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

September 3, 2015

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, but involve important heuristic refinements. Simulation experiments show that the proposed approximations are effective for both large-scale and small-scale queueing systems in both the efficiency-driven regime and the quality-and-efficiency-driven regime for a significant range of the traffic intensity ρ and the abandonment rate θ parameters, roughly for $\rho > 1.02$ and $\theta < 2.0$.

keywords: queues, queues with customer abandonment, queueing 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 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 [8, 25]. However, data analysis from service systems has also shown that the distributions of the service and patience times are often not nearly exponential [3]. Data analysis also has suggested that in some cases the arrival process might not be well modeled by a Poisson process; see [1, 13, 15, 16, 38].

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).

Based on many-server heavy-traffic limits. Our primary source for these new engineering approximations is the body of many-server heavy-traffic (MSHT) Gaussian process limits for the time-varying non-Markovian $G_t/GI/n_t+GI$ queueing model in [19, 20, 21, 36], especially [21], which includes the restriction to exponential (M) service. These theorems were established for systems that alternate between underloaded or QD (quality-driven, [8]) and overloaded or ED (efficiency-driven) regimes; the system was assumed to be never in the critically loaded or QED (quality-and-efficiency-driven) regime. Thus, the limits provide an alternating one-sided view. Nevertheless, simulation experiments showed that the approximations are remarkably effective for difficult time-varying systems that are mostly not nearly critically loaded.

The limit theorems in [19, 20, 21] are for the more general time-varying model. For service systems, there is also 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 [9, 29], as reviewed in [10]. Indeed, that is the common approach used to staff service systems in practice.

The MSHT limit theorems in [19, 20, 21, 36] apply to our stationary model as a special case. We investigate the direct application of these Gaussian proocess limits, but we find that the direct approximations are ineffective when the system is near critical loading, i.e., 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.

Consistent with the one-sided approach to the MSHT limits in [21], the refined Gaussian appropximations are only effective in giving a one-sided view. For $\rho > 1$ ($\rho < 1$), the approximations are effective for describing the steady-state number in queue (in service), but not the other. Since, we primarily want to describe queue lengths and waiting times, we primarily focus on heavily loaded models with $\rho > 1$. 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 is not to 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.

Even though we focus on heavily loaded models with $\rho > 1$, as observed previously, e.g. [35, 36], these are practical cases that often occur in practice, because the abandonment always keeps the system stable. 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 approximatioons attractive alternatives even for the Markovian $M/M/n + M$ Erlang-A model, as in [8, 25]; e.g., for quick approximations it may not be necessary to solve any birth-and-death equations. (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.)

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 heavily studied, e.g., see [17, 33, 31] 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, unlike 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 [6], 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 [14] and there has been much work since then. The limits in [21] can be viewed as generalizations.

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 [14, 11] and then continuing with stationary models with abandonment [8, 4, 5], and then time-varying Markov models in [23] and non-Markovian models in [19, 20, 21]; 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. Exact results for the $M/M/n + GI$ model were established in [37], which led to further studies that provided better understanding, such as [24]. 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 [35]. As we have confirmed in our simulation experiments, these approximations from [35] 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 [35, 36] 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 [21] 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 [23] have also been developed by [26]. These Gaussian and skewness closure approximations are remarkably effective, even for critically loaded models, but so far they are restricted to Markovian models.

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 [19] and obtained as a limit in the MSHT functional weak law of large numbers (FWLLN) in [20], 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 functional central limit theorem (FCLT) in [21] (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

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

Since these direct Gaussian approximations are ineffective except when the system is significantly overloaded, e.g., when $\rho > 1.2$, we investigate a simple refinement based on truncation that was proposed in (1.2) of [21], but never tested; indeed that is the purpose of the present paper. As in (1.2) of [21], the first step is to make the obvious refinement for the queue length, letting

$$Q_n(\infty) \equiv (X_n(\infty) - n)^+ \approx (nX(\infty) + \sqrt{n}\hat{X}(\infty)) - n)^+. \tag{3}$$

The remaining approximations involve variations on this truncation idea. Thus these approximations based on the FCLT for the $G/M/n+GI$ model are called the *truncated Gaussian approximations* (TGAs); see §3 for the details.

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 and §5, we present results of numerical examples to test the performance of the Gaussian approximations. We conclude in in §9. Additional material, including results revealing the limitations of our approximations, appear in the appendix [22].

2 Many-Server Heavy-Traffic Limits

The MSHT limits are obtained by considering a sequence of 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 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 $\bar{F}$ and $\bar{G}$ be the complementary cumulative distribution functions (ccdfs) of F and G. Thus the traffic intensity is $\rho_n = \lambda_n/s_n\mu = \lambda/s \equiv \rho$ for all n.

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

$$\left(N_n(t) - n\lambda\, t\right)/\sqrt{n} \Rightarrow c_\lambda \mathcal{B}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty, \tag{4}$$

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 [2, 21, 32] 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 elapsed head-of-line waiting time 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.

FWLLN 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 [20], we have the joint convergence for the LLN-scaled functions

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

where the limit is a vector of deterministic functions, specified in [36, 19]. We next summarize the steady state performance.

