Probability in the Engineering and Informational Sciences, page 1 of 23, 2020.
doi:10.1017/S0269964820000509

SET-VALUED PERFORMANCE APPROXIMATIONS FOR THE $GI/GI/K$ QUEUE GIVEN PARTIAL INFORMATION

YAN CHEN AND WARD WHITT

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA

E-mails: yc3107@columbia.edu; ww2040@columbia.edu

In order to understand queueing performance given only partial information about the model, we propose determining intervals of likely values of performance measures given that limited information. We illustrate this approach for the mean steady-state waiting time in the $GI/GI/K$ queue. We start by specifying the first two moments of the interarrival-time and service-time distributions, and then consider additional information about these underlying distributions, in particular, a third moment and a Laplace transform value. As a theoretical basis, we apply extremal models yielding tight upper and lower bounds on the asymptotic decay rate of the steady-state waiting-time tail probability. We illustrate by constructing the theoretically justified intervals of values for the decay rate and the associated heuristically determined interval of values for the mean waiting times. Without extra information, the extremal models involve two-point distributions, which yield a wide range for the mean. Adding constraints on the third moment and a transform value produces three-point extremal distributions, which significantly reduce the range, producing practical levels of accuracy.

Keywords: bounds, extremal queues, mean waiting time, multi-server queues, performance approximations, queues

1. INTRODUCTION

Despite many significant research contributions in queueing theory over the years, what Kingman [34] wrote 50 years ago largely remains true today:

> It is a fair criticism of the theory of queues as it has been developed over the years that, even in the simple cases for which explicit analytical solutions can be found, these solutions are too complicated to be of practical use. It has been argued elsewhere [33] that the criticism is to be met to some degree by the analysis of situations where robust approximations exist, such as that of "heavy traffic." It is, however, important to know how accurately such approximations represent the true solution, and the significance of inequalities for the various quantities of interest thus become apparent.

Just as Kingman [34] did, we consider this problem for the $GI/GI/K$ queue, which is a K-server queue with unlimited waiting room and service in order of arrival by the

 0269-9648/19 $25.00

first available server, where the interarrival times and service times come from independent sequences of independent and identically distributed (i.i.d.) random variables distributed as U and V with general cumulative distribution functions (cdf's) F and G. We are especially interested in the performance impact of the variability of these underlying cdf's F and G (both assumed to have finite first two moments). To describe the extent of the variability independent of the mean, we let c_a^2 and c_s^2 be the squared coefficient of variation (scv, variance divided by the square of the mean) of U and V. We start by considering the special case $K = 1$, but it is significant that our approach extends directly to $K > 1$.

The complication is well illustrated by the formula for the mean of steady-state waiting time W (before starting service) for $K = 1$,

$$E[W] = \sum_{k=1}^{\infty} \frac{E[S_k^+]}{k} < \infty, \tag{1.1}$$

where $[x]^+ \equiv \max\{x, 0\}$, S_k is the kth partial sum of k i.i.d. random variables distributed as $X \equiv V - U$ and $\equiv$ means equality by definition; see Sect. X.2 of [4] and Sect. 8.5 of [12]. Formula (1.1) is mathematically elegant but not convenient for computation. Formula (1.1) is reviewed in Ref. [2], which is devoted to algorithms to compute $E[W]$ and $P(W > t)$ when $K = 1$ for general F and G based on alternative integral representations. In general, simulation remains an attractive method, although it applies to only one specified model, does not yield the insight of formulas, and is a relatively time-consuming numerical procedure. Unfortunately, the situation is much worse for $K > 1$ because there is no analog of (1.1) that has been found for $K > 1$.

A candidate simple and insightful approximation formula for $E[W]$ is provided by the heavy-traffic approximation (HTA). Choose measuring units by setting $E[U] = 1$, so that $E[V] = \rho K$, where ρ is the traffic intensity. Then, the second moments are $E[U^2] = c_a^2 + 1$ and $E[V^2] = \rho^2 K^2 (c_s^2 + 1)$. In this context, the HTA for the mean with $K \geq 1$ is

$$E[W] \approx \frac{\rho^2 (c_a^2 + c_s^2)}{2(1 - \rho)}. \tag{1.2}$$

For $K = 1$, the HTA in (1.2) is obtained by combining the $M/GI/1$ Pollaczek–Khintchine exact formula for the special case of a Poisson arrival process, where $c_a^2 = 1$, with the heavy-traffic limit in [30]. The extension to $K > 1$ was provided in Refs. [5,24,25,36]. (We do not consider the many-server heavy-traffic scaling in [22] or [18].) The limit shows that the approximation is asymptotically correct in the sense that

$$E[W] = \text{HTA} + o(1 - \rho) \quad \text{as } \rho \uparrow 1, \tag{1.3}$$

where $o(x)$ is a quantity $h(x)$ such that $h(x)/x \to 0$ as $x \to 0$.

In this setting, the problem posed in [34] can be expressed as follows: How accurate is formula (1.2)? First, it is well known that, consistent with (1.3), the accuracy of (1.2) improves as ρ increases; for asymptotic theory, see Thm. XIII.6.7 of [4]. More generally that question is answered for the case $K = 1$, at least in part, by the large literature on bounds for $E[W]$, given the partial specification by the parameter 4-tuple

$$(E[U], E[U^2], E[V], E[V^2]) \equiv (1, c_a^2, \rho, c_s^2), \tag{1.4}$$

starting from Refs. [14,16,31,34,48] and continuing with Refs. [6,9] and the many references therein.

Unfortunately, however, this program has not yet been very successful. As shown by Table 1 in Ref. [9], the range of possible values of the mean $E[W]$ in the $GI/GI/1$ model given the first two moments of U and V is quite wide and so is of limited value. (We elaborate on this important point in Sect. 2 of the Appendix [7].) Consequently, we need to use more information about F and G. However, relatively little is known about the impact of additional information, beyond the early results for the $GI/M/1$ model in [44,45], [35] and queues with phase-type distributions in [26,27]. Almost nothing is known about the case $K > 1$, but it is known that the range given the first two moments of U and V is even wider; see Refs. [15,21].

The purpose of this paper is to investigate a new approach to obtain useful set-valued approximations for the steady-state mean $E[W]$ given partial information about the two underlying distributions, which addresses two difficulties: (i) producing a useful smaller set of possible values than is possible with the first two moments and (ii) applying to the challenging multi-server case $K > 1$ as well as $K = 1$.

We show that it is possible to address both of these difficulties by taking an indirect approach. Instead of focusing directly on the mean, we first focus on the (asymptotic) decay rate of the steady-state waiting-time distribution because it is more tractable. For the decay rate, we apply recent tight upper and lower bounds established in [8] by applying the theory of Tchebycheff systems. We use the extremal models for the decay rate to create likely intervals for the mean $E[W]$. So far, this indirect approach is heuristic, because while we produce valid performance guarantees for the decay rate, we do not produce performance guarantees for the associated mean. Nevertheless, we are able to produce useful intervals of likely values for the mean, given only two more parameters for each of the underlying distributions: the third moment and a single value of its Laplace transform.

This study is not only useful for understanding simple approximations such as (1.2). It is also useful when we fit a specific model to data, such as a $Ph/Ph/K$ model with phase-type distributions, and then compute the exact value of the mean $E[W]$ for that model by simulation or a numerical algorithm. Even in that case, it is natural to ask how the performance depends on limited information.

Here is how the rest of the paper is organized: In Section 2, we give an overview of the supporting theory and our approximation procedure. In Section 3, we review the supporting theory for the decay rate. In Section 4, we apply that the theory to develop the set-valued approximations for the mean, as outlined above. In Section 5, we conduct simulation experiments to evaluate the procedure. The tables in Section 5 show that the heuristic procedure above with an appropriate parameter choice is effective, but we regard it as a proof of concept rather than the final word. In Section 6, we draw conclusions. We give a concise summary of the approximation procedure in Section 6.1. Additional supporting material appears in Ref. [7].

2. OVERVIEW

In Section 2.1, we explain the theoretical basis in terms of the decay rate. In Section 2.2, we give a quick overview of our proposed procedure.

2.1. Applying Tchebycheff Systems to the Asymptotic Decay Rate

In order to make progress, we propose a new approach based on the asymptotic decay rate. To do so, we restrict attention to the light-tailed case, where the service-time cdf G has

finite moments of all orders. We then typically have

$$P(W > t) \sim \alpha e^{-\theta_W t} \quad \text{as } t \to \infty, \tag{2.1}$$

where $f(t) \sim g(t)$ as $t \to \infty$ means that $f(t)/g(t) \to 1$. Then, we call θ_W the (asymptotic) decay rate and have the rough approximations $E[W] \approx \alpha/\theta_W \approx 1/\theta_W$; for example, see Ref. [3]. The key observation is that, in great generality, but under regularity conditions, the decay rate θ_W in (2.1) is attained as the unique positive real root of an equation involving the Laplace transforms of U and V, for example, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$, see Ref. [3]. In particular, the equation for the decay rate is

$$\hat{f}(s)\hat{g}(-s) = 1. \tag{2.2}$$

In this light-tailed setting, in Ref. [8] we have shown that the theory of Tchebycheff (T) systems from Ref. [28], as used in Refs. [17,20,23,26,27,41,44,45], can be applied to determine extremal models (yielding tight upper and lower bounds) on the decay rate θ_W above. The T-system theory is a refinement of the classical moment problem which provides conditions for identifying the probability measure P obtaining the supremum of an integral $\int_a^b \phi\, dP$ given a finite number of integral constraints of the form $\int_a^b u_i\, dP = m_i$. The extremal probability measures have a prescribed form if the set of functions $\{\phi, u_0, \ldots, u_m\}$ have a special form; see Sect. 2 of [8] for a concise review suitable for the results here. As indicated by Lemma 2.2 of [8], the relevant T system here arises when the functions ϕ and u_i are either moments or values of the Laplace transform. That structure suffices for the asymptotic decay rate by virtue of (2.2). In Section 3, we review the extremal results for the decay rate that we will apply. See (3.3) for the extension to $K > 1$.

