider a range of scale n from $n = 10$ to $n = 500$, with the traffic intensity ρ and number of server n satisfying the relation (25) with $\beta = -1$, which are: $n = 10$ ($\rho = 1.31$), $n = 50$ ($\rho = 1.14$), $n = 100$ ($\rho = 1.10$), $n = 200$ ($\rho = 1.07$), and $n = 500$ ($\rho = 1.022$). (The last case is at the edge of the range of effectiveness for TGA.)

Table 13 compares TGA to the hazard-rate scaling (HRS) approximation from [33] and the exact and approximate values from [44] for the first H_2 patience distribution with parameter $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. For this "easy" example, Table 13 shows that all methods are effective for large scale, but TGA is consistently appreciably better for the mean $E[V]$, especially for small scale.

However, Tables 14 and 15 show that the performance of TGA degrades significantly for the last two more challenging patience distributions, while the HRS approximation from [33] continues to be effective. However, even for these challenging examples, TGA performs better than HRS from [33] and Z&M from [44] in the case of $n = 10$ (small scale). Indeed, for $n = 10$, the TGA values are reasonable, whereas Z&M is off by a factor of 10.5 for the mean $E[V]$ in Table 14 and HRS is off by a factor of 15.0 for the mean $E[V]$ in Table 15. The is partly explained by the scaling, because for $n = 10$ the QED scaling in (25) makes $\rho = 1.31$ even though $\beta = -0.5$.

We conclude that all these approximations have advantages, which at least partly depends on whether the QED or ED regime is most natural.

Table 13: A comparison of TGA and DGA approximations for the probability of delay PoD, the mean wait $E[V]$ and the probability of abandonment PoA to the results in Table 3 of [33] for the $M/M/n+H_2$ model with $\rho = 1+1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and H_2 patience with $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. The exact results and approximations Z&M come from [44].

	PoD				E[V]				PoA			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.795	0.741	0.781	0.792	11.386	12.8341	14.641	10.103	0.276	0.321	0.268	0.245
50	0.784	0.759	0.778	0.798	5.699	6.030	6.421	5.438	0.140	0.151	0.138	0.140
100	0.781	0.763	0.777	0.793	4.148	4.472	4.320	4.036	0.103	0.108	0.102	0.103
200	0.780	0.767	0.777	0.791	2.996	3.084	3.186	2.956	0.0743	0.0771	0.0738	0.0747
500	0.778	0.770	0.776	0.794	1.932	1.968	2.010	1.931	0.0481	0.0492	0.0478	0.0489

Table 14: A comparison of TGA and DGA approximations for the probability of delay PoD, the mean wait $E[V]$ and the probability of abandonment PoA to the results in Table 6 of [33] for the $M/M/n + H_2$ model with $\rho = 1 + 1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and H_2 patience with $(p, \theta_1, \theta_2) = (0.9, 1, 200)$.

	PoD				E[V]				PoA			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.7909	0.3139	0.7821	0.7462	11.8884	1.1349	16.2358	9.9089	0.2763	0.4268	0.2677	0.2584
50	0.7158	0.3242	0.709	0.6014	4.5527	0.5277	5.405	3.4147	0.1455	0.1838	0.1428	0.1211
100	0.6663	0.327	0.6587	0.5789	2.7067	0.377	3.0844	1.2267	0.1087	0.1313	0.1072	0.0761
200	0.6063	0.329	0.5979	0.584	1.4942	0.2686	1.6453	0.6486	0.0808	0.0936	0.08	0.0613
500	0.5213	0.3309	0.5127	0.5858	0.6201	0.1711	0.6577	0.3495	0.0541	0.0596	0.0538	0.0495

Table 15: A comparison of TGA and DGA approximations for the probability of delay PoD, the mean wait $E[V]$ and the probability of abandonment PoA to the results in Table 9 of [33] for the $M/M/n + GI$ model with $\rho = 1 + 1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and increasing patience-time hazard rate.

	PoD				E[V]				PoA			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.4454	0.7408	0.2913	0.6310	0.8557	1.1349	12.8341	1.1201	0.3312	0.3209	0.3297	0.4033
50	0.4041	0.7586	0.3559	0.6647	0.5277	6.4212	0.565	0.9411	0.1679	0.1507	0.1658	0.2839
100	0.4152	0.7633	0.3859	0.6783	0.4786	4.4716	0.4877	0.8368	0.1215	0.108	0.1202	0.2243
200	0.4351	0.7667	0.4169	0.6917	0.4107	3.0844	0.4191	0.7245	0.087	0.0771	0.0863	0.1690
500	0.4688	0.7698	0.4588	0.7079	0.3347	1.9681	0.3404	0.5779	0.0553	0.0492	0.055	0.1078