Theorem 2.1 (Theorem 3.1 in [36] and Theorem 4.1 in [18])**.** *The stationary fluid model with capacity s arising as the MSHT FWLLN limit of stationary $G/GI/n + GI$ models, where model n has mean service time $E[S] = 1$ and 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 $\rho \leq 1$, then for $x \geq 0$,

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

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

$$B(\infty) = \int_0^\infty b(x)\, dx = s, \quad b(x) = s\mu \bar{G}(x) \quad \textit{for} \quad x \geq 0,$$

$$v(\infty) = w(\infty) = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad q(x) = \lambda \bar{F}(x) \quad \textit{for} \quad 0 \leq x \leq w(\infty),$$

$$Q(\infty) = \int_0^\infty q(x)\, dx = \lambda \int_0^{w(\infty)} \bar{F}(x) dx.$$

FCLT limits for the $G/M/n+GI$ queue. As in [21], we now restrict to the $G/M/n+GI$ model having M service. To provide the FCLT limits, let the CLT-scaled processes be

$$\hat{X}_n(t) = \frac{X_n(t) - nX(t)}{\sqrt{n}}, \quad \hat{Q}_n(t) = \frac{Q_n(t) - nQ(t)}{\sqrt{n}}, \quad \hat{B}_n(t) = \frac{B_n(t) - nB(t)}{\sqrt{n}},$$
$$\hat{W}_n(t) = \sqrt{n}\left(W_n(t) - w(t)\right), \quad \hat{V}_n(t) = \sqrt{n}\left(V_n(t) - v(t)\right), \tag{6}$$

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 (5), i.e., the deterministic fluid functions in [19]. 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 [21]) by simply letting $\lambda(t) = \lambda$ and $s(t) = s = 1$. (As in Theorem 2.1, we let the limiting fluid model have capacity 1, so that there are n servers in model n.)

Theorem 2.2 (MSHT FCLT limits for the $G/M/n+GI$ queues)**.** *Consider the sequence of time-varying $G/M/n+GI$ queues having s_n servers. Under regularity conditions in [21], including appropriate initial convergence at time* 0*,*

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

(a) When $\rho < 1$, i.e. in an underloaded (UL) interval, $\hat{Q} = \hat{W} = \hat{V} = 0$, and $\hat{B} \overset{\mathrm{d}}{=} \hat{X}$ 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

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

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

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

*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"). In addition, $H(t,u)$, $I_i(t)$, $K_i(t)$ are deterministic analytic functions given in §**??** of the appendix. Both $\hat{X}(t)$ and $\hat{W}(t)$ are zero mean Gaussian processes with variance process*

$$\sigma_W^2(t) \;=\; H^2(t,0)\left(\int_0^t H^2(u,0)I^2(u)\,du + Var\left(\hat{W}(0)\right)\right), \tag{7}$$

$$\begin{aligned}
\sigma_X^2(t) &= \int_{t-w(t)}^t \lambda\,\bar{F}(t-s)\left((c_\lambda^2-1)\bar{F}(t-s)+1\right)ds\\
&\quad + \lambda^2\bar{F}^2(w(t))\sigma_W^2(t) + Var\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2.
\end{aligned} \tag{8}$$

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

$$\sigma_X^2 \equiv \lambda(c_\lambda^2 - 1)\int_0^\infty \bar{G}^2(s)\,ds + \lambda\int_0^\infty \bar{G}(s)\,ds. \tag{9}$$

(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

$$\sigma_W^2 \equiv \frac{(c_\lambda^2 - 1) + 2\rho}{2\rho^2 f(w(\infty))} \quad \textit{and} \quad \sigma_X^2 \equiv \sigma_W^2 + \lambda\int_0^{w(\infty)} \bar{F}(u)\left(1 + (c_\lambda^2 - 1)\bar{F}(u)\right)du, \tag{10}$$

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.

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

Inserting it to (8), we get the remaining limit

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

where $\lim_{t\to\infty} \mathrm{Var}\left(\hat{X}(0)\right) \cdot \left(\bar{F}_w(t)\right)^2 = 0$ 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,

$$\lim_{t\to\infty} \hat{X}(t) = \lim_{t\to\infty} \lim_{n\to\infty} \hat{X}_n(t) = \lim_{n\to\infty} \lim_{t\to\infty} \hat{X}_n(t) = \lim_{n\to\infty} \hat{X}_n(\infty). \tag{11}$$

We conjecture that (11) 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 [11, 8, 37].

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 [27] and references there, as shown for the $M/M/n$ model in [14].

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 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

$$X_n \approx X_n^{DGA} \equiv nX + \sqrt{n}\hat{X}, \tag{12}$$

$$B_n \approx B_n^{DGA} \equiv nB + \sqrt{n}\hat{B}, \quad Q_n^{DGA} \equiv nQ + \sqrt{n}\hat{Q}, \tag{13}$$

$$W_n \approx W_n^{DGA} \equiv w + \frac{1}{\sqrt{n}}\hat{W}, \tag{14}$$

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,

$$PoD_n \approx PoD_n^{DGA} \equiv \mathbb{P}\left(W_n^{DGA} > 0\right) \quad \text{and} \tag{15}$$

$$PoA_n \approx PoA_n^{DGA} \equiv \mathbb{P}\left(W_n^{DGA} > A\right), \tag{16}$$