2.2. Application to Reveal Likely Intervals for the Mean Waiting Time

We now briefly describe the basic procedure for generating likely intervals for the mean $E[W]$ in the $GI/GI/1$ model. We start by specifying the basic parameter vector $(1, c_a^2, \rho, c_s^2)$ in (1.4). Next, we specify a reference base $GI/GI/1$ model with those parameters, depending on the pair of cdf's (F, G). For that reference model, we determine the third moments $m_{a,3}$ of F and $m_{s,3}$ of G and the asymptotic decay rate $\theta_W \equiv \theta_W(F, G)$.

Now, we come to the heuristic steps. In order to apply Theorem 3.2 for the decay rate to generate an interval of likely decay rates, we need to specify a finite upper bound on the support and a Laplace transform value for each of the distributions F and G. Let the support bounds be M_a for F and ρM_s for G. Let the arguments of the Laplace transforms be μ_a for F and μ_s for G, so that we are specifying $\hat{f}(\mu_a)$ with $\mu_a > 0$ for F and $\hat{g}(-\mu_s)$ with $0 < \mu_s < s^*$ for G, where s^* is a theoretical limit specified in Assumption 3.1.

The remaining parameter vector (μ_a, M_a, μ_s, M_s) in addition to (1.4) is specified to obtain an effective heuristic interval of likely values for the mean $E[W]$. For that purpose, we use two positive tuning parameters ϵ and R. We emphasize that this is a heuristic step, so some judgment is required. For understanding, it is good to carry out the procedure for a few candidate values of ϵ and R.

First, Theorem 3.2 for the decay rate requires finite support bounds for F and G. Thus, our approach is to choose support bounds that should not affect the results much. With that rough goal in mind, we let ϵ be a small value such as $\epsilon = 0.001$. Then, we choose M_a and M_s to satisfy

$$P(U > M_a E[U]) = P(U > M_a) = P(V > M_s E[V]) = P(V > \rho M_s) = \epsilon. \tag{2.3}$$

Next, we specify the arguments μ_a and μ_s. We examine all possible orderings compared to θ_W. After study, we suggest one of the two orderings:

$$\mu_s, \mu_a \le \theta_W \quad \text{and} \quad \mu_s \le \theta_W \le \mu_a, \tag{2.4}$$

where θ_W is the reference base decay rate. Theorem 3.2 implies that the interval of possible decay rates decreases as these arguments approach the base decay rate θ_W. Thus, for any given ordering, we let

$$\mu \equiv \theta_W/R \quad \text{if } \mu \le \theta_W \quad \text{and} \quad \mu \equiv R\theta_W \quad \text{if } \mu \ge \theta_W \tag{2.5}$$

for suitable R, such as $R = 20$. The exact interval of possible decay rates decreases to the single base value θ_W as $R \downarrow 1$. Because there is no simple relation between the mean $E[W]$ and the decay rate θ_W, it is important that we not try to make the interval for the decay rate too short. Accordingly, it is important that R not be too small.

3. REVIEW OF EXTREMAL MODELS FOR THE ASYMPTOTIC DECAY RATE

In this section, we review the supporting theory for the decay rate determined in Ref. [8]. In Section 3.1, we provide additional background on the decay rate θ_W in (2.1). In Section 3.2, we exhibit the two-point extremal models given two moments and finite support for U and V. In Section 3.3, we exhibit the three-point extremal models given three moments, a Laplace transform value and a support bound for U and V. In Section 3.4, we establish the results for the case of unbounded support.

3.1. Background on the Decay Rate

To increase the level of generality, instead of (2.1), we can let θ_W be defined by the critical exponent in the Kingman-Lundberg bound for the $GI/GI/1$ queue, as in Ref. [32] and Sect. XIII.5 of [4], defined by

$$\theta_W \equiv \inf\{x \ge 0 : P(W > t) \le e^{-xt},\ t \ge 0\}, \tag{3.1}$$

so that large waiting times correspond to small values of θ_W. Under regularity conditions, θ_W in (3.1) coincides with the asymptotic decay rate studied in large-deviations theory, defined by

$$\theta_W \equiv \lim_{x\to\infty} \frac{-\log P(W > x)}{x}. \tag{3.2}$$

We assume that a strictly positive infimum exists in (3.1) and a strictly positive limit exists in (3.2), which requires that the service-time V must have a finite moment generating function $E[e^{sV}]$ for some $s > 0$. (We obtain $\theta_W = \infty$ if $P(V - U \le 0) = 1$ and thus $P(W = 0) = 1$.) Thus, we are considering the light-tail case as in the discussion of exponential change of measure in Chap. XIII in [4], large deviation limits in Corollary 1 in Sect. 1.2 of [19] and approximations in [3]. More about the asymptotic decay rate can be found in discussions of the caudal characteristic curve of queues in [39] and effective bandwidths in [11,29,46] and references therein.

As stated in the Introduction section, for our queueing application, the key observation is that, under regularity conditions, the asymptotic decay rate θ_W in (2.1), (3.1), or (3.2) is attained as the unique positive real root of Eq. (2.2) involving the Laplace transforms of U and V, for example, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$. Equivalently, as in Sect. XIII.1 of [4],

$\kappa_F(s) + \kappa_G(-s) = 0$, where $\kappa_F(s) \equiv \log(\hat{f}(s))$ is the cumulant generating function. (The function $\hat{g}(-s) \equiv E[e^{sV}]$ for $s > 0$ is the moment generating function (mgf).)

Indeed, it is well known that the distribution of W depends on $V - U$, which has Laplace transform $\hat{f}(-s)\hat{g}(s)$. Moreover, Chapter II.5 of [13] shows that the distribution of W can be characterized by all complex roots of equations related to (2.2).

Given the simple structure in (2.2), the extremal result and alternative ones follow from the theory of T systems, as reviewed in Ref. [8]. To state the results, we impose some technical conditions. In contrast to the mean $E[W]$, which is finite for all models given the partial moment information in (1.4), as can be seen from Sect. X.2 of [4], the decay rate is not well defined for all these models. Hence, in order to establish the extremal results for the decay rate in (3.1) given the partial moment information in (1.4), we make the following assumption.

ASSUMPTION 3.1 (finite moment generating function): *Assume that there exists* s^*, $0 < s^* \leq \infty$, *such that the service-time cdf* G *has a finite moment generating function* $\hat{g}(-s) = \int_0^\infty e^{sx}\, dG(x)$ *for all* s, $0 < s < s^*$.

In general, we need to impose additional regularity conditions to have the limit for the decay rate in (3.2) be well-defined, as can be seen from Corollary 1 and Prop. 2 in [19] and Thms. 2.1, 5.5, and 5.3 in Chap. XIII in [4]. Instead of adding additional assumptions, we allow the decay rate to be defined by (3.1). It coincides with (3.2) when the limit exists.

We still need extra conditions for (2.2) to have a solution; see Example 5 in Sect. 3 and Thm. 5 in Sect. 7 of [3]. However, no extra condition is needed when G has support in $[0, M_s]$ because then $E[e^{tV}] \leq e^{tM_s}$ for all $t > 0$, so that $s^* = \infty$ in Assumption 3.1.

As indicated in Ref. [3], the asymptotic decay rate also is well defined for the more general $GI/GI/K$ model. We have fixed $E[U] = 1$. If instead we had fixed $E[V] = 1$, then $\theta_W(K) = K\theta_W(1)$, as in (5) of [3], where $U(K) = U(1)/K$ to keep ρ fixed. Since we fix $E[U] = 1$, we get $\theta_W(K) = \theta_W(1) \equiv \theta_W$. (As a sanity check, this can easily be verified for the $P(W > t|W > 0) = e^{-\theta_W t}$ in the $M/M/K$ model; see Thm. 9.1 in Sect. III.9 on p. 108 of [4].) However, we must adjust the service-time V to maintain $\rho = E[V]/KE[U]$. Thus, we leave U independent of K, but we let $V(K) = KV(1)$. Thus, the finite support of $V(K)$ becomes $[0, \rho K M_s]$, the pth moment of $E[V(K)^p] = K^p E[V(1)^p]$ and the Laplace transforms are related to $\hat{g}_{V(K)}(s) = \hat{g}_{V(1)}(Ks)$. This implies that we can apply the extremal distributions for $K = 1$ to directly obtain the corresponding extremal distributions for $K > 1$: If $V^*(K)$ is the extremal random variable as a function of K, then $V^*(K) = KV^*(1)$.

In Ref. [3], it was observed that the extension to $K > 1$ in (5) there was proved for the $GI/PH/K$ by Neuts and Takahashi [40]. A continuity result in Thm. 3.1 of [8] implies that result applies to the general $GI/GI/K$ model. If the decay rate θ_W is well defined for the $GI/GI/1$ model with (U, V) having cdf's (F, G) where $E[U] = 1$, then it is well defined in the associated $GI/GI/K$ model with (U, KV) with the same cdf F and

$$\theta_W(K) = \theta_W(1) \equiv \theta_W \quad \text{for } K > 1. \tag{3.3}$$

3.2. Two-Point Extremal Distributions Given Only Two Moments

We first consider the classical case in which we specify two moments. Let $\mathcal{P}_2(m, m^2(c^2 + 1), M)$ be the set of all cdf's with mean m, support $[0, mM]$ and second moment $m^2(c^2 + 1)$, where c^2 is the scv with $c^2 + 1 < M < \infty$. (The last property ensures that the set $\mathcal{P}_2(m, m^2(c^2 + 1), M)$ is non-empty.) The extremal distributions for the decay rate will be the extremal distributions P_{U}^* and P_{L}^* for T systems in Sect. 2.2.1 of [8].

In this classical setting, the extremal distributions P_{U}^* and P_{L}^* are special two-point distributions. The set of two-point distributions is a one-dimensional parametric family. In particular, any two-point distribution with mean m, scv c^2, and support mM has probability mass $c^2/(c^2+(b-1)^2)$ at mb, and mass $(b-1)^2/(c^2+(b-1)^2)$ on $m(1-c^2/(b-1))$ for $1+c^2 \le b \le M$.