8.3 Underloaded Models

In this section we present the performance of the underloaded $M/M/n+M$ in QED regime, where β defined in (25) is fixed at 0.5. During UL intervals, it is easy to check that PoD and PoA are 0 in the MSHT limits. The fluid and diffusion limit of waiting time in underloaded intervals implies that $w(\infty) = v(\infty) = 0$ and $\sigma_W = 0$. In order to improve the performance, we replace $w(\infty) = 0$ by $\mathrm{E}\left[W^{TGA}\right] = \mathrm{E}\left[Q^{TGA}\right]/\mu$ and $\sigma_W = 0$ by $\mathrm{Var}\left(W^{TGA}\right) = \mathrm{Var}\left(Q^{TGA}\right) + \mathrm{E}\left[Q^{TGA}\right]$, then design $P_D^{TGA}(n)$ and $P_A^{TGA}(n)$ as

$$\begin{aligned} P_D^{TGA}(n) &= \bar{\Phi}\left(-a'_W(n)\right), \\ P_A^{TGA}(n) &= \int_0^\infty \bar{\Phi}\left(a'_W(n)\left(\frac{x}{w}-1\right)\right) f(x)\mathrm{d}x, \end{aligned} \tag{26}$$

where $a'_W(n) = \sqrt{n}\mathrm{E}[W^{TGA}]/\mathrm{Var}\left(\hat{W}^{TGA}\right)$.

The main idea of (26) is to use $W^{TGA} = \sum_{i=1}^{Q^{TGA}} S_i$ to replace the zero waiting time; here S_i is the processing time of the i^{th} customers. We omit the impact of abandonments since according to the numerical results on underloaded intervals, the probability of abandonments are about 10^{-2}. To verify (26), it is suffice to prove that

$$\mathrm{E}\left[W^{TGA}\right] = \mathrm{E}\left[Q^{TGA}\right]/\mu, \quad \text{and} \quad \mathrm{Var}\left(W^{TGA}\right) = \mathrm{Var}\left(Q^{TGA}\right) + \mathrm{E}\left[Q^{TGA}\right].$$

Proof. It is obvious for the expression of $\mathrm{E}[W^{TGA}]$ so we only focus on $\mathrm{Var}\left(W^{TGA}\right)$ here.

$$\begin{aligned} \mathrm{E}\left[\left(W^{TGA}\right)^2\right] &= \mathrm{E}\left[\sum_{i=1}^{Q^{TGA}} S_i^2 + 2\sum_{j<i}^{Q^{TGA}} S_i S_j\right] = \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{Q^{TGA}} S_i^2 + 2\sum_{j<i}^{Q^{TGA}} S_i S_j \middle| Q^{TGA}\right]\right] \\ &= \mathrm{E}\left[2Q^{TGA} + Q^{TGA}(Q^{TGA}-1)\right] = \mathrm{E}\left[Q^{TGA}\right] + \mathrm{E}\left[\left(Q^{TGA}\right)^2\right], \end{aligned}$$

which implies the expression of $\mathrm{Var}(W^{TGA})$. □

Table 16 shows the performance of an underloaded $M/M/n+M$ model with $\rho = 0.95$. In particular, with parameters $\lambda = 100$, $\rho = 0.95$ and $0.1 \le \theta \le 4.0$. Table 16 shows good performance of TGA for the means of X_n and B_n in all cases and for the variances of X_n and B_n with $0.5 \le \theta \le 2.0$, but poor performance otherwise.

Table 16: The performance for underloaded models: a comparison between simulation estimates and exact numerical values for the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $n = 100, \rho = 0.95$ and $0.1 \leq \theta \leq 2$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.00E+2	9.50E+1	same	9.80E+1	9.50E+1	same	9.64E+1	9.50E+1	same
rel. err.		5%			3%			1%	
Var(X)	2.18E+2	9.50E+1	same	1.56E+2	9.50E+1	same	1.20E+2	9.50E+1	same
rel. err.		56%			39%			21%	
E[B]	9.44E+1	9.50E+1	9.31E+1	9.40E+1	9.50E+1	9.31E+1	9.36E+1	9.50E+1	9.31E+1
rel. err.		1%	1%		1%	1%		2%	0%
Var(B)	4.91E+1	9.50E+1	5.31E+1	5.03E+1	9.50E+1	5.31E+1	5.11E+1	9.50E+1	5.31E+1
rel. err.		94%	8%		89%	6%		86%	4%
E[Q]	5.78E+0	0.00E+0	1.89E+0	4.06E+0	0.00E+0	1.89E+0	2.88E+0	0.00E+0	1.89E+0
rel. err.		100%	67%		100%	53%		100%	34%
Var(Q)	1.04E+2	0.00E+0	1.59E+1	5.67E+1	0.00E+0	1.59E+1	3.22E+1	0.00E+0	1.59E+1
rel. err.		100%	85%		100%	72%		100%	51%
E[V]	6.16E-2	0.00E+0	1.89E-2	4.37E-2	0.00E+0	1.89E-2	3.14E-2	0.00E+0	1.89E-2
rel. err.		100%	69%		100%	57%		100%	40%
Var(V)	1.13E-2	0.00E+0	1.78E-3	6.24E-3	0.00E+0	1.78E-3	3.60E-3	0.00E+0	1.78E-3
rel. err.		100%	84%		100%	72%		100%	51%
PoD	4.49E-1	0.00E+0	6.73E-1	4.06E-1	0.00E+0	6.73E-1	3.64E-1	0.00E+0	6.73E-1
rel. err.		100%	50%		100%	66%		100%	85%
PoA	6.09E-3	NaN	1.88E-3	1.07E-2	NaN	4.70E-3	1.51E-2	NaN	9.38E-3
rel. err.		NaN	69%		NaN	56%		NaN	38%