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

$$\begin{aligned} Q_n^{TGA} &\equiv \left(X_n^{DGA} - s_n\right)^+ = \sqrt{n}\sigma_X \left(\frac{\hat{X}}{\sigma_X} + \frac{\sqrt{n}(X-s)}{\sigma_X}\right)^+ \equiv \sqrt{n}\sigma_X \left(\mathcal{Z} \vee -a_X(n)\right) + n(X-s), \\ B_n^{TGA} &\equiv \left(X_n^{DGA} \wedge s_n\right)^+ = \sqrt{n}\sigma_X \left(\left(\mathcal{Z} \wedge -a_X(n)\right) + \frac{\sqrt{n}}{\sigma_X}X\right)^+, \end{aligned} \tag{17}$$

where $\mathcal{Z}$ is a standard Gaussian random variable, σ_X is given in (10), and

$$a_X(n) = \sqrt{n}(\rho_n - 1)/\sigma_X. \tag{18}$$

In the same spirit, we truncate the DGAs for waiting times (14) to obtain their TGAs.

$$\begin{aligned} W_n^{TGA} &= V_n^{TGA} = \left(W_n^{DGA}\right)^+ = \left(w + \frac{\hat{W}}{\sqrt{n}}\right)^+ = w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right) \\ &= w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right)^+ \end{aligned} \tag{19}$$

where σ_W is given in (10) and

$$a_W(n) = \sqrt{n}w/\sigma_W. \tag{20}$$

Again, the means, variances and distributions are then approximated in the obvious way, while the PoD and PoA are natural analogs of (15) and (16), i.e.,

$$\begin{aligned} PoD_n &\approx PoD_n^{TGA} \equiv \mathbb{P}\left(W_n^{TGA} > 0\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > 0\right) \\ &= \Phi\left(-a_W(n)\right), \end{aligned} \tag{21}$$

$$\begin{aligned} PoA_n &\approx PoA_n^{TGA} \equiv \mathbb{P}\left(W_n^{TGA} > A\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > A\right) \\ &= \int_0^\infty \Phi\left(a_W(n)\left(\frac{x}{w} - 1\right)\right) f(x)\mathrm{d}x, \end{aligned} \tag{22}$$

where $\Phi(\cdot)$ is the ccdf of standard normal distribution and $f(\cdot)$ is the probability density function (pdf) of patience time.

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

$$(X_n^{TGA}, E[X_n^{TGA}], \mathrm{Var}(X_n^{TGA}), PoA^{TGA}, PoD^{TGA}) = (X_n^{DGA}, E[X_n^{DGA}], \mathrm{Var}(X_n^{DGA}), PoA^{DGA}, PoD^{DGA})$$

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. Because of the M service, the service-completion process is a Poisson process with rate $\mu s(t)$, which yields Brownian FCLT limit, corresponding to the BM $\mathcal{B}_2 = \mathcal{B}_s$ in Theorem 2.2 (See Theorem 4.2 in [21] for details; also see (4.5) and (6.64) there). When service becomes GI, the prelimit of the service-completion process is much more complicated (no longer Poisson). We capture the nonexponential service distribution using the first and second moments. Specifically, we scale the BM $\mathcal{B}_2$ by the service-time squared coefficient of variation (scv, variance divided by the square of the mean) c_s, which will replace the "1" by "c_s" in the numerator of (10), yielding

$$\sigma^2_{W_G} \quad = \quad \frac{(c^2_\lambda - 1) + (c_s + 1)\rho}{2s\rho^2 f(w)}. \tag{23}$$

By replacing σ_W by σ_{W_G} in (10), (17)–(19), we obtain TGA-G. It is significant that TGA-G reduces to TGA when service is M, for which $c^2_s = 1$.

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 [35]. That algorithm was developed to treat more general models, 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 12 in the appendix. We thus start by showing the experimental results for $\rho = 1.2$ with abandonment rates 0.5, 1.0 and 2.0 in Table 1, where $M(m)$ refers to an exponential distribution with mean m. For the lower two abandonments rates ($\theta \leq 1.0$), 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 5%. However, for $\theta = 2.0$, we see that TGA provides some improvement.

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 [11, 8]. Table 2 shows the main performance measures for the three abandonment rates $\theta = 0.5$, 1.0 and 2.0. Table 2 shows that DGA performs poorly in this case, but TGA provides dramatic improvement, having all errors less than 10%.

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. However, we do point out that, even for the $M/M/n+M$

Table 1: $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ with $\lambda = 100\rho = 120$ and $\theta = 0.5$, 1 and 2

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
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%	

model, the TGA approximation degrades as a function of the parameter pair (ρ, θ) as either ρ decreases further toward 1.0 or as θ moves away from the range $(0, 2.0]$. 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 PoD and PoA for TGA was 7%. On the other hand, Table 15 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., $E[Q] = 1.47$ and $E[W] = 0.017$ for $\theta = 4$.

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 [21]. For these and all other non-Markov models, we will use simulation to estimate the exact values of the performance measures. We describe our simulation methodology in §8.