Let subscripts a and s denote sets for the interarrival and service times, respectively. Let F_0 and F_u (G_0 and G_u) be the two-point extremal interarrival-time (service-time) cdf's corresponding to P_{L}^* and P_{U}^*, respectively, in the space $\mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ ($\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$) from Sect. 2.2.1 of [8]. (Recall our convention that $E[U]=1$ and $E[V]=\rho$. Hence, the support of V is $[0, \rho M_s]$.)

Consequently, F_0 has probability mass $c_a^2/(1+c_a^2)$ at 0 and probability mass $1/(c_a^2+1)$ at $m(c_a^2+1)$, while F_u has mass $c_a^2/(c_a^2+(M_a-1)^2)$ at the upper bound of the support, M_a, and mass $(M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $m(1-c_a^2/(M_a-1))$.

We are especially interested in the map

$$\theta_W : \mathcal{P}_{a,2}(1, 1+c_a^2, M_a) \times \mathcal{P}_{s,2}(\rho, \rho^2(1+c_s^2), M_s) \to \mathbb{R}, \tag{3.4}$$

where $0<\rho<1$ and $\theta_W(F,G)$ is the asymptotic decay rate of the steady-state waiting time $W(F,G)$ with interarrival-time cdf $F \in \mathcal{P}_{a,2}(1, 1+c_a^2, M_a)$ and service-time cdf $G \in \mathcal{P}_{s,2}(\rho, \rho^2(1+c_s^2), M_s)$. In Ref. [8] we also consider case in which one cdf is specified, in which case it need not have bounded support, but we do not discuss those cases here.

THEOREM 3.1 (two-point extremal distributions for the decay rate, Thm. 3.2 of [8: *)] Let F_0, F_u, G_0, and G_u be the two-point extremal cdf's for the $GI/GI/1$ queue defined above.*

For all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$,

$$\theta_W(F_0, G_u) \le \theta_W(F,G) \le \theta_W(F_u, G_0). \tag{3.5}$$

Based on Theorem 3.1, the overall extremal $GI/GI/1$ models are thus (F_0, G_u) and (F_u, G_0). Our assumption that the distributions have bounded support plays an important role. That is evident from the following elementary proposition.

PROPOSITION 3.1 (limits as the support increases): *Under the assumptions of Theorem 3.1, for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$,*

$$\theta_W(F, G_u) \to 0 \quad \text{as } M_s \to \infty, \tag{3.6}$$

while

$$\theta_W(F_u, G) \to \theta_W(F_1, G) \quad \text{as } M_a \to \infty, \tag{3.7}$$

where F_1 is the cdf of the unit point mass on 1, associated with the $D/GI/1$ model.

REMARK 3.1 (the decay rates of other steady-state distributions): *Analogs of Theorem 3.1 (and the later Theorem 3.2) hold for the steady-state continuous-time queue length and workload because there are simple relations among all these decay rates. That follows from Thm. 6, Prop. 9, and Prop. 2 of [19]. For the workload, the decay rate is the same; for the queue length, $\theta_Q = \hat{g}(-\theta_W)$.*

REMARK 3.2 (comparison to the mean): *The extremal model* (F_0, G_u) *in Theorem 3.1 yielding the smallest decay rate coincides with the conjectured upper bound model for the mean* $E[W]$*, but the extremal model* (F_u, G_0) *in Theorem 3.1 yielding the largest decay rate does not coincide with the lower bound for the mean; see Sect. 2.4.1 of [9].*

3.3. Laplace Transform Constraints to Reduce the Range

We now add additional constraints on the cdf's F and G. In order to apply Lemma 2.2 of [8], we add constraints on higher moments and transform values of F and G. In particular, following [17,44], we add a third moment and a value of the Laplace transform. With (2.2) in mind, we now impose constraints on the Laplace transform $\hat{f}(s)$ at $s = \mu_a > 0$ and on the reciprocal of the mgf, $1/\hat{g}(-s)$, at $s = \mu_s$, $0 < \mu_s < s^*$, for s^* in Assumption 3.1.

For the new extremal distributions, let $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ be the subset of F in $\mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ having specified third moment $m_{a,3}$ and Laplace transform value $\hat{f}(\mu_a)$. Since we are working with the mgf $\hat{g}(-s)$ for $s > 0$, let $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ be the subset of G in $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), M_s)$ having specified third moment $m_{s,3}$ and mgf value $\hat{g}(-\mu_s)$ at μ_s for $0 < \mu_s < s^*$. (Recall that $s^* = +\infty$ if G has bounded support.)

Let $F_{\rm L}$ and $F_{\rm U}$ ($G_{\rm L}$ and $G_{\rm U}$) be the three-point extremal interarrival-time (service-time) cdf's corresponding to $P_{\rm L}^*$ and $P_{\rm U}^*$, respectively, in the space $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ ($\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$) based on Sect. 2.2.1 of [8]. (Recall our convention that $E[U] = 1$ and $E[V] = \rho$.) In particular, $F_{\rm L}$ ($F_{\rm U}$) is the unique element of $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ with support on the set $\{0, x_1, x_2\}$ (on the set $\{x_1, x_2, M_a\}$) for $0 < x_1 < x_2 < M_a$, while $G_{\rm L}$ ($G_{\rm U}$) is the unique element of $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ with support on the set $\{0, \bar{x}_1, \bar{x}_2\}$ (on the set $\{\bar{x}_1, \bar{x}_2, \rho M_s\}$) for $0 < \bar{x}_1 < \bar{x}_2 < \rho M_s$.

Again, in Ref. [8] we also consider case in which one cdf is specified, in which case it need not have bounded support, but we do not discuss those cases here.

THEOREM 3.2 (three-point extremal distributions for the decay rate, Thm. 3.3 of [8]): *Let* $F_{\rm L}$*,* $F_{\rm U}$*,* $G_{\rm L}$*, and* $G_{\rm U}$ *be the three-point extremal cdf's for the* $GI/GI/1$ *queue defined above.*

For all $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ *with* $\mu_a > 0$ *and* $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ *with* $\mu_s > 0$*, the decay rate* $\theta_W(F, G)$ *is well defined as the unique positive solution of (2.2). Moreover, for all* (F, G) *in these sets, the following four pairs of lower and upper bounds for* $\theta_W(F, G)$ *are valid:*

$$\begin{aligned}
&(i) \quad \theta_W(F_{\rm L}, G_{\rm U}) \le \theta_W(F, G) \le \theta_W(F_{\rm U}, G_{\rm L}) \quad \text{if } \mu_s, \mu_a \le \theta_W,\\
&(ii) \quad \theta_W(F_{\rm U}, G_{\rm U}) \le \theta_W(F, G) \le \theta_W(F_{\rm L}, G_{\rm L}) \quad \text{if } \mu_s \le \theta_W \le \mu_a,\\
&(iii) \quad \theta_W(F_{\rm U}, G_{\rm L}) \le \theta_W(F, G) \le \theta_W(F_{\rm L}, G_{\rm U}) \quad \text{if } \theta_W \le \mu_s, \mu_a,\ \mu_s < s^*,\\
&(iv) \quad \theta_W(F_{\rm L}, G_{\rm L}) \le \theta_W(F, G) \le \theta_W(F_{\rm U}, G_{\rm U}) \quad \text{if } \mu_a \le \theta_W \le \mu_s < s^*. \qquad \textbf{(3.8)}
\end{aligned}$$

The bounds on θ_W *get tighter as* μ_a *and* μ_s *move closer to* $\theta_W(F, G)$*. The bounds coincide with* θ_W *when* $\mu_a = \theta_W$ *in* (a) *and* $\mu_s = \theta_W$ *in* (b)*.*

REMARK 3.3 (choice of the Laplace transform values): *The final conclusion of Theorem 3.2 has important practical implications. It shows that, for any given model with a specified decay rate, the range of possible decay rate values consistent with the partial information becomes smaller as the arguments of the Laplace transforms become closer to the final decay rate.*

3.4. Extending the Extremal Models to Unbounded Support

The T-system theory and the Markov–Krein theorem extend to unbounded support intervals as shown by Karlin and Studden [28] and as indicated in Refs. [17,20]. The extension is easy if the extremal distribution places no mass on the upper endpoint. Then, the same extremal distribution holds for all larger support bounds, including the unbounded interval $[0, \infty)$.

First, in the setting of the two-point extremal distributions in Theorem 3.1, the extremal cdf's F_0 and G_0 have support on $\{0, x\}$ for appropriate x and so remain valid if we increase M_a and M_s. (The x depends on the cdf.)

Similarly, in the setting of the three-point extremal distributions in Theorem 3.1, the extremal cdf's $F_{\rm L}$ and $G_{\rm L}$ have support on $\{0, x_1, x_2\}$ for appropriate x_1 and x_2 and so remain valid if we increase M_a and M_s. (Again, the points x_1 and x_2 depend on the cdf.)

Consequently, we need to make no adjustments for truncation provided we use the following special case of (3.8):

$$\begin{aligned} \theta_W(F_{\rm L}, G_{\rm L}) \le \theta_W(F, G) \quad \text{for} \quad \mu_a \le \theta_W \le \mu_s < s^*, \\ \theta_W(F_{\rm L}, G_{\rm L}) \ge \theta_W(F, G) \quad \text{for} \quad \mu_s \le \theta_W \le \mu_a. \end{aligned} \tag{3.9}$$

This recipe also eliminates the need to consider multiple cases.

We state the result formally in the following corollary. To simplify, we make the following stronger assumption.

Assumption 3.2 (uniformly good cdf G): *In addition to Assumption 3.1, assume that, for the service-time cdf G, Eq. (2.2) has a finite solution for all $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1)$.*

Note that Assumption 3.2 is satisfied by the M, H_k, and E_k distributions considered here and many others, but we need to avoid pathological examples like Example 5 of [3].