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	9.50E+1	9.50E+1	same	9.38E+1	9.50E+1	same	9.28E+1	9.50E+1	same
rel. err.		0%			1%			2%	
Var(X)	9.50E+1	9.50E+1	same	7.78E+1	9.50E+1	same	6.64E+1	9.50E+1	same
rel. err.		0%			22%			43%	
E[B]	9.31E+1	9.50E+1	9.31E+1	9.26E+1	9.50E+1	9.31E+1	9.21E+1	9.50E+1	9.31E+1
rel. err.		2%	0%		3%	1%		3%	1%
Var(B)	5.16E+1	9.50E+1	5.31E+1	5.16E+1	9.50E+1	5.31E+1	5.12E+1	9.50E+1	5.31E+1
rel. err.		84%	3%		84%	3%		86%	4%
E[Q]	1.92E+0	0.00E+0	1.89E+0	1.21E+0	0.00E+0	1.89E+0	7.23E-1	0.00E+0	1.89E+0
rel. err.		100%	1%		100%	57%		100%	161%
Var(Q)	1.69E+1	0.00E+0	1.59E+1	8.28E+0	0.00E+0	1.59E+1	3.82E+0	0.00E+0	1.59E+1
rel. err.		100%	6%		100%	92%		100%	315%
E[V]	2.13E-2	0.00E+0	1.89E-2	1.38E-2	0.00E+0	1.89E-2	8.60E-3	0.00E+0	1.89E-2
rel. err.		100%	11%		100%	37%		100%	120%
Var(V)	1.93E-3	0.00E+0	1.78E-3	9.79E-4	0.00E+0	1.78E-3	4.76E-4	0.00E+0	1.78E-3
rel. err.		100%	8%		100%	81%		100%	273%
PoD	3.17E-1	0.00E+0	6.73E-1	2.68E-1	0.00E+0	6.73E-1	2.21E-1	0.00E+0	6.73E-1
rel. err.		100%	112%		100%	151%		100%	205%
PoA	2.02E-2	NaN	1.87E-2	2.54E-2	NaN	3.70E-2	3.04E-2	NaN	7.26E-2
rel. err.		NaN	7%		NaN	46%		NaN	139%

9 Simulation Methodology

We used simulation to estimate the exact values for all non-Markovian models. We now provide extra details about our simulation methodology. For $n = 100$, we estimated all the performance measures using 2000 independent replications over the time interval $[0, T]$ with $T = 100$, starting empty in each case. To have statistical precision for all the steady-state estimates, we need to ensure, first, that the system has approximately reached the steady state before sampling, and, second, that enough sampled data are collected to give reasonable accuracy, which we judge by using 95% confidence intervals. We discuss these issues in turn.

9.1 From Transient to Steady State

To avoid bias caused by the initial transient starting empty, we do not collect data from an initial portion of each run. We stop sampling at time $0.95T = 95$ We also eliminate a final portion so that we can observe the waiting times experienced by all arrivals in the main measurement interval; i.e., to avoid abnormal zeroes in sampled potential waiting times, which results from that some virtual customers' (potential) waiting times not being sampled at the end of simulation at $T = 100$. In particular, for $n = 100$ we use the data in $[40, 95]$ from each run over $[0, 100]$ to estimate the steady-state performance functions.

To illustrate the initial transient and when it tends to disappear, we show an example, using the $H_2(105^{-1}, 2)/H_2(1, 2)/100 + H_2(2, 2)$ model. Figure 12 shows plots of the transient (time-dependent) mean and variance functions for the queue length. Figure 12 shows that the performance is close to steady state after time 20. To be safe, we use the data in $[40, 95]$ to estimate the steady-state performance functions.