Table 2: $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ with $\lambda = 100\rho = 105$ and $\theta = 0.5$, 1 and 2

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E$[X]$ rel. err.	1.52E+2	1.50E+2 1%	same	1.22E+2	1.20E+2 2%	same	1.11E+2	1.10E+2 1%	same
Var(X) rel. err.	9.25E+2	1.05E+3 16%	same	3.47E+2	4.20E+2 21%	same	1.81E+2	2.10E+2 16%	same
E$[Q]$ rel. err.	5.22E+1	4.99E+1 4%	5.08E+1 2%	2.30E+1	1.99E+1 14%	2.17E+1 6%	1.27E+1	9.94E+0 22%	1.21E+1 5%
Var(Q) rel. err.	8.99E+2	1.05E+3 19%	9.41E+2 7%	3.05E+2	4.20E+2 37%	3.11E+2 2%	1.35E+2	2.10E+2 56%	1.33E+2 1%
E$[W]$ rel. err.	5.14E-1	4.88E-1 5%	4.96E-1 3%	2.29E-1	1.95E-1 15%	2.12E-1 8%	1.28E-1	9.80E-2 23%	1.18E-1 7%
Var(W) rel. err.	8.59E-2	1.00E-1 19%	8.97E-2 7%	2.93E-2	4.00E-2 36%	2.97E-2 1%	1.31E-2	2.00E-2 53%	1.27E-2 3%
PoD rel. err.	9.67E-1	9.39E-1 3%	same	8.90E-1	8.35E-1 6%	same	8.03E-1	7.56E-1 6%	same
PoA rel. err.	4.97E-2	4.80E-2 4%	same	5.47E-2	5.09E-2 7%	same	6.04E-2	5.60E-2 8%	same
	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E$[X]$ rel. err.	1.05E+2	1.05E+2 0%	same	1.01E+2	1.02E+2 1%	same	9.86E+1	1.01E+2 3%	same
Var(X) rel. err.	1.05E+2	1.05E+2 0%	same	6.85E+1	5.24E+1 23%	same	5.00E+1	2.61E+1 48%	same
E$[Q]$ rel. err.	7.03E+0	4.92E+0 30%	7.01E+0 0%	3.88E+0	2.36E+0 40%	4.22E+0 8%	2.12E+0	1.13E+0 47%	2.65E+0 25%
Var(Q) rel. err.	5.92E+1	1.05E+2 77%	5.71E+1 3%	2.57E+1	5.24E+1 103%	2.50E+1 3%	1.10E+1	2.61E+1 139%	1.13E+1 3%
E$[W]$ rel. err.	7.22E-2	4.90E-2 32%	6.91E-2 4%	4.09E-2	2.40E-2 42%	4.18E-2 2%	2.32E-2	1.20E-2 48%	2.65E-2 14%
Var(W) rel. err.	5.88E-3	1.00E-2 70%	5.49E-3 7%	2.64E-3	5.00E-3 89%	2.42E-3 9%	1.18E-3	2.50E-3 112%	1.10E-3 6%
PoD rel. err.	7.00E-1	6.88E-1 2%	same	5.92E-1	6.33E-1 6%	same	4.87E-1	5.95E-1 22%	same
PoA rel. err.	6.70E-2	6.43E-2 3%	same	7.40E-2	7.59E-2 3%	same	8.07E-2	9.31E-2 16%	same

5.1 Non-Poisson arrivals

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 stardard 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 shows the experimental results for five values of c_λ^2 with $0.25 \leq c_\lambda^2 \leq 4.0$. 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 $LN(\lambda^{-1}, 1)$, we see that the interarrival-time distribution matters beyond its mean and variance.

Table 3: $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$

	$c_\lambda^2 = 0.25$			$c_\lambda^2 = 0.5$			$c_\lambda^2 = 1$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X] rel. err.	1.11E+2 ±1.16E-1	1.10E+2 1%	same	1.11E+2 ±1.25E-1	1.10E+2 1%	same	1.11E+2 ±1.44E-1	1.10E+2 1%	same
Var(X) rel. err.	1.16E+2 ±2.62E+1	1.31E+2 13%	same	1.38E+2 ±2.83E+1	1.57E+2 14%	same	1.80E+2 ±3.27E+1	2.10E+2 17%	same
E[Q] rel. err.	1.15E+1 ±1.04E-1	9.94E+0 14%	1.12E+1 3%	1.20E+1 ±1.10E-1	9.94E+0 17%	1.15E+1 5%	1.27E+1 ±1.24E-1	9.94E+0 22%	1.21E+1 5%
Var(Q) rel. err.	9.22E+1 ±3.18E+0	1.31E+2 42%	9.23E+1 0%	1.07E+2 ±3.62E+0	1.57E+2 48%	1.06E+2 0%	1.33E+2 ±4.46E+0	2.10E+2 58%	1.33E+2 0%
E[W] rel. err.	1.18E-1 ±1.05E-3	9.80E-2 17%	1.10E-1 7%	1.23E-1 ±1.10E-3	9.80E-2 20%	1.13E-1 8%	1.28E-1 ±1.22E-3	9.80E-2 23%	1.18E-1 7%
Var(W) rel. err.	9.46E-3 ±1.05E-3	1.29E-2 36%	9.03E-3 5%	1.07E-2 ±1.10E-3	1.52E-2 42%	1.03E-2 4%	1.30E-2 ±1.22E-3	2.00E-2 54%	1.27E-2 2%
PoD rel. err.	8.22E-1 ±3.02E-3	8.06E-1 2%	same	8.08E-1 ±3.15E-3	7.86E-1 3%	same	7.80E-1 ±3.31E-3	7.56E-1 3%	same
PoA rel. err.	5.60E-2 ±6.45E-4	5.36E-2 4%	same	5.83E-2 ±6.70E-4	5.49E-2 6%	same	6.01E-2 ±6.99E-4	5.75E-2 4%	same