Corollary 3.1 (extension to unbounded support): *Consider the setting of Theorem 3.2 extended by allowing unbounded support, that is, $M_a = M_s = \infty$.*

(a) For any $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1))$ satisfying Assumption 3.2, the decay rate $\theta_W(F, G)$ is well defined as the unique positive solution of (2.2). Moreover, if $\mu_a \le \theta_W$, then

$$\theta_W(F_{\rm L}, G) \le \theta_W(F, G) \tag{3.10}$$

for all $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a)$.

(b) For any $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s)$ satisfying Assumption 3.2, the decay rate $\theta_W(F, G)$ is well defined as the unique positive solution of (2.2). Moreover, if $\theta_W \le \mu_s < s^*$, then

$$\theta_W(F, G_{\rm L}) \ge \theta_W(F, G) \tag{3.11}$$

for all $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1)$.

(c) For all (F, G) such that Assumption 3.2 holds, the decay rate $\theta_W(F, G)$ is well defined as the unique positive solution of (2.2) and (3.9) holds.

4. APPLICATION TO PRODUCE A PRACTICAL RANGE FOR THE MEAN

We now elaborate in Section 2.2 and more carefully describe how we apply the theoretical results for the decay rate in Section 3 to develop a practical way to identify intervals of

likely values for the mean steady-state waiting time given the basic moment parameters in (1.4) and the additional parameters introduced in Theorems 3.1 and 3.2. This analysis is heuristic because we have no explicit relation between the decay rate and the mean, but the general idea is that the mean should be decreasing in the decay rate.

We start in Section 4.1 by discussing the support bounds used in Theorems 3.1 and 3.2. Then, in Section 4.2, we indicate how we can obtain the extremal models with three-point distributions derived in Theorem 3.2.

4.1. Choosing the Support Bounds M_a and ρM_s for $GI/GI/1$

Before considering the support bounds, we emphasize that the range of possible values for the mean $E[W]$ in the $GI/GI/1$ model given only the first two moments of U and V tends to be remarkably wide. That is shown in Table 1 of [9] and in Tables 1–4 in [7]. The relative errors tend to increase in c_a^2 but decrease in ρ and c_s^2; see Sect. 2 of [7].

4.1.1. Starting with a model or data In order to apply Theorem 3.2, we need support bounds on F and G. Hence, starting from a specific model or data with unbounded U and V, we suggest choosing the support bounds M_a and ρM_s that tend to not affect the mean too much. In particular, we suggest choosing M_a and M_s so that the tail probability is suitably small, that is, so that

$$P(U > M_a E[U]) = P(U > M_a) = P(V > M_s E[V]) = P(V > \rho M_s) = \epsilon \tag{4.1}$$

for a suitably small ϵ such as 0.001; see Sect. 3 of [7] for more discussion.

For our numerical experiments, we start with standard M, E_k, and H_2 distributions, which are determined by their first two moments. For M, $c^2 = 1$; for E_k, $c^2 = 1/k < 1$; for H_2 distributions, $c^2 \geq 1$. For H_2, we assume balanced means to reduce the number of parameters from 3 to 2, as in (3.7) of [42]. We suggest using a simple exponential approximation based on the asymptotic decay rates of these distributions, which are well-defined. Thus, we choose M_s so that

$$\epsilon = P(V/E[V] > M_s) \approx e^{-\theta_V M_s}, \tag{4.2}$$

where θ_V is the asymptotic decay rate of V.

For M, the decay rate of $V/E[V]$ is $\theta_V = 1$; for E_k, the scv is $1/k$, while the decay rate of $V/E[V]$ is $\theta_V = k$, so we let $\theta_V = 1/c_s^2$ for $c_s^2 \geq 0.01$, and $\theta(\rho, c_s^2) = 100$ for $c_s^2 \leq 0.01$ to avoid the deterministic case with $c_s^2 = 0$. Our examples use $c_s^2 = 0.5$, for which $\theta_V = 2$. In the case of H_2 with balanced means, by (37) in [42], the asymptotic decay rate of $V/E[V]$ is

$$\theta_V = 1 - \sqrt{(c_s^2 - 1)/(c_s^2 + 1)}. \tag{4.3}$$

Our examples use $c_s^2 = 4.0$, for which we use $\theta_V = (1 - \sqrt{3/5}) = 0.2254$.

We now see how the extremal UB model $F_0/G_u/1$ and the LB model $F_u/G_0/1$ for the decay rate from Theorem 3.1 apply to the mean $E[W]$ with $K = 1$ when we introduce the parameters M_a and M_s following the prescription above. Table 1 shows the results for five cases: $(c_a^2, c_s^2) = (1.0, 1.0)$, $(4.0, 4.0)$, $(0.5, 0.5)$, $(4.0, 0.5)$, and $(0.5, 4.0)$. (We show more results for other traffic intensities in Sect. 4 of [7].) We show two candidate support bounds for each case, based on $\epsilon = 0.01$ and 0.001 in (4.1). For comparison, Table 1 shows the HTA and the tight UB and LB given only the moments as well as the values of the mean with the support bounds.

Table 1 shows that the range decreases as the traffic intensity increases and as the support bounds decrease. For $\rho = 0.7$, the tight UB is not too far above the HTA approximation, but the tight LB tends to be far below. The mean for the $F_u/G_0/1$ model with M_a is significantly larger than the tight LB, but still the final range is very large, except for the one case $(c_a^2, c_s^2) = (0.5, 4.0)$. Note that the relative error is only about 5% for $\rho = 0.7$ in that good case.

To obtain these estimates of $E[W]$ and later ones, we use simulation. We implement standard Monte-Carlo simulation to estimate the sample mean of the steady-state waiting time with a run length (number of arrivals) $N = 5 \times 10^8$ and 20 independent replications for the model $F_u/G_0/1$, but it helps to use an efficiency-improvement algorithm for the $F_0/G_u/1$ model with the atom at the upper support bound, as discussed in Ref. [9]. We implement the [38] simulation algorithm with total simulation length $T = 1 \times 10^7$ and 20 independent replications for the model $F_0/G_u/1$. We can construct 95% confidence interval by using statistical t-test. The worst-case confidence interval length for Monte-Carlo simulation achieves 10^{-3} level which happens at the highest ρ, while the worst-case confidence interval length for the [38] simulation is around 10^{-4} level. (See Ref. [9] for more discussion.)

REMARK 4.1 (starting with HTAs): *An alternative approach for obtaining the support bounds is to use HTAs. In addition to the HTA for the mean in (1.2), we can use the associated HTA for the decay rate,*

$$\theta_W \approx \frac{2(1-\rho)}{\rho(c_a^2 + c_s^2)}, \tag{4.4}$$

which is obtained by combining the $M/M/1$ exact formula $\theta_W = (1-\rho)/\rho$ with the heavy-traffic asymptotic expansion established in [1]; that is,

$$\theta_W(\rho) = \frac{2(1-\rho)}{c_a^2 + c_s^2} + C(1-\rho)^2 + O(1-\rho)^3 \quad \text{as } \rho \uparrow 1, \tag{4.5}$$

where C is an (explicit) function of the first three moments of the mean-1 random variables U and V/ρ. Related asymptotics and approximations for the $GI/GI/s$ and $BMAP/GI/1$ models are established in [3,10] and Corollary 3 of [19].

To show that we could also start from the HTAs for the mean $E[W]$ in (1.2) and for the decay rate θ_W in (4.4) instead of the exact models based on E_2 and H_2 distributions, again using the case of balanced means to reduce the H_2 parameters from 3 to 2. Table 2 compares the exact values of θ_W and $E[W]$ to these HTAs. Table 2 shows that the HTA in (1.2) overestimates the exact value when $c_a^2 = 0.5$, which is consistent with the refinement in (44) and (45) of [43].

We found that the support bounds determined by the HTA are similar to those for the exact model. Overall, we found that, with the procedure based on (4.1), the support bounds reduce the range of possible value in all cases, but not greatly, so that the range is still very wide in most cases. The tables also show that the cases differ dramatically. The relative errors are remarkably small for $(c_a^2, c_s^2) = (0.5, 4.0)$ but remarkably large for $(c_a^2, c_s^2) = (4.0, 0.5)$.

4.2. Determining the Extremal Models from Theorem 3.2

We now investigate how we can apply Theorem 3.2 to obtain a better indication of typical values of the mean $E[W]$. Paralleling the two-moment case discussed above, we assume that we are given the first three moments of the underlying cdf's F and G. We also assume

TABLE 1. Comparing bounds for $E[W]$ using $F_u/G_0/1$ (UB) and $F_0/G_u/1$ (LB) with (M_a, M_s) from Section 4.1 using $\epsilon = 0.001$ and 0.01 in (4.1)

	ρ	Tight LB	$M_a = 9$	$M_a = 7$	HTA (1.2)	$M_s = 7$	$M_s = 9$	Tight UB
$c_a^2 = c_s^2 = 1$	0.50	0.000	0.122	0.162	0.500	0.810	0.821	0.846
	0.70	0.467	0.970	1.130	1.633	2.025	2.036	2.071
	0.90	3.600	7.265	7.596	8.100	8.564	8.579	8.620
			$M_a = 39.9$	$M_a = 31.1$		$M_s = 31.1$	$M_s = 39.9$	
$c_a^2 = c_s^2 = 4$	0.50	0.750	1.013	1.097	2.000	3.419	3.430	3.470
	0.70	2.917	4.303	4.748	6.533	8.384	8.394	8.441
	0.90	15.750	28.924	30.239	32.400	34.658	34.671	34.721
			$M_a = 4.5$	$M_a = 3.5$		$M_s = 31.1$	$M_s = 39.9$	
$c_a^2 = 0.5, c_s^2 = 4$	0.50	0.750	0.957	0.988	1.125	1.263	1.270	1.289
	0.70	2.917	3.464	3.494	3.675	3.841	3.851	3.875
	0.90	15.750	17.973	17.993	18.225	18.408	18.427	18.470
			$M_a = 39.9$	$M_a = 31.1$		$M_s = 3.5$	$M_s = 4.5$	
$c_a^2 = 4, c_s^2 = 0.5$	0.50	0.000	0.000	0.000	1.125	2.556	2.559	2.595
	0.70	0.058	0.342	0.450	3.675	5.524	5.533	5.583
	0.90	1.575	9.075	11.988	18.225	20.469	20.486	20.546
			$M_a = 4.5$	$M_a = 3.5$		$M_s = 3.5$	$M_s = 4.5$	
$c_a^2 = 0.5, c_s^2 = 0.5$	0.50	0.000	0.000	0.000	0.250	0.377	0.388	0.414
	0.70	0.058	0.410	0.530	0.817	0.966	0.982	1.017
	0.90	1.575	3.613	3.771	4.050	4.207	4.229	4.295