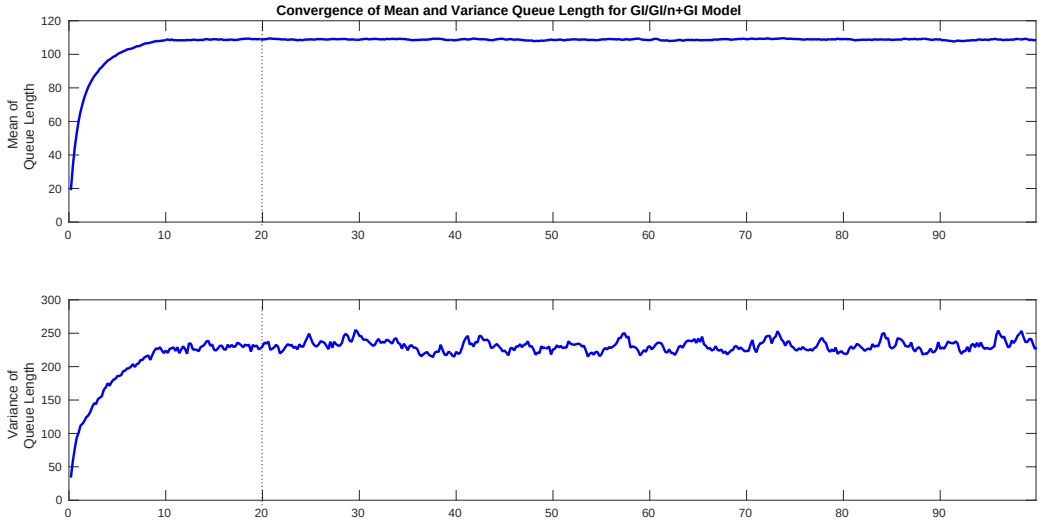

Figure 12: Simulation estimates of the mean and variance of the queue length as a function of time in the $H_2/H_2/n + H_2$ model to show the approach to steady state.

9.2 The Sampling Procedure

To determine the potential waiting times at time t, which are for an arrival with unlimited patience that would arrive at time t (the usual virtual waiting time, modified to include unlimited

patience), we generate virtual customers that do not affect the other customers. In particular, in the r^{th} simulation replication, $1 \le r \le R$, we periodically generate virtual arrivals at deterministic times $t_1, t_k, \ldots, t_{N_v}$ with $t_k \equiv k\Delta t$ and $\Delta t = 0.1$, $1 \le k \le N_v \equiv \lfloor T/\Delta t \rfloor$. The virtual customers have the same waiting time distribution. They abandon as if they are the real customers but they will not be removed from the queue if they abandon. They still wait in queue until their turn to enter service so that we can record their virtual waiting time as potential waiting times. We do not allow them to enter service, so these virtual customers do not affect the system dynamics. We use indicator variables $\eta^a_{r,k}$ and $\eta^d_{r,k}$ to record if the virtual arrival at t_k on the r^{th} path abandons and is delayed, namely

$$\eta^d_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > 0, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \eta^a_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > A_k, \\ 0, & \text{otherwise} \end{cases}, \tag{27}$$

where A_k is the patience time of the k^{th} virtual arrival, $V_r(k) \equiv E_r(k) - A_r(k)$ records the potential waiting time at time t_k, $A_r(k) \equiv k \cdot \Delta t$ and $E_r(k)$ are the times at which the k^{th} virtual customer arrives and enters service.

For the number of customers in queue and service at each time, We sample the continuous-time queue-length process and number of busy servers at discrete time points $t_1, t_2, \ldots, t_{N_v}$, denoted by $Q_r(k)$ and $B_r(k)$. Here we make sure to exclude the virtual arrivals.

9.3 Constructing Confidence Intervals

All estimates of target performance measures and corresponding confidence intervals are based on assuming i.i.d. samples, which is justified because we take a single estimate from each of the $R = 2000$ independent samples.

To illustrate how we construct these estimators, we use the queue-length process Q for an example. On the r^{th} path, we sample values for the queue length at $N = 551$ evenly-spaced time points in the interval $[0.4T, 0.95T] = [40, 95]$, denoted by $Q_{r,1}, \ldots, Q_{r,N}$. To construct the confidence intervals for $\mathrm{E}[Q]$ and $\mathrm{E}[Q^2]$, first, for each replication $r = 1, 2, \ldots, R$, we let

$$\widetilde{Q}_r \equiv \frac{1}{N}\sum_{l=1}^{N} Q_{r,l} \quad \text{and} \quad \widetilde{Q}^{(2)}_r \equiv \frac{1}{N}\sum_{l=1}^{N} (Q_{r,l})^2. \tag{28}$$

Even though the random variables being averaged in (28) are typically dependent, these are valid estimators for the true mean $\mathrm{E}[Q]$ and second moment $\mathrm{E}[Q^2]$. Experience shows that the average of these $N = 551$ values has lower variance than a single observation from the end of the run.