	$M, c_\lambda^2 = 1$			$c_\lambda^2 = 2$			$c_\lambda^2 = 4$		
Perf.	Exact	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X] rel. err.	1.05E+2	1.05E+2 0%	same	1.12E+2 ±1.69E-1	1.10E+2 2%	same	1.13E+2 ±2.17E-1	1.10E+2 3%	same
Var(X) rel. err.	1.05E+2	1.05E+2 5%	same	2.60E+2 ±3.88E+1	3.15E+2 21%	same	4.08E+2 ±5.02E+1	5.25E+2 29%	same
E[Q] rel. err.	7.03E+0	4.92E+0 30%	7.01E+0 0%	1.39E+1 ±1.40E-1	9.94E+0 29%	1.31E+1 6%	1.61E+1 ±1.73E-1	9.94E+0 38%	1.50E+1 7%
Var(Q) rel. err.	5.92E+1	1.05E+2 86%	5.71E+1 2%	1.79E+2 ±5.81E+0	3.15E+2 76%	1.82E+2 2%	2.61E+2 ±8.52E+0	5.25E+2 101%	2.75E+2 6%
E[W] rel. err.	7.22E-2	4.90E-2 32%	6.91E-2 4%	1.38E-1 ±1.36E-3	9.80E-2 29%	1.28E-1 7%	1.56E-1 ±1.63E-3	9.80E-2 37%	1.45E-1 7%
Var(W) rel. err.	5.88E-3	1.00E-2 79%	5.49E-3 2%	1.68E-2 ±1.36E-3	2.95E-2 76%	1.72E-2 3%	2.32E-2 ±1.63E-3	4.86E-2 110%	2.57E-2 11%
PoD rel. err.	7.00E-1	6.88E-1 4%	same	7.51E-1 ±3.46E-3	7.16E-1 5%	same	7.26E-1 ±3.76E-3	6.72E-1 8%	same
PoA rel. err.	6.70E-2	6.68E-2 1%	same	6.46E-2 ±7.69E-4	6.22E-2 4%	same	7.29E-2 ±8.58E-4	7.02E-2 4%	same

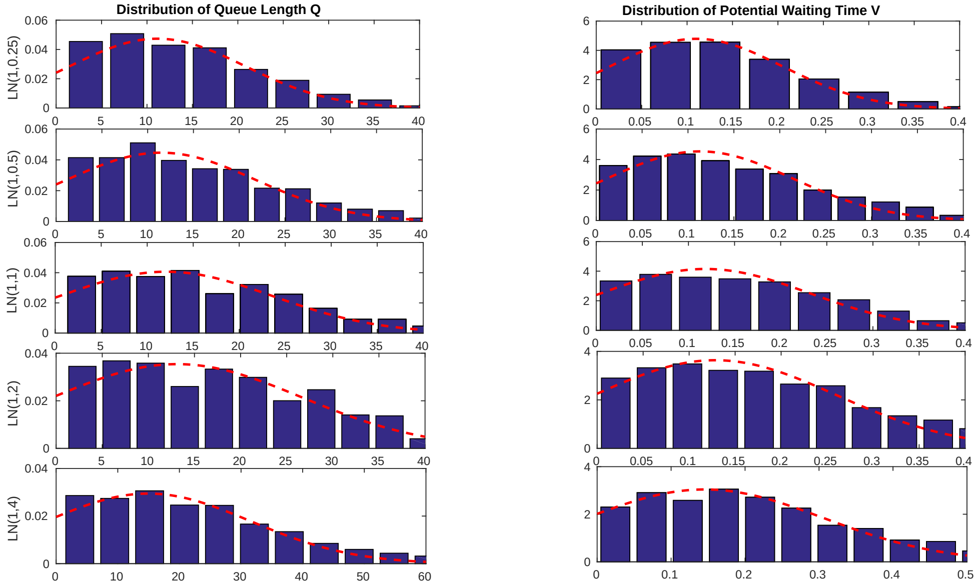

Figure 1: TGA Approximating Distributions of Q_n and V_n for the $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$

Second, Table 3 shows that, just like Table 2, TGA performs consistently well, whereas DGA does not. Figure 1 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) $\{\Gamma(t), t \geq\}$, e.g., see [7]. 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 $\{\Gamma(t), t \geq 0\}$ that is a 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 yild 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) $c_\lambda^2 = 1.5$, where

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

See Proposition 1 in [30] and also see [12].

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 DGA, 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 [35, 36]. 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. 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.

Table 4: $MMPP(\lambda^{-1}, c_\lambda^2)/M/100 + M(\theta^{-1})$ with $(\lambda, \rho, \theta) = (105, 1.05, 0.5)$.