TABLE 2. Decay rates and mean values: exact compared to the HT approximations in (1.2) and (4.4)

ρ	Exact θ_W	Approximate θ_W	Exact $E[W]$	HTA	ρ	Exact θ_W	Approximate θ_W	Exact $E[W]$	HTA
	$c_a^2 = c_s^2 = 0.5$					$c_a^2 = c_s^2 = 4$			
0.5	2.00	2.00	0.195	0.250	0.5	0.244	0.250	2.02	2.00
0.7	0.857	0.857	0.725	0.817	0.7	0.106	0.107	6.61	6.53
0.9	0.222	0.222	3.92	4.05	0.9	0.0278	0.0278	32.6	32.4
	$c_a^2 = 4, c_s^2 = 0.5$					$c_a^2 = 0.5, c_s^2 = 4$			
0.5	0.826	0.444	0.882	1.13	0.5	0.311	0.444	1.05	1.13
0.7	0.260	0.190	3.37	3.68	0.7	0.153	0.190	3.56	3.68
0.9	0.0537	0.049	18.0	18.2	0.9	0.0458	0.0494	18.0	18.2

that we have determined the support bounds and a reference decay rate θ_W associated with a candidate model as in Section 4.1 or with the aid of the HTA in Remark 4.1. While determining those quantities, it is natural to also determine the associated mean $E[W]$, which would be the usual direct approximation. It is a reference to check the set-valued approximation.

To obtain exact results from Theorem 3.2, we should have the moments and decay rate of the truncated distribution with the support bounds, but for simplicity, we simply apply the moments and decay rate determined for the original base model without support bounds. We found that the impact of that simplifying assumption tends to be negligible.

It now remains to determine the extremal distributions themselves. For the specified parameters, it suffices to solve the equations characterizing the extremal models. First, we can solve the system of equations provided by the T-system theory by using a nonlinear equation solver (we used MATLAB). Second, a convenient way to calculate the extremal distributions approximately (to any desired accuracy) is to assume finite support and apply linear programming to minimize (or maximize) the Laplace transform given the constraints. We can let the support be $\{kM_a/n : 0 \le k \le n\}$, so that the only variables are the probabilities p_k assigned to the points $x_k \equiv kM_a/n$. As in Thm. 2.1 of [8], there will necessarily be five-point extremal distributions given the four constraints using this approach. The solution converges to the three-point solution for the original support set $[0, M_a]$ as $n \to \infty$. Moreover, we can see that the optimal solution does not depend on the argument of the Laplace transform provided that the sign of $\mu - \theta_W$ does not change.

4.3. Choosing the Laplace Transform Arguments

Our proposed method is based on Theorem 3.2 in Section 3.3. We start with a concrete model determined by the pair of cdf's (F, G), which typically have unbounded support. We first calculate (i) the decay rate θ_W for that model by solving for the unique positive root of the single equation (2.2) involving the Laplace transforms $\hat{f}$ and $\hat{g}$ of F and G and (ii) four parameters from each of the underlying cdf's F and G: the first three moments and one argument of each Laplace transform.

We have two alternatives for each of the arguments μ_s of $\hat{g}(-s)$ and μ_a of $\hat{f}(s)$: either $\le \theta_W$ or $\ge \theta_W$. Our experiments indicate that we can set

$$\mu \equiv \theta_W/R \quad \text{if } \mu \le \theta_W \quad \text{and} \quad \mu \equiv R\theta_W \quad \text{if } \mu \ge \theta_W \tag{4.6}$$

for suitable R, for example, $R \in \{1, 5, 10, 20\}$. We find that it is better to have $\mu_s \leq \theta_W$. For the concrete model, we also directly calculate the mean steady-state waiting time $E[W]$, but the goal is to determine a set of likely values of that mean given *any* model with the two-moment parameters in (1.4) and the small set of additional parameters.

Because we are working with distributions with unbounded support, we must be careful about Assumption 3.1. Hence, in the implementation, we do not allow $\mu_s > s^*$. Thus, if we are considering one of the cases with $\mu_s \geq \theta_W$, then we first check to see if $R\theta_W > s^*$ for our largest value of R, which we take to be $R = 20$. If it is, then we create alternative values of μ_s in the interval (θ_W, s^*). In particular, we use

$$\mu_s \equiv \theta_W + \left(\frac{R}{25}\right)(s^* - \theta_W), \quad R = 5k, \quad 1 \leq k \leq 4, \tag{4.7}$$

so that the values of R remain in $\{5, 10, 15, 20\}$, but all values are within the interval (θ_W, s^*). However, cases (i) and (ii) in (3.8) that we recommend present no difficulties.

4.3.1. An illustration To illustrate, Table 3 shows the explicit numerical values of the three-point extremal distributions $F_{\rm L}, G_{\rm L}$ and $F_{\rm U}, G_{\rm U}$ obtained in the case $c_a^2 = c_s^2 = 4, \rho = 0.7$ with $R \in \{1, 5, 10, 20\}$.

For these extremal models, we must determine the associated decay rates. Figure 1 plots the extremal Laplace transforms $\hat{f}(s)$ and $1/\hat{g}(-s)$ for UB (LHS) and LB (RHS) for the case $c_a^2 = c_s^2 = 4$ and $\rho = 0.7$. The curves intersect at the decay rate θ_W. The decay rate for $R = 1$ is 0.106, while for $R = 20$, it is 0.098 for the UB and 0.110 for the LB.

4.3.2. Specification details From Theorem 3.2 and Corollary 3.1, we see that we have five candidate ways to set the positive arguments of the Laplace transform $\hat{f}(s)$ and the mgf $\hat{g}(-s)$: the four cases with bounded support in (3.8) and the single composite version with unbounded support in (3.9). These alternatives have advantages and disadvantages. First, the finite support bounds in (3.8) require truncation, so we either must calculate new parameters for the truncated model with bounded support or use the parameters of the original distributions with unbounded support without altering them. In addition, we must choose among the four alternatives in (3.8).

The alternative with unbounded support in (3.9) is appealing because it requires no truncation and we need not choose among four cases. On the other hand, it uses different parameter specifications for the minimum and maximum, which can distort the results, leading to anomalies such as the lower bound for the mean exceeding the upper bound.

We performed extensive experiments to test these alternatives and deduced that it is better to use the finite support bounds in (3.8) provided that the support bounds are chosen to have negligible impact, as in Section 4.1. In particular, we found that the parameters were not significantly altered by the truncation. For example, for the $E_2/H_2/1$ model with $\rho = 0.7$, the second and third moments of V with truncation were $s_2 = 2.44, s_3 = 20.19$, and $s_2 = 2.45, 20.58$ without truncation.) Hence, our procedure for the mean $E[W]$ uses the parameters taken directly from the base model with unbounded support or the HTAs, but then applies the results in (3.8) with the constructed support bounds.

It still remains to select one of the four alternatives in (3.8). From our experiments, we conclude that a good robust approximation is obtained by doing all four cases, and using the minimum of the four lower bounds for $E[W]$ for the final lower bound, and the maximum of the four upper bounds for $E[W]$ as the final upper bound. However, that requires more computational effort. Hence, we also propose a way to select one of the four alternatives.

TABLE 3. Numerical examples of extremal distributions in the case $c_a^2 = c_s^2 = 4, \rho = 0.7$ with $R \in \{1, 5, 10, 20\}$

$R=1$	F			G			$R=5$	F			G		
$F_L/G_L/1$	q_1	q_2	q_3	p_1	p_2	p_3	$F_L/G_L/1$	q_1	q_2	q_3	p_1	p_2	p_3
	0.620	0.370	1.04×10^{-02}	0.677	0.317	6.08×10^{-03}		0.526	0.459	1.57×10^{-02}	0.656	0.336	7.69×10^{-03}
	y_1	y_2	y_3	x_1	x_2	x_3		y_1	y_2	y_3	x_1	x_2	x_3
	0	2.21	17.6	0	1.93	14.4		0.0	1.65	15.5	0	1.78	13.4
$F_U/G_U/1$	q_1	q_2	q_3	p_1	p_2	p_3	$F_U/G_U/1$	q_1	q_2	q_3	p_1	p_2	p_3
	0.956	0.0433	2.88×10^{-04}	0.965	0.0345	1.73×10^{-04}		0.936	0.0639	4.30×10^{-04}	0.963	0.0370	2.12×10^{-04}
	y_1	y_2	y_3	x_1	x_2	x_3		y_1	y_2	y_3	x_1	x_2	x_3
	0.587	9.86	39.9	0.440	7.86	27.9		0.505	7.99	39.9	0.431	7.54	27.9
$R=10$	F			G			$R=20$	F			G		
$F_L/G_L/1$	q_1	q_2	q_3	p_1	p_2	p_3	$F_L/G_L/1$	q_1	q_2	q_3	p_1	p_2	p_3
	0.451	0.530	1.87×10^{-02}	0.654	0.338	7.88×10^{-03}		0.358	0.621	2.14×10^{-02}	0.653	0.339	7.97×10^{-03}
	y_1	y_2	y_3	x_1	x_2	x_3		y_1	y_2	y_3	x_1	x_2	x_3
	0.0	1.37	14.6	0	1.76	13.4		0	1.13	14.0	0	1.75	13.3
$F_U/G_U/1$	q_1	q_2	q_3	p_1	p_2	p_3	$F_U/G_U/1$	q_1	q_2	q_3	p_1	p_2	p_3
	0.917	0.0828	5.02×10^{-04}	0.962	0.0374	2.17×10^{-04}		0.891	0.108	5.62×10^{-04}	0.962	0.0376	2.20×10^{-04}
	y_1	y_2	y_3	x_1	x_2	x_3		y_1	y_2	y_3	x_1	x_2	x_3
	0.439	6.97	39.9	0.430	7.50	27.9		0.360	6.08	39.9	0.429	7.48	27.9