To get the overall estimators of $\mathrm{E}[Q]$ and second moment $\mathrm{E}[Q^2]$, and their CI's, we use the R independent samples $\widetilde{Q}_1, \ldots, \widetilde{Q}_R$ $(\widetilde{Q}^{(2)}_1, \ldots, \widetilde{Q}^{(2)}_R)$ to compute the sample mean and sample variance of the queue length and its second moment in the usual way, i.e.,

$$\bar{Q}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}_r \quad \text{and} \quad S^2_Q(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}_r - \bar{Q}(R)\right)^2, \tag{29}$$

$$\bar{Q}^{(2)}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}^{(2)}_r \quad \text{and} \quad S^2_{Q^{(2)}}(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}^{(2)}_r - \bar{Q}^{(2)}(R)\right)^2. \tag{30}$$

The random variables $\bar{Q}(R)$ and $\bar{Q}^{(2)}(R)$ in (29) our our final estimators for the true mean $\mathrm{E}[Q]$ and second moment $\mathrm{E}[Q^2]$.

As usual, the $(1-100\alpha\%)$-confidence intervals for the mean and second moment of the queue length are

$$\left[\bar{Q}(R) - z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}, \bar{Q}(R) + z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}\right], \quad \text{and} \tag{31}$$

$$\left[\bar{Q}^{(2)}(R) - z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}, \bar{Q}^{(2)}(R) + z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}\right], \tag{32}$$

where z_α is the α-percentile of the standard Gaussian distribution. Since we use 95% CI's, we use $\alpha = 0.025$.

Since $Var(Q) = E[Q^2] - (E[Q])^2$, we estimate the variance by

$$\bar{V}(R) \equiv \bar{Q}^{(2)}(R) - \left(\bar{Q}(R)\right)^2. \tag{33}$$

We then approximate the CI halfwidth of the the variance by the CI halfwidth of the second moment. We thus roughly estimate the CI of the variance as

$$\left[\bar{V}(R) - z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}, \bar{V}(R) + z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}\right]. \tag{34}$$

We discuss this approximation further with the numerical example below.

For the probability of abandonment (similar procedure for the probability of delay), we sample values for the indicator at $N = 551$ evenly-spaced time points in the interval $[0.4T, 0.95T]$ on the r^{th} run, denoted by $\eta_{r,1}^a, \ldots, \eta_{r,N}^a$, and we let $\widetilde{P}_r^a \equiv (1/N)\sum_{l=1}^N \eta_{r,l}^a$, for $r = 1, 2, \ldots, R$. The $(1-100\alpha\%)$-confidence interval is

$$\left[\bar{P}^a(R) - z_{\alpha/2}\sqrt{\frac{S_a^2(R)}{R}}, \bar{P}^a(R) + z_{\alpha/2}\sqrt{\frac{S_a^2(R)}{R}}\right],$$

where

$$\bar{P}^a(R) \equiv \frac{1}{R}\sum_{r=1}^R \widetilde{P}_r^a \quad \text{and} \quad S_a^2(R) \equiv \frac{1}{R-1}\sum_{r=1}^R \left(\widetilde{P}_r^a - \bar{P}^a(R)\right)^2.$$

To substantiate our procedures and verify that we obtain adequate statistical precision, we compare the estimated performance measures of $M/M/n+M$ model to corresponding exact solutions, which are calculated by the same algorithms of [41]. The Table 17 shows that the procedures are sound and the the statistical precision is adequate.

We use Table 17 to elaborate on the approximate CI for the variance. To do so, we focus on the queue length in the case $\theta = 0.1$. Notice that the CI for the mean is 30.0 ± 0.52, so that the relative halfwidth for the mean is 1.7%. However, by squaring the upper and lower limits, we see that a rough symmetric CI for $(E[Q])^2$ is 900 ± 30, using the gap at the upper limit, so that the relative halfwidth for the square of the mean is 3.3%. Our direct estimate of the second moment is $613 + (30)^2 = 1513$ and our direct estimate of its CI is 1513 ± 43.6, so that the relative halfwidth is 2.8%. Our approximation thus estimates the CI of the variance as 613 ± 43.6, so that the approximate relative halfwidth is 7.1%, which we judge to be conservative.