	$c_\lambda^2 = 1.5$		$c_\lambda^2 = 2$		$c_\lambda^2 = 4$		$c_\lambda^2 = 10$	
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%

Table 5: $M/M/100 + LN(\theta^{-1}, c_{ab}^2)$ with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

	$LN(2, 0.25)$			$LN(2, 0.5)$		
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: Impact of the abandonment distribution beyond its mean and variance: 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] rel. err.	1.05E+2	1.05E+2 0%	same	1.32E+2 ±2.79E-1	1.36E+2 3%	same
Var(X) rel. err.	1.05E+2	1.05E+2 0%	same	3.58E+2 ±7.27E+1	3.18E+2 11%	same
E[Q] rel. err.	7.03E+0	4.92E+0 30%	7.01E+0 0%	3.24E+1 ±2.67E-1	3.65E+1 13%	3.66E+1 13%
Var(Q) rel. err.	5.92E+1	1.05E+2 77%	5.71E+1 3%	3.32E+2 ±1.74E+1	3.18E+2 4%	3.07E+2 8%
E[W] rel. err.	7.22E-2	4.90E-2 32%	6.91E-2 4%	3.15E-1 ±2.54E-3	3.53E-1 12%	3.54E-1 12%
Var(W) rel. err.	5.88E-3	1.00E-2 70%	5.49E-3 7%	2.98E-2 ±2.54E-3	2.82E-2 6%	2.73E-2 9%
PoD rel. err.	7.00E-1	6.88E-1 2%	same	9.48E-1 ±2.19E-3	9.82E-1 4%	same
PoA rel. err.	6.70E-2	6.43E-2 3%	same	5.08E-2 ±7.74E-4	6.04E-2 19%	same

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 an impact on the queueing system beyond its mean and variance, just as for the interarrival-time distribution. Approximations of distributions of Q_n and W_n are given in Figure 2.

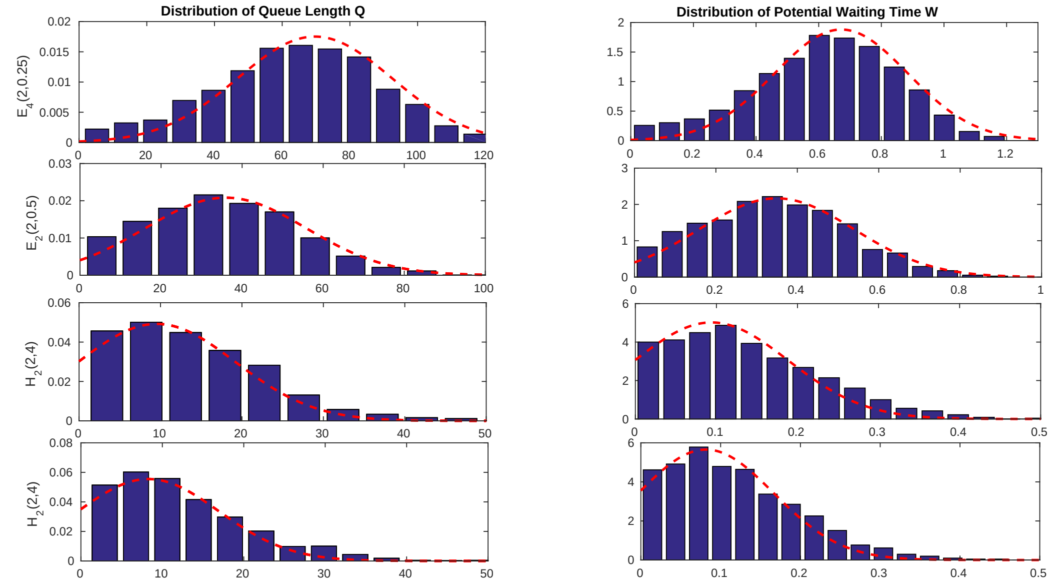

Figure 2: TGA Approximating Distributions of Q_n and W_n for the $M/M/100+LN(\theta^{-1},c_{ab}^2)$ with $(\lambda,\rho,\theta)=(100,1.05,0.5)$

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 [28] for more details.

Table 7: $M(\lambda^{-1})/PH(1, c_s^2)/100 + M(\theta^{-1})$ with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$

	$c_s^2 = 0.25$				$c_s^2 = 0.5$			
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%		

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 3 shows the corresponding approximations of the distributions.

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

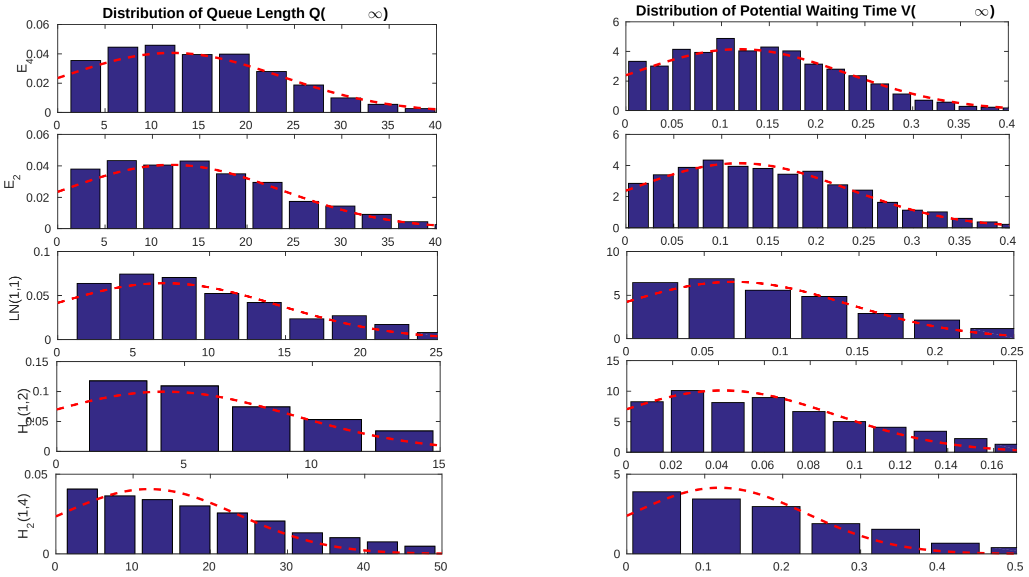

Figure 3: Approximating the distributions of Q_n and V_n for $M(105^{-1})/PH/100 + M(2)$ with c_s^2 ranges from 0.25, 0.5, 1, 2, to 4