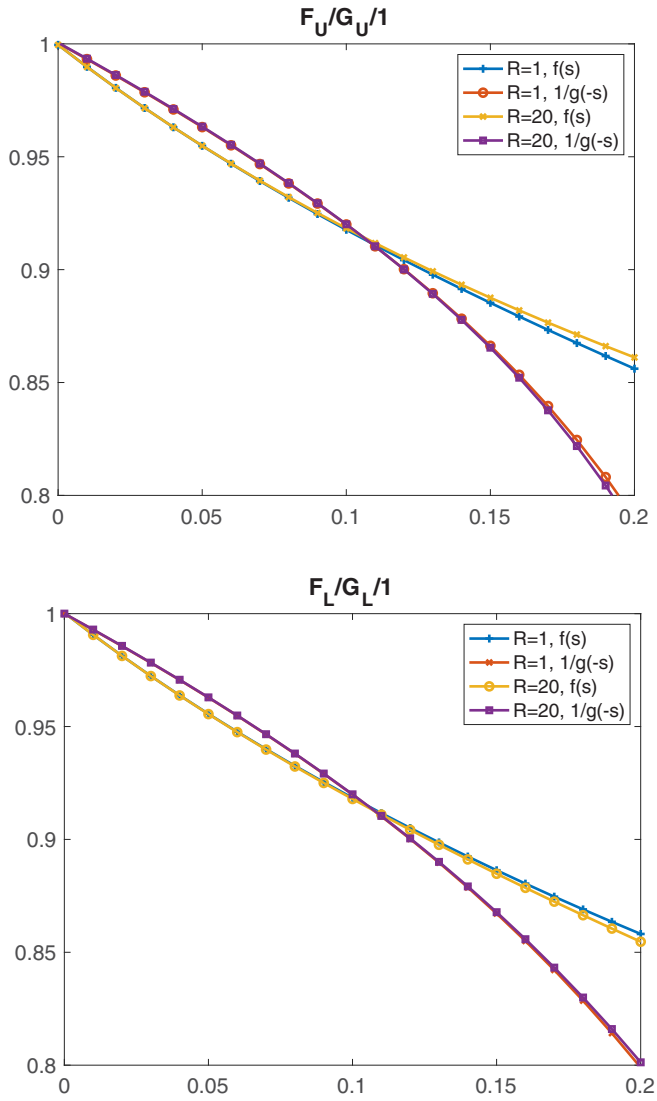

FIGURE **1.** Display of $\hat{f}(s)$ and $1/\hat{g}(-s)$ for UB (LHS) and LB (RHS) for the case $c_a^2 = c_s^2 = 4$ and $\rho = 0.7$: the decay rate for $R = 1$ is 0.106 and for $R = 20$ in UB is 0.098 and in LB is 0.110.

We first observe that $F_{\rm L}$ ($F_{\rm U}$) in (3.8) of Theorem 3.2 is the natural analog of F_0 (F_u) from Theorem 3.1, having 0 (M_a) as one of the mass points. Thus, case (i) in (3.8) is the natural choice. Our experiments indicate that cases (i) and (ii) are good. We examine all four cases for the models we consider in our experiments.

5. NUMERICAL EXPERIMENTS

In this section, we report the results of experiments to evaluate the proposed heuristic set-valued approximations. We start in Section 5.1 by looking at the $M/M/1$ base model.

TABLE 4. Bounds for θ_W (exact) and $E[W]$ (approximate) for $\rho = 0.7$ and $c_a^2 = c_s^2 = 1$ based on $M/M/1$ (For reference, exact values for $M/M/1$ are $\theta_W = (1-\rho)/\rho = 0.4286$ and $E[W] = \rho^2/(1-\rho) = 1.63$.)

Case	θ_W			$E[W]$			Case	θ_W			$E[W]$		
(3.8)	$R=5$	10	20	$R=5$	10	20	(3.8)	$R=5$	10	20	$R=5$	10	20
(i)	0.426	0.425	0.425	1.67	1.67	1.68	(ii)	0.421	0.418	0.415	1.59	1.62	1.68
	0.432	0.432	0.439	1.65	1.65	1.56		0.434	0.437	0.446	1.53	1.56	1.61
(iii)	0.422	0.417	0.409	1.71	1.72	1.71	(iv)	0.426	0.424	0.418	1.61	1.60	1.57
	0.434	0.436	0.436	1.65	1.63	1.62		0.431	0.432	0.429	1.60	1.61	1.63

In Section 5.2, we report the numerical results of our application of this method to the $GI/GI/1$ queue. Finally, in Section 5.3, we report the numerical results for the $GI/GI/2$ queue. More appears in Ref. [7].

5.1. The $M/M/1$ Reference Case: $c_a^2 = c_s^2 = 1$ with $\rho = 0.7$

To start, Table 4 below shows the results for all four cases associated with the $M/M/1$ reference base model with $\rho = 0.7$ and three possible values of R in (4.6).

From the analytical formulas $\theta_W = (1-\rho)/\rho = 0.4286$ and $E[W] = \rho^2/(1-\rho) = 1.63$, we see that θ_W ($E[W]$) is strictly decreasing (increasing) in ρ. Of course, we are considering a large collection of models with $c_a^2 = c_s^2 = 1$, not simply $M/M/1$, but it is our reference case from which we extract parameters.

Consistent, with Theorem 3.2, Table 4 shows that the decay rate associated with the UB (LB) for $E[W]$ is decreasing (increasing) in R in each case, while the reverse order tends to hold for $E[W]$ too. There are minor exceptions in cases (iii) and (iv) because we get the decay rates from the original $M/M/1$ model.

From Table 4, we obtain the composite bounds for $E[W]$ based on all four cases. With $R = 20$, the composite bounds are

$$\min_{1\le i\le 4}\{E[W_{LB,i}(R=20)]\} = 1.56 < E[W] = 1.63 < 1.71 = \max_{1\le i\le 4}\{E[W_{UB,i}(R=20)]\}. \tag{5.1}$$

Notice that the interval $[1.56, 1.71]$ in (5.1) is not too different from the intervals $[1.56, 1.68]$ in case (i) with $\mu_s, \mu_a < \theta_W$ and $[1.61, 1.68]$ in case (ii) with $\mu_s < \theta_W < \mu_a$. On the other hand, the LB 1.62 for $E[W]$ in case (iii) is too large, while the UB 1.57 for $E[W]$ in case (iv) is too small. Thus, we tentatively conclude that it is better to have $\mu_s \le \theta_W$. For this case, the choice of μ_s seems to be more important than μ_a. We tentatively conclude that the cases (i) and (ii) in (3.8) are both consistently effective for the $M/M/1$ base model, while the other alternatives are not.

5.2. Non-Exponential $GI/GI/1$ Base Models

We now extend the study to the four models with $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the H_2 and E_2 distributions. Table 5 shows the approximate upper bounds (top) and lower bounds (bottom) for $E[W]$ with $\rho = 0.7$ and $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the E_2 and H_2 models in each of the four cases in (3.8) of Theorem 3.2 for three values in R in (4.6). The cases are labeled at the left by the base model. (The exact values of $E[W]$ for $H_2/H_2/1$, $H_2/E_2/1$, $E_2/H_2/1$, and $E_2/E_2/1$ are 6.61, 3.37, 3.56, and 0.725, respectively.)

Table 5 reinforces the conclusions about Table 4 for the case $c_a^2 = c_s^2 = 1$ based on the $M/M/1$ model. Table 5 shows that the UB exceeds the LB for all models and all values of

TABLE 5. Approximate upper and lower bounds for $E[W]$ for $\rho = 0.7$ and $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the E_2 and H_2 models in each of the four cases in (3.8) of Theorem 3.2 for three values in R in (4.6) (The exact values of $E[W]$ for $H_2/H_2/1$, $H_2/E_2/1$, $E_2/H_2/1$, and $E_2/E_2/1$ are 6.61, 3.37, 3.56, and 0.725.)

Model	(i)			(ii)			(iii)			(iv)		
	$R=5$	10	20	$R=5$	10	20	$R=5$	10	20	$R=5$	10	20
H_2/H_2	6.93	6.94	6.73	6.28	6.19	7.20	6.93	7.08	7.20	6.72	6.72	6.66
	6.53	6.52	6.12	6.49	6.44	6.41	6.70	6.56	6.47	6.26	6.25	6.21
H_2/E_2	3.57	3.61	3.63	3.92	4.19	4.33	3.57	3.60	3.63	3.57	3.60	3.63
	3.06	3.08	3.06	2.95	2.82	2.69	3.06	3.08	3.06	3.06	3.08	3.06
E_2/H_2	3.62	3.68	3.68	3.53	3.54	3.56	3.51	3.51	3.52	3.52	3.52	3.49
	3.52	3.55	3.51	2.95	2.82	2.69	3.59	3.59	3.57	3.53	3.53	3.53
E_2/E_2	0.738	0.738	0.729	0.721	0.719	0.734	0.766	0.767	0.762	0.701	0.689	0.673
	0.737	0.733	0.704	0.642	0.625	0.642	0.730	0.730	0.721	0.736	0.738	0.753

R in case (i) with $\mu_a, \mu_a \le \theta_W$, while this good property holds for case (ii) except for the case $c_a^2 = c_s^2 = 4.0$ based on the $H_2/H_2/1$ model, but it holds there as well for $R = 20$. In contrast, cases (iii) and (iv) perform significantly worse. In case (iii), the LB exceeds the UB for the case $c_a^2 = 0.5$, $c_s^2 = 4.0$ based on the $E_2/H_2/1$ model. In case (iv), the LB exceeds the UB for the case $c_a^2 = 0.5$, $c_s^2 = 4.0$ based on the $E_2/H_2/1$ model.