Table 17: A comparison between simulation estimates and exact numerical values for the $M(102^{-1})/M(1)/100 + M(\theta^{-1})$ model, confirming the validity of both algorithms

	$\theta = 0.1$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
Perf. Meas.	Exact	Sim.	Exact	Sim.	Exact	Sim.	Exact	Sim.	Exact	Sim.
E[X] rel. err.	1.52E+2	1.52E+2 ±6.75E-1	1.22E+2	1.22E+2 ±2.78E-1	1.11E+2	1.11E+2 ±1.46E-1	1.05E+2	1.05E+2 ±8.47E-2	1.01E+2	1.01E+2 ±5.61E-2
Var(X) rel. err.	9.25E+2	9.07E+2 ±2.12E+2	3.47E+2	3.46E+2 ±7.03E+1	1.81E+2	1.80E+2 ±3.32E+1	1.05E+2	1.05E+2 ±1.79E+1	6.85E+1	6.81E+1 ±1.12E+1
E[Q] rel. err.	5.22E+1	5.20E+1 ±6.67E-1	2.30E+1	2.31E+1 ±2.62E-1	1.27E+1	1.27E+1 ±1.26E-1	7.03E+0	7.01E+0 ±6.15E-2	3.88E+0	3.90E+0 ±3.06E-2
Var(Q) rel. err.	8.99E+2	8.82E+2 ±7.88E+1	3.05E+2	3.06E+2 ±1.55E+1	1.35E+2	1.34E+2 ±4.53E+0	5.92E+1	5.91E+1 ±1.41E+0	2.57E+1	2.58E+1 ±4.55E-1
E[W] rel. err.	5.14E-1	5.12E-1 ±6.52E-3	2.29E-1	2.30E-1 ±2.57E-3	1.28E-1	1.28E-1 ±1.24E-3	7.22E-2	7.20E-2 ±6.16E-4	4.09E-2	4.11E-2 ±3.13E-4
Var(W) rel. err.	8.59E-2	8.41E-2 ±7.58E-3	2.93E-2	2.94E-2 ±1.50E-3	1.31E-2	1.31E-2 ±4.45E-4	5.88E-3	5.88E-3 ±1.42E-4	2.64E-3	2.64E-3 ±4.69E-5
PoD rel. err.	9.67E-1	9.67E-1 ±1.93E-3	8.90E-1	8.91E-1 ±2.91E-3	8.03E-1	8.03E-1 ±3.17E-3	7.00E-1	6.99E-1 ±3.11E-3	5.92E-1	5.95E-1 ±2.81E-3
PoA rel. err.	4.97E-2	4.98E-2 ±8.36E-4	5.47E-2	5.50E-2 ±8.51E-4	6.04E-2	6.07E-2 ±8.43E-4	6.70E-2	6.65E-2 ±8.29E-4	7.40E-2	7.40E-2 ±8.34E-4

10 Conclusion

In this paper we have developed and evaluated approximations for the key steady-state performance measures in the heavily loaded stationary $G/GI/n + GI$ model. These approximations are remarkably simple and easy to implement, even though they have a relatively complicated basis in many-server heavy-traffic (MSHT) limits. As can be seen from §3.2, the basic truncated Gaussian approximations (TGA) for $\rho > 1$ can be expressed in terms of (i) the deterministic fluid approximating pair $(w, Q) \equiv (w(\infty), Q(\infty))$ in Theorem 2.1 (b), (ii) the limiting variance pair $(\sigma_W, \sigma_X) \equiv (\sigma_W(\infty), \sigma_X(\infty))$ in Theorem 2.3 (b) and (iii) the standard (mean-0, variance-1) Gaussian cdf Φ and pdf ϕ. The TGA-G refinement in §3.3 only requires an adjustment to σ_W using the service-time scv c_s^2. The approximate probability of abandonment PoA in (23) also requires a numerical integration with the full patience pdf f. Most of these approximate performance measures are easier to compute than corresponding quantities in the Markov $M/M/n + M$ Erlang-A model; see [30, 41]. For that model, the advantage of such simple approximations was previously emphasized by [11, 40, 30].

The basis for these simple approximations is the collection of MSHT limits for the $G/GI/n+GI$ model in the efficiency-driven (ED) regime in [24, 25, 26, 42], reviewed in §2. These MSHT limits explain why the approximations require that the scale (number of servers) n and the load (traffic intensity) ρ be suitably large, and that the abandonment rate be not too high.

It is significant that the approximations go beyond a direct application of the MSHT limits. We developed the approximations in §3. After presenting the direct DGA Gaussian approximations in §3.1, we applied truncation to obtain the refined TGA approximations in §3.2 and then subsequently, in §3.3, we heuristically modified the MSHT limit in [26] for non-exponential GI service to obtain the final TGA-G approximations, which coincide with TGA for exponential service times.

In §§4-7 we report results of extensive simulations studying the approximations. These experiments show that, for large scale with $n = 100$, the approximations are effective for a significant range of the traffic intensity (ρ) and the abandonment rate (θ) parameters, roughly for $\rho > 1.02$ and $\theta < 2.0$. After first comparing the approximations to exact numerical results for the Markov

$M/M/n+M$ model in §4, we carefully examined the impact of non-Markov elements for the arrival process (including a non-renewal MMPP example) and the patience distribution in §5 and for the service time distribution in §§6.1 and 6.2. In §7 we showed that these approximations also remain effective for smaller scale, assuming that the remaining parameters are adjusted appropriately. In §9 we described the simulation methodology. Additional details are provided in an online appendix.

