Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Extremal $GI/GI/1$ Queues Given Two Moments

Yan Chen
Industrial Engineering and Operations Research, Columbia University, yc3107@columbia.edu
Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu

This paper studies upper bounds for the mean (steady-state and transient) waiting time in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions. For distributions with support on bounded intervals, we show that the upper bounds (with one distribution given and overall) are attained at distributions with support on at most three points. The proof exploits fixed point theory and optimization theory in addition to standard stochastic theory for the model. We then apply relatively tractable numerical algorithms to identify the optimal distributions within that class. For the overall upper bound with unbounded support sets, we propose a simple approximation formula and provide a numerical comparison of the approximations and bounds, showing that the new approximate bound is very accurate.

Key words: GI/GI/1 queue, tight bounds, extremal queues, bounds for the mean steady-state mean waiting time, moment problem

History: March 16, 2019

1. Introduction

In this paper we address a long-standing open problem for the classical $GI/GI/1$ queueing model: determining a tight upper bound for the mean steady-state waiting time, given the first two moments of the interarrival-time and service-time distributions; see Daley et al. (1992), especially §10, Wolff and Wang (2003) and references therein.

1.1. The $GI/GI/1$ Model

The $GI/GI/1$ single-server queue has unlimited waiting space and the first-come first-served service discipline. There is a sequence of independent and identically distributed (i.i.d.) service times $\{V_n : n \geq 0\}$, each distributed as V with cumulative distribution function (cdf) G, which is independent of a sequence of i.i.d. interarrival times $\{U_n : n \geq 0\}$ each distributed as U with cdf F. With the understanding that a 0^{th} customer arrives at time 0 to find an empty system, V_n is the service time of customer n, whle U_n is the interarrival time between customers n and $n+1$.

Let U have mean $\mathbb{E}[U] \equiv \lambda^{-1} \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $\mathbb{E}[V] \equiv \tau \equiv \rho$ and scv c_s^2, where $\rho \equiv \lambda\tau < 1$, so that the model is stable. (Let $\equiv$ denote equality by definition.)

Let W_n be the waiting time of customer n, i.e., the time from arrival until starting service, assuming that the system starts empty with $W_0 \equiv 0$. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

$$W_{n+1} = [W_n + V_n - U_n]^+, \quad n \geq 0, \tag{1}$$

where $x^+ \equiv \max\{x, 0\}$. Let W be the steady-state waiting time. It is also well known that $W_n \stackrel{\text{d}}{=} \max\{S_k : 0 \leq k \leq n\}$ and $W \stackrel{\text{d}}{=} \max\{S_k : k \geq 0\}$, where $\stackrel{\text{d}}{=}$ denotes equality in distribution, $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_k - U_k$, $k \geq 1$; e.g., see §§X.1-X.2 of Asmussen (2003) or (13) in §8.5 of Chung (2001). It is also known that, under the specified finite moment conditions, W_n and W are proper random variables with finite means, given by

$$E[W_n] = \sum_{k=1}^{n} \frac{\mathbb{E}[S_k^+]}{k} < \infty \quad \text{and} \quad E[W] = \sum_{k=1}^{\infty} \frac{\mathbb{E}[S_k^+]}{k} < \infty. \tag{2}$$

1.2. Classical Results: Exact, Approximate and Bounds

For the $M/GI/1$ special case, when the interarrival time has an exponential distribution, we have the classical Pollaczek-Khintchine formula

$$E[W] = \frac{\tau\rho(1+c_s^2)}{2(1-\rho)} = \frac{\rho^2(1+c_s^2)}{2(1-\rho)}. \tag{3}$$

A natural commonly used approximation for the $GI/GI/1$ model, inspired by (3), which we call the heavy-traffic approximation, because it is motivated by the early heavy-traffic limit in Kingman (1961), is

$$E[W] \equiv E[W(\rho, c_a^2, c_s^2)] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}. \tag{4}$$

The most familiar upper bound (UB) on $E[W]$ is the Kingman (1962) bound,

$$E[W] \le \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}, \tag{5}$$

which is known to be asymptotically correct in heavy traffic (as $\rho \to 1$).

A better UB depending on these same parameters was obtained by Daley (1977). in particular, the Daley (1977) UB replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$, i.e.,

$$E[W] \le \frac{\rho^2([(2-\rho)c_a^2/\rho] + c_s^2)}{2(1-\rho)}. \tag{6}$$

Note that $(2-\rho)/\rho < 1/\rho^2$ because $\rho(2-\rho) < 1$ for all ρ, $0 < \rho < 1$.

In contrast to the tight UB that we study, the tight lower bound (LB) for the steady-state mean has been known for a long time; see Stoyan and Stoyan (1974), §5.4 of Stoyan (1983), §V of Whitt (1984b), Theorem 3.1 of Daley et al. (1992) and references there:

$$E[W(LB)] = \frac{\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}. \tag{7}$$

The LB is attained asymptotically at a deterministic interarrival time with the specified mean and at any three-point service-time distribution that has all mass on nonnegative-integer multiples of the deterministic interarrival time. The service part follows from Ott (1987). (All service-time distributions satisfying these requirements yield the same mean.)

1.3. Motivation: Approximations for Non-Markovian Open Queueing Networks

Our original interest in the bounds was primarily motivated by parametric-decomposition approximations for non-Markovian open networks of single-server queues, as in Whitt (1983b), where each queue is approximated by a $GI/GI/1$ queue partially characterized by the parameter vector

$(\lambda, c_a^2, \tau, c_s^2)$, obtained by solving traffic rate equations for the arrival rate λ at each queue and after solving associated traffic variability equations to generate an approximating scv c_a^2 of the arrival process. Because the internal arrival processes are usually not renewal and the interarrival distribution is not known, there is no concrete $GI/GI/1$ model to analyze more carefully. To gain some insight into these approximations (not yet addressing the dependence among interarrival times), It is natural to regard such approximations for the $GI/GI/1$ model as set-valued functions, applying to all models with the same parameter vector $(\lambda, c_a^2, \tau, c_s^2)$.

For the special case of the $GI/M/1$ model with bounded interval of support for the interarrival-time cdf F, the extremal $GI/M/1$ models were studied in Whitt (1984b), where intervals of bounded support were also used together with the theory of Tchebychev systems; as in Karlin and Studden (1966). (The focus in Whitt (1984b) was on the mean steady state number in system, but it is easily seen that the extremal interarrival-time distributions are the same for the mean steady-state waiting time.) For the $GI/M/1$ model, the extremal distributions are two-point distributions.

Let $\mathcal{P}_{2,2}(M) \equiv \mathcal{P}_{2,2}(m_1, c^2, M)$ be the set of all two-point distributions with mean m_1 and second moment $m_2 = m_1^2(c^2+1)$ with support in $[0, m_1 M]$. The set $\mathcal{P}_{2,2}(M)$ is a one-dimensional parametric family. Any element has probability mass $c_a^2/(c^2+(b-1)^2)$ at $m_1 b$, and mass $(b-1)^2/(c_a^2+(b-1)^2)$ on $m_1(1-c_a^2/(b-1))$ for $1+c^2 \le b \le M$. The cases $b = 1+c^2$ and $b = M$ constitute the two extremal distributions.

For $GI/M/1$, the UB interarrival-time cdf with mean m_1 and second moment $m_2 = m_1^2(c_a^2+1)$ with support in $[0, m_1 M_a]$, referred to here as F_0, arises for $b = 1 + c^2$. In particular, F_0 has probability mass $c_a^2/(1+c_a^2)$ at 0 and probability mass $1/(c_a^2+1)$ at $(m_2/m_1) = m_1(c_a^2+1)$.

The corresponding LB interarrival-time cdf, referred to here as F_u, arises for $b = M_a$. In particular, F_u has probability mass $c_a^2/(c_a^2+(M_a-1)^2)$ at the upper bound of the support, $m_1 M_a$, and mass $(M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $m_1(1-c_a^2/(M_a-1))$. (For the interarrival time, we scale, i.e., choose measuring units for time, so that $m_1 = 1$.) We use the notation G_0 and G_u for the corresponding service-time cdf's G with mean ρ and support $[0, \rho M_s]$.

That technical approach and the basic results were first established by Rolski (1972) and Holtzman (1973), and then elaborated on by Eckberg (1977) and Johnson and Taaffe (1993). Since the range of possible values is quite large, while the distributions that attain the bounds are unusual (two-point distributions), the papers Klincewicz and Whitt (1984), Whitt (1984c) and Johnson and Taaffe (1990a) focused on reducing the range by imposing shape constraints. In this paper we do not consider shape constraints.

1.4. Related Literature

The literature on bounds for the $GI/GI/1$ queue is well reviewed in Daley et al. (1992) and Wolff and Wang (2003), so we will be brief. The use of optimization to study the bounding problem for queues seems to have begun with Klincewicz and Whitt (1984) and Johnson and Taaffe (1990b). Bertsimas and Natarajan (2007) provides a tractable semi-definite program as a relaxation model for solving steady-state waiting time of $GI/GI/c$ to derive bounds, while Osogami and Raymond (2013) bounds the transient tail probability of $GI/GI/1$ by a semi-definite program.

Several researchers have studied bounds for the more complex many-server queue. In addition to Bertsimas and Natarajan (2007), Gupta et al. (2010) and Gupta and Osogami (2011) investigate the bounds and approximations of the $M/GI/c$ queue. Gupta et al. (2010) explains why two moment information is insufficient for good accuracy of steady-state approximations of $M/GI/c$. Gupta and Osogami (2011) establishes a tight bound for the $M/GI/K$ in light traffic. Finally, Li and Goldberg (2017) establish bounds for $GI/GI/c$ intended for the many-server heavy-traffic regime.

1.5. Organization

In §2 we obtain our main result, Theorem 1, which shows that there exist extremal interarrival-time and service-time cdf's that have support on at most three points when the interarrival-time and service-time cdf's have bounded support. The proof of Theorem 1 is based on a new line of reasoning. It draws on the Lindley recursion, the Kakutani fixed point theorem, the general

moment problem and the duality theory of linear programming. Theorem 1 We obtain an analog of Theorem 1 for the transient mean $E[W_n]$ in §3. We identify more structure of the extremal distributions in a large class of special cases in §4.

We start our numerical studies in §5 by introducing a multinomial formulation for the transient mean $E[W_n]$ over the product space of the two sets of three-point distributions. We use that multinomial representation to formulate a non-convex nonlinear program for the overall UB, which we solve by applying sequential quadratic programming (SQP) as discussed in Ch. 18 of Nocedal and Wright (1999). The SQP algorithm converges at a local optimum, so we apply it with randomly selected initial conditions. We found that all local optima for the overall UB are two-point distributions and that the best local optimum always has interarrival-time cdf F_0. See §5.3 for our final conclusions. In §6 we do a careful simulation study over the product space of two-point distributions. Since the two-point distributions form a one-parameter family, we are able to expose more of the structure of the mean waiting times. Finally, in §7 we draw conclusions.

We provide additional supporting material in the e-companion (EC). We start in §EC.2 by providing the postponed part of the proof of Theorem 1 in §2. In §EC.3 we provide postponed proofs for §4. In §EC.4 we present an alternative way to identify extremal distributions by combining Tchebycheff systems with our proof of Theorem 1. In §EC.5 we prove Theorem 6 establishing a new upper bound formula. We discuss the extension to unbounded support in §EC.6. Finally, we present additional tables and plots in the e-companion. In Chen and Whitt (2018) we develop and evaluate algorithms for the conjectured tight overall UB for $E[W]$ in §5.3 here. In Chen and Whitt (2019) we obtain new results for the LB with finite support and other constraints on the interarrival-time cdf F.

2. Reduction to Three-Point Distributions

In this section we show that it suffices to consider interarrival-time and service-time cdfs with support on at most three points in our search for upper bounds on the steady-state mean waiting time $E[W]$.

let $\mathcal{P}_n$ be the set of all probability measures on a subset of $\mathbb{R}$ with specified first n moments. The set $\mathcal{P}_n$ is a convex set, because the convex combination of two probability measures is just the mixture; i.e., for all p, $0 \le p \le 1$,

$$P_{mix,p} \equiv pP_1 + (1-p)P_2 \in \mathcal{P}_n \quad \text{if} \quad P_1 \in \mathcal{P}_n \quad \text{and} \quad P_2 \in \mathcal{P}_n, \tag{8}$$

because the n^{th} moment of the mixture is the mixture of the n^{th} moments, which is just the common value of the components. let $\mathcal{P}_{n,k}$ be the subset of probability measures in $\mathcal{P}_n$ that have support on at most k points.

We use the scv to parameterize, so let $\mathcal{P}_2 \equiv \mathcal{P}_2(m, c^2)$ be the set of all cdf's with mean m and second moment $m^2(c^2+1)$ where $c^2 < \infty$. Let $\mathcal{P}_2(M) \equiv \mathcal{P}_2(m, c^2, M)$ be the subset of all cdf's in $\mathcal{P}_2$ with support in the closed interval $[0, mM]$ having mean m and second moment $m^2(c^2+1)$ where $c^2 + 1 < M < \infty$. (The last property ensures that the set $\mathcal{P}_2(M)$ is non-empty.) Let subscripts a and s denote sets for the interarrival and service times, respectively.

We are interested in the map

$$w : \mathcal{P}_{a,2}(1, c_a^2) \times \mathcal{P}_{s,2}(\rho, c_s^2) \to \mathbb{R}, \tag{9}$$

where $0 < \rho < 1$ and

$$w(F, G) \equiv E[W(F, G)] \tag{10}$$

for W being a random variable with the distribution of the steady-state waiting time in the $GI/GI/1$ queue with interarrival-time cdf $F \in \mathcal{P}_{a,2}$ and service-time cdf $G \in \mathcal{P}_{s,2}$.

The function w in (10) has explicit form in (2) and an algorithm is given in Abate et al. (1993), but that algorithm has an analytic property of the transform that is not suitable for the present problem.) Note that the mean interarrival time is 1 and the mean service time is ρ, so that the traffic intensity is ρ.

THEOREM 1. (*reduction to a three-point distribution*) *Consider the class of $GI/GI/1$ queues with interarrival times $\{U_n\}$ distributed as U with cdf $F \in \mathcal{P}_{a,2}$ and service times $\{V_n\}$ distributed as*

V with cdf $G \in \mathcal{P}_{s,2}$ where $0 < \rho < 1$ and the sets $\mathcal{P}_{a,2}$ and $\mathcal{P}_{s,2}$ are nonempty. The function $w : \mathcal{P}_{a,2} \times \mathcal{P}_{s,2} \to \mathbb{R}$ in (9) is continuous. Hence, the following suprema are attained as indicated:

(a) For any specified $G \in \mathcal{P}_{s,2}$ and $1 + c_a^2 \le M_a < \infty$, there exists $F^*(G) \in \mathcal{P}_{a,2,3}(M_a)$ such that

$$w_a^{\uparrow}(G) \equiv \sup\{w(F,G) : F \in \mathcal{P}_{a,2}(M_a)\} = \sup\{w(F,G) : F \in \mathcal{P}_{a,2,3}(M_a)\} = w(F^*(G),G). \tag{11}$$

(b) For any specified $F \in \mathcal{P}_{a,2}$ and $1 + c_s^2 \le M_s < \infty$, there exists $G^*(F) \in \mathcal{P}_{s,2,3}(M_s)$ such that

$$w_s^{\uparrow}(F) \equiv \sup\{w(F,G) : G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w(F,G) : G \in \mathcal{P}_{s,2,3}(M_s)\} = w(F,G^*(F)). \tag{12}$$

(c) For any given (M_a, M_s) with $1 + c_a^2 \le M_a < \infty$ and $1 + c_s^2 \le M_s < \infty$, there exists (F^{**}, G^{**}) in $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$ such that

$$\begin{aligned} w^{\uparrow} &\equiv \sup\{w(F,G) : F \in \mathcal{P}_{a,2}(M_a), G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w(F,G) : F \in \mathcal{P}_{a,2,3}(M_a), G \in \mathcal{P}_{s,2,3}(M_s)\} \\ &= w(F^{**},G^{**}) = w_a^{\uparrow}(G^{**}) = w_s^{\uparrow}(F^{**}). \end{aligned} \tag{13}$$

REMARK 1. (uniqueness) There is no claim of uniqueness in Theorem 1. Indeed, the $M/GI/1$ formula in (3) implies that there is no uniqueness in case (b) when F is exponential; see Remark EC.1 for more discussion.

We give a brief sketch of the proof in §2.2, and then give more details in §EC.2. Since the proof draws on results for the moment problem, we review that next.

2.1. The Moment Problem for Distributions with Compact Support

Our problem can be approached via the classical theory for the moment problem, as in Lasserre (2010), Smith (1995) and references therein. Some simplification can be gained by considering continuous functions on a compact metric space domain, so that suprema and infima are attained.

For the general moment problem, let $\mathcal{P}_n \equiv \mathcal{P}_n(\mathcal{C})$ be the set of all probability measures on a compact subset $\mathcal{C}$ of $\mathbb{R}$ with specified first n moments, where the k^{th} moment of P is defined as $\int x^k \, dP$. Assume that $\mathcal{P}_n$ is not empty and let $\mathcal{P}_n$ be endowed with the topology of weak

convergence, as determined by the Prohorov or Lévy metric, as in §3.2 and §11.3 of Whitt (2002). let $\mathcal{P}_{n,k}$ be the subset of probability measures in $\mathcal{P}_n$ that have support on at most k points in $\mathcal{C}$.

The following is a generalization of a standard result in linear programming (LP), stating that the supremum (or infimum) is attained at a basic feasible solution or an extreme point. The set of extreme points of the set $\mathcal{P}_n$ is the subset $\mathcal{P}_{n,n+1}$. In fact, our proof of Theorem 1 will only use the LP version where $\mathcal{C}$ has finite support.

THEOREM 2. (*a version of the classic moment problem*) *Let $\phi : \mathcal{C} \to \mathbb{R}$ be a continuous function, where $\mathcal{C}$ is a compact subset of $\mathbb{R}$. Assume that $\mathcal{P}_n$ is not empty. Then there exists $P^* \in \mathcal{P}_{n,n+1}$ such that*

$$\sup\{\int_0^M \phi\, dP : P \in \mathcal{P}_n\} = \sup\{\int_0^M \phi\, dP : P \in \mathcal{P}_{n,n+1}\} = \sum_{k=1}^{n+1} \phi(t_k) P^*(\{t_k\}), \tag{14}$$

where $\{t_k : 1 \le k \le n+1\}$ is the support of P^*.

.

Proof. First, because the support $\mathcal{C}$ is a compact subset of $\mathbb{R}$ and the set $\mathcal{P}_n$ is not empty by assumption, the space $\mathcal{P}_n$ is a compact metric space with the usual topology of convergence in distribution, as a consequence of Prohorov's theorem; e.g., Theorem 11.6.1 of Whitt (2002). (In general, the set of all probability measures on a compact metric space with the usual topology of weak convergence is itself a compact metric space; see Theorem II.6.4 of Parthasarathy (1967).)

Second, because the function ϕ is continuous, we can apply the continuous mapping theorem as in §3.4 of Whitt (2002) to deduce that the induced map $\phi : \mathcal{P}_n \to \mathbb{R}$ defined by

$$\phi(P) \equiv \int_0^b \phi\, dP \tag{15}$$

is continuous as well. Hence, the induced map in (15) is a continuous bounded real-valued function on a compact metric space, so that the supremum in (14) is attained. Then the theory for the classical moment problem implies that it is attained in $\mathcal{P}_{n,n+1}$; see §2 of Smith (1995). ∎

2.2. Sketch of the Proof of Theorem 1

We now outline the proof of part (a); see §EC.2 for more details. The proof of part (b) is very similar, aided by using a reverse-time argument; see Step 5 in §EC.2. Then (c) is a well known consequence of both (a) and (b); e.g., see Lemma EC.1 in the e-companion to Whitt and You (2018). So consider (a).

Let $W(F,G)$ be the steady-state waiting time for any specified interarrival-time cdf $F \in \mathcal{P}_{a,2}(M_a)$ and service-time cdf $G \in \mathcal{P}_{s,2}$. Let $\mathcal{E} \equiv \mathcal{E}(G)$ be the set of cdfs F^* that attain the supremum

$$E[W(F^*,G)] = \sup\{E[W(F,G)] : F \in \mathcal{P}_{a,2}(M_a)\} \tag{16}$$

for any given cdf $G \in \mathcal{P}_{s,2}$. We know that the supremum must be attained, i.e., $\mathcal{E} \neq \emptyset$, because we are maximizing a continuous function over a compact metric space. (However, we are not claiming uniqueness for F^* in (16).) Our goal here is to show that

$$\mathcal{E} \bigcap \mathcal{P}_{a,2,3}(M_a) \neq \emptyset. \tag{17}$$

To show that F^* in (16) can be chosen in $\mathcal{P}_{a,2,3}$, i.e., to establish (17), we introduce a new line of reasoning for this problem. In particular, we exploit fixed point theory and optimization theory. The basis is the classical Lindley recursion for the waiting time in (1). It is well known that the distribution of the steady-state waiting time $W(F,G)$ is the unique solution to the stochastic fixed-point equation

$$W(F,G) \stackrel{\mathrm{d}}{=} [W(F,G) + V - U]^+, \tag{18}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution, while the three random variables on the right are independent with the distributions of V and U being G and F, respectively.