Table 17 in [7] displays the corresponding rates obtained in deriving the extremal distributions used for the mean $E[W]$ in Table 5. That table confirms Theorem 3.2, just like Table 4. (Again there are minor discrepancies because we get the decay rates from the original models.)

We offer two possible explanations for the better performance of cases (i) and (ii) in (3.8) of Theorem 3.2. First, since large waiting times tend to be caused by large service times and short interarrival times (leading to clumps of arrivals), we should pin down $E[W]$ most effectively from parameters with case (ii) with $\mu_s < \theta_W < \mu_a$ as in (4.6). A second consideration is the nature of the distribution itself. Given an E_k distribution that has a pdf h with $h(0) = 0$, large values of μ are not likely to help much. In contrast, a more variable H_2 distribution could be helped by additional specification wherever it appears. Thus, cases (iii) and (iv) with $c_a^2 = 0.5$ involving an E_2 arrival process are likely to not perform well, as we have seen.

5.3. Examples for Multi-Server Queues

We now discuss experiments for $K > 1$. For ease of applications, it is significant that we can apply the result for $K = 1$ to derive the decay rate. To apply the results for $K = 1$ to $K > 1$, we use the same extremal interarrival-time distribution, but multiply the extremal service-time random variable by K. We then can apply simulation to estimate $E[W]$ just as before.

Table 6 shows the approximate upper and lower bounds for $E[W]$ obtained by this method for three values of ρ in $\{0.5, 0.7, 0.9\}$ and the five pairs of variability parameters (c_a^2, c_s^2) from $\{0.5, 1.0, 4.0\}$ in case (ii) of (3.8) in Theorem 3.2 for $R \in \{5, 10, 20\}$. Table 6 confirms that the procedure extends directly to $GI/GI/K$ queues with $K > 1$.

To illustrate the procedure for larger K, Table 7 shows set-valued approximations for $E[W]$ in the $M/M/10$ and $E_2/E_2/10$ models for $\rho \in \{0.7, 0.9\}$.

From readily available algorithms for $M/M/10$, we see that the exact values of $E[W]$ for $\rho = 0.7$ and 0.9 are 0.519 and 6.03, respectively, which fall right in the middle of the interval

TABLE 6. The improved UB and LB for $E[W]$ in $GI/GI/2$ for $(c_a^2, c_s^2) \in \{(1,1),(4.0,4.0),(4.0,0.5),(0.5,4,0),(0.5,0.5)\}$, $\rho \in \{0.5, 0.7, 0.9\}$, and $R \in \{5, 10, 20\}$

$\rho = 0.5$	$c_a^2 = c_s^2 = 1$			$\rho = 0.7$	$c_a^2 = c_s^2 = 1$			$\rho = 0.9$	$c_a^2 = c_s^2 = 1$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.353	0.405	0.427	UB	1.34	1.39	1.41	UB	7.69	7.69	7.71
LB	0.290	0.262	0.251	LB	1.30	1.31	1.33	LB	7.67	7.62	7.61
$\rho = 0.5$	$c_a^2 = c_s^2 = 0.5$			$\rho = 0.7$	$c_a^2 = c_s^2 = 0.5$			$\rho = 0.9$	$c_a^2 = c_s^2 = 0.5$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.129	0.152	0.162	UB	0.590	0.606	0.608	UB	3.68	3.70	3.66
LB	0.092	0.087	0.086	LB	0.531	0.522	0.534	LB	3.64	3.66	3.64
$\rho = 0.5$	$c_a^2 = c_s^2 = 4$			$\rho = 0.7$	$c_a^2 = c_s^2 = 4$			$\rho = 0.9$	$c_a^2 = c_s^2 = 4$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	1.34	1.44	1.68	UB	5.29	5.37	5.76	UB	30.6	30.4	31.6
LB	1.30	1.27	1.21	LB	5.58	5.54	5.49	LB	30.9	30.7	30.8
$\rho = 0.5$	$c_a^2 = 4, c_s^2 = 0.5$			$\rho = 0.7$	$c_a^2 = 4, c_s^2 = 0.5$			$\rho = 0.9$	$c_a^2 = 4, c_s^2 = 0.5$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	1.33	1.49	1.59	UB	3.64	3.78	4.02	UB	17.9	17.9	18.1
LB	0.356	0.286	0.230	LB	2.65	2.56	2.43	LB	17.5	17.5	17.6
$\rho = 0.5$	$c_a^2 = 0.5, c_s^2 = 4$			$\rho = 0.7$	$c_a^2 = 0.5, c_s^2 = 4$			$\rho = 0.9$	$c_a^2 = 0.5, c_s^2 = 4$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.540	0.548	0.556	UB	2.56	2.56	2.58	UB	16.6	16.6	17.0
LB	0.588	0.591	0.593	LB	2.73	2.74	2.72	LB	16.7	16.7	16.4

TABLE 7. The set-valued approximations of $E[W]$ in $M/M/10$ (upper) and $E_2/E_2/10$ (lower) using case (ii) of (3.8) for $\rho = 0.7$ (left) and $\rho = 0.9$ (right)

$\rho = 0.7$	θ_W			$E[W]$			$\rho = 0.9$	θ_W			$E[W]$		
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.421	0.418	0.415	0.520	0.523	0.539		0.111	0.111	0.110	5.97	6.05	6.07
	0.434	0.437	0.446	0.524	0.520	0.469		0.111	0.111	0.111	6.01	5.94	5.94
$\rho = 0.7$	θ_W			$E[W]$			$\rho = 0.9$	θ_W			$E[W]$		
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.842	0.833	0.825	0.176	0.177	0.179		0.222	0.221	0.221	2.76	2.71	2.74
	0.880	0.889	0.893	0.162	0.162	0.161		0.222	0.223	0.223	2.73	2.74	2.73

$[LB, UB]$ in each case. In contrast, the HTA in (1.2) are 1.633 and 8.10, which seriously overestimates the mean for $K = 10$. However, it is well known that the HTA, which tends to be good for $K = 1$, typically overestimates the mean for $K > 1$; for example, see Ref. [47] and references therein.

6. CONCLUSIONS

In this paper, we investigated how a few additional constraints on an interarrival time U with cdf F and a service times V with cdf G in the $GI/GI/K$ queue can help understand the quality of simple approximations for steady-state performance measures given partial information provided by the first two moments of U and V as specified by the parameter 4-tuple $(1, c_a^2, \rho, c_s^2)$ in (1.4). The idea is to obtain an interval of likely values for performance measures given the partial information. This problem is interesting and important for the

relatively well-understood $GI/GI/1$ model, but it is even more important for the challenging cases with $K > 1$.

As a theoretical basis, we applied extremal models yielding tight upper and lower bounds on the (asymptotic) decay rate of the steady-state waiting-time tail probability, recently established in [8] by applying the theory of Tchebycheff systems. In order to be able to apply Lemma 2.2 of [8], we require that the additional constraints correspond to higher moments and Laplace transform values of F and G.

The decay rate is defined as in (2.1), (3.1), or (3.2). We reviewed the extremal models for the decay rate in the $GI/GI/1$ queue in Section 3. It is significant that these results extend to the $GI/GI/K$ queue. Moreover, we chose scaling with $E[U] = 1$ that made $\theta_W(K) = \theta_W(1) \equiv \theta_W$, so that the extremal distributions for $K > 1$ are simple modifications of the extremal distributions for $K = 1$.

In Section 4, we showed that we can apply the theoretical results for the decay rate in Section 3 to develop a practical way to identify intervals of likely values for the mean steady-state waiting time $E[W]$ given the basic moment parameters in (1.4) and the additional parameters introduced in Theorems 3.1 and 3.2, namely support bounds, the third moments, and values of the Laplace transform. We conducted extensive numerical experiments to study our proposed approach. We found that the proposed method based on cases (i) and (ii) in (3.8) of Theorem 3.2 is consistently effective for a range of base $GI/GI/K$ models. This performance is illustrated in Section 5. For example, with these bounds, Table 5 shows that the maximum error of the midpoint of each interval in case (i) is less than 10% for all four models. We emphasize that this good performance in our estimates of $E[W]$ depends critically on the extra parameters introduced in Theorem 3.2. With only the parameters in (1.4), the range is usually very wide, as shown in Table 1 of [9] and Sect. 2 of [7].

Overall, we contributed to a better understanding of simple queueing approximations such as (1.2) in typical $GI/GI/K$ cases. Our investigation supported extensive experience that the HTA in (1.2) tends to be quite good for $K = 1$, with the understanding that there is a wide range of possibilities. On the other hand, we found that (1.2) seriously overestimates the true value for $K = 10$, which already was a motivation for the many-server approximations for multi-server queues; for example, see Table I of [22]. So far, the method for $K > 1$ has only been studied for small values of K, for example, $K \leq 10$.

6.1. Summary: An Overall Recommended Procedure

While the general idea of set-valued approximations given partial information is relatively simple and natural, there are many possibilities for the specific implementation. We have investigated many, so that the full story is somewhat complicated. Thus, we give a simple final recipe based on Theorem 3.2.

(i) Start with a concrete $GI/GI/K$ model. It suffices to start with $K = 1$, even if interest is in $K > 1$. By (3.3), the extremal model is independent of K if we use the scaling with $E[U] = 1$. (The method has been tested for $K \leq 10$.)

(ii) From that start, obtain the Laplace transforms and first three moments of each of the two underlying distributions F and G. Use the scaling with $E[U] = 1$. From those, obtain the reference decay rate θ_W and mean $E[W]$. The decay rate θ_W is obtained as the unique solution to the Laplace transform Eq. (2.2), given the Laplace transforms of F and G, as depicted in Figure 1. This step produces the basic parameters $(1, c_a^2, \rho, c_s^2)$ in (1.4) of one case, but also the tools we need to go further.