10.1 The Impact of Model Features on System Performance

It is well known that congestion tends to be increasing in λ and decreasing in μ, n and θ. The way that congestion depends on these basic parameters is quite well understood. We refer to [43] for a careful sensitivity study of performance in the Erlang-A model, which shows that performance is quite sensitive to small percentage changes in the arrival rate λ or the service rate μ (and thus the traffic intensity ρ), but is relatively insensitive to small changes in the abandonment rate θ.

The MSHT limits help expose the structure as well, as discussed for Markov models in [14, 11, 40]. First, the MSHT scaling shows that, for n not too small, performance depends on n and ρ primarily through the single parameter $(1-\rho)\sqrt{n} \equiv \beta$, with $\beta < 0$ corresponding to the overloaded case. Second, §4 of [40] shows for the $M/M/n+M$ model that performance depends on n and θ primarily through n/θ for large values of that ratio.

The fluid limits for the general $G/GI/n+GI$ model in Theorem 2.1 are very useful for exposing the primary impact of model elements upon performance. Theorem 2.1 shows that performance primarily depends on the arrival process only via its rate and on the service-time distribution only via its mean. In contrast, when $\rho > 1$, the fluid limit depends on the full patience-time cdf F, but then only on the value of its inverse F^{-1} at the limiting fluid abandonment rate $(\rho-1)/\rho$. the head-of-line waiting time $w(\infty)$ is determined by $w(\infty) = F^{-1}((\rho-1)/\rho)$. From a practical engineering perspective, we see that this characteristic of the patience cdf is critical, not the mean or variance. Table 6 illustrated that the patience cdf matters beyond its mean and variance. The fluid approximation shows that, if the full patience cdf F gets larger in stochastic order, i.e., if the function $\bar{F}$ increases so that customers are more patient, then congestion should increase, just as in the $M/M/n+M$ model.

Moreover, As we have observed in §5, the general stationary G arrival process only affects the diffusion limit in Theorem 2.3 though the asymptotic variance parameter c_λ^2 appearing in the assumed FCLT for the arrival process in (5). That was illustrated in Table 3 when we displayed results for models that have $M(1)$ and $LN(1,1)$ arrival processes. Theorem 2.3 shows that congestion tends to increase appproximately proportional to c_λ^2. In contrast, the patience-time cdf F affects the diffusion limit in a complicated way.

A main conclusion in [41, 42] was that the steady state performance of the $M/GI/n+GI$ model tends to be nearly insensitive to the service-time distribution beyond its mean. Our experiments confirm that conclusion for the main performance measures considered, e.g., for the mean values of the steady-state queue length and waiting time, but we show that the variance and full distribution depend significantly on the service-time distribution beyond its mean. Moreover, our refined TGA-G approximation successfully captures that effect.

Because Theorem 2.3 is for the $G/M/s+GI$ model, we do not directly see the impact of the service-time distribution, but the fluid limits suggests that it is not so great. Our heuristic refinement TGA-G depends on the service-time distribution only through its first two moments. In particular, the variance $\sigma_{W_G}^2$ in §3.3 is increasing in c_s^2. Because the approximation is quite effective, we conclude that the scv c_s^2 captures much of the impact.

10.2 The Impact of Model Features on the Accuracy of the Approximations

Much of this paper has been devoted to carefully examining the accuracy of the proposed engineering approximations. First, for the $M/M/n+M$ model, Tables 1 and 2 and Figures 1-3 show that the accuracy of the approximation is quite good for large scale, which we take to be $n = 100$, provided that the model is reasonably overloaded (OL), as specified by our OL condition $\rho > 1.02$ and $\theta < 2.0$. Given that performance depends on n and ρ for n not too small primarily through the single parameter $(1-\rho)\sqrt{n} \equiv \beta$, with $\beta < 0$ corresponding to the overloaded case, it is noteworthy that for $n = 100$ our OL condition $\rho > 1.02$ corresponds to $\beta < -0.2$. For large n,

Clearly, the system ceases to be OL as ρ decreases toward critical loading (CL), characterized by $\rho = 1$, and as θ increases above 2.0, as shown in Figures 1 and 3. The degradation of performance for $\theta = 4$ and 10 is shown in Figure 21 in the appendix. These figures show that, for $\rho > 1$, a high abandonment rate θ degrades approximation accuracy the most.

Tables 3 and 4 and Figure 4 show that the accuracy is quite insensitive to the arrival process variability as characterized by c_λ^2. Table 5 and Figure 6 show that the same is true for the variability of the patience distribution as characterized by c_{ab}^2 to a large extent, but there is some degradation as c_{ab}^2 gets small.