The key initial step is to reformulate our goal in (17) as a fixed point problem.

Step 1. Characterization as a fixed point. Let the cdf G of the service-time V be given and fixed. Let U_F denote the random variable U with cdf F. Given (18), for any F^* satisfying (16),

we can identify a subset of the F^* in $\mathcal{E}$ by formulating a fixed point problem over $\mathcal{P}_{a,2}(M_a)$. First, observe that

$$\begin{aligned} E[W(F^*,G)] &= \sup\{E[(W(F_1,G)+V-U_{F_1})^+]: F_1 \in \mathcal{P}_{a,2}(M_a)\} \\ &= \sup\{E[(W(F_1,G)+V-U_{F_2})^+]: F_1 = F_2 \in \mathcal{P}_{a,2}(M_a)\}, \end{aligned} \tag{19}$$

where the three random variables $W(F_1,G)$, V and U_{F_2} are mutually independent and $W(F_1,G)$ has the steady-state distribution associated with (F_1,G). In particular, note that $W(F_1,G)$ is the steady-state waiting time when the interarrival time has cdf F_1, but U_{F_2} has the cdf F_2, so that we must also require that $F_1 = F_2$. Also note that the second supremum in (19) is over both F_1 and F_2, subject to the constraint.

We approach the reformulated optimization in (19) by first ignoring the constraint $F_1 = F_2$ and the optimization over F_1. Afterwards, we impose those conditions by finding a fixed point. Hence, we consider the optimization problem

$$\zeta(F_1) \equiv \sup\{E[(W(F_1,G)+V-U_{F_2})^+]: F_2 \in \mathcal{P}_{a,2}(M_a)\}, \tag{20}$$

where the three random variables $W(F_1,G)$, V and U_{F_2} are mutually independent and $W(F_1,G)$ has the steady-state distribution associated with (F_1,G). We next impose the requirement that $F_1 = F_2$.

To impose the requirement that $F_1 = F_2$, we construct a map η mapping the space $\mathcal{P}_{a,2}(M_a)$ into the set $2^{\mathcal{P}_{a,2}(M_a)}$ of all subsets of $\mathcal{P}_{a,2}(M_a)$. Let $\eta(F_1)$ be the set of all cdfs F_2 attaining the supremum of the function ζ in (20). Let $\mathcal{P}^*_{a,2}$ be the subset of all fixed points of the map η; i.e.,

$$\mathcal{P}^*_{a,2} \equiv \{F \in \mathcal{P}_{a,2}(M_a) : F \in \eta(F)\}. \tag{21}$$

Clearly, if $\bar{F} \in \mathcal{P}^*_{a,2}$, then

$$\sup\{E[(W(\bar{F},G)+V-U_{F_2})^+]: F_2 \in \mathcal{P}_{a,2}(M_a)\} = E[(W(\bar{F},G)+V-U_{\bar{F}})^+]. \tag{22}$$

However, we still need to optimize over $\bar{F}$ in (22), which corresponds to optimizing over F_1 in (19) as well as F_2. Let F^* be a cdf that attains the supremum over $\bar{F}$ in (22). That supremum over

$\bar{F}$ will be attained because it is for a continuous function over a compact set. That cdf F^* also satisfies (19) and so must be in $\mathcal{E}$. Hence, we have shown that

$$\mathcal{E} \bigcap \mathcal{P}^*_{a,2} \neq \emptyset. \tag{23}$$

Thus we will achieve our goal by proving that the set $\mathcal{P}^*_{a,2}$ is a nonempty subset of $\mathcal{P}_{a,2,3}(M_a)$. However, note that we cannot claim set containment between $\mathcal{E}$ and $\mathcal{P}^*_{a,2}$ in either direction.

The Rest of the proof: Steps 2-5. We complete the proof for case (a) in Steps 2-4. In Step 2 we prove that the set $\mathcal{P}^*_{a,2}$ is nonempty by applying the Kakutani fixed point theorem. To apply the Kakutani fixed point theorem, we first restrict attention to cdf's F with finite support. We then treat the general case by a limiting argument in Step 4. In Step 3 we show that, given that F has the assumed finite support, $\mathcal{P}^*_{a,2}$ is a subset of $\mathcal{P}_{a,2,3}(M_a)$; i.e., the fixed points must be three-point distributions. We do two supporting asymptotic arguments in Step 4. We then extend the result to case (b) in Step 5. We now elaborate on Step 3. For the remaining technical details, see §EC.2.

Step 3(a). Applying Theorem 2. To show that $\mathcal{P}^*_{a,2}$ is a subset of $\mathcal{P}_{a,2,3}(M_a)$, we exploit Theorem 2, which is only an ordinary linear program (LP) with the finite support. To do so, we write (20) in the form of (14). In particular, for G the fixed cdf of the service time V and H the cdf of the waiting time $W(F_1, G)$ with finite mean, we can write

$$\sup\{E[(W(F_1,G)+V-U^+_{F_2}] : F_2 \in \mathcal{P}_{a,2}(M_a)\} = \sup\{\int_0^{M_a} \phi(u)\, dF_2 : F_2 \in \mathcal{P}_{a,2}(M_a)\} \tag{24}$$

for ϕ expressed as the double integral

$$\phi(u) \equiv \int_0^\infty \int_0^\infty (x+v-u)^+ \, dG(v)\, dH(x), \quad 0 \le u \le M_a. \tag{25}$$

Next observe that ϕ in (25) is a bounded continuous real-valued function of u because the cdfs G and H have bounded mean. Hence, we can apply Theorem 2 to deduce that, for any pair of cdf's (G,H) of (V,W), we may take $F_2 \in \mathcal{P}_{a,2,3}(M_a)$ in the optimization in (24).

Step 3(b). Uniqueness in the LP. We next prove that, when we take $F_1 = F^*$ for F_1 in (24) and $F^* \in \mathcal{P}^*_{a,2}$, the LP in (24) restricted to finite support always has a unique solution. Thus,

there is no solution other than the one that must be in $\mathcal{P}_{a,2,3}(M_a)$. Hence, we have shown that $\mathcal{P}^*_{a,2} \subseteq \mathcal{P}_{a,2,3}(M_a)$. (Note that we are showing that the LP in (24) has a unique solution for a particular F_1 in $\mathcal{P}^*_{a,2}$. That does not mean that (16) necessarily has a unique solution.)

To establish that uniqueness of the solution of the LP in (24), we apply duality theory for linear programs. The objective of the dual problem is to find the vector $\lambda^* \equiv (\lambda^*_0, \lambda^*_1, \lambda^*_2)$ that attains the infimum

$$\gamma(m_1, m_2) \equiv \inf_{\lambda \equiv (\lambda_0, \lambda_1, \lambda_2)} \{\lambda_0 + \lambda_1 m_1 + \lambda_2 m_2\}, \tag{26}$$

where $m_i \equiv E[U^i]$, $i = 1, 2$ and λ_i are the decision variables (which are unconstrained), such that

$$\psi(u) \equiv \lambda_0 + \lambda_1 u + \lambda_2 u^2 \geq \phi(u) \quad \text{for all} \quad u \in \mathcal{F} \tag{27}$$

where $\mathcal{F}$ is the support of F and

$$\phi(u) \equiv \int_0^\infty \int_0^\infty (x + v - u)^+ dH(x) dG(v) = \int_0^\infty (x - u)^+ d\Gamma(x) \tag{28}$$

where Γ is the cdf of $W + V$, as in (25).

In particular, we establish uniqueness by showing that the dual LP has a non-degenerate optimal solution; i.e., we apply the following lemma; e.g., see pp. 1128-1129 of Appa (2002).

LEMMA 1. (*non-degeneracy and uniqueness in LP*) *A standard LP has a unique optimal solution if and only if its dual has a nondegenerate optimal solution.*

It is easy to see that both the primal LP and the dual LP have feasible solutions. Hence, we see that the dual LP problem has at least one optimal solution. Given that the dual has an optimal solution, we show that the dual has a non-degenerate optimal solution by showing that the dual does not have a degenerate optimal solution. To do so, we first determine the structure of the function ϕ in (25) for case (a), which requires regularity conditions on the service-time cdf G. We later in Step 4 relax the regularity condition on G by doing a limiting argument.

Step 3(c). Regularity conditions on G. In particular, we assume that G is a distribution in $\mathcal{P}_{s,2}$ with rational Laplace transform, as in Smith (1953) or §II.5.10 of Cohen (1982). Following

Cohen (1982), we say that the random variable or its cdf G is in K_n. That implies that cdf G has a positive density and that the cdf H of W has a positive density except for an atom at 0. Those properties in turn imply that $W+V$ has a positive density. In particular, we use the following lemma.

LEMMA 2. *If $V+W$ has cdf Γ with*

$$\Gamma(x) = \int_0^x \gamma(y)\,dy \quad \textit{for} \quad x \ge 0, \tag{29}$$

then the function ϕ in (25) can be expressed as

$$\phi(u) = \int_0^\infty (x-u)^+\gamma(x)\,dx, \quad u \ge 0. \tag{30}$$

Hence, $\phi(0) = E[W+V]$ and the first two derivatives of ϕ in (25) exist for $u>0$ and satisfy

$$\begin{aligned} \dot{\phi}(u) &\equiv \frac{d\phi(t)}{dt}(u) = \Gamma(u) - 1 < 0 \quad \textit{and} \\ \ddot{\phi}(u) &\equiv \frac{d\dot{\phi}(t)}{dt}(u) = \gamma(u) > 0. \end{aligned} \tag{31}$$

Thus, ϕ is continuous. If in addition γ is strictly positive on $[0, M_a]$, as occurs when the cdf G of V is in K_n, then ϕ is strictly decreasing and strictly convex on $[0, M_a]$.

Proof. To calculate the derivatives, we apply the Leibniz integral rule for differentiation of integrals of integrable functions that are differentiable almost everywhere. Observe that the derivative of $(x-u)^+\gamma(x)$ with respect to u is $-\gamma(x)$ for $u<x$. That implies that

$$\dot{\phi}(u) = -\int_u^\infty \gamma(x)\,dx = \Gamma(u) - 1. \tag{32}$$

The rest follows directly. ∎

This uniqueness argument for (24) shows that the LP under the regularity conditions on service time cdf G always has a unique solution, so the solution must be in $\mathcal{P}_{a,2,3}(M_a)$. But it does not imply that the set $\mathcal{P}^*_{a,2}$ contains only a single element.

Step 5. part (b). Instead of (20) in part (a), we now have

$$\zeta(G_1) \equiv \sup\{E[(W(F,G_1) + V_{G_2} - U)^+] : G_2 \in \mathcal{P}_{s,2}(M_s)\}, \tag{33}$$

where the three random variables $W(F,G_1)$, V_{G_2} and U are mutually independent and $W(F,G_1)$ has the steady-state distribution associated with (F,G_1). We reduce the proof to the proof of part (a) by focusing on $M_s - v$ instead of v. Instead of the function ϕ in (25), we work with

$$\phi_s(v) \equiv E[(W + M_s - v - U)^+], \tag{34}$$

and we show that it has the same structure as ϕ, so we can use the rest of the proof for part (a).

The reverse-time construction can be confusing, so we explain briefly. The starting point is the formulation of the moment problem in §2.1. We change the underlying measure from the distribution of V to the distribution of $M_s - V$, so the new k^{th} moment becomes $\hat{m}_k \equiv E[(M_s - V)^k]$. When we view the function to $\phi_s(v)$ in (34) as an expectation with respect to the distribution of $M_s - V$, we change the underlying measure, which causes the objective function of the dual in (26) to change to $\lambda_0 + \lambda_1\hat{m}_1 + \lambda_2\hat{m}_2$. On the other hand, the function $\psi(x) \equiv \lambda_0 + \lambda_1 x + \lambda_2 x^2$ remains unchanged. Lemma EC.4 shows that ϕ_s has essentially the same structure as ϕ in Lemma 2. That completes our sketch of the proof; see §EC.2 for the remaining details.

3. An Analog of Theorem 1 for the Transient Mean

We now show that we can also obtain three-point extremal distributions for the transient mean $E[W_n]$ in the $GI/GI/1$ model. Paralleling (9) and (10), let $w_n : \mathcal{P}_{a,2} \times \mathcal{P}_{s,2} \to \mathbb{R}$ be defined by

$$w_n(F,G) \equiv E[W_n(F,G)], \tag{35}$$

using the formula in (2).

THEOREM 3. (*reduction to a three-point distribution for the transient mean*) *In the setting of Theorem 1, the function w_n in (35) is continuous and the domain is a compact metric space. Hence, the following suprema are attained as indicated:*

(*a*) *For any specified $G \in \mathcal{P}_{s,2}$, there exists $F^*(G) \in \mathcal{P}_{a,2,3}(M_a)$ such that*

$$w^{\uparrow}_{a,n}(G) \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}(M_a)\} = \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a)\} = w(F^*(G),G). \tag{36}$$

(*b*) *For any specified* $F \in \mathcal{P}_{a,2}$, *there exists* $G^*(F) \in \mathcal{P}_{s,2,3}(\rho M_s)$ *such that*

$$w^{\uparrow}_{a,n}(F) \equiv \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2}(\rho M_s)\} = \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2,3}(\rho M_s)\} = w(F, G^*(F)). \quad (37)$$

(*c*) *There exists* (F^{**}, G^{**}) *in* $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(\rho M_s)$ *such that*

$$w^{\uparrow}_n \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}(M_a), G \in \mathcal{P}_{s,2}(\rho M_s)\} = \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a), G \in \mathcal{P}_{s,2,3}(\rho M_s)\}$$
$$= w_n(F^{**}, G^{**}) = w^{\uparrow}_{a,n}(G^{**}) = w^{\uparrow}_{s,n}(F^{**}). \quad (38)$$

Proof of Theorem 3. Just as for Theorem 1, we only prove part (a) in detail. Based on (2), we first express the mean as

$$E[W_n] = \sum_{k=1}^{n} k^{-1} \int_0^{kM_a} \int_0^{\infty} (v-u)^+ \, dP_{S^a_k}(u) \, dP_{S^s_k}(v)$$
$$= \int_0^{M_a} \phi_{a,n}(u) \, dF(u), \quad (39)$$

where S^a_k is the partial sum of the first k i.i.d interarrival times each with cdf F, where F has mean 1 and support on $[0, M_a]$ with $M_a \geq c_a^2 + 1$, S^s_k is the partial sum of the first k i.i.d. service times each with cdf G, where G has mean ρ and finite scv c_s^2, and

$$\phi_{a,n}(u) \equiv \sum_{k=1}^{n} k^{-1} \int_0^{(k-1)M_a} \int_0^{\infty} (v-x-u)^+ \, dP_{S^a_{k-1}}(x) \, dP_{S^s_k}(v), \quad (40)$$

provided that the common cdf, say F_1, of the $k-1$ i.i.d interarrival times that produces the partial sum S^a_{k-1} coincides with the cdf of U_k, say F_2. Thus, paralleling our treatment of the steady-state waiting time, we have created a framework in which we can apply a fixed-point argument.

In particular, paralleling (20), we define $\zeta(F_1)$ with $\zeta : \mathcal{P}_{a,2}(M_a) \to \mathbb{R}$ defined by

$$\zeta(F_1) \equiv \sup\{E[(W_n(F_1, G; F_2)] : F_2 \in \mathcal{P}_{a,2}(M_a)\}, \quad (41)$$

where $W_n(F_1, G; F_2)$ is understood to be the transient waiting time, starting empty, where the $k-1$ i.i.d. interarrival times making up S^a_{k-1} each have cdf F_1, the k i.i.d. service times making up S^s_k

each have cdf G and the one unspecified interarrival time has cdf F_2. Paralleling Step 1 in the proof of Theorem 1, let $\eta(F_1)$ be the set of maximizers of (41). Let $\mathcal{P}^*_{a,2}(M_a)$ be the set of all fixed points of the map $\eta : \mathcal{P}_{a,2}(M_a) \to 2^{\mathcal{P}_{a,2}(M_a)}$, i.e.,

$$\mathcal{P}^*_{a,2} \equiv \{F \in \mathcal{P}_{a,2}(M_a) : F \in \eta(F)\}. \tag{42}$$

A minor modification of the proof of Theorem 1 (which also accounts for the need to optimize over F_1 in (41)) shows that $\mathcal{P}^*_{a,2}$ is nonempty and always contains an element of $\mathcal{P}_{a,2,3}(M_a)$. The rest of the argument can now follow Steps 2-5 of the proof of Theorem 1. For that purpose, we initially restrict attention to interarrival-time cdf's F with finite support. For Step 3b, we initially assume that the cdf G is in K_n, and observe that implies that the cdf of the sum S^s_k also is in K_n. That in turn implies that $S^s_k - S^a_{k-1}$ has a density. Hence, we can apply Lemma 2 as in Step 3b of the proof of Theorem 1. We now have (30) and (31), where γ is the density of $S^s_k - S^a_{k-1}$, which is a finite sum of translates of the density of S^s_k. The rest of the proof is essentially the same. Hence the proof is complete. ∎

4. More Structure of the Extremal Distributions

We now show that the fixed point framework used to prove Theorem 1 also can be applied to establish more properties of the extremal distributions. For this purpose, we exploit the special form of the LP in (24) and (25) and the dual in (26)-(28) for (a). That depends on the structure of the function ϕ in (28) for (a) and its analog ϕ_s in (34) for case (b). That in turn depends on the three-point fixed point cdfs F^* in $\mathcal{P}^*_{a,2}$ in (a) and G^* in $\mathcal{P}^*_{s,2}$ in (b).

We have developed two different sufficient conditions, the first in the form of unimodal distributions in the $GI/GI/1$ model and the second in the form of Tchebycheff systems. We describe the first here and the second in §EC.4. As in our proof of Theorem 1, we first impose regularity conditions on the fixed distribution, which is the service-time cdf G in (a). In particular, we assume that G is in K_n for (a).

THEOREM 4. (*more structure of the extremal distribution*) *Consider the setting of Theorem 1. Let* $F^*(G) \in \mathcal{P}_{a,2,3}(M_a), G^*(F) \in \mathcal{P}_{a,2,3}(M_s)$ *be elements in the set of fixed points* $\mathcal{P}^*_{a,2}$ *in (21) for (a) and* $\mathcal{P}^*_{s,2}$ *for (b), respectively.*

(a) For any fixed service-time cdf G *on* $[0,\infty)$ *in* $K_n \bigcap \mathcal{P}_{s,2}$, *let* $\gamma = \ddot{\phi}$ *be the pdf of* $W(F^*, G) + V_G$ *in Lemma 2. (i) If* M_a *is sufficiently large, then* M_a *is not contained in the support of* F^*. *(ii) If* γ *is unimodal for* F^*, *then* $F^* \in \mathcal{P}_{a,2,2}$.

(b) For any fixed interarrival-time cdf F *on* $[0,\infty)$ *in* $K_n \bigcap \mathcal{P}_{a,2}$, *let* $\theta = \ddot{\phi}_s$ *be the pdf of* $W(F, G^*) + M_s - U_F$ *in (EC.15) and (EC.16), which also depends on* G. *(i) If* θ *is unimodal for* $G^*(F) \in \mathcal{P}_{s,2,3}(M_s)$, *then* $G^* \in \mathcal{P}_{s,2,2}$. *(ii) If* F *has a strictly monotone density, then* θ *is strictly monotone for all* $G \in \mathcal{P}_{s,2,3}(M_s)$ *and* $w^{\uparrow}(F) = w(F, G_0)$.

We give the proof of Theorem 4 in §EC.3.

COROLLARY 1. (*the* $H_k/GI/1$ *model*) *In the setting of Theorem 4, if* F *is* H_k, *then* $w^{\uparrow}(F) = w(F, G_0)$ *in (b).*

.

Corollary 1 extends the results for the $K_2/GI/1$ model covered by Theorem 11 in §V of Whitt (1984b) and Whitt (1984a).