$0.25 \leq \theta \leq 2.0$. More generally, we observe that the performance of TGA-G is acceptable for a range of the abandonment rate $\theta \in [0.1, 2]$. However, just as for the $M/M/n + M$ model, the approximation accuracy starts to degrade when θ increases; see the appendix for more examples.

Table 8: $M(\lambda^{-1})/H_2(1,2)/100 + M(\theta^{-1})$ with $\lambda = 100$, $\rho = 105$ and $0.5 \leq \theta \leq 2$

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

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: $0.25 \leq c_s^2 \leq 4.0$.

Just as in Table 7, Table 9 shows that the mean values such as EQ 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 14 in the appendix.

Table 9: $H_2(\lambda^{-1}, 2)/PH/100 + H_2(1/\theta, 2)$ with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

	$c_s^2 = 0.25$			$c_s^2 = 0.5$		
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 such as DGA and TGA-G should perform better as the scale increases. Thus, we considered the base case with $n = 100$ as a good case, 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

$$\sqrt{n}\,(1 - \rho_n) \approx \beta, \quad -\infty < \beta < \infty, \tag{24}$$

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 as we change n. In particular, we now consider smaller scaling n, but we also change the value of ρ_n so that the QoS factor remains fixed at $\beta = -0.5$. 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 corresponds to traffic intensities $\rho_n = 1.11, 1.16, 1.22$.

Table 10: $H_2(\lambda^{-1})/H_2(1,2)/n + H_2(\theta^{-1}, 2)$ with $\rho = 1 - \beta/\sqrt{n}$, $\lambda = n\rho$

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

Table 10 shows that TGA-G remains effective for systems with fewer servers. Complementing Table 10, Figure 4 shows that TGA-G can be used to approximate the distributions of key performance measures. Figure 4 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.

8 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,

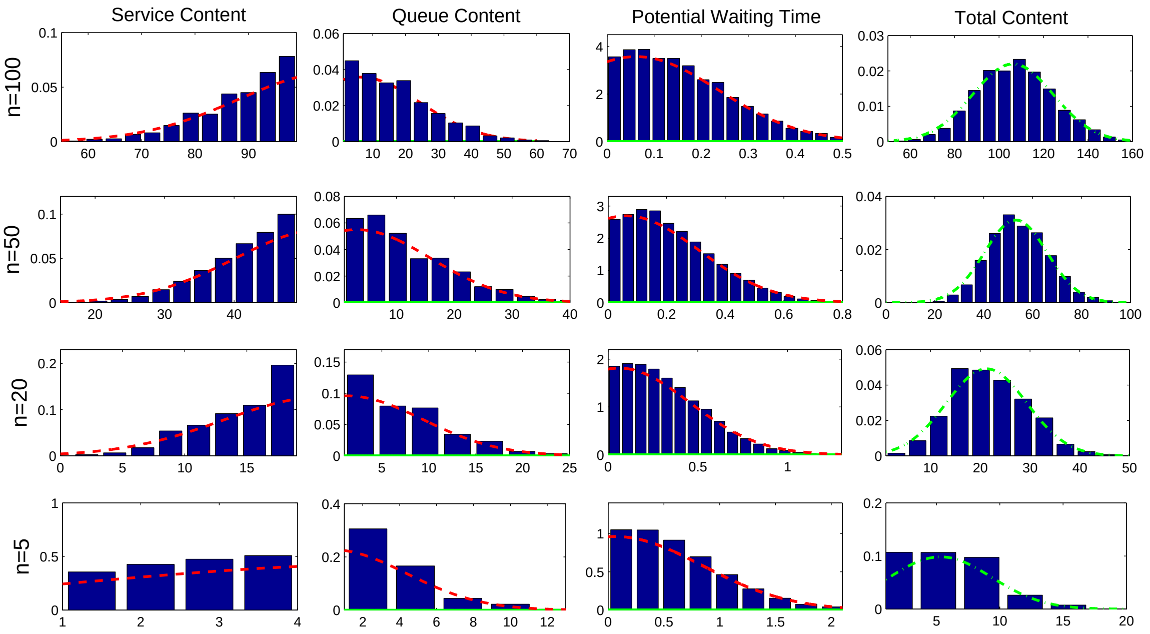

Figure 4: Approximating the distributions of B_n, Q_n, V_n and X_n for $H_2(\lambda^{-1}, 2)/M/n + H_2(\theta^{-1}, 2)$ with $\theta = 0.5$.

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.

8.1 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 5 shows plots of the transient (time-dependent) mean and variance functions for the queue length. Figure 5 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.