(iii) Obtain the associated support bounds M_a for F and $\rho K M_s$ for G. Choose these to have negligible impact on the tail probability, as specified in (4.1) for $K = 1$. Use $\epsilon = 0.001$, as for the more conservative case in Table 1.

(iv) Choose values of the Laplace transforms using the initial decay rate θ_W and (4.6) depending on the parameter R, and possibly (4.7) if required. We suggest looking at three cases with $R \in \{5, 10, 20\}$ to show a range, but if we had to pick only one, then we would choose $R = 20$ because it is the most conservative choice.

(v) Next, numerically determine the extremal distributions $(F_{\rm L}, F_{\rm U}, G_{\rm L}, G_{\rm U})$ specified by Theorem 3.2, as indicated in Section 4.2. We suggest using a nonlinear equation solving algorithm as in MATLAB.

(vi) Then, calculate the decay rate θ_W for each extremal model. Again, the decay rate can be computed from (2.2), using the easily constructed Laplace transforms of the three-point extremal distributions in $(F_{\rm L}, F_{\rm U}, G_{\rm L}, G_{\rm U})$; for example, $\hat{f}_{\rm L}(s) = \sum_{i=1}^{3} e^{-sx_i} p_i$, when the three mass points are (x_1, x_2, x_3) with associated positive probabilities (p_1, p_2, p_3) for $F_{\rm L}$.

(vii) Finally, the extremal values of the mean $E[W]$ (and other performance measures of interest) can be estimated by doing simulation of the extremal queueing models. The mean and decay rate for the base model provides a consistency check for this final step.

6.2. Directions for Future Research

There are many directions for future research. First, it remains to expose the precise relation between $E[W]$ and θ_W. (There is useful theory in Sect. II.5 of [13]. Some numerical work appears in Ref. [3].) Second, it remains to explore the approximation for other performance measures such as the tail probability $P(W > t)$. We expect even better results for large t, but then worse results for $t = 0$; see Ref. [3]. Thm. 1 of [6] shows that tight upper and lower bounds can be obtained directly fir higher moments $E[W^k]$ for $K = 1$, but it remains to consider $K > 1$.

Third, there are many opportunities for further work with $K > 1$, including related to many-server heavy-traffic scaling in [22]. It remains to develop and study procedures for large values of K. There is also opportunity for improved rare-event simulation for the extremal queues with $K > 1$ paralleling [38] used for $K = 1$ in [9]; see Ref. [37] for some. Finally, we think that there is great potential for applying this general approach for obtaining set-valued approximations given partial model information to other stochastic models.

Acknowledgment

Research support was received from NSF grant CMMI 1634133.

References

1. Abate, J. & Whitt, W. (1994). A heavy-traffic expansion for the asymptotic decay rates of tail probabilities in multi-channel queues. *Operations Research Letters* 15: 223–230.
2. Abate, J., Choudhury, G.L., & Whitt, W. (1993). Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Archiv fur Elektronik und Ubertragungstechnik* 47(5/6): 311–321.
3. Abate, J., Choudhury, G.L., & Whitt, W. (1995). Exponential approximations for tail probabilities in queues, I: Waiting times. *Operations Research* 43(5): 885–901.

4. Asmussen, S. (2003). *Applied probability and queues*, 2nd ed. New York: Springer.
5. Borovkov, A.A. (1965). Some limit theorems in the theory of mass service, II. *Theory of Probability and Its Applications* 10: 375–400.
6. Chen, Y. & Whitt, W. (2019). Extremal $GI/GI/1$ queues given two moments: Exploiting Tchebycheff systems. Submitted for publication, Columbia University. http://www.columbia.edu/~ww2040/allpapers.html.
7. Chen, Y. & Whitt, W. (2020). Appendix to set-valued performance approximations for the $GI/GI/K$ queue given partial information. Columbia University. http://www.columbia.edu/~ww2040/allpapers.html.
8. Chen, Y. & Whitt, W. (2020). Extremal models for the $GI/GI/K$ waiting-time tail-probability decay rate. *Operations Research Letters* 48: 770–776.
9. Chen, Y. & Whitt, W. (2020). Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue. *Queueing Systems* 94: 327–356.
10. Choudhury, G.L. & Whitt, W. (1994). Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/G/1$ queue. *Stochastic Models* 10(2): 453–498.
11. Choudhury, G.L., Lucantoni, D., & Whitt, W. (1996). Squeezing the most out of ATM. *IEEE Transactions on Communications* 44(2): 203–217.
12. Chung, K.L. (2001). *A course in probability theory*, 3rd ed. New York: Academic Press.
13. Cohen, J.W. (1982). *The single server queue*, 2nd ed. Amsterdam: North-Holland.
14. Daley, D.J. (1977). Inequalities for moments of tails of random variables, with queueing applications. *Zeitschrift fur Wahrscheinlichkeitsetheorie Verw Gebiete* 41: 139–143.
15. Daley, D.J. (1997). Some results for the mean waiting-time and workloads in the $GI/GI/k$ queue. In J.H. Dshalalow (ed.), *Froniers in queueing: Models and applicatiions in science and engineering*. Boca Raton, FL: CRC Press, pp. 35–59.
16. Daley, D.J., Kreinin, A.Y., & Trengove, C. (1992). Inequalities concerning the waiting-time in single-server queues: A survey. In U.N. Bhat and I.V. Basawa (eds), *Queueing and related models*. Oxford: Clarendon Press, pp. 177–223.
17. Eckberg, A.E. (1977). Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Mathematics of Operations Research* 2(2): 135–142.
18. Gamarnik, D. & Goldberg, D.A. (2013). Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Annals of Applied Probability* 23(6): 2382–2419.
19. Glynn, P.W. & Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* 31: 131–156.
20. Gupta, V. & Osogami, T. (2011). On Markov-Krein characterization of the mean waiting time in $M/G/K$ and other queueing systems. *Queueing Systems* 68: 339–352.
21. Gupta, V., Dai, J., Harchol-Balter, M., & Zwart, B. (2010). On the inapproximability of $M/G/K$: Why two moments of job size distribution are not enough. *Queueing Systems* 64: 5–48.
22. Halfin, S. & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3): 567–588.
23. Holtzman, J.M. (1973). The accuracy of the equivalent random method with renewal inouts. *Bell System Technical Journal* 52(9): 1673–1679.
24. Iglehart, D.L. & Whitt, W. (1970). Multiple channel queues in heavy traffic, I. *Advances in Applied Probability* 2(1): 150–177.
25. Iglehart, D.L. & Whitt, W. (1970). Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability* 2(2): 355–369.
26. Johnson, M.A. & Taaffe, M.R. (1991). An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queieing Systems* 8(1–2): 129–148.
27. Johnson, M.A. & Taaffe, M.R. (1993). Tchebycheff systems for probability analysis. *American Journal of Mathematical and Management Sciences* 13(1–2): 83–111.
28. Karlin, S. & Studden, W.J. (1966). *Tchebycheff systems: With applications in analysis and statistics*, vol. 137. New York: Wiley.
29. Kelly, F.P. (1996). Notes on effective bandwidths. In F.P. Kelly, S. Zachary, and I. Ziedins (eds), *Stochastic networks: Theory and applications*. Oxford: Clarendon Press, pp. 141–168.
30. Kingman, J.F.C. (1961). The single server queue in heavy traffic. *Proceedings of the Cambridge Philosophical Society* 77: 902–904.
31. Kingman, J.F.C. (1962). Inequalities for the queue $GI/G/1$. *Biometrika* 49(3/4): 315–324.
32. Kingman, J.F.C. (1964). A martingale inequality in the theory of queues. *Proceedings of the Cambridge Philosophical Society* 59: 359–361.

33. Kingman, J.F.C. (1966). The heavy traffic approximation in the theory of queues. In W.L. Smith and W.E. Wilkinson (eds), *Proceedings of the Symposium on Congestion Theory*. Chael Hill, NC: The University of North Carolina Press, pp. 137–159.
34. Kingman, J.F.C. (1970). Inequalities in the theory of queues. *Journal of the Royal Statistical Society: Series B* 32(1): 102–110.
35. Klincewicz, J. & Whitt, W. (1984). On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Technical Journal* 63(1): 115–138.
36. Kollerstrom, J. (1974). Heavy traffic theory for queues with several servers. *Journal of Applied Probability* 11(3): 544–552.
37. Minh, D.L. (1989). Simulating $GI/G/k$ queues in heavy traffic. *Management Science* 33(9): 1192–1199.
38. Minh, D.L. & Sorli, R.M. (1983). Simulating the $GI/G/1$ queue in heavy traffic. *Operations Research* 31(5): 966–971.
39. Neuts, M.F. (1986). The caudal characteristic curve of queues. *Advances in Applied Probability* 18: 221–254.
40. Neuts, M.F. & Takahashi, Y. (1981). Asymptotic behavior of stationary distributions in the $GI/PH/C$ queue with heterogeneous servers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57: 441–452.
41. Rolski, T. (1972). Some inequalities for $GI/M/n$ queues. *Zastosowania Matematyki Applicationes Mathematicae* 13(1): 43–47.
42. Whitt, W. (1982). Approximating a point process by a renewal process: Two basic methods. *Operations Research* 30: 125–147.
43. Whitt, W. (1983). The queueing network analyzer. *Bell Laboratories Technical Journal* 62(9): 2779–2815.
44. Whitt, W. (1984). On approximations for queues, I. *AT&T Bell Laboratories Technical Journal* 63(1): 115–137.
45. Whitt, W. (1984). On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal* 63(1): 163–175.
46. Whitt, W. (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multiclass queues. *Telecommunication Systems* 2: 71–107.
47. Whitt, W. (2004). A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52(6): 922–941.
48. Wolff, R.W. & Wang, C. (2003). Idle period approximations and bounds for the $GI/G/1$ queue. *Advances in Applied Probability* 35(3): 773–792.