5. Numerical Support for the Overall Upper Bound

In this section we combine Theorem 1 (c) with numerical optimization for the transient mean $E[W_n]$ to deduce the form of the overall upper bound. In §5.1 we show that results for the transient mean imply corresponding results for the steady-state mean. In §5.2 we formulate an optimization problem for the transient mean based on a multinomial representation. Then in §5.3 we draw conclusions from this numerical study. We provide further support with simulations for two-point distributions in §6.

5.1. From $E[W_n]$ to $E[W]$

We now show that it suffices to consider the transient mean $E[W_n]$ for the three-point distributions and finite n in order to treat $E[W]$. Theorem 1 implies that we only need consider three-point distributions. Theorems 5 and 3 together provide a new proof of Theorem 1, but that does not help greatly, because our proof of Theorem 3 was largely based on the proof of Theorem 1,

THEOREM 5. (*reduction to the transient mean*) *Consider the $GI/GI/1$ queues in Theorem 1.*

(*a*) *For any specified $G \in \mathcal{P}_{s,2}$, if there exists $F_n \in \mathcal{P}_{a,2,3}(M_a)$ such that*

$$w_n(F_n, G) = w_{a,n}^{\uparrow}(G) \equiv \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2,3}(M_a)\} \quad \textit{for all} \quad n \geq 1, \tag{43}$$

then the sequence $\{F_n : n \geq 1\}$ is tight, so that there exists a convergent subsequence. Moreover, if F is the limit of any convergent subsequence, then F is in $\mathcal{P}_{a,2,3}(M_a)$ and F is optimal for $E[W(F, G)]$, i.e., $w_a^{\uparrow}(G) = w(F, G)$ for the steady-state mean.

(*b*) *For any specified $F \in \mathcal{P}_{a,2}$, if there exists $G_n \in \mathcal{P}_{s,2,3}(M_s)$ such that*

$$w_n(F, G_n) = w_{s,n}^{\uparrow}(F) \equiv \sup\{w_n(F, G) : G \in \mathcal{P}_{s,2,3}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{44}$$

then the sequence $\{G_n : n \geq 1\}$ is tight, so that there exists a convergent subsequence. Moreover, if G is the limit of any convergent subsequence, then G is in $\mathcal{P}_{s,2,3}(M_s)$ and G is optimal for $E[W(F, G)]$, i.e., $w_s^{\uparrow}(F) = w(F, G)$ for the steady-state mean.

(*c*) *If there exists (F_n, G_n) in $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$ such that*

$$w_n(F_n, G_n) = w_n^{\uparrow} \equiv \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2,3}(M_a), G \in \mathcal{P}_{s,2,3}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{45}$$

then the sequence $\{(F_n, G_n) : n \geq 1\}$ is tight, so that there exists a convergent subsequence. Moreover, if (F, G) is the limit of any convergent subsequence, then (F, G) is in $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$ and the pair (F, G) is optimal for $E[W]$, i.e., $w^{\uparrow} = w(F, G)$ for the steady-state mean.

Proof. We only prove (c), because the others are proved in the same way. As observed before, because the support sets $[0, M_a]$ and $[0, \rho M_s]$ are compact intervals, the spaces $\mathcal{P}_{a,2}(M_a)$, $\mathcal{P}_{s,2}(M_s)$ and their product are compact metric spaces, as are the spaces $\mathcal{P}_{a,2,3}(M_a)$, $\mathcal{P}_{s,2,3}(M_s)$ and their product, because they are closed subsets. Hence the tightness follows, which implies that there exists a convergent subsequence by Prohorov's theorem in §11.6 of Whitt (2002) and the limit (F, G) of any such subsequence $\{(F_{n_k}, G_{n_k}) : k \geq 1\}$ must remain in the space $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$. Suppose that (F', G') is another candidate pair of cdf's in $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$. By the assumed optimality, we must have $w_{n_k}(F_{n_k}, G_{n_k}) \geq w_{n_k}(F', G')$ for all k. Then, by continuity, using §X.6 of Asmussen (2003) again, we conclude that $w^{\uparrow} = w(F, G)$ for the steady-state mean. ∎

By the same reasoning, an analog of Theorem 5 holds for two-point distributions.

COROLLARY 2. *In the setting of Theorem 5, (i) if $F_n \in \mathcal{P}_{a,2,2}(M_a)$ for all n in (a), then $F \in \mathcal{P}_{a,2,2}(M_a)$; if $G_n \in \mathcal{P}_{s,2,2}(M_s)$ for all n in (b), then $G \in \mathcal{P}_{s,2,2}(M_s)$; if $(F_n, G_n) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$ for all n in (c), then $(F, G) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$.*

Proof. The same argument applies because $\mathcal{P}_{2,2}(M)$ is a closed subset of $\mathcal{P}_{2,3}(M)$. ∎

.

5.2. The Multinomial Representation for the Transient Mean $E[W_n]$

We can represent the transient mean in (2) in terms of two independent multinomial distributions. Let the cdf G in $\mathcal{P}_{s,2,3}$ with specified mean ρ and scv c_s^2 be parameterized by the vector of mass points $\mathbf{v} \equiv (v_1, v_2, v_3)$ and the vector of probabilities $\mathbf{p} \equiv (p_1, p_2, p_3)$. For every positive integer k, define a multinomial probability mass function on the vector of nonnegative integers $\mathbf{k} \equiv (k_1, k_2, k_3)$ by

$$P_k(\mathbf{p}) \equiv \frac{k! p_1^{k_1} p_2^{k_2} p_3^{k_3}}{k_1! k_2! k_3!}, \tag{46}$$

where it is understood that $\mathbf{k}\mathbf{e}' \equiv k_1 + k_2 + k_3 = k$. Similarly, let the cdf F in $\mathcal{P}_{a,2,3}$ with specified mean 1 and scv c_a^2 be parameterized by the vector of mass points $\mathbf{u} \equiv (u_1, u_2, u_3)$ and probabilities $\mathbf{q} \equiv (q_1, q_2, q_3)$ on the vector of nonnegative integers $\mathbf{w} \equiv (w_1, w_2, w_3)$, so that

$$Q_k(\mathbf{q}) \equiv \frac{k! q_1^{w_1} q_2^{w_2} q_3^{w_3}}{w_1! w_2! w_3!}, \tag{47}$$

where it is understood that $\mathbf{we}' \equiv w_1 + w_2 + w_3 = k$.

Then, from (2),

$$E[W_n] = \sum_{k=1}^{n} \frac{1}{k} \sum_{(\mathbf{k},\mathbf{w})\in\mathcal{I}} \max\{0, \sum_{i=1}^{3}(k_i v_i - w_j u_j)\} P_k(\mathbf{p}) Q_k(\mathbf{q}), \tag{48}$$

where $\mathcal{I}$ is the set of all pairs of vectors $(\mathbf{k},\mathbf{w})$ with both $\mathbf{ke}' \equiv k_1 + k_2 + k_3 = k$ and $\mathbf{we}' \equiv w_1 + w_2 + w_3 = k$.

For any given n and any given distributions G in $\mathcal{P}_{s,2,3}$ parameterized by the pair $(\mathbf{v},\mathbf{p})$ and F in $\mathcal{P}_{a,2,3}$ parameterized by the pair $(\mathbf{u},\mathbf{q})$, we can calculate the transient mean $E[W_n]$ by calculating the sum in (48). We can easily evaluate $E[W_n]$ for candidate cases provided that n is not too large.

Next, for the overall optimization over $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$, we write

$$\sup\{E[W_n(\mathbf{v},\mathbf{p},\mathbf{u},\mathbf{q})] : ((\mathbf{v},\mathbf{p}),(\mathbf{u},\mathbf{q})) \in \mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)\}, \tag{49}$$

using (48). We now write this optimization problem in a more conventional way, from which we see that the optimization is a form of non-convex nonlinear program. In particular, we write for the means $m_1 \equiv E[U] \equiv 1$, $m_2 \equiv E[U^2] \equiv m_1^2(c_a^2+1)$, $s_1 \equiv E[V] \equiv \rho$ and $s_2 \equiv E[V^2] \equiv s_1^2(c_a^2+1)$,

$$\begin{aligned}
&\text{maximize} \sum_{k=1}^{n} \frac{1}{k} \sum_{\sum k_i = k, \sum_j w_j = k} \max(\sum_i k_i v_i - \sum_j w_j u_i, 0) P(k_1,k_2,k_3) Q(w_1,w_2,w_3) \\
&\text{subject to } \sum_{j=1}^{3} u_j q_j = m_1, \quad \sum_{j=1}^{3} u_j^2 q_j = (1+c_a^2) m_1^2, \\
&\qquad \sum_{j=1}^{3} v_j p_j = s_1, \quad \sum_{j=1}^{3} v_j^2 p_j = (1+c_s^2) s_1^2, \\
&\qquad \sum_{j=1}^{3} p_j = \sum_{k=1}^{3} q_k = 1, \\
&\qquad M_s \ge v_j \ge 0,\ M_a \ge u_j \ge 0,\ p_j \ge 0,\ q_j \ge 0, \quad 1 \le j \le 3.
\end{aligned} \tag{50}$$

We solved this non-convex nonlinear program in (50) by applying sequential quadratic programming (SQP) as discussed in Chapter 18 of Nocedal and Wright (1999). In particular, we applied the Matlab variant of SQL, which is a second-order method, implementing Schittkowski's NLPQL Fortran algorithm. This algorithm converges at a local optimum. Since the algorithm is not guaranteed to reach a global optimum, we run the algorithm for a large collection of uniform randomly chosen initial conditions.

We found that the local optimum solution is usually $(F_0, G_{u,n})$, where $G_{u,n}$ is a two-point distribution that converges to G_u as $n \to \infty$. In the rare cases that we obtain a different solution, we found that it is always in $\mathcal{P}_{a,2,2} \times \mathcal{P}_{s,2,2}$. Moreover, in these cases, we can find a different initial condition for which $(F_0, G_{u,n})$ is the local optimum, and that $E[W(F_0, G_{u,n})]$ is larger than for other local optima.

5.3. The Numerical Conclusion about the Overall Upper Bound

From extensive numerical experiments, which draw on our mathematical results, we conclude that the extremal UB interarrival-time cdf F_0 for $GI/M/1$ also holds for all $GI/GI/1$, but the extremal service-time distribution is more complicated because it depends on both n and M_s. In summary, Theorem 1 and our numerical results support the following conjecture about the overall tight upper bound.

CONJECTURE 1. (*the tight upper bound*)

(*a*) *Given any parameter vector* $(1, c_a^2, \rho, c_s^2)$ *and a bounded interval of support* $[0, \rho M_s]$ *for the service-time cdf* G, *where* $M_s \geq c_s^2 + 1$, *the pair* (F_0, G_u) *attains the tight UB of the steady-state mean* $E[W]$, *while a pair* $(F_0, G_{u,n})$ *attains the tight UB of the transient mean* $E[W_n]$, *where* $G_{u,n}$ *is a two-point distribution with* $G_{u,n} \Rightarrow G_u$ *as* $n \to \infty$.

(*b*) *When* G *has an unbounded interval of support* $[0, \infty)$, *the tight UB of* $E[W]$ *is not attained directly, but is obtained asymptotically in the limit as* $M_s \to \infty$ *in part (a). The extremal service-time cdf* G_u *is asymptotically deterministic with the given mean* ρ *as* $M_s \to \infty$, *but that deterministic distribution does not have parameter* c_s^2 *if* $c_s^2 \neq 0$. *Moreover, the mean* $E[W(M_s)]$ *does not approach the mean in the associated extremal* $GI/D/1$ *queue as* $M_s \to \infty$.

Let G_{u^*} in $E[W(F, G_{u^*})]$ be shorthand for the limit of $E[W(F, G_u)]$ as $M_s \to \infty$ as in Conjecture 1 (b). We obtain an UB for $E[W(F_0, G_{u^*})]$, assuming Conjecture 1

THEOREM 6. (*an UB for* $E[W(F_0, G_{u^*})]$) *For the* $GI/GI/1$ *queue with parameter four-tuple* $(1, c_a^2, \rho, c_s^2)$, *if* $E[W(F_0, G_{u^*})]$ *is the tight UB as claimed in Conjecture 1, then*

$$E[W(F_0, G_{u^*})] \leq \frac{2(1-\rho)\rho/(1-\delta)c_a^2 + \rho^2 c_s^2}{2(1-\rho)} < \frac{\rho(2-\rho)c_a^2 + \rho^2 c_s^2}{2(1-\rho)}, \tag{51}$$

where $\delta \in (0,1)$ *and* $\delta = \exp(-(1-\delta)/\rho)$.

We call formula (51) the "new UB," but because it relies on Conjecture 1 it is only verified numerically so far. Formula (51) draws on §10 of Daley et al. (1992); it is based on Conjecture III on p. 211 of Daley et al. (1992), but we show in §EC.6.3 that the conjecture is actually not correct. We prove Theorem 6 in §EC.5.

Counterexamples were constructed in §V of Whitt (1984b), drawing on Whitt (1984a), and in §8 of Wolff and Wang (2003) that contradict corresponding conjectures that analogs of Conjecture 1 hold when one distribution is fixed.

Tables 1 and 2 compare the numerically computed values of the conjectured tight UB, $E[W(F_0, G_{u^*})]$, drawing on Chen and Whitt (2018), to the heavy-traffic approximation (HTA) in (4), the new upper bound in (51), the Daley (1977) bound in (6) and the Kingman (1962) bound in (5) over a range of ρ for the scv pairs $(c_a^2, c_s^2) = (4.0, 4.0)$ and $(0.5, 0.5)$. In order to focus on the variability independent of the traffic intensity ρ, we display the scaled mean waiting time values $(1-\rho)E[W]/\rho^2$, which are constant for the heavy-traffic approximation in (4), being equal to $(c_a^2 + c_s^2)/2$. simulation algorithm, discussed in §6.1. Tables EC.2 and EC.3 in the e-companion give comparable results for the mixed pairs $(c_a^2, c_s^2) = (4.0, 0.5)$ and $(0.5, 4.0)$, while Tables EC.4-EC.7 show all the unscaled values.

In these tables we also show the value of δ in the new UB (51) and the maximum relative error (MRE) between the UB approximation and the tight UB. The MRE over all four cases was 5.7%. which occurred for $c_a^2 = c_s^2 = 0.5$ and $\rho = 0.5$.

We also display the lower bound (LB) in (7), which is far less than the other values, indicating the wide range of possible values. The extremely low LB occurs because it is associated with the $D/GI/1$ model, which is approached by the F_u extremal distribution as the support limit $M_a \to \infty$ for any c_a^2. Notice that the LB is actually 0 for many cases with low traffic intensity; that occurs if and only if $P(V \le U) = 1$. Hence, the LB looks especially bad for the case $(c_a^2 = 4.0, c_s^2 = 0.5)$ in Table EC.2, because it is the same as for the case $(c_a^2 = 0.5, c_s^2 = 0.5)$ in Table 2 and even for $(c_a^2 = 0.0, c_s^2 = 0.5)$ in the $D/GI/1$ model. We discuss the LB in Chen and Whitt (2019).

Table 1 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$.

ρ	Tight LB	HTA (4)	Tight UB	new UB (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	4.000	38.001	38.002	0.000	0.00%	40.000	202.000
0.20	0.000	4.000	18.078	18.112	0.007	0.19%	20.000	52.000
0.30	0.833	4.000	11.661	11.731	0.041	0.60%	13.333	24.222
0.40	1.250	4.000	8.640	8.722	0.107	0.94%	10.000	14.500
0.50	1.500	4.000	6.940	7.020	0.203	1.15%	8.000	10.000
0.60	1.667	4.000	5.883	5.946	0.324	1.07%	6.667	7.556
0.70	1.786	4.000	5.168	5.216	0.467	0.93%	5.714	6.082
0.80	1.875	4.000	4.662	4.693	0.629	0.67%	5.000	5.125
0.90	1.944	4.000	4.287	4.302	0.807	0.35%	4.444	4.469
0.95	1.974	4.000	4.134	4.142	0.902	0.18%	4.211	4.216
0.98	1.990	4.000	4.052	4.055	0.960	0.07%	4.082	4.082
0.99	1.995	4.000	4.025	4.027	0.980	0.04%	4.040	4.041

From this analysis, we see that conjectured new UB (51) is an excellent approximation for the conjectured UB $E[W(F_0, G_{u^*})]$. Moreover, we see that there is significant improvement going from the Kingman (1962) bound in (5) to the Daley (1977) bound in (6) to the new UB in (51). We also see that the heavy-traffic approximation is consistent with the UBs in all cases. Moreover, all the approximations are asymptotically correct as $\rho \uparrow 1$. The heavy-traffic approximation in (4) tends to be much closer to the UB than the lower bound, which shows that the overall MRE can be large and that the heavy-traffic approximation tends to be relatively conservative, as usually is desired in applications.

6. A Systematic Study Over All Two-Point Distributions

The optimization in §5 supports Conjecture 1, but not as strongly as we would like. A more convincing conclusion from §5 is that it suffices to reduce the search for an optimum to the smaller subset of two-point distributions, i.e., to the product space $\mathcal{P}_{a,2,2} \times \mathcal{P}_{s,2,2}$. This space is relatively

Table 2 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 0.5$.

ρ	Tight LB	HTA (4)	Tight UB	new UB (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	0.500	4.750	4.750	0.000	0.00%	5.000	25.250
0.20	0.000	0.500	2.252	2.264	0.007	0.54%	2.500	6.500
0.30	0.000	0.500	1.432	1.466	0.041	2.36%	1.667	3.028
0.40	0.000	0.500	1.049	1.090	0.107	3.82%	1.250	1.813
0.50	0.000	0.500	0.827	0.878	0.203	5.72%	1.000	1.250
0.60	0.000	0.500	0.708	0.743	0.324	4.71%	0.833	0.944
0.70	0.036	0.500	0.623	0.652	0.467	4.53%	0.714	0.760
0.80	0.125	0.500	0.569	0.587	0.629	2.95%	0.625	0.641
0.90	0.194	0.500	0.530	0.538	0.807	1.38%	0.556	0.559
0.95	0.224	0.500	0.514	0.518	0.902	0.65%	0.526	0.527
0.98	0.240	0.500	0.505	0.507	0.960	0.27%	0.510	0.510
0.99	0.245	0.500	0.503	0.503	0.980	0.14%	0.505	0.505

easy to analyze because each of the sets $\mathcal{P}_{a,2,2}$ and $\mathcal{P}_{s,2,2}$ is one-dimensional, as indicated in §1.3. The G_0 counterexample from §8 of Wolff and Wang (2003) also falls in this set.

6.1. Simulation Experiments

To analyze the mean waiting times for the two-point interarrival-time and service-time distributions, we primarily use stochastic simulation. (We also verify for lower traffic intensities by applying the multinomial representation in §5.2 for finite n.)

We study various simulation approaches in Chen and Whitt (2018). For the transient mean $E[W_n]$, we use direct numerical simulation, but for the steady-state simulations we mostly use the simulation method in Minh and Sorli (1983) that exploits the representation of $E[W]$ in terms of the steady-state idle time I and the random variable I_e that has the associated equilibrium excess distribution, i.e.,

$$E[W] = -\frac{E[X^2]}{2E[X]} - E[I_e] = -\frac{E[X^2]}{2E[X]} - \frac{E[I^2]}{2E[I]} = \frac{\rho^2 c_s^2 + c_a^2 + (1-\rho)^2}{2(1-\rho)} - \frac{E[I^2]}{2E[I]}; \tag{52}$$

which is also used in Wolff and Wang (2003). For each simulation experiment, we perform multiple (usually $20-40$) i.i.d. replications. Within each replication we look at the long-run average after deleting an initial portion to allow the system to approach steady state if deemed helpful. It is well known that obtaining good statistical accuracy is more challenging as ρ increases, e.g., see Whitt (1989), but that challenge is largely avoided by using (52). There is also a well known issue of one long run versus multiple replications, e.g., see Whitt (1991).

We do not report confidence intervals for all the individual results, but we did do a careful study of the statistical precision. To illustrate, Table 3 compares the 95% confidence intervals associated with estimates of the steady-state mean $E[W(F_0, G_u)]$ for the parameter triple $(\rho, c_a^2, c_s^2) = (0.5, 4.0, 4.0)$ obtained by making the statistical t test to multiple replications of runs of various length. The table compares the standard simulation for various run lengths N (number of arrivals) and the Minh and Sorli (1983) algorithm for various run lengths T (length of time, over which we average the observed idle periods) and numbers of replications n. (See Chen and Whitt (2018) for more discussion.)

Table 3 Confidence interval halfwidths for estimates of the steady-state mean $E[W(F_0, G_u)]$ for the parameter triple $(\rho, c_a^2, c_s^2) = (0.5, 4.0, 4.0)$

	Monte Carlo simulation			Minh and Sorli simulation		
replications	$N = 1E+05$	$N = 1E+06$	$N = 1E+07$	$T = 1E+05$	$T = 1E+06$	$T = 1E+07$
20	6.64E-02	2.45E-02	8.01E-03	1.58E-03	4.81E-04	1.55E-04
40	5.59E-02	1.27E-02	4.22E-03	1.20E-03	3.20E-04	9.89E-05
60	3.69E-02	1.20E-02	4.23E-03	8.44E-04	2.88E-04	8.03E-05
80	3.52E-02	1.17E-02	3.72E-03	7.54E-04	2.27E-04	9.55E-05
100	2.61E-02	9.94E-03	3.13E-03	6.06E-04	2.02E-04	7.20E-05

6.2. The Impact of the Interarrival-Time Distribution

Figure 1 reports simulation results for $E[W_{20}]$ (left) and $E[W]$ (right) in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 30$. (The maximum 95% confidence interval was less than 10^{-4}.) We focus on

the impact of b_a (for F) in the permissible range $[5,30]$ for six values of b_s (for G) ranging from 5 to 30. (Recall that the parameter b was defined in §1.3.)

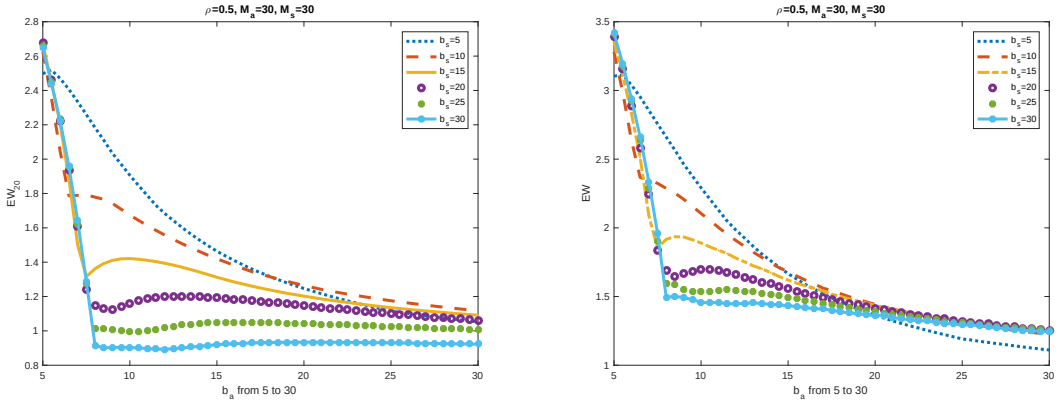

Figure 1 Simulation estimates of the transient mean $E[W_{20}]$ (left) and the steady-state mean $E[W]$ (right) as a function of b_a for six cases of b_s the in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 30$.

Figure 1 shows that the mean waiting times tend to be much larger at the extreme left, which is associated with $b_a = 5$ or F_0. However, we see some subtle behavior. For example, for $b_s = 20$, we clearly see that the mean is not monotonically decreasing in b_a, but nevertheless, F_0 is clearly optimal.

On the other hand, a close examination of the extreme case $b_s = 5$ shows that the largest value of b_a does not occur for $b_a = 5$, but in fact occurs at a slightly higher value. That turns out to be the counterexample. In particular, Tables 4 and 5 present detailed simulation estimates of $E[W]$ and $E[W_{20}]$. In both Tables 4 and 5 we see that the maximum mean waiting time value in the first row, i.e., over b_a when $b_s = 5$ is not attained at $b_a = 5.0$, but is instead attained at $b_a = 5.25$. For emphasis, in each case we highlight both the maximum entry in the first row and the maximum entry in the table. Therefore, for that service-time distribution (which is G_0), the extremal inter-arrival time is not F_0.

Note that F_0 is optimal for all other b_s and the difference between $\max\{E[W(F,G_0)] : F\} - E[W(F_0,G_0)]$ is very small. Moreover, consistent with Conjecture 1, the overall UB is attained at the pair (F_0, G_u). Finally, note that the difference across each row tends to be greater than the difference across each column.

Table 4 Simulation estimates of $E[W]$ as a function of b_a and b_s when $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = 7 < M_s = 10$.

$b_s \backslash b_a$	5.00	5.25	5.50	5.75	6.00	6.25	6.50	6.75	7.0
5.0	3.110	**3.134**	3.117	3.083	3.040	2.997	2.950	2.910	2.863
5.5	3.179	3.026	3.019	3.009	2.975	2.938	2.901	2.860	2.823
6.0	3.191	3.065	2.932	2.907	2.905	2.876	2.844	2.809	2.767
7.0	3.181	3.067	2.942	2.797	2.748	2.720	2.713	2.691	2.670
8.0	3.195	3.056	2.934	2.810	2.664	2.611	2.591	2.564	2.553
9.0	3.239	3.092	2.931	2.792	2.663	2.525	2.472	2.467	2.449
10.0	**3.282**	3.142	2.986	2.812	2.640	2.507	2.367	2.350	2.349

Table 5 Simulation estimates of $E[W_{20}]$ as a function of b_a and b_s when $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = 7 < M_s = 10$.

$b_s \backslash b_a$	5.00	5.25	5.50	5.75	6.00	6.25	6.50	6.75	7.00
5.0	2.497	**2.530**	2.518	2.497	2.469	2.439	2.406	2.371	2.335
5.5	2.557	2.414	2.420	2.422	2.402	2.378	2.351	2.320	2.288
6.0	2.561	2.447	2.328	2.318	2.328	2.312	2.290	2.266	2.239
7.0	2.549	2.447	2.331	2.204	2.165	2.149	2.154	2.150	2.132
8.0	2.556	2.430	2.319	2.208	2.074	2.029	2.021	2.010	2.007
9.0	2.598	2.456	2.310	2.183	2.068	1.937	1.895	1.903	1.898
10.0	**2.626**	2.506	2.353	2.188	2.043	1.921	1.786	1.779	1.789

6.3. The Impact of the Service-Time Distribution

Figure 1 also shows the impact of the service-time distribution, but that impact is more complicated. For $E[W]$ with $b_s = 0.5$, we see that the curve crosses the other curves in the middle. We now investigate what is the optimal value of b_s over $[1 + c_s^2, M_s]$ for $E[W_n]$ and $E[W]$. For that purpose, Figure 2 plots the values of $E[W_{10}]$ (left) and $E[W_{20}]$ (right) as a function of b_s in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$, $M_s = 300$ and $b_a = (1 + c_a^2)$. For Figure 2, we use the optimization in §5

with a numerical method to directly compute a good finite truncation of objective in the nonlinear program (50). For these cases, we find $b_s^*(10) = 35.10$ and $b_s^*(20) = 41.12$.

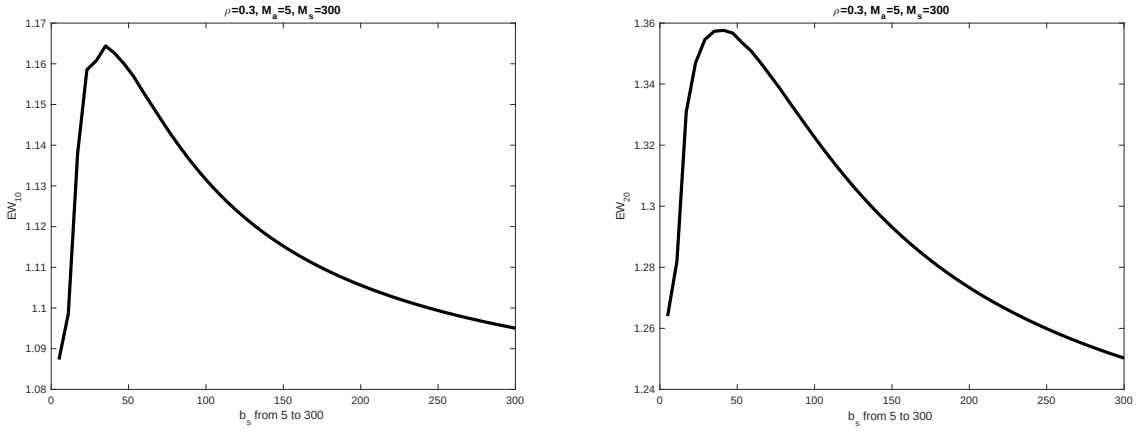

Figure 2 The transient mean waiting time $E[W_n]$ for $n = 10, 20$ as a function of b_s up to $M_s = 300$. $b_s^*(10) = 35.10, b_s^*(20) = 41.12$.

As a function of b_s, the transient mean waiting time $E[W_n]$ is approximately first increasing and then decreasing at all traffic levels. Therefore, for each n, there exists $b_s^*(n)$ such that $E[W(F_0, b_s^*(n))] \geq E[W(F_0, b_s); F \in \mathcal{P}_{a,2,2}]$. Another important observation is that $b_s^*(n)$ is a function of n and $b_s^*(20) > b_s^*(10)$ under traffic level $\rho = 0.3$.

Now we investigate the extremal $b_s^*(n)$ as a function of n. Figure 3 shows $E[W_n]$ as a function of n for the light traffic $\rho = 0.2$ (left) and $\rho = 0.3$ (right). Figure 3 shows that $b_s^*(n)$ tends to be increasing with n given $b_a = (1 + c_a^2)$, but is not uniformly so. In particular, for $\rho = 0.3$ on the right, we see a dip at $n = 15$.

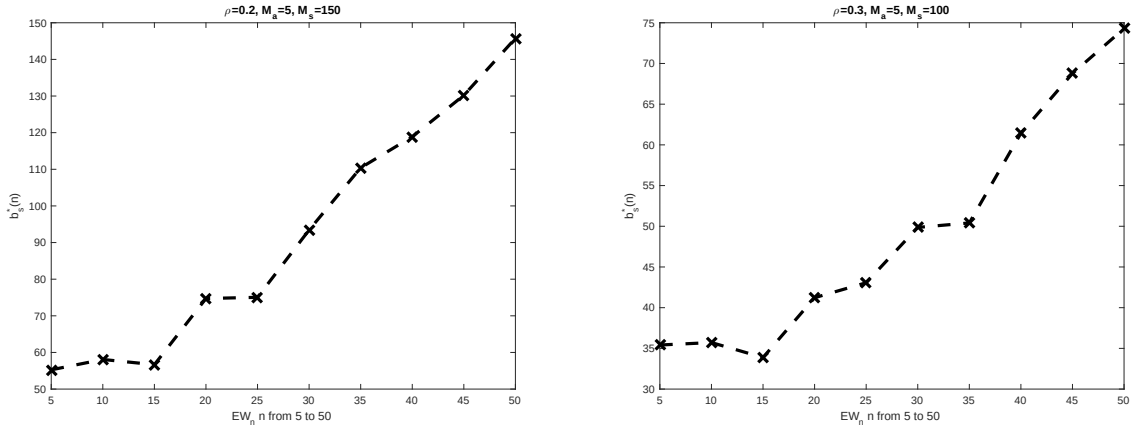

Figure 3 Performance of $b_s^*(n)$ associated with $E[W_n]$ for $5 \leq n \leq 50$.

Nevertheless, the upper bound queue over $\mathcal{P}_{a,2,2} \times \mathcal{P}_{s,2,2}$ for transient mean waiting time $E[W_n]$ is $F_0/G_{b_s^*(n)}/1$ with $b_s^*(n)$ primarily increasing with n.

We next directly examine the steady-state mean waiting time $E[W]$ for set $b_a = (1 + c_a^2)$ and $M_s = 100$. We use Minh and Sorli (1983) method with simulation length over a time interval of length 10^6 and 40 i.i.d. replications. (The maximum 95% confidence interval was again less than 10^{-4}.) To illustrate, Figure 4 shows the results for the traffic levels $\rho = 0.3$ (left) and $\rho = 0.9$ (right).

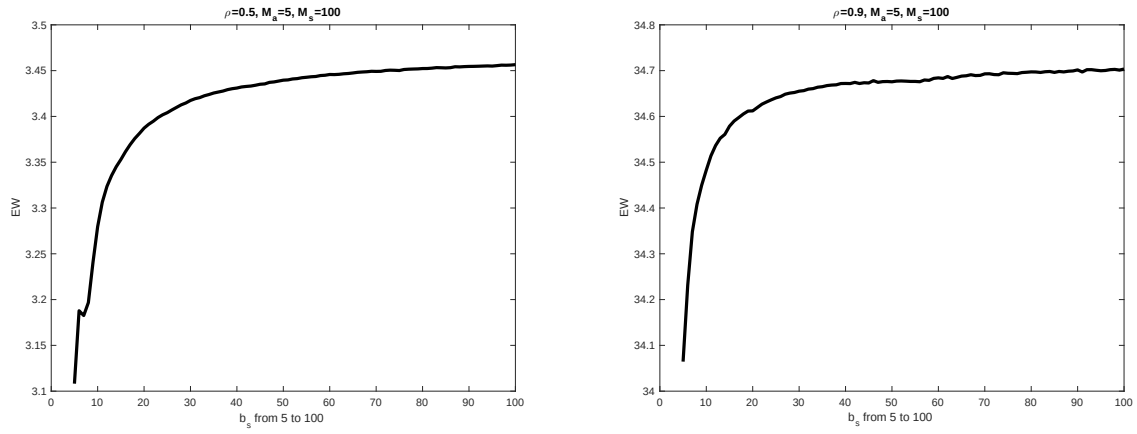

Figure 4 $E[W(F_0, G)]$ for $G \in \mathcal{P}_{s,2,2}$ as a function of b_s given $b_a = (1 + c_a^2)$.

Just as in Figure 3, Figure 4 shows that the steady-state mean $E[W]$ is eventually increasing in b_s, given $b_a = (1 + c_a^2)$, strongly supporting the conclusion that the upper bound is attained at

(F_0, G_u). Hence, the optimal b_s is M_s. Since $E[W_n] \to E[W]$, we must also have $b_s^*(n) \to b_s^* = M_s$ as $n \to \infty$.

7. Conclusions

We have established new results about tight upper bounds for the mean steady-state waiting time in the $GI/GI/1$ model given the first two moments of the interarrival time and service time, specified by the parameter vector $(1, c_a^2, \rho, c_s^2)$. Theorem 1 in §2 shows that the upper bounds (overall and with one distribution fixed) are attained at distributions with support on at most three points. Theorem 3 in §3 provides an analog of Theorem 1 for the transient mean in §3. Theorem 4 in §4 exposes additional structure of the extremal distributions when one distribution is given.

In the rest of the paper, including the e-companion, we applied numerical methods to further identify the extremal distributions. From a practical engineering perspective, we have addressed the important question about the tight upper bound. The combination of mathematical and numerical results strongly supports Conjecture 1 in §5.3, which states that the overall upper bound is attained by $E[W(F_0, G_{u^*})]$, i.e., at the extremal two-point distributions, modified by a limit, as many have thought. However, because the analysis is partly numerical, it still remains to provide a mathematical proof. We also provided a new upper bound analytical formula (51), which is a valid bound under Conjecture 1. Drawing on algorithms to compute $E[W(F_0, G_{u^*})]$ in Chen and Whitt (2018), Tables 1 and 2 illustrate that the new UB formula is quite accurate, providing significantly improvement over previous bounds.

There are many remaining problems for research. In addition to providing a full mathematical proof of Conjecture 1, it remains to identify the extremal distributions with one distribution given, as in parts (a) and (b) of Theorem 1, that go beyond Theorem 4. It also remains to establish similar results for other models. The method of proof here can be adapted to other settings, as illustrated by the proof of Theorem 3 for the transient mean.

Acknowledgments

This research was supported by NSF CMMI 1634133.

References

Abate, J., G. L. Choudhury, W. Whitt. 1993. Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Archiv fur Elektronik und bertragungstechnik* **47**(5/6) 311–321.

Appa, G. 2002. On the uniqueness of solutions to linear programs. *Journal of the Operational Research Society* **53** 1127–1132.

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.

Berge, C. 1963. *Topological Spaces*. Macmillan, New York. (English translation of the 1959 French edition).

Bertsimas, D., K. Natarajan. 2007. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems* **56** 27–39.

Border, K. C. 1985. *Fixed Point Theorems with Application to Economics and Game Theory*. Cambridge University Press, New York.

Chen, Y., W. Whitt. 2018. Algorithms for the upper bound mean waiting time in the $GI/GI/1$ extremal queue. Submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chen, Y., W. Whitt. 2019. On the lower bound mean waiting time in the $GI/GI/1$ extremal queue. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chung, K. L. 2001. *A Course in Probability Theory*. 3rd ed. Academic Press, New York.

Cohen, J. W. 1982. *The Single Server Queue*. 2nd ed. North-Holland, Amsterdam.

Daley, D. J. 1977. Inequalities for moments of tails of random variables, with queueing applications. *Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete* **41** 139–143.

Daley, D. J., A. Ya. Kreinin, C.D. Trengove. 1992. Inequalities concerning the waiting-time in single-server queues: a survey. U. N. Bhat, I. V. Basawa, eds., *Queueing and Related Models*. Clarendon Press, 177–223.

Eckberg, A. E. 1977. Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Mathematics of Operations Research* **2**(2) 135–142.

Gupta, V., J. Dai, M. Harchol-Balter, B. Zwart. 2010. On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems* **64** 5–48.

Gupta, V., T. Osogami. 2011. On Markov-Krein characterization of the mean waiting time in $M/G/K$ and other queueing systems. *Queueing Systems* **68** 339–352.

Halfin, S. 1983. Batch delays versus customer delays. *Bell Laboratories Technical Journal* **62**(7) 2011–2015.

Holtzman, J. M. 1973. The accuracy of the equivalent random method with renewal inouts. *Bell System Techn ical Journal* **52**(9) 1673–1679.

Johnson, M. A., M. R. Taaffe. 1990a. Matching moments to phase distributions: Density function shapes. *Stochastic Models* **6**(2) 283–306.

Johnson, M. A., M. R. Taaffe. 1993. Tchebycheff systems for probability analysis. *American Journal of Mathematical and Management Sciences* **13**(1-2) 83–111.

Johnson, M. A., M.R. Taaffe. 1990b. Matching moments to phase distributions: nonlinear programming approaches. *Stochastic Models* **6**(2) 259–281.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* **8**(3) 457–459.

Karlin, S., W. J. Studden. 1966. *Tchebycheff Systems; With Applications in Analysis and Statistics*, vol. 137. Wiley, New York.

Kingman, J. F. C. 1961. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* **77** 902–904.

Kingman, J. F. C. 1962. Inequalities for the queue $GI/G/1$. *Biometrika* **49**(3/4) 315–324.

Klincewicz, J. G., W. Whitt. 1984. On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Technical Journal* **63**(1) 139–161.

Lasserre, J. B. 2010. *Moments, Positive Polynomials and Their Applications*. Imperial College Press.

Li, Y., D. A. Goldberg. 2017. Simple and explicit bounds for multii-server queues with universal $1/(1-\rho)$ and better scaling. ArXiv:1706.04628v1.

Minh, D. L., R. M. Sorli. 1983. Simulating the $GI/G/1$ queue in heavy traffic. *Operations Research* **31**(5) 966–971.

Nocedal, J., S. J. Wright. 1999. *Numerical Optimization*. Springer, New York.

Osogami, T., R. Raymond. 2013. Analysis of transient queues with semidefinite optimization. *Queueing Systems* **73** 195–234.

Ott, T. J. 1987. Simple inequalties for the $D/G/1$ queue. *Opperations Research* **35**(4) 589–597.

Parthasarathy, K. R. 1967. *Probability Measures on a Metric Space*. Academic Press, New York.

Rolski, T. 1972. Some inequalities for $GI/M/n$ queues. *Zast. Mat.* **13**(1) 43–47.

Ross, S. M. 1996. *Stochastic Processes*. 2nd ed. Wiley, New York.

Smith, J. 1995. Generalized Chebychev inequalities: Theory and application in decision analysis. *Operations Research* **43** 807–825.

Smith, W. 1953. On the distribution of queueing times. *Mathematical Proceedings of the Cambridge Philosophical Society* **49**(3) 449–461.

Stoyan, D. 1983. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley and Sons, New York. Translated and edited from 1977 German Edition by D. J. Daley.

Stoyan, D., H. Stoyan. 1974. Inequalities for the mean waiting time in single-line queueing systems. *Engineering Cybernetics* **12**(6) 79–81.

Whitt, W. 1983a. Comparing batch delays and customer delays. *Bell Laboratories Technical Journal* **62**(7) 2001–2009.

Whitt, W. 1983b. The queueing network analyzer. *Bell Laboratories Technical Journal* **62**(9) 2779–2815.

Whitt, W. 1984a. Minimizing delays in the $GI/G/1$ queue. *Operations Research* **32**(1) 41–51.

Whitt, W. 1984b. On approximations for queues, I. *AT&T Bell Laboratories Technical Journal* **63**(1) 115–137.

Whitt, W. 1984c. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal* **63**(1) 163–175.

Whitt, W. 1989. Planning queueing simulations. *Management Science* **35**(11) 1341–1366.

Whitt, W. 1991. The efficiency of one long run versus independent replications in steady-state simulation. *Management Science* **37**(6) 645–666.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Whitt, W., W. You. 2018. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research* **66** 100–120.

Wolff, R. W., C. Wang. 2003. Idle period approximations and bounds for the $GI/G/1$ queue. *Advances in Applied Probability* **35**(3) 773–792.

e-Companion to "Extremal $GI/GI/1$ Queues" by Y. Chen and W. Whitt

EC.1. Overview

In this appendix to the main paper, we provide postponed proofs and then we present additional tables and plots. First, in §EC.2 we complete the proof of Theorem 1 showing the existence of three-point extremal queues. In §EC.3 we present the proof of Theorem 4 identifying explicit extremal distributions under an extra monotonicity assumption. In §EC.4 we present an alternative way to identify the explicit extremal distributions by means of Tchebycheff systems. Then in §EC.5 we prove Theorem 6, which establishes the new UB formula given Conjecture 1. In §EC.6 we discuss the extension to unbounded support.

We next provide additional numerical results. First, in §EC.7 we present additional numerical comparisons of the bounds and approximations, supplementing Tables 1 and 2 in §7. §EC.8 we present numerical values of $E[W_n(F_0, G_u)]$ from the optimization and optimal search in §5 that complement Table EC.8. In §EC.9 we present additional counterexamples to strong conclusons with one distribution fixed. In §EC.10 we present additional numerical results for the upper bound of the steady-state mean $E[W]$ when one distribution is deterministic, further supplementing §6.

EC.2. More on the Proof of Theorem 1.

We have given the first step of the proof of part (a) in §2.2. We now elaborate on Steps 2-4 for part (a) and give the proof for (b) and thus (c) in Step 5 here.

Step 2. Existence by the Kakutani Fixed Point Theorem. We now show that the set $\mathcal{P}^*_{a,2}$ of fixed points in (21) is nonempty. Recall that $\mathcal{P}^*_{a,2}$ is the set of all fixed points of the map $\eta : \mathcal{P}_{a,2}(M_a) \to 2^{\mathcal{P}_{a,2}(M_a)}$, where $\eta(F_1)$ is the set of all maximizers of $\zeta(F_1)$ in (20). For this purpose, we apply the Kakutani fixed point theorem; e.g., see Kakutani (1941) and Border (1985), so we state it here.

THEOREM EC.1. (*Kakutani fixed point theorem*) *If S is a non-empty compact and convex subset of some Euclidean space $\mathbb{R}^d$ and $\psi : S \to 2^S$ is a set-valued function with a closed graph such that*

$\psi(x)$ is non-empty and convex for all $x \in S$, then the map ψ has a fixed point, i.e., there exists $x \in S$ such that $x \in \psi(x)$.

In order to be able to work within the Euclidean space $\mathbb{R}^d$, we first restrict attention to the set of probability measures with finite support in $[0, M_a]$; that is homeomorphic to a convex compact subset of $\mathbb{R}^n$. We use an asymptotic argument to get the the entire set $\mathcal{P}_{a,2}(M_a)$ in Step 4. Thus, for $k \geq 3$, let $\mathcal{P}^e_{a,2,k+1}$ be the subset of cdf's F with support

$$\mathcal{S}_{k+1} \equiv \{x_1, \ldots, x_{k+1} : 0 \leq x_1 < \cdots < x_{k+1} \leq x_u\} = \mathcal{S}^e_{k+1} \equiv \{jM_a/k : 0 \leq j \leq k\}.$$

The space $\mathcal{P}^e_{a,2,k+1}$ is homeomorphic to a non-empty compact and convex subset of $\mathbb{R}^{k+1}$. (If desired, we can let $k = 2^l$, making the subsets indexed by l nested, $\mathcal{S}^e_{l+1} \subseteq \mathcal{S}^e_{(l+1)+1}$.) Hence, we can apply the Kakutani fixed point theorem to show that the set of fixed points $\mathcal{P}^*_{a,2}$ in (21) is nonempty when we restrict F to $\mathcal{P}^e_{a,2,k+1}$.

To apply Theorem EC.1, we let ψ in Theorem EC.1 be η, where $\eta(F_1)$ is the set of all maximizers of $\zeta(F_1)$ in (20). Thus, we need to show that $\eta(F_1)$ has a closed graph and that $\eta(F_1)$ is nonempty and convex for each F_1. Recall that a set-valued function ψ is said to have a closed graph (or be upper-hemicontinuous) if for all sequences $\{(x_n, y_n) : n \geq 1\}$ such that $y_n \in \psi(x_n)$ for all n, $x_n \to x$ and $y_n \to y$, we also have $y \in \psi(x)$.

To show that η has a closed graph, we apply the Berge maximum theorem, e.g., Berge (1963), a version of which we state here.

THEOREM EC.2. (*Berge maximum theorem*) *Let S be a compact metric spaces; let $w : S \times S \to \mathbb{R}$ be a continuous function; let $w^\uparrow(x_1) \equiv \sup\{w(x_1, x_2) : x_2 \in \S\}$; and let $\eta : S \to 2^S$ be the set of $x_2 \in S$ such that $w(x_1, x_2) = w^\uparrow(x_1)$. Then η has a closed graph (is upper-hemicontinuous), $\eta(x_1)$ is nonempty, compact and $w^\uparrow : S \to \mathbb{R}$ is continuous.*

.

To establish the continuity condition in our context, we use the continuity of the mean steady-state waiting time as a function of the interarrival-time cdf F within the set $\mathcal{P}_{a,2}(M_a)$ with specified finite first two moments, see §X.6 of Asmussen (2003).

It remains to show that $\eta(F_1)$ is convex for each F_1 when $\eta(F_1)$ is the set of all maximizers of $\zeta(F_1)$ in (20), but that convexity follows from the linearity in F_2 of the integral in (24). The set $\eta(F)$ is also nonempty because we are maximizing a continuous function over a compact metric space. Hence the proof of Step 2 is complete.

Step 3a. Application of Theorem 2 in case (a). We show that $\mathcal{P}^*_{a,2}$ is a subset of $\mathcal{P}_{a,2,3}(M_a)$ by exploiting Theorem 2. Most of the proof of this step has been completed in the sketch in §2.2, but we repeat for clarity. We first write (20) in the form of (14). In particular, for G the fixed cdf of the service time V and H the cdf of a candidate waiting time $W(F^*, G)$ with finite mean where $F^* \in \mathcal{P}^*_{a,2}$, we can write

$$\sup\{E[(W(F^*,G)+V-U_{F_2})^+] : F_2 \in \mathcal{P}_{a,2}\} = \sup\{\int_0^{M_a} \phi(u)\, dF : F \in \mathcal{P}_{a,2}(M_a)\} \tag{EC.1}$$

for ϕ expressed as the double integral

$$\phi(u) \equiv \int_0^\infty \int_0^\infty (x+v-u)^+ \, dG(v)\, dH(x), \quad 0 \le u \le M_a. \tag{EC.2}$$

Next observe that ϕ in (EC.2) is a bounded continuous real-valued function of u because the cdf H has bounded mean. Hence, we can apply Theorem 2 to deduce that, for any pair of cdf's (G,H) of (V,W), we may take $F_2 \in \mathcal{P}_{a,2,3}(M_a)$.

Step 3b. Uniqueness via duality. We now show that, for any F^* in $\mathcal{P}^*_{a,2}$, the optimal solution in (EC.1) is unique, so that the fixed point solution F^* necessarily is in $\mathcal{P}_{a,2,3}(M_a)$. To do so, we impose regularity conditions on the two cdf's F and G, but later in Steps 4b and 4c we show that these regularity conditions can be relaxed. At first, our conclusion will depend on these regularity conditions.

To establish the uniqueuess, we consider the dual problem associated with the optimization in Theorem 2 as in (3) of Smith (1995). In particular, we are focusing on ϕ in (EC.2). The objective of the dual problem is to find the vector $\lambda^* \equiv (\lambda_0^*, \lambda_1^*, \lambda_2^*)$ that attains the infimum

$$\gamma(m_1,m_2) \equiv \inf_{\lambda \equiv (\lambda_0,\lambda_1,\lambda_2)} \{\lambda_0 + \lambda_1 m_1 + \lambda_2 m_2\}, \tag{EC.3}$$

where $m_i \equiv E[U^i]$, $i = 1, 2$ and λ_i are the decision variables (which are unconstrained), such that

$$\psi(u) \equiv \lambda_0 + \lambda_1 u + \lambda_2 u^2 \geq \phi(u) \quad \text{for all} \quad u \in \mathcal{F} \tag{EC.4}$$

where $\mathcal{F}$ is the support of F and

$$\phi(u) \equiv \int_0^\infty \int_0^\infty (x + v - u)^+ dH(x) dG(v) = \int_0^\infty (x - u)^+ d\Gamma(x) \tag{EC.5}$$

where Γ is the cdf of $W + V$, as in (EC.2). We see that the constraints produce quadratic functions $\psi(u)$ in (EC.4) that are required to dominate $\phi(u)$ in (EC.2).

To apply the Kakutani fixed point theorem in Step 2, we initially restrict attention to cdf's F with finite support, which we assume contain the endpoints 0 and M_a. Hence, we impose this condition on the cdf F, so that the optimization in Step 3 and the dual above are actually standard LP's. Thus, to establish uniqueness of optimal solution in the LP of step 3, we apply the following lemma; e.g., see pp. 1128-9 of Appa (2002).

LEMMA EC.1. (*non-degeneracy and uniqueness in LP*) *A standard LP has a unique optimal solution if and only if its dual has a non-degenerate optimal solution.*

To show that the dual problem has a non-degenerate optimal solution, we first determine the structure of the function ϕ in (EC.2) for case (a), which is where we introduce the regularity condition on the cdf G in Lemma 2.

For our detailed proof of uniqueness in part (a), we use Lemma 2, showing that, under regularity conditions, the function ϕ in (EC.2) is continuous, strictly positive, strictly decreasing and strictly convex. We start knowing that both the dual LP and the primal LP (EC.1) have feasible solutions in each optimization of Step 3, and thus they both have at least one optimal solution. Recall that we are working with standard LP's, where the cdf F has finite support set $\mathcal{F}$. But note that the support set $\mathcal{F}$ always contains the two endpoints, which we have assumed are 0 and M_a. First, if $M_a = m_2$, then the primal has the unique feasible, and thus optimal, solution F_0. So henceforth assume that $M_a > m_2$ as well.

We will show that the primal LP (EC.1) has a unique solution by applying Lemma EC.1 and showing that all optimal solutions of the dual cannot be degenerate. That implies that the dual has at least one non-degenerate optimal solution. Hence, we will show that we cannot have the optimal λ_i^* be 0 for any i. First, we must have $\lambda_0 \geq \phi(0) > 0$, so we cannot have $\lambda_0^* = 0$.

Next, suppose that $\lambda_1 = 0$. In this setting, with $\lambda_0^* > 0$ and $\lambda_1^* = 0$, if $\lambda_2^* \geq 0$, then ψ can intersect ϕ only at 0, which cannot correspond to a feasible solution of the primal. On the other hand, if $\lambda_2^* < 0$, then ψ can only intersect ϕ at the two endpoints (without crossing), but that does not correspond to a feasible solution of the primal, assuming that $M_a > m_2$. Hence, we cannot have a degenerate optimal solution with $\lambda_1^* = 0$.

Finally, suppose that $\lambda_2^* = 0$, which makes ψ linear. Then again ψ can only meet ϕ at the two endpoints without crossing, but that does not correspond to a feasible solution of the primal, assuming that $M_a > m_2$ (as we have done).

Step 4. Two Asymptotic Arguments.

We now complete the proof of Theorem 1 (a) by carrying out two asymptotic arguments.

Step 4a. The first asymptotic argument. For each $k \geq 2$, let $F(k)$ be a fixed point with support $\mathcal{S}_{k+1}^e$. Since all these cdf's have common finite first two moments, the sequence $\{F(k) : k \geq 2\}$ is necessarily tight, so that there exists a subsequence $\{F(k_j) : j \geq 1\}$ such that $F(k_j) \Rightarrow F^*$ as $j \to \infty$. Moreover, since the cdf's have finite second moments, we have convergence of the associated steady-state waiting times $W_{k_j} \Rightarrow W^*$ and moments $E[W_{k_j}] \to E[W^*]$ as $j \to \infty$, again by virtue of §X.6 of Asmussen (2003). The limit then yields the desired fixed point in $\mathcal{P}_{a,2}(M_a)$. To summarize, we have seen that $F(k_j) \in \mathcal{P}_{a,2}(M_a)$ for all j and that

$$E[W_{k_j}] = E[(W_{k_j} + V - U_{F(k_j)})^+] \quad \text{for all} \quad j \geq 1,$$

where the three random variables on the right are regarded as mutually independent. Then the validity extends to the limit as $j \to \infty$, giving $F^* \in \mathcal{P}_{a,2}(M_a)$ and

$$E[W^*] = E[(W^* + V - U_{F^*})^+] \quad \text{for all} \quad j \geq 1,$$

where the three random variables on the right are again regarded as mutually independent. ∎

Step 4b. The second asymptotic argument. So far, we have established the existence of the fixed point in $\mathcal{P}_{a,2,3}(M_a)$ for all fixed G in $K_n \bigcap \mathcal{P}_{s,2}$, where K_n are the distributions with rational Laplace transform, as in Smith (1953) or §II.5.10 of Cohen (1982). (This regularity condition was used in Lemma 2 to guarantee that ϕ is strictly positive, strictly decreasing and strictly convex, which in turn was used in the uniqueness proof in Step 3c.) We use the following basic lemma to extend the result beyond that class.

LEMMA EC.2. (*a dense subset*) *The subset $K_n \bigcap \mathcal{P}_{s,2}$ is a dense subset of $\mathcal{P}_{s,2}$.*

Proof. Observe that any point mass on the positive halfline can be expressed as the limit of Erlang E_n distributions (which are in K_n) with fixed mean and variance approaching 0 as $n \to \infty$. Thus, any distribution with finite support is the limit of finite mixtures of E_n distributions (which also are in K_n). Since arbitrary distributions can be expressed as limits of distributions with finite support, we see that the conclusion holds. ∎

Hence, we can apply essentially the same argument as in Step 4b to prove that the result can be extended to an arbitrary cdf G in $\mathcal{P}_{s,2}$. For any fixed G in $\mathcal{P}_{s,2}$ and $n \geq 1$, let G_n be a cdf in $K_n \bigcap \mathcal{P}_{s,2}$ such that $G_n \Rightarrow G$ as $n \to \infty$. Let F_n be a fixed point in $\mathcal{P}_{a,2,3}(M_a)$ associated with G_n for each $n \geq 1$. Since, the sequence $\{F_n : n \geq 1\}$ is tight, it contains a convergent subsequence with limit F^*, which is in $\mathcal{P}_{a,2,3}(M_a)$ because it is compact. As in Step 4b, that limiting F^* is the fixed point associated with the limiting G.

Step 5. Application of Theorem 2 in case (b). We now treat case (b). The first two steps are essentially the same, but there are some differences in the third step. We reduce the proof to case (a) by using a reverse-time representation.

Step 5a. A Reverse-Time Representation. Instead of (20), we have

$$\zeta(G_1) \equiv \sup\{E[(W_1 + V - U)^+] : G_V \in \mathcal{P}_{s,2}(M_s)\}, \tag{EC.6}$$

where G_V is understood to be the cdf of V, W_1 is the steady-state waiting time associated with G_1 and the three variables W_1, V and U in (20) are taken to be mutually independent. Now we

introduce a reverse-time construction, i.e., focusing on $M_s - v$ instead of v. In particular, we write (EC.6) in the form of (14) after changing the underlying measure from the cdf G of V to the cdf G_{M_s-V} of $M_s - V$. The optimization becomes

$$\sup\{E[(W+V-U)^+] : G_V \in \mathcal{P}_{s,2}(M_s)\} = \sup\{\int_0^{M_s} \phi_s(v)\, dG_{M_s-v} : G_{M_s-v} \in \mathcal{P}_{s,2}(M_s)\} \quad \text{(EC.7)}$$

for ϕ_s expressed as the double integral

$$\phi_s(v) \equiv \int_0^\infty \int_0^\infty (x + M_s - v - u)^+ \, dF(u)\, dH(x), \quad 0 \le v \le M_s. \quad \text{(EC.8)}$$

A tricky part is the value of the moments. The formula for a moment has a fixed form for any measure. Hence, when we change the measure, we necessarily change the values of the moments. In our context, new values for the moments are

$$\hat{m}_k \equiv E[(M_s - V)^k], \quad k \ge 1. \quad \text{(EC.9)}$$

Since $E[V] = \rho$, we obtain $\hat{m}_1 = M_s - \rho$ and $\hat{m}_2 = M_s^2 - 2\rho M_s + \rho^2(c_s^2 + 1)$.

Hence, paralleling (26)-(28), the new dual problem is

$$\gamma(m_1, m_2) \equiv \inf_{\lambda \equiv (\lambda_0, \lambda_1, \lambda_2)} \{\lambda_0 + \lambda_1 \hat{m}_1 + \lambda_2 \hat{m}_2\}, \quad \text{(EC.10)}$$

where $\hat{m}_i \equiv E[(M_s - V)^i]$, $i = 1, 2$ and λ_i are the decision variables (which are unconstrained), such that

$$\psi(v) \equiv \lambda_0 + \lambda_1 v + \lambda_2 v^2 \ge \phi_s(v) \quad \text{for all} \quad v \in \mathcal{G} \quad \text{(EC.11)}$$

where $\mathcal{G}$ is the support of G and $\phi_s(v)$ is in (EC.8)

Next observe that, as before, ϕ_s in (EC.8) is a bounded continuous real-valued function of v because and the cdf H has bounded mean. Hence, we can apply Theorem 2 to deduce that, for any pair of cdf's (F, H) of (U, W), we may take $G \in \mathcal{P}_{s,2,3}(M_s)$.

We now exhibit the structure of the function ϕ_s and show that it has the same essential structure as ϕ in part (a). First, we write

$$\phi_s(v) = E[(W + M_s - v - U)^+] = E[(X - v)^+] \quad \text{for} \quad X \equiv M_s + W - U. \quad \text{(EC.12)}$$

We use the following basic result in our analysis.

LEMMA EC.3. *For any $GI/GI/1$ queue with $F \in \mathcal{P}_{a,2}$, $G \in \mathcal{P}_{s,2}$ and $\rho < 1$, $P(W = 0) > 0$.*

We now characterize the structure of ϕ_s under regularity conditions imposed on the interarrival-time cdf F.

LEMMA EC.4. *If, as occurs when the cdf F of U is in K_n, (i) the cdf F is differentiable with a strictly positive pdf f that can be expressed as*

$$f(u) = \int_0^u \dot{f}(x)\, dx, \quad u \geq 0, \tag{EC.13}$$

where $\dot{f}$ is integrable, and (ii) W has a cdf H with $H(0) > 0$ and

$$H(x) = H(0) + \int_0^x h(w)\, dw \quad x \geq 0, \tag{EC.14}$$

where h is strictly positive and integrable over the halfline, then ϕ_s in (EC.12) can be expressed as

$$\phi_s(v) = H(0)E[(M_s - U - v)^+] + \int_0^\infty h(w)E[(w + M_s - U - v)^+]\, dw > 0, \tag{EC.15}$$

so that the first two derivatives of ϕ_s in (EC.12) and (EC.15) exist for $v > 0$ and, with satisfy

$$\begin{aligned} \dot{\phi}_s(v) &= \Theta(v) - 1 = -H(0)F(M_s - v) - \int_0^\infty h(w)F(w + M_s - v)\, dw < 0, \\ \ddot{\phi}_s(v) &= \theta(v) = H(0)f(M_s - v) + \int_0^\infty h(w)f(w + M_s - v)\, dw > 0, \quad v \geq 0, \end{aligned} \tag{EC.16}$$

where $\Theta(v) \equiv P(W + M_s - U \leq v)$ and $F^c \equiv 1 - F$, so that ϕ_s is strictly positive, strictly decreasing and strictly convex on $[0, \rho M_s]$. Moreover, from (EC.16) we see that if f is strictly decreasing, then $\ddot{\phi}_s(v)$ is strictly increasing as well.

Proof. After carefully treating the atom, we can apply the same proof as for Lemma 2. The first line of (EC.16) follows from Lemma EC.3. Given that $P(W = 0) > 0$ and U has support on the entire halfline, it is clear that $\Theta(v) < 1$ for all $v > 0$, so that $\dot{\phi}_s(v) < 0$ for all $v \leq M_s$. ∎

Given that the structure of the dual problem is the same as for the previous one, we can use our proof of uniqueness in case (a). That completes the proof of Theorem 1. ∎

REMARK EC.1. (the special case of $M/GI/1$ for case (b))

For $M/GI/1$ in case (b), we already know the answer. In this case, we know that the $E[W]$ is insensitive to the service-time cdf beyond its first two moments. For any steady-state waiting time W with cdf H, let its Laplace transform be

$$\hat{h}(s) \equiv E[e^{-sW}] = \int_0^\infty e^{-sw}\, dH(w), \tag{EC.17}$$

noting that (EC.17) includes a term for the atom $H(0)$.

For this case, we show that ϕ_s is a relatively simple function. In particular, since

$$\begin{aligned}\int_0^v (v-u)e^{-u}du &= v - ve^{-v} - \int_0^v ue^{-u}du \\ &= v - ve^{-v} + ve^{-v} + e^{-v} - 1 = v + e^{-v} - 1,\end{aligned} \tag{EC.18}$$

$$\phi_s(v) = e^{-M_s}\hat{h}(1)e^v + E[W] + M_s - v - 1, \tag{EC.19}$$

so that, in addition to $\phi_s(u) > 0$ from (EC.15), we have

$$\begin{aligned}\dot{\phi}_s(v) &= e^{-M_s}\hat{h}(1)e^v - 1 < 0, \quad \text{and} \\ \ddot{\phi}_s(v) &= e^{-M_s}\hat{h}(1)e^v > 0.\end{aligned} \tag{EC.20}$$

Hence, $\phi_s(v)$ is a linear combination of $\{1, v, e^v\}$. Therefore, the system $\{1, v, v^2, \phi_s(v)\}$ is a T-system for any steady-state waiting time distribution W. So we can deduce that either G_0 or G_u must be contained in the fixed point set $\mathcal{P}^*_{s,2}$.

Finally, we note that even though the mean steady-state waiting time $E[W]$ depends on G only via its first two moments in the $M/GI/1$ model, the full distribution of W depends on the full service-time distribution, being uniquely characterized, as can be seen from the Pollaczek-Khintchine transform for $M/GI/1$. Thus, in the fixed-point iteration, there is a unique optimum, but that fixed-point iteration also depends on the distribution of W to begin the iteration.

EC.3. Proof of Theorem 4.

We first give a detailed proof of case (a). By Theorem 1 (a), we know that there is an extremal distribution with at most three points in its support. We want to further reduce the possibilities to a two-point distribution or even to the natural candidate F_0.

For the following, we assume that the service-time cdf G has the regularity conditions assumed in Lemma 2 and the interarrival-time cdf a fixed point solution obtained from the proof of Theorem 1, i.e., $F^* \in \mathcal{P}^*_{a,2}(M_a)$ in (21).

LEMMA EC.5. *Suppose that $G \in K_n$ and $F^* \in \mathcal{P}^*_{a,2}(M_a)$ in (21). If M_a is sufficently large, then M_a is not in the support of F^*.*

Proof. We prove that M_a cannot be part of an optimal solution if M_a is suitably large by showing that the associated dual for (a) in (26)-(28) cannot have M_a in a solution. We first observe that the objective function in (26) is independent of M_a, provided only that the choice of M_a is consistent with the specified moments. Hence, we will be focusing on the constraints in (27). The function ϕ in (27) depends on M_a and so do the optimal solution $(\lambda_0, \lambda_1, \lambda_2)$. Henceforth, these are understood to be the optimal values.

For each M_a, the support of $F^* \equiv F^*(M_a)$ can be identified by the roots of the equation

$$\psi(r) \equiv \lambda_0 + \lambda_1 r + \lambda r^2 = \phi(r), \tag{EC.21}$$

where ψ and λ_i correspond to the optimal solution. We use the knowledge of the sign of these optimal values. In particular, $\lambda_0 \geq \phi(0) = E[W(F^*, G)] + \rho > 0$ and $\lambda_1 < 0 < \lambda_2$.

By Theorem 1, equation (EC.21) has at most three roots. Because the mean of F^* is necessarily 1, the least root $r_1(M_a)$ satisfies $0 \leq r_1(M_a) < 1$ for all M_a. We will assume that M_a is also a root and show that leads to a contradiction when M_a is sufficiently large.

Given that $r_1(M_a) < 1$ and M_a are roots of (EC.21), we can look at the difference

$$\begin{aligned} \psi(r_1(M_a)) - \psi(M_a) &= \lambda_1(M_a)(r_1(M_a) - M_a) + \lambda_2(M_a)(r_1(M_a)^2 - M_a^2) \\ &= \phi(r_1(M_a)) - \phi(M_a), \end{aligned} \tag{EC.22}$$

so that, after dividing by $M_a - r_1(M_a)$,

$$\begin{aligned} -\lambda_1(M_a) &= \lambda_2(M_a)(r_1(M_a) + M_a) + \frac{\phi(r_1(M_a)) - \phi(M_a)}{M_a - r_1(M_a)} \\ &\le \lambda_2(M_a)(1 + M_a) + \frac{\phi(0)}{M_a - 1}. \end{aligned} \tag{EC.23}$$

As a consequence of (EC.23),

$$\limsup_{M_a \to \infty} \{-\lambda_1(M_a)/\lambda_2(M_a)M_a\} \le 1. \tag{EC.24}$$

We now develop a contradiction with (EC.24). To do so, we consider the quadratic equation

$$\psi(x) - \phi(M_a) = 0, \quad x \ge 0. \tag{EC.25}$$

The solutions of this quadratic equation are

$$x = \frac{-\lambda_1(M_a) \pm \sqrt{\lambda_1(M_a)^2 - 4\lambda_2(M_a)(\lambda_0(M_a) - \phi(M_a))}}{2\lambda_2(M_a)}. \tag{EC.26}$$

Since we have tentatively assumed that M_a is in the optimal solution, M_a is necessarily a root of this quadratic equation. Since $\phi(x) \downarrow 0$, the quadratic equation in (EC.25) has two real roots, one at M_a and one greater than M_a. but we cannot have any roots less than M_a. Hence, we must have $-\lambda_1(M_a)/2\lambda_2(M_a) \ge M_a$, which implies that

$$\liminf_{M_a \to \infty} \{-\lambda_1(M_a)/\lambda_2(M_a)M_a\} \ge 2, \tag{EC.27}$$

but that contradicts (EC.24). ∎

We next introduce conditions on the pdf γ of $W(F^*, G) + V_G$ that guarantee that there must be a two-point extremal distribution.

LEMMA EC.6. *Suppose that $G \in K_n$ and $F^* \in \mathcal{P}^*_{a,2}(M_a)$ in (21). If γ is unimodal, then the support of F^* must be in $\mathcal{P}_{a,2,2}$.*

Proof. We first show that, if γ is unimodal, then the extremal distribution cannot have support $\{0, x, M_a\}$ for some x, $0 < x < M_a$. Afterwards, we show that the support of F^* must be in $\mathcal{P}_{a,2,2}$ if γ is unimodal.

Step 1. We now show that if γ is unimodal, then F^* cannot have support $\{0, x_1, M_a\}$ for some x, $0 < x_1 < M_a$. Suppose that F^* does have support $\{0, x_1, M_a\}$ for some x_1, $0 < x_1 < M_a$. Because 0 is in the support of F^*, we have $\psi(0) = \phi(0)$ and $\dot{\psi}(0) \geq \dot{\phi}(0)$. We consider the three alternatives for the second derivative.

Case (i). If $\ddot{\psi}(0) > \ddot{\phi}(0)$, $\dot{\psi}(x) - \dot{\phi}(x)$ is a first increasing, then decreasing and finally increasing function between $[0, x_1]$. Therefore, there exists at least two zeros for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, x_1]$. But F^* has support on M_a, so there exists at least one zero for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in (x_1, M_a]$. There are at least three zeros for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, M_a]$, which contradicts with γ being unimodal.

Case (ii). If $\ddot{\psi}(0) = \ddot{\phi}(0)$, then unimodal γ implies there exists at most one zero for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in (0, M_a]$. We observe the $\dot{\psi}(x) - \dot{\phi}(x)$ is a first decreasing and then increasing function between $[0, x_1]$. So there exists at least one zero for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, x_1]$. But F^* has support on M_a, so we obtain at least two zeros for $\ddot{\phi}(x) = \ddot{\psi}(x)$ for $x \in (0, M_a]$. Hence, that also leads to a contradiction.

Case (iii). Finally, if $\ddot{\psi}(0) < \ddot{\phi}(0)$, then $\dot{\psi}(x) - \dot{\phi}(x)$ is a first decreasing and then increasing function between $[0, x_1]$. So there exists at least one zero for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, x_1]$. F^* has support on M_a leads to at least two zeros $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, M_a]$. But if γ is unimodal with $\ddot{\psi}(0) < \ddot{\phi}(0)$, then there exists at most one zero for $\ddot{\phi}(x) = \ddot{\psi}(x)$ when $x \in [0, M_a]$. That also leads to a contradiction.

Step 2. We now turn to the second part. we now show that the support of F^* must be in $\mathcal{P}_{a,2,2}$ if γ is unimodal. For the second part, we assume that F^* is attained at the three points x_i with $0 \leq x_1 < x_2 < x_3 < M_a$. (The argument is essentially the same in the other case: $0 < x_1 < x_2 < x_3 \leq M_a$.) We will show that γ cannot be unimodal. To do so, we observe that if γ is unimodal, then the

equation $c-\gamma(x)=0$ cannot have at least three zeros in either the open interval $[0,M_a)$ or $(0,M_a]$ (including at most one endpoint) for some $c>0$. However, if F^* is attained at the three points x_i with $0\le x_1<x_2<x_3<M_a$, then $D(x)\equiv\psi(x)-\phi(x)$ must have four extreme points in the open interval $(0,M_a)$: the two minima x_2 and x_3 and the two maxima in the intervals (x_1,x_2) and (x_2,x_3). These four extreme points of $D(x)\equiv\psi(x)-\phi(x)$ are attained at the points x satisfying $\dot{D}(x)=0$. Next observe that, since $D(x)\ge 0$ for $0\le x\le M_a$, that as a consequence, $\dot{D}(x)$ must have three extreme points in $(0,M_a)$. Since $\gamma(x)=\ddot{D}(x)$, that implies that the the equation $c-\gamma(x)=0$ must have three zeros, and so cannot be unimodal. ∎

We conclude by giving a brief sketch of the proof of Theorem 4 (b). First, we are so far unable to prove an analog of Lemma EC.5 for (b), so that does not appear in (b). The problem in (b) differs from the problem in (a), because in (b) the objective function depends strongly on M_s, as can be seen from (EC.9) and (EC.10).

However, the proof of Lemma EC.6 extends directly to (b) with the time reversal. Thus, we can conclude that, if f is unimodal, then $G^*\in\mathcal{P}_{s,2,2}$. Moreover, we see that, if f is strictly monotone decreasing, then θ in by (EC.16) in Lemma EC.4 is strictly monotone increasing, which implies that either G_0 or G_u must be an extremal distribution.

LEMMA EC.7. *In the setting of Theorem 1* (b)*, suppose that $F\in K_n$ and $G^*\in\mathcal{P}^*_{s,2}(M_s)$ in the analog of* (21)*. If θ is monotone, then G_0 is an extremal distribution.*

Proof. We first show that either G_0 or G_u must be an extreme point. Then we show that it can only be G_0. For the first step, the argument is a minor variant of the second step in the proof of Lemma EC.6 except we have one less point. We now assume that G^* is attained at two interior points, i.e., at the two points x_i with $0<x_1<x_2<M_s$. Then $D(x)\equiv\psi(x)-\phi_s(x)$ must have three extreme points in the open interval $(0,M_s)$: the two minima x_1 and x_2 and the maximum in the interval (x_1,x_2). These three extreme points of $D(x)\equiv\psi(x)-\phi(x)$ are attained at the points x satisfying $\dot{D}(x)=0$. Next observe that, since $D(x)\ge 0$ for $0\le x\le M_s$, that as a consequence, $\dot{D}(x)$ must have two extreme points in $(0,M_s)$. Since $\theta(x)=\ddot{D}(x)$, that implies that the the equation

$c-\theta(x)=0$ must have two zeros, and so θ cannot be monotone. Hence there must be only one interior point. Finally, we observe that the extremal cdf must be G_0 by looking at the derivative at the one interior point x, which must be positive, preventing another root at M_s. ∎

EC.4. Tchebycheff Systems and Two-Point Extremal Distributions

A variant of the proof of Theorem 1 can yield two-point extremal distributions if we can show that the function ϕ in (EC.2) for case (a) together with the basic functions 1, u and u^2 is a T (Tchebycheff) system.

In particular, for these results we apply the Markov-Krein theorem from Karlin and Studden (1966), Johnson and Taaffe (1993), Gupta and Osogami (2011). The functions $\{f_0,...,f_n\}$ form a *Tchebycheff* system over $[a,b]$ provided the Tchebycheff determinants are strictly positive whenever $a\le x_0<x_1,...,x_{n-1}<x_n\le b$.

THEOREM EC.3. (*Markov-Krein*) *If* $\{f_0,...,f_n\}$ *and* $\{f_0,...,f_n,\phi\}$ *are T systems on the interval* $[0,M]$, *then there exists unique extremal distributions* μ_L *and* μ_U *of* $m=\{1,m_1,...,m_n\}$ *such that infimum and supremum of the following two moment problems,*

$$\inf_{\mu\in\mathcal{D}}\{\mathbb{E}[\phi(u)]:\mathbb{E}[f_i(u)]=m_i,i=0,1,2,...,n\},$$

$$\sup_{\mu\in\mathcal{D}}\{\mathbb{E}[\phi(u)]:\mathbb{E}[f_i(u)]=m_i,i=0,1,2,...,n\}$$

are attained. For the case $n=2$ *with* $f_i(u)\equiv u^i$, *the extremal distributions are* F_0 *and* F_u *in* $\mathcal{P}_{2,2}$. $\mathcal{D}$ *is the set includes all non-negative probability measures.*

In order to apply the Markov-Krein theorem to our problem, it remains to show that the assumed T-system property holds. Note that the functions 1, u, u^2 and $-(x-u)^+$ do *not* form a T system, because the function $-(x-u)^+$ is piecewise linear, but in (a) the integration in $\phi(u)$ with respect to a positive density can help. The following is a direct consequence of Theorem EC.3 and the proof of Theorem 1.

THEOREM EC.4. (*Further reduction to the classic extremal two-point distributions*) *In the setting of Theorem 1, let* $F^*(G)\in\mathcal{P}_{a,2,3}(M_a),G^*(F)\in\mathcal{P}_{s,2,3}(M_s)$ *be fixed point solutions from Theorem 1.*

Let $\phi(u) = E[(W(F^*(G), G)] + V_G - u)^+]$ *for* $u \in [0, M_a]$ *and* $\phi_s(v) = E[(W(F, G^*(F)) + M_s - v - U_F)^+]$ *for* $v \in [0, \rho M_s]$.

(*a*) *For any specified* $G \in \mathcal{P}_{s,2}$, *if* $-\phi(u)$ *in* (28) *with* $\{1, u, u^2\}$ *consists of a T system, then*

$$w_a^{\uparrow}(G) \equiv \sup\{w(F,G) : F \in \mathcal{P}_{a,2}\} = \sup\{w(F,G) : F \in \mathcal{P}_{a,2,2}\} = w(F_0, G). \tag{EC.28}$$

(*b*) *For any specified* $F \in \mathcal{P}_{a,2}$, *if* $-\phi_s(v)$ *in* (EC.12) *with* $\{1, v, v^2\}$ *consists of a T system, then*

$$w_s^{\uparrow}(F) \equiv \sup\{w(F,G) : G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w(F,G) : G \in \mathcal{P}_{s,2,2}(M_s)\} = w(F, G_u). \tag{EC.29}$$

(*c*) *If both conditions in* (*a*) *and* (*b*) *are satisfied, then*

$$\begin{aligned} w^{\uparrow} &\equiv \sup\{w(F,G) : F \in \mathcal{P}_{a,2}(M_a), G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w(F,G) : F \in \mathcal{P}_{a,2,2}(M_a), G \in \mathcal{P}_{s,2,2}(M_s)\} \\ &= w(F_0, G_u). \end{aligned} \tag{EC.30}$$

EC.5. Proof of Theorem 6

In this section we prove Theorem 6, which provides an UB for $E[W]$ in the conjectured $F_0/G_{u^*}/1$ extremal $GI/GI/1$ queue. The notation G_{u^*} means the limit of G_u as $M_s \to \infty$.

Following §10 of Daley et al. (1992), we concentrate on the class $\mathcal{P}_{a,2} \times \mathcal{P}_{s,2}$ and attempt to determine the best choices of functions $a(\rho), b(\rho)$ such that

$$E[W] \le \frac{a(\rho)c_a^2 + b(\rho)c_s^2}{2(1-\rho)}. \tag{EC.31}$$

We apply Delay's decomposition in the subsequent Theorem EC.5 to $\lim_{M_s \to \infty} E[W(F, G_u)]$ to obtain

$$\lim_{M_s \to \infty} E[W(F, G_u)] = E[W(F, D)] + \lim_{M_s \to \infty} E[W(D, G_u)] = E[W(F, D)] + \frac{c_s^2}{2(1-\rho)}. \tag{EC.32}$$

Consequently, $b(\rho) \ge b_{LB}(\rho) = 1$. From (EC.32), the lower bound of $a(\rho)$ can be given by

$$a(\rho) \ge a_{LB}(\rho) = \inf_{c_a^2 > 0} \{\frac{2(1-\rho)}{c_a^2} \sup_{F \in \mathcal{P}_{a,2}} E[W(F, D)]\}. \tag{EC.33}$$

The $a_{LB}(\rho)$ is the best choice (if it exists) when set $b(\rho)=1$. The $a_{LB}(\rho)$ and $b_{LB}(\rho)$ can give a new upper bound for $GI/GI/1$, so that we obtain

$$E[W(F,G)] \le E[W(F_0,G_{u^*})] \le \frac{a_{LB}(\rho)c_a^2+c_s^2}{2(1-\rho)} \le \frac{a(\rho)c_a^2+b(\rho)c_s^2}{2(1-\rho)}. \tag{EC.34}$$

Now we are left to determine the $a_{LB}(\rho)$. At this point we focus on the candidate bounding system $F_0/GI/1$, so we obtain a proof only for this case. We obtain an alternative representation in Chen and Whitt (2018), which we state here. In particular, we can convert the queue $F_0/GI/1$ into $D/RS(V,p)/1$ where $RS(V,p)=\sum_{k=1}^{N(p)} V_k$ is a random sum of i.i.d. variables distributed as V, $N(p)$ is a geometric random variable on the positive integers having $E[(N(p))]=1/p$ with $1/p=1+c_a^2$. Here is the specific lemma:

LEMMA EC.8. (*Theorem 1 in Chen and Whitt (2018)*) *For the $F_0/GI/1$ model with service time V having mean ρ and scv c_s^2, the mean steady-state waiting time can be expressed as*

$$\begin{aligned}\mathbb{E}[W(F_0(p)/GI/1)] &= \mathbb{E}[W(D(1/p)/RS(V,p)/1)] + (\mathbb{E}[N(p)]-1)\mathbb{E}[V] \\ &= \mathbb{E}[W(D(1/p)/RS(V,p)/1)] + \rho(1-p)/p \\ &= \mathbb{E}[W(D(1/p)/RS(V,p)/1)] + \rho c_a^2. \end{aligned} \tag{EC.35}$$

Proof. The F_0 interarrival time means that a random number of arrivals, distributed as $N(p)$, arrive at deterministic intervals with deterministic value $1/p=c_a^2+1$. So the model has batch arrivals. The result in (EC.35) follows from Halfin (1983) or Theorem 1 of Whitt (1983a), which states that the delay of an arbitrary customer in the batch is distributed the same as the delay of the last customer in the batch when the batch-size distribution is geometric. Because $\mathbb{E}[W(D(1/p)/RS(V,p)/1)]$ is the expected delay of the first customer in a batch, we need to add the second term in (EC.35) to get the delay of the last customer in the batch; e.g., see §III of Whitt (1983a). ∎

Hence, we apply Lemma EC.8 to write

$$E[W(F_0,G)] = E[W(D,RS(V,p))] + \rho c_a^2. \tag{EC.36}$$

For the rest, we use a stochastic comparison argument involving convex stochastic order, as in §9.5 of Ross (1996) or in §1.7 and Chapter 5 of Stoyan (1983). Let convex order be denoted by $\leq_c$. In particular, consider an $F_0/GI/1$ system for which $S \leq_c S'$ where S' denotes a exponential random variable with mean $E[S]$. Then for two sequences of i.i.d. variables $\{S_n\}$ and $\{S'_n\}$,

$$S_1 + ... + S_{N(p)} \leq_c S'_1 + ... + S'_{N(p)}. \tag{EC.37}$$

However, the righthand side is distributed as an exponential random variable with mean $N(p)E[S]$, where $N(p)$ is a geometric random variable with mean $E[N(p)] = 1 + c_a^2$. Hence, we obtain

$$(S_1 + ... + S_{N(p)})/E[N(p)] \leq_c S'. \tag{EC.38}$$

Consequently,

$$\begin{aligned}(1+c_a^2)^{-1}W(D, RS(V,p)) &=_d W((1+c_a^2)D, S_1 + ... + S_{N(p)}) \\ &=_d W(D, (S_1 + .. + S_{N(p)})/(1+c_a^2)) \\ &\leq_c W(D, S') = W(D, M). \end{aligned} \tag{EC.39}$$

Hence,

$$(1+c_a^2)^{-1}E[W(D, RS(V,p))] \leq EW[(D, M)] = \delta\rho/(1-\delta). \tag{EC.40}$$

where $\delta = \exp(-(1-\delta)/\rho)$.

Finally, combine (EC.33), (EC.36) and (EC.40) to obtain

$$\begin{aligned} a_{LB}(\rho) &= \inf_{c_a^2>0} \frac{2(1-\rho)\sup_{F\in\mathcal{P}_{a,2}} E[W(F,D)]}{c_a^2} \\ &= \inf_{c_a^2>0} \frac{2(1-\rho)E[W(F_0,D)]}{c_a^2} \leq \inf_{c_a^2>0} \{2\rho(1-\rho) + \frac{(1+c_a^2)\delta\rho/(1-\delta)2(1-\rho)}{c_a^2}\} \\ &\to \frac{\rho(2-2\rho)}{1-\delta} (\text{as } c_a^2 \to \infty). \end{aligned} \tag{EC.41}$$

So $a_{LB}(\rho) \leq \rho(2-2\rho)/(1-\delta)$ and

$$E[W(F_0, G_{u^*})] \leq \frac{a_{LB}(\rho)c_a^2 + c_s^2}{2(1-\rho)} \leq \frac{2(1-\rho)\rho/(1-\delta)c_a^2 + \rho^2 c_s^2}{2(1-\rho)}. \tag{EC.42}$$

■

EC.6. Extension to Unbounded Intervals of Support

In this section we discuss what happens when we increase the intervals of support $[0, M_a]$ and $[0, \rho M_s]$. Throughout this section we assume that the UB for finite support has been shown to be (F_0, G_u). We ask what happens as we let $M_a \to \infty$ and $M_s \to \infty$.

EC.6.1. Unbounded Support for the Interarrival Time

First, for the interarrival-time cdf F, the cdf F_0 is optimal for the UB for all M_a, and thus remains optimal as $M_a \to \infty$. In contrast, for the lower bound, which we mostly do not discuss here, the extremal interarrival-time cdf is F_u, which places positive mass on M_a. Then the extremal interarrival-time cdf $F_u \equiv F_u(M_a)$ converges to the deterministic distribution with mean 1 as $M_a \to \infty$, which of course has $c_a^2 = 0$, which is likely to be inconsistent with the specified parameter. Nevertheless, the mean waiting time converges to the value $E[W(D, G)]$ of the associated $D/GI/1$ model, as we saw in Tables 1-2. Moreover, as discussed in Theorem 3.1 of Daley et al. (1992), that yields the well-known tight LB.

EC.6.2. Unbounded Support for the Service Time

The situation is more complicated when we let $M_s \to \infty$ for the upper bound. Just as for the interarrival-time cdf F_u, the service-time cdf $G_u \equiv G_u(M_s)$ converges to the deterministic cdf with the mean ρ of G_u as $M_s \to \infty$. However, the mean waiting time fails to converge to the mean waiting time of the associated $GI/D/1$ queue.

We propose two approaches to this problem. The first way is to exploit the representation in terms of the idle time in (52), as was done in Minh and Sorli (1983) and Wolff and Wang (2003). It turns out that the mean idle time does converge as $M_s \to \infty$. We discuss this approach in Chen and Whitt (2018). The second approach is to exploit the Daley decomposition from §10 of Daley et al. (1992), which we discuss next.

EC.6.3. The Daley Decomposition and Conjectures

We now discuss a decomposition for the mean steady-state waiting time $E[W]$ and three conjectures in §10 of Daley et al. (1992). The decomposition appears in equation (10.2) of Daley et al. (1992), where it is attributed to unpublished by D. J. Daley in 1984. We state it in the following theorem. Let G_{u^*} be shorthand for the limit $E[W(F,G_u)]$ as $M_s \to \infty$ and let D_m denote a deterministic cdf with mass 1 on m.

THEOREM EC.5. (*the Daley decomposition in* (10.2) *of* Daley et al. (1992)) *Consider the* $GI/GI/1$ *model with specified interarrival-time cdf* $F \in \mathcal{P}_{a,2}(1,c_a^2)$ *and unspecified service-time cdf* $G \in \mathcal{P}_{s,2}(\rho,c_s^2,M_s)$. *As* $M_s \to \infty$,

$$\begin{aligned} E[W(F,G_{u*})] \equiv \lim_{M_s\to\infty} E[W(F,G_u(M_s))] &= E[W(F,D_\rho)] + E[W(D_1,G_{u*})] \\ &= E[W(F,D_\rho)] + \frac{\rho^2 c_s^2}{2(1-\rho)}. \end{aligned} \tag{EC.43}$$

Proof. We only give a brief overview. We do a regenerative analysis to compute the mean waiting time, looking at successive busy cycles starting empty. We exploit the classic result that the steady-state mean waiting time is the expected sum of the waiting times over one cycle divided by the expected length of one cycle; e.g., see §3.6 and §3.7 of Ross (1996).

As M_s increases, the two-point cdf $G_u \equiv G_u(M_s)$ necessarily places probability of order $O(1/M_s^2)$ on M_s and the rest of the mass on a point just less than the mean service time, ρ. For very large M_s, there will be only rarely, with probability of order $O(1/M_s^2)$, a large service time of order $O(M_s)$. In the limit, most customers never encounter this large service time, so that we get a contribution to the overall mean $E[W]$ corresponding to $E[W(F,D_\rho)]$ in the first term on the right in (EC.43).

On the other hand, the total impact of the very large waiting time of order M_s is roughly the area of the triangle with height $O(M_s)$ and width $O(M_s)$, which itself is $O(M_s^2)$. When combined with the $O(1/M_s^2)$ probability, this produces an additional $O(1)$ impact on the steady-state mean, which is given by the second term on the right in (EC.43). Moreover, because we can use a law-of-large-numbers argument to treat this large service time, the asymptotic impact of that large

service time is independent of the interarrival-time cdf beyond its mean, so we can substitute D_1 for the original interarrival-time cdf F with mean 1 in the second term. ∎

Conjecture 1 shows that $\sup\{E[W(GI,D)]:F\} = E[W(F_0,D)]$. Hence, we can apply Theorem EC.5 to obtain the following corollary, which verifies (more strongly supports) Conjectures I and II on p. 209 of Daley et al. (1992).

COROLLARY EC.1. (*decomposition of the upper bound*) *For the $GI/GI/1$ model with unspecified interarrival-time cdf $F \in \mathcal{P}_{a,2}(1,c_a^2)$ and unspecified service-time cdf $G \in \mathcal{P}_{s,2}(\rho,c_s^2,M_s)$,*

$$\lim_{M_s\to\infty} \sup\{E[W(F,G)]:F,G\} = E[W(F_0,G_u*)] = E[W(F_0,D_\rho)] + \frac{\rho^2 c_s^2}{2(1-\rho)}. \qquad \text{(EC.44)}$$

Table EC.1 provides a numerical verification of Corollary EC.1 (and thus also Theorem EC.5). Table EC.1 reports simulation results using 20 i.i.d. replications, each with run length with 10^7. We show results for the four cases with $c_a^2 = 0.5, 4.0$ and $c_s^2 = 0.5, 4.0$ across a wide range of ρ.

Table EC.1 A comparison of two algorithms for computing $E[W(F_0,G_{u^*})]$ in the four cases $c_a^2 = 0.5, 4.0$ and $c_s^2 = 0.5, 4.0$

	$c_a^2 = c_s^2 = 4$		$c_a^2 = c_s^2 = 1/2$		$c_a^2 = 4, c_s^2 = 1/2$		$c_a^2 = 1/2, c_s^2 = 4$	
ρ	Daley's Reduction	$F_0/G_{u^*}/1$	Daley's Reduction	$F_0/G_{u^*}/1$	Daley's Reduction	$F_0/G_{u^*}/1$	Daley's Reduction	$F_0/G_{u^*}/1$
0.10	0.422	0.422	0.053	0.053	0.403	0.403	0.072	0.072
0.20	0.904	0.904	0.113	0.113	0.816	0.816	0.200	0.200
0.30	1.500	1.499	0.184	0.184	1.275	1.274	0.409	0.409
0.40	2.304	2.304	0.280	0.280	1.837	1.835	0.746	0.746
0.50	3.469	3.471	0.414	0.414	2.596	2.595	1.289	1.289
0.60	5.296	5.295	0.638	0.638	3.719	3.709	2.213	2.213
0.70	8.439	8.442	1.017	1.017	5.582	5.563	3.875	3.875
0.80	14.91	14.92	1.821	1.822	9.310	9.293	7.422	7.422
0.90	34.73	34.72	4.294	4.295	20.53	20.53	18.47	18.47
0.95	74.52	74.62	9.281	9.284	43.00	43.00	40.87	40.87
0.98	194.7	194.6	24.29	24.27	109.3	110.5	108.3	108.3
0.99	394.0	394.5	49.32	49.27	221.3	223.0	220.9	220.8

Our numerical results in Tables 1 and 2 also show that, while the UB approximation in (51) is an excellent approximation, it is not exact, which contradicts Conjecture III on p. 211 of Daley et al.

(1992). On the positive side, Corollary EC.1 provides the basis for an effective way to compute the overall upper bound $E[W]$.

EC.7. Numerical Comparison of the Bounds and Approximations

We now supplement Tables EC.2,EC.3 by making numerical comparions for the scaled means $(1-\rho)E[W]/\rho^2$ in two other cases: $(c_a^2, c_s^2) = (4.0, 0.5)$, $(0.5, 4.0)$. Tables EC.4-EC.7 then present the corresponding unscaled values.

Table EC.2 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = 4.0$ and $c_s^2 = 0.5$

ρ	Tight LB	HTA (4)	Tight UB	new UB (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	2.250	36.251	36.252	0.000	0.00%	38.250	200.250
0.15	0.000	2.250	22.934	22.946	0.001	0.05%	24.917	89.139
0.20	0.000	2.250	16.328	16.362	0.007	0.21%	18.250	50.250
0.25	0.000	2.250	12.436	12.493	0.020	0.45%	14.250	32.250
0.30	0.000	2.250	9.911	9.981	0.041	0.71%	11.583	22.472
0.35	0.000	2.250	8.161	8.239	0.070	0.96%	9.679	16.577
0.40	0.000	2.250	6.890	6.972	0.107	1.16%	8.250	12.750
0.45	0.000	2.250	5.933	6.014	0.152	1.35%	7.139	10.127
0.50	0.000	2.250	5.190	5.270	0.203	1.51%	6.250	8.250
0.55	0.000	2.250	4.606	4.677	0.261	1.53%	5.523	6.862
0.60	0.000	2.250	4.133	4.196	0.324	1.50%	4.917	5.806
0.65	0.000	2.250	3.744	3.799	0.393	1.45%	4.404	4.984
0.70	0.036	2.250	3.418	3.466	0.467	1.39%	3.964	4.332
0.75	0.083	2.250	3.145	3.184	0.546	1.23%	3.583	3.806
0.80	0.125	2.250	2.912	2.943	0.629	1.06%	3.250	3.375
0.85	0.162	2.250	2.710	2.734	0.716	0.86%	2.956	3.018
0.90	0.194	2.250	2.537	2.552	0.807	0.59%	2.694	2.719
0.95	0.224	2.250	2.384	2.392	0.902	0.31%	2.461	2.466
0.98	0.240	2.250	2.301	2.305	0.960	0.17%	2.332	2.332
0.99	0.245	2.250	2.275	2.277	0.980	0.09%	2.290	2.291

Table EC.3 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = 0.5$ and $c_s^2 = 4.0$

ρ	Tight LB	HTA (4)	Tight UB	new UB (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	2.250	6.500	6.500	0.000	0.00%	6.750	27.000
0.15	0.000	2.250	4.833	4.837	0.001	0.07%	5.083	13.111
0.20	0.000	2.250	4.002	4.014	0.007	0.30%	4.250	8.250
0.25	0.500	2.250	3.506	3.530	0.020	0.68%	3.750	6.000
0.30	0.833	2.250	3.182	3.216	0.041	1.08%	3.417	4.778
0.35	1.071	2.250	2.959	2.999	0.070	1.32%	3.179	4.041
0.40	1.250	2.250	2.799	2.840	0.107	1.47%	3.000	3.563
0.45	1.389	2.250	2.678	2.721	0.152	1.58%	2.861	3.235
0.50	1.500	2.250	2.577	2.628	0.203	1.91%	2.750	3.000
0.55	1.591	2.250	2.516	2.553	0.261	1.45%	2.659	2.826
0.60	1.667	2.250	2.458	2.493	0.324	1.40%	2.583	2.694
0.65	1.731	2.250	2.413	2.444	0.393	1.26%	2.519	2.592
0.70	1.786	2.250	2.373	2.402	0.467	1.23%	2.464	2.510
0.75	1.833	2.250	2.333	2.367	0.546	1.41%	2.417	2.444
0.80	1.875	2.250	2.319	2.337	0.629	0.74%	2.375	2.391
0.85	1.912	2.250	2.299	2.310	0.716	0.48%	2.338	2.346
0.90	1.944	2.250	2.280	2.288	0.807	0.32%	2.306	2.309
0.95	1.974	2.250	2.264	2.268	0.902	0.15%	2.276	2.277
0.98	1.990	2.250	2.255	2.257	0.960	0.06%	2.260	2.260
0.99	1.995	2.250	2.253	2.253	0.980	0.03%	2.255	2.255

Table EC.4 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 4.0$ and $c_s^2 = 4.0$

ρ	Tight LB	HTA (4)	Tight UB	UB Approx (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	0.044	0.422	0.422	0.000	0.00%	0.444	2.244
0.15	0.000	0.106	0.653	0.654	0.001	0.05%	0.706	2.406
0.20	0.000	0.200	0.904	0.906	0.007	0.19%	1.000	2.600
0.25	0.042	0.333	1.182	1.187	0.020	0.40%	1.333	2.833
0.30	0.107	0.514	1.499	1.508	0.041	0.60%	1.714	3.114
0.35	0.202	0.754	1.868	1.883	0.070	0.79%	2.154	3.454
0.40	0.333	1.067	2.304	2.326	0.107	0.94%	2.667	3.867
0.45	0.511	1.473	2.829	2.859	0.152	1.06%	3.273	4.373
0.50	0.750	2.000	3.470	3.510	0.203	1.15%	4.000	5.000
0.55	1.069	2.689	4.272	4.321	0.261	1.13%	4.889	5.789
0.60	1.500	3.600	5.295	5.352	0.324	1.07%	6.000	6.800
0.65	2.089	4.829	6.632	6.698	0.393	1.00%	7.429	8.129
0.70	2.917	6.533	8.441	8.520	0.467	0.93%	9.333	9.933
0.75	4.125	9.000	11.014	11.102	0.546	0.80%	12.000	12.500
0.80	6.000	12.800	14.917	15.017	0.629	0.67%	16.000	16.400
0.85	9.208	19.267	21.484	21.597	0.716	0.53%	22.667	22.967
0.90	15.750	32.400	34.721	34.843	0.807	0.35%	36.000	36.200
0.95	35.625	72.200	74.621	74.755	0.902	0.18%	76.000	76.100
0.98	95.550	192.080	194.557	194.702	0.960	0.07%	196.000	196.040
0.99	195.525	392.040	394.533	394.684	0.980	0.04%	396.000	396.020

Table EC.5 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 4.0$ and $c_s^2 = 0.5$

ρ	Tight LB	HTA (4)	Tight UB	UB Approx (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	0.025	0.403	0.403	0.000	0.00%	0.425	2.225
0.15	0.000	0.060	0.607	0.607	0.001	0.05%	0.660	2.360
0.20	0.000	0.113	0.816	0.818	0.007	0.21%	0.913	2.513
0.25	0.000	0.188	1.036	1.041	0.020	0.45%	1.188	2.688
0.30	0.000	0.289	1.274	1.283	0.041	0.71%	1.489	2.889
0.35	0.000	0.424	1.538	1.553	0.070	0.96%	1.824	3.124
0.40	0.000	0.600	1.837	1.859	0.107	1.16%	2.200	3.400
0.45	0.000	0.828	2.184	2.214	0.152	1.35%	2.628	3.728
0.50	0.000	1.125	2.595	2.635	0.203	1.51%	3.125	4.125
0.55	0.000	1.513	3.096	3.144	0.261	1.53%	3.713	4.613
0.60	0.000	2.025	3.720	3.777	0.324	1.50%	4.425	5.225
0.65	0.000	2.716	4.519	4.586	0.393	1.45%	5.316	6.016
0.70	0.058	3.675	5.583	5.662	0.467	1.39%	6.475	7.075
0.75	0.188	5.063	7.077	7.165	0.546	1.23%	8.063	8.563
0.80	0.400	7.200	9.317	9.417	0.629	1.06%	10.400	10.800
0.85	0.779	10.838	13.055	13.168	0.716	0.86%	14.238	14.538
0.90	1.575	18.225	20.546	20.668	0.807	0.59%	21.825	22.025
0.95	4.037	40.613	43.033	43.168	0.902	0.31%	44.413	44.513
0.98	11.515	108.045	110.479	110.667	0.960	0.17%	111.965	112.005
0.99	24.008	220.523	222.971	223.167	0.980	0.09%	224.483	224.503

Table EC.6 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 0.5$ and $c_s^2 = 4.0$

ρ	Tight LB	HTA (4)	Tight UB	UB Approx (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	0.025	0.072	0.072	0.000	0.00%	0.075	0.300
0.15	0.000	0.060	0.128	0.128	0.001	0.07%	0.135	0.347
0.20	0.000	0.113	0.200	0.201	0.007	0.30%	0.213	0.413
0.25	0.042	0.188	0.292	0.294	0.020	0.68%	0.313	0.500
0.30	0.107	0.289	0.409	0.414	0.041	1.08%	0.439	0.614
0.35	0.202	0.424	0.558	0.565	0.070	1.32%	0.599	0.762
0.40	0.333	0.600	0.746	0.757	0.107	1.47%	0.800	0.950
0.45	0.511	0.828	0.986	1.002	0.152	1.58%	1.053	1.191
0.50	0.750	1.125	1.289	1.314	0.203	1.91%	1.375	1.500
0.55	1.069	1.513	1.692	1.716	0.261	1.45%	1.788	1.900
0.60	1.500	2.025	2.212	2.244	0.324	1.40%	2.325	2.425
0.65	2.089	2.716	2.913	2.950	0.393	1.26%	3.041	3.129
0.70	2.917	3.675	3.875	3.923	0.467	1.23%	4.025	4.100
0.75	4.125	5.063	5.250	5.325	0.546	1.41%	5.438	5.500
0.80	6.000	7.200	7.422	7.477	0.629	0.74%	7.600	7.650
0.85	9.208	10.838	11.075	11.129	0.716	0.48%	11.263	11.300
0.90	15.750	18.225	18.470	18.530	0.807	0.32%	18.675	18.700
0.95	35.625	40.613	40.871	40.932	0.902	0.15%	41.088	41.100
0.98	95.550	108.045	108.307	108.373	0.960	0.06%	108.535	108.540
0.99	195.525	220.523	220.783	220.853	0.980	0.03%	221.018	221.020

Table EC.7 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 0.5$ and $c_s^2 = 0.5$

ρ	Tight LB	HTA (4)	Tight UB	UB Approx (51)	δ	MRE	Daley (6)	Kingman (5)
0.10	0.000	0.006	0.053	0.053	0.000	0.00%	0.056	0.281
0.15	0.000	0.013	0.082	0.082	0.001	0.11%	0.088	0.301
0.20	0.000	0.025	0.113	0.113	0.007	0.54%	0.125	0.325
0.25	0.000	0.042	0.146	0.148	0.020	1.35%	0.167	0.354
0.30	0.000	0.064	0.184	0.189	0.041	2.36%	0.214	0.389
0.35	0.000	0.094	0.228	0.235	0.070	3.16%	0.269	0.432
0.40	0.000	0.133	0.280	0.291	0.107	3.82%	0.333	0.483
0.45	0.000	0.184	0.342	0.357	0.152	4.43%	0.409	0.547
0.50	0.000	0.250	0.414	0.439	0.203	5.72%	0.500	0.625
0.55	0.000	0.336	0.515	0.540	0.261	4.62%	0.611	0.724
0.60	0.000	0.450	0.637	0.669	0.324	4.71%	0.750	0.850
0.65	0.000	0.604	0.800	0.837	0.393	4.45%	0.929	1.016
0.70	0.058	0.817	1.017	1.065	0.467	4.53%	1.167	1.242
0.75	0.188	1.125	1.312	1.388	0.546	5.42%	1.500	1.563
0.80	0.400	1.600	1.822	1.877	0.629	2.95%	2.000	2.050
0.85	0.779	2.408	2.646	2.700	0.716	1.99%	2.833	2.871
0.90	1.575	4.050	4.295	4.355	0.807	1.38%	4.500	4.525
0.95	4.037	9.025	9.284	9.344	0.902	0.65%	9.500	9.512
0.98	11.515	24.010	24.271	24.338	0.960	0.27%	24.500	24.505
0.99	24.008	49.005	49.265	49.336	0.980	0.14%	49.500	49.503

EC.8. The UB Transient Mean from the Optimization and Numerical Search

To illustrate our results, we report results from a further experiment in which we performed a numerical search over the candidate two-point service-time distributions $G_{u,n}$ for the mean waiting time $E[W_n(F_0, G_{u,n})]$ as a function of n using the multinomial exact representation in §5.2 for a class of models ($\rho = \{0.1, ..., 0.9\}$, $c_a^2 = \{1/2, 4\}, c_s^2 = \{1/2, 4\}, M_a = M_s = 10$), and $n = 1, 5, ..., 50$. For all these cases, we first found by the optimization that the local optimum was obtained at $(F_0, G_{u,n})$. We then conducted the search to carefully identify the optimal values among these candidate $G_{u,n}$. (See the next section for details.). Table EC.8 presents numerical results for the case $c_a^2 = c_s^2 = 4.0$ for a range of n and ρ. Tables EC.9-EC.11 present results for the other three cases $(c_a^2, c_s^2) = (4.0, 0.5)$, $(0.5, 4.0)$ and $(0.5, 0.5)$).

Table EC.8 Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization and numerical search for $c_a^2 = c_s^2 = 4.0$

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.080	0.160	0.240	0.320	0.400	0.489	0.579	0.668	0.758
5	0.269	0.538	0.813	1.095	1.414	1.777	2.140	2.505	2.882
10	0.357	0.716	1.102	1.525	2.056	2.634	3.228	3.869	4.555
15	0.386	0.778	1.220	1.744	2.410	3.137	3.949	4.832	5.776
20	0.395	0.804	1.281	1.871	2.626	3.508	4.499	5.602	6.808
25	0.399	0.814	1.313	1.948	2.781	3.782	4.933	6.242	7.693
30	0.400	0.820	1.332	1.999	2.896	3.992	5.291	6.794	8.508
35	0.400	0.822	1.343	2.032	2.979	4.163	5.590	7.270	9.185
40	0.400	0.824	1.349	2.056	3.040	4.299	5.846	7.696	9.858
45	0.400	0.824	1.354	2.072	3.088	4.411	6.067	8.075	10.423
50	0.400	0.825	1.356	2.084	3.126	4.505	6.260	8.421	11.002

Of course, we witness the well known property that $E[W_n]$ is increasing in n, c_a^2 and c_s^2. We also see that $E[W_n]$ tends to be slightly smaller for the pair $(0.5, 4.0)$ than for the pair $(4.0, 0.5)$, but

these are similar, as suggested by the HT limit. In support of the corresponding result for $E[W]$, we see convergence well before the final $n = 50$ for the lower traffic intensities.

We also report optimization results for $E[W_n]$ from (50) for the special cases of the $GI/D/1$ and $D/GI/1$ models with $(c_a^2 = 4.0, M_a = 100)$ and $(c_s^2 = 4.0, M_s = 100)$, respectively, in Tables EC.12 and EC.13. For the $GI/D/1$ model, the optimization terminates with the same extremal two-point cdf F_0. For the $D/GI/1$ model, as in Tables 1-2, we perform an additional search to identify the optimal distribution $G_{u,n}$ for each n.

We now supplement Table EC.8 with corresponding numerical values of $E[W(F_0, G_{u,n})]$ obtained from the SQP optimization followed by a detailed numerical search to find the best possible two-point service cdf $G_{u,n}$. Tables EC.9-EC.11 present corresponding results for the cases $(c_a^2, c_s^2) = (4.0, 0.5)$, $(0.5, 4.0)$ and $(0.5, 0.5)$.

Table EC.9 Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 4.0$ and $c_s^2 = 0.5$

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.080	0.160	0.240	0.320	0.400	0.481	0.563	0.644	0.725
5	0.269	0.538	0.807	1.078	1.356	1.638	1.920	2.207	2.499
10	0.357	0.714	1.073	1.447	1.831	2.241	2.702	3.203	3.740
15	0.386	0.772	1.167	1.590	2.074	2.621	3.225	3.902	4.660
20	0.395	0.792	1.206	1.679	2.228	2.860	3.603	4.449	5.411
25	0.399	0.799	1.230	1.730	2.324	3.039	3.888	4.893	6.053
30	0.400	0.803	1.242	1.759	2.393	3.169	4.118	5.262	6.615
35	0.400	0.805	1.248	1.779	2.439	3.268	4.306	5.579	7.114
40	0.400	0.805	1.252	1.791	2.474	3.347	4.460	5.857	7.567
45	0.400	0.806	1.254	1.800	2.498	3.408	4.591	6.102	7.982
50	0.400	0.806	1.256	1.806	2.517	3.458	4.702	6.319	8.364

Table EC.10 Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 0.5$ and $c_s^2 = 4.0$

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.033	0.082	0.147	0.220	0.305	0.400	0.500	0.600	0.700
5	0.051	0.147	0.303	0.515	0.780	1.097	1.465	1.874	2.301
10	0.051	0.151	0.331	0.607	0.982	1.458	2.043	2.723	3.477
15	0.051	0.152	0.335	0.636	1.075	1.654	2.400	3.301	4.338
20	0.051	0.152	0.337	0.647	1.122	1.779	2.648	3.744	5.033
25	0.051	0.152	0.337	0.652	1.148	1.864	2.836	4.097	5.624
30	0.051	0.152	0.337	0.653	1.163	1.923	2.981	4.392	6.141
35	0.051	0.152	0.337	0.654	1.172	1.965	3.096	4.642	6.600
40	0.051	0.152	0.337	0.655	1.177	1.995	3.190	4.857	7.015
45	0.051	0.152	0.337	0.655	1.181	2.018	3.268	5.046	7.395
50	0.051	0.152	0.337	0.655	1.183	2.034	3.333	5.214	7.744

Table EC.11 Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 0.5$ and $c_s^2 = 0.5$

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.033	0.069	0.106	0.145	0.187	0.230	0.274	0.317	0.361
5	0.050	0.106	0.171	0.248	0.347	0.472	0.626	0.802	1.008
10	0.050	0.107	0.176	0.265	0.386	0.557	0.793	1.096	1.483
15	0.050	0.107	0.176	0.268	0.398	0.590	0.872	1.271	1.813
20	0.050	0.107	0.176	0.268	0.402	0.606	0.917	1.388	2.067
25	0.050	0.107	0.176	0.268	0.404	0.615	0.943	1.471	2.273
30	0.050	0.107	0.176	0.268	0.404	0.619	0.961	1.533	2.446
35	0.050	0.107	0.176	0.268	0.405	0.622	0.973	1.580	2.593
40	0.050	0.107	0.176	0.268	0.405	0.623	0.982	1.616	2.722
45	0.050	0.107	0.176	0.268	0.405	0.624	0.988	1.645	2.834
50	0.050	0.107	0.176	0.268	0.405	0.624	0.993	1.668	2.935

Of course, we witness the well known property that $E[W_n]$ is increasing in n, c_a^2 and c_s^2. We also see that $E[W_n]$ tends to be slightly smaller for the pair $(0.5, 4.0)$ than for the pair $(4.0, 0.5)$, but these are similar, as suggested by the HT limit. In support of the corresponding result for $E[W]$, we see convergence well before the final $n = 50$ for the lower traffic intensities.

We also report optimization results for $E[W_n]$ from (50) for the special cases of the $GI/D/1$ and $D/GI/1$ models with $(c_a^2 = 4.0, M_a = 100)$ and $(c_s^2 = 4.0, M_s = 100)$, respectively, in Tables EC.12 and EC.13. For the $GI/D/1$ model, the optimization terminates with the same extremal two-point

cdf F_0. For the $D/GI/1$ model, as in Tables 1-2, we perform an additional search to identify the optimal $b_s^*(n)$ for each n. To sum up, these tables support Conjecture 1.

Table EC.12 Numerical values of $E[W_n]$ in the extremal $GI/D/1$ model with $M_a = 100$, $c_a^2 = 4.0$ and $c_s^2 = 0.0$

$\rho \backslash n$	10	15	20	25	30	35	40	45	50
0.10	0.357	0.386	0.395	0.398	0.400	0.400	0.400	0.400	0.400
0.15	0.536	0.579	0.593	0.598	0.599	0.600	0.600	0.600	0.600
0.20	0.714	0.772	0.791	0.797	0.800	0.802	0.803	0.804	0.804
0.25	0.893	0.965	0.988	1.001	1.009	1.012	1.013	1.014	1.015
0.30	1.071	1.158	1.194	1.217	1.228	1.234	1.237	1.239	1.240
0.35	1.250	1.353	1.413	1.447	1.463	1.474	1.480	1.484	1.486
0.40	1.428	1.562	1.648	1.691	1.719	1.737	1.748	1.756	1.760
0.45	1.607	1.785	1.896	1.958	2.002	2.028	2.047	2.060	2.069
0.50	1.785	2.022	2.159	2.251	2.310	2.353	2.383	2.405	2.421
0.55	1.977	2.274	2.447	2.572	2.656	2.720	2.765	2.800	2.827
0.60	2.183	2.539	2.762	2.922	3.042	3.129	3.200	3.253	3.296
0.65	2.398	2.814	3.100	3.305	3.466	3.590	3.689	3.770	3.836
0.70	2.622	3.106	3.461	3.724	3.931	4.102	4.242	4.358	4.456
0.75	2.859	3.423	3.847	4.182	4.451	4.674	4.865	5.029	5.171
0.80	3.101	3.757	4.262	4.673	5.017	5.309	5.562	5.784	5.982
0.85	3.350	4.108	4.707	5.205	5.631	6.005	6.336	6.632	6.900
0.90	3.611	4.481	5.186	5.784	6.306	6.773	7.194	7.579	7.933

Table EC.13 Numerical values of $E[W_n]$ in the extremal $D/GI/1$ model with $M_s = 10$, $c_a^2 = 0.0$ and $c_s^2 = 4.0$

$\rho \backslash n$	10	15	20	25	30	35	40	45	50
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.15	0.012	0.025	0.025	0.012	0.012	0.012	0.012	0.012	0.025
0.2	0.048	0.058	0.058	0.048	0.048	0.048	0.048	0.048	0.058
0.25	0.091	0.115	0.115	0.091	0.091	0.091	0.091	0.091	0.115
0.3	0.174	0.195	0.195	0.174	0.174	0.174	0.174	0.174	0.195
0.35	0.272	0.300	0.301	0.274	0.274	0.274	0.274	0.274	0.301
0.4	0.407	0.441	0.445	0.418	0.419	0.419	0.419	0.419	0.447
0.45	0.568	0.620	0.631	0.601	0.602	0.603	0.603	0.603	0.640
0.5	0.764	0.833	0.862	0.844	0.848	0.851	0.852	0.853	0.892
0.55	0.985	1.086	1.142	1.139	1.154	1.162	1.168	1.171	1.219
0.6	1.241	1.382	1.472	1.514	1.547	1.569	1.585	1.595	1.642
0.65	1.520	1.728	1.860	1.951	2.017	2.064	2.099	2.125	2.176
0.7	1.837	2.121	2.319	2.462	2.574	2.659	2.728	2.783	2.840
0.75	2.183	2.563	2.843	3.035	3.223	3.362	3.477	3.575	3.658
0.8	2.536	3.038	3.422	3.673	3.978	4.186	4.365	4.520	4.657
0.85	2.924	3.568	4.068	4.371	4.826	5.128	5.394	5.632	5.844
0.9	3.317	4.110	4.747	5.120	5.755	6.171	6.545	6.886	7.200

EC.9. Additional Counterexamples When One Distribution is Given

In this section we report additional experiments to provide more counterexamples when one distribution is given. Recall that strong evidence has already been given in Tables 4 and 5. For the steady-state mean $E[W]$, we use simulation method in Minh and Sorli (1983) with simulation length T^* =1E+06 and 20 i.i.d. replications to compute $E[W]$ for the case $\rho = 0.5$, $c_a^2 = 4$, and $c_s^2 = 4$ with $b_a \in [1 + c_a^2, M_a]$ (LHS of the following Figure EC.1). For the RHS of Figure EC.1, we use Monte Carlo simulation method with N =5E+07 and report average results based on 20 identical independent replications for studying the effects of b_s on $E[W]$ for different cases of b_a. It is already known that when $b_a = (1 + c_a^2)$, the $E[W]$ is increasing with b_s.

Figure EC.1 shows simulation estimates of the steady-state mean $E[W]$ as a function of b_a in $[(1+c_a^2), M_a = 7]$ for $b_s = 5$, i.e., for G_0 (left) and as a function of b_s in $[(1+c_s^2), M_s = 20]$ for various b_a (right). The optimal values of b_s as a function of b_a, denoted by $b_s^*(b_a)$, are: $b_s^*(10) = 5.0, b_s^*(15) = 8, b_s^*(20) = 11, b_s^*(25) = 18, b_s^*(30) = 20$.

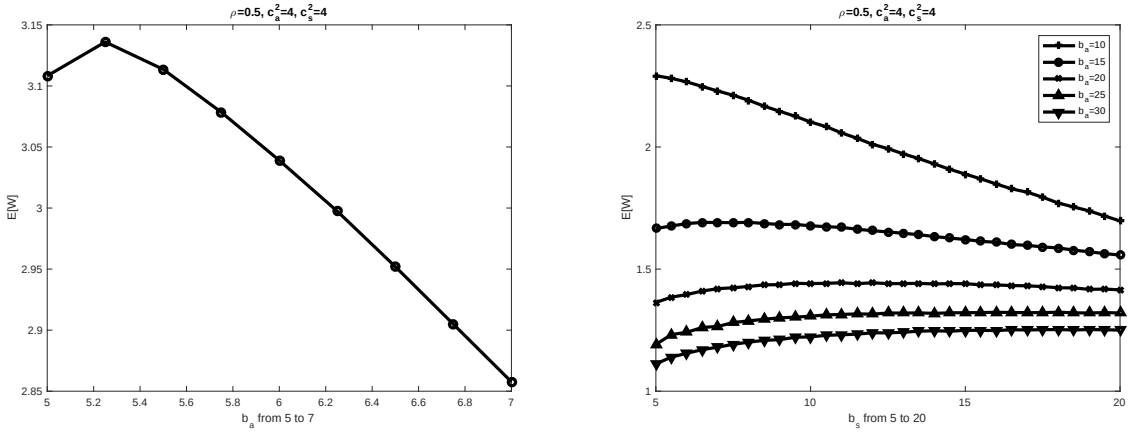

Figure EC.1 Simulation estimates of the steady-state mean $E[W]$ as a function of b_a in $[(1 + c_a^2), M_a = 7]$ for $b_s = 5$, i.e., for G_0 (left) and as a function of b_s in $[(1 + c_s^2), M_s = 20]$ for various b_a (right).

The plot on the left in Figure EC.1 dramatically shows the counterexample from Wolff and Wang (2003)); it shows that the maximum is not attained at F_0 when the service-time cdf is G_0. The plot on the right shows the more complex behavior that is possible for b_s (the service-time cdf G) as a

function of b_a (the interarrival-time cdf F). When $b_a = 5$ (F_0), we see that the mean is increasing in b_s, but when $b_a > 5$, we see more complicated behavior. For the three cases $b_a = 15, 20, 25$, there exists $b_s^*(b_a) \in (1 + c_s^2, M_s)$ such that the extremal service-time cdf is neither associated with b_s on the left (G_0) nor with b_s on the right (G_u).

EC.10. When One Distribution is Deterministic

We have already looked at the $GI/D/1$ and $D/GI/1$ models in Tables EC.12 and EC.13. They showed the transient mean waiting times $E[W_n]$ as a function of n and ρ resulting from the optimization in §5. For all those cases, the transient mean was maximized at $(F_0, G_{u,n})$. We now consider the steady-state mean $E[W]$.

For $D/GI/1$ and $GI/D/1$, we implement the same simulation search for different cases of b_a, b_s throughout traffic level from $\rho = 0.1$ to $\rho = 0.9$. We use Monte Carlo simulation method with $N = 1E+07$ and report average of 20 identical independent replications. Tables EC.14 and EC.15 present results that are consistent with optimization results for transient mean waiting time that the upper bounds of $D/GI/1$ and $GI/D/1$ of steady-state mean and transient mean are attained by G_u and F_0.

Table EC.14 Simulation search for $GI/D/1$ over b_a with mean 1 arrival

$b_a \backslash \rho$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5.0	0.400	0.804	1.242	1.770	2.469	3.496	5.171	8.50	18.41
5.5	0.000	0.450	0.964	1.536	2.262	3.307	5.006	8.34	18.30
6.0	0.000	0.000	0.626	1.271	2.040	3.102	4.812	8.19	18.26
6.5	0.000	0.000	0.206	0.965	1.795	2.896	4.627	8.02	18.01
7.0	0.000	0.000	0.000	0.600	1.526	2.674	4.436	7.83	17.95
7.5	0.000	0.000	0.000	0.163	1.224	2.436	4.232	7.65	17.71
8.0	0.000	0.000	0.000	0.000	0.875	2.182	4.017	7.46	17.50
8.5	0.000	0.000	0.000	0.000	0.468	1.909	3.802	7.26	17.49
9.0	0.000	0.000	0.000	0.000	0.000	1.612	3.573	7.09	17.19
9.5	0.000	0.000	0.000	0.000	0.000	1.277	3.337	6.88	17.05
10.0	0.000	0.000	0.000	0.000	0.000	0.899	3.084	6.68	16.83

Table EC.15 Simulation search for $D/GI/1$ over b_s with mean 1 arrival

$b_s \backslash \rho$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.000	0.058	0.195	0.447	0.893	1.670	3.114	6.23	16.00
11	0.004	0.064	0.200	0.457	0.903	1.682	3.129	6.24	16.02
12	0.007	0.067	0.205	0.462	0.911	1.691	3.141	6.26	16.04
13	0.008	0.068	0.210	0.469	0.918	1.702	3.151	6.27	16.05
14	0.009	0.070	0.211	0.474	0.924	1.709	3.160	6.28	16.06
15	0.010	0.073	0.216	0.476	0.929	1.714	3.167	6.29	16.07
16	0.011	0.075	0.218	0.481	0.934	1.721	3.174	6.29	16.08
17	0.011	0.076	0.221	0.484	0.938	1.726	3.179	6.30	16.09
18	0.011	0.077	0.223	0.487	0.941	1.730	3.184	6.31	16.10
19	0.011	0.079	0.224	0.490	0.945	1.734	3.189	6.31	16.10
20	0.012	0.080	0.227	0.492	0.948	1.737	3.193	6.32	16.11

To sum up, for the transient mean waiting time $E[W_n]$, the numerical experiments show that there exists $b_a^* = (1 + c_a^2)$ and $b_s^*(n)$ such that the $\sup\{E[W_n(F,G) : F, G \in \mathcal{P}_{a,2,2} \times \mathcal{P}_{s,2,2}]\}$ is attained. We find that $b_s^*(n)$ is not strictly increasing, but that there exists an n_0 after which it is increasing In all cases, we find that $G_{u,n} \Rightarrow G_u$ as $n \to \infty$. For the steady-state mean waiting time $E[W]$, the UB is attained when b_a^* is $(1 + c_a^2)$ and $b_s^* = M_s$. Hence, the UB for the steady-state mean waiting time is attained at (F_0, G_u).