8.2 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, $1 \leq r \leq R$, we periodically generate virtual arrivals at deterministic

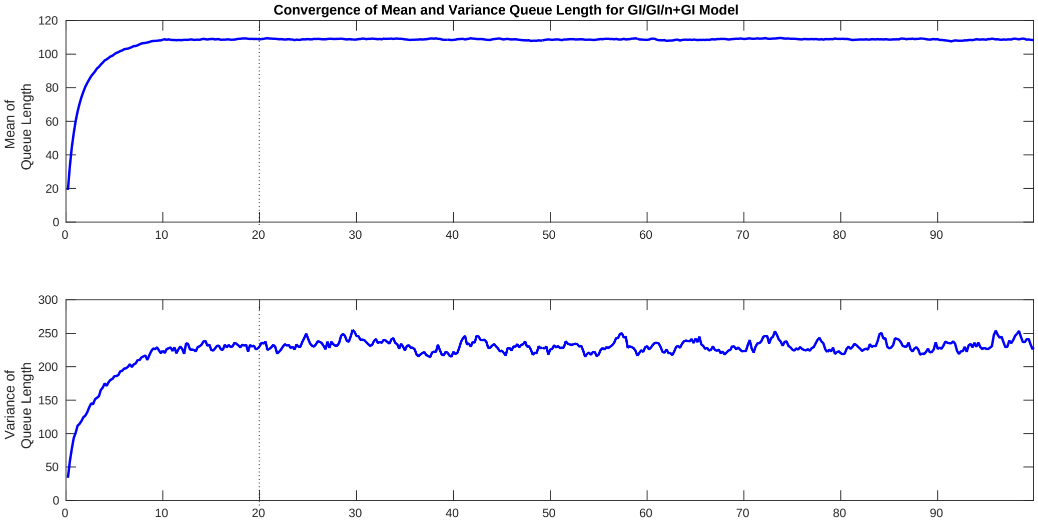

Figure 5: Converging to steady state of the $H_2/H_2/n + H_2$ Model

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

$$\eta^d_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > 0, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \eta^a_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > A_k, \\ 0, & \text{otherwise} \end{cases}, \tag{25}$$

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.

8.3 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

$$\widetilde{Q}_r \equiv \frac{1}{N}\sum_{l=1}^{N} Q_{r,l} \quad \text{and} \quad \widetilde{Q}^{(2)}_r \equiv \frac{1}{N}\sum_{l=1}^{N} (Q_{r,l})^2. \tag{26}$$

Even though the random variables being averaged in (26) are typically dependent, these are valid estimators for the true mean $E[Q]$ and second moment $E[Q^2]$. 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 $E[Q]$ and second moment $E[Q^2]$, and their CI's, we use the R independent samples $\widetilde{Q}_1, \ldots, \widetilde{Q}_R$ ($\widetilde{Q}_1^{(2)}, \ldots, \widetilde{Q}_R^{(2)}$) to compute the sample mean and sample variance of the queue length and its second moment in the usual way, i.e.,

$$\bar{Q}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}_r \quad \text{and} \quad S_Q^2(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}_r - \bar{Q}(R)\right)^2, \tag{27}$$

$$\bar{Q}^{(2)}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}_r^{(2)} \quad \text{and} \quad S_{Q^{(2)}}^2(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}_r^{(2)} - \bar{Q}^{(2)}(R)\right)^2. \tag{28}$$

The random variables $\bar{Q}(R)$ and $\bar{Q}^{(2)}(R)$ in (27) our our final estimators for the true mean $E[Q]$ and second moment $E[Q^2]$.

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

$$\left[\bar{Q}(R) - z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}, \bar{Q}(R) + z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}\right], \quad \text{and} \tag{29}$$

$$\left[\bar{Q}^{(2)}(R) - z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}, \bar{Q}^{(2)}(R) + z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}\right], \tag{30}$$

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

$$\bar{V}(R) \equiv \bar{Q}^{(2)}(R) - \left(\bar{Q}(R)\right)^2. \tag{31}$$

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

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

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 [35]. The Table 11 shows that the procedures are sound and the the statistical precision is adequate.

Table 11: Comparison between Exact Values for the $M(102^{-1})/M(1)/100+M(\theta^{-1})$ with Simulation Estimates

	$\theta = 0.1$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
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

We use Table 11 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.

9 Conclusion

In this paper we have developed and evaluated approximations for the key steady-state performance measures in the stationary $G/GI/n+GI$ model. These approximations are based on many-server heavy-traffic (MSHT) limits in [19, 20, 21, 36], but also involve important heuristic refinements. These 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.

After reviewing the underlying MSHT limits in §2, 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 [21] 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 §8 we described the simulation methodology. Additional details are provided in an online appendix.

A main conclusion in [35, 36] 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.

There are many good directions for future research. It remains to provide better theoretical justification, especially for the non-exponential service-time distributions. It also remains to develop effective approximations in other ranges of the parameters.

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

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

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] Liu, Y., Whitt, W. and Yu, Y. (2015). Appendix to: Approximations for heavily-loaded $G/GI/n + GI$ queues. Http://yunanliu.wordpress.ncsu.edu/files/2014/02/LiuWhittYuApprox090215app.pdf.

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

[24] 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.

[25] Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: The palm/erlang-a queue, with applications to call centers. In D., S. and Fhnrich, K.-P. (eds.), *Advances in Services Innovations*. Springer, pp. 17–48.

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

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

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

[29] 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.

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

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

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

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

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

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

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

[37] 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.

[38] 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.