10.3 Directions for Future Research

There are many good directions for future research. First, exact performance measures are still needed for the $M/GI/n+GI$ model and more general models with GI service. Second, MSHT limits are still needed for the $G/GI/n+GI$ models with abandonment and GI service, going beyond the established MSHT fluid limits. More generally, It remains to provide better theoretical justification for the TGA-G approximation with non-exponential service-time distributions in §3.3 and/or even better simple approximations, if possible. It also remains to develop effective approximations for other ranges of the parameters. It remains to find and study new approximations such as those recently proposed in [16]; they seem to effectively cope with rapidly changing patience hazard rate functions caused by delay annnouncements over time.

Acknowledgement

We thank the National Science Foundation for support: NSF grants CMMI 1362310 (first and third authors) and CMMI 1066372 and 1265070 (second author).

References

[1] Avramidis, A. N., Deslauriers, A. and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Sci* 50:896–908.

[2] Baccelli, F. and Hebuterne, G. (1981). On queues with impatient custoemrs. In *Performance 1981*. North-Holland, Amsterdam, pp. 159–179.

[3] Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition.

[4] Brandt, A. and Brandt, M. (1999). On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* 35:1–18.

[5] Brandt, A. and Brandt, M. (2002). Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system. *Queueing Systems* 41:73–94.

[6] 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.

[7] Dai, J. G. and He, S. (2010). Customer abandonment in many-server queues. *Mathematics of Operations Research* 35(2):347–362.

[8] Dai, J. G., He, S. and Tezcan, T. (2010). Many-server diffusion limits for $G/Ph/n+GI$. *The Annals of Applied Probability* 20:1854–1890.

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

[10] Fischer, W. and Meier-Hellstern, K. (1992). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 18:149171.

[11] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management* 4:208–227.

[12] Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37:84–97.

[13] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16:13–39.

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

[15] He, B., Liu, Y. and Whitt, W. (2015). Staffing a service system with non-Poisson nonstationary arrivals. Working paper, Department of Industrial Engineering and Operations Research, Columbia University.

[16] Huang, J., Mandelbaum, A., Zhang, H. and Zhang, J. (2015). Refined models for efficiency-driven queues with applications to delay announcements and staffing. Working paper.

[17] Ibrahim, R., L'Ecuyer, P., Regnard, N. and Shen, H. (2012). On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference* 2012:256–267.

[18] Iglehart, D. L. (1965). Limit diffusion approximations for the many-server queue and the repairman problem. *Journal Applied Probability* 2:355–369.

[19] Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 17:307–318.

[20] Kang, W. and Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* 20(6):2204–2260.

[21] Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper Management* 16(3):464–480.

[22] Li, A. and Whitt, W. (2014). Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation* 80:82–101.

[23] Liu, Y. and Whitt, W. (2011). Large-time asymptotics for the $G_t/M_t/s_t+GI$ many-server fluid queue with abandonmens. *Queueing Systems* 67:145–182.

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

[25] Liu, Y. and Whitt, W. (2012). A many-server fluid limit for the $G_t/GI/s_t+GI$ queueing model experiencing periods of overloading. *Operations Research Letters* 40:307–312.

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

[27] Liu, Y., Whitt, W. and Yu, Y. (2016). Appendix to: Approximations for heavily-loaded $G/GI/n+GI$ queues. Http://yunanliu.wordpress.ncsu.edu/files/2014/02/LiuWhittYuApprox021616app.pdf.

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

[29] Mandelbaum, A. and Zeltyn, S. (2004). The impact of customers' patience on delay and abandonment: some emempirical-driven experiemnts with the $M/M/n+G$ queue. *OR Spectrum* 26:377–411.

[30] Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In Spath, D. and Fahnrich, K. (eds.), *Advances in Services Innovations*. Springer, pp. 17–48.

[31] Massey, W. A. and Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75:243–277.

[32] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65:325–364.

[33] Reed, J. and Tezcan, T. (2012). Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Operations Research* 60:981–995.

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

[35] Whitt, W. (1991). The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3):307–314.

[36] Whitt, W. (1992). Asymptotic formulas for Markov processes with applications to simulation. *Operations Research* 40(2):279–291.

[37] Whitt, W. (1993). Approximations for the $GI/G/m$ queue. *Production and Operations Management* 2(2):114–161.

[38] Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Ther Application to Queues*. Springer.

[39] Whitt, W. (2004). A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52(6):922–941.

[40] Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50:1449–1461.

[41] Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Sci* 51(2):221–235.

[42] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research* 54:37–54.

[43] Whitt, W. (2006). Sensitivity of performance in the Erlang A model to changes in the model parameters. *Operations Research* 54:247–260.

[44] Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51:361–402.

[45] Zhang, X., Hong, L. J. and Glynn, P. W. (2014). Timescales in modeling call center arrivals. Working paper, Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology.