The impact of a heavy-tailed service-time distribution upon the $\mathrm{M}/\mathrm{GI}/s$ waiting-time distribution

Ward Whitt

AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA
E-mail: wow@research.att.com

Received 5 January 1999; revised 6 January 2000

By exploiting an infinite-server-model lower bound, we show that the tails of the steady-state and transient waiting-time distributions in the $\mathrm{M}/\mathrm{GI}/s$ queue with unlimited waiting room and the first-come first-served discipline are bounded below by tails of Poisson distributions. As a consequence, the tail of the steady-state waiting-time distribution is bounded below by a constant times the sth power of the tail of the service-time stationary-excess distribution. We apply that bound to show that the steady-state waiting-time distribution has a heavy tail (with appropriate definition) whenever the service-time distribution does. We also establish additional results that enable us to nearly capture the full asymptotics in both light and heavy traffic. The difference between the asymptotic behavior in these two regions shows that the actual asymptotic form must be quite complicated.

Keywords: heavy-tailed distributions, subexponential distributions, tail probability asymptotics, multi-server queues, $\mathrm{M}/\mathrm{GI}/s$ queue, waiting times, existence of finite moments

1. Introduction

There recently has been great interest in the performance of queues with heavy-tailed (or long-tailed) service-time distributions. Asymptotic results for the tail of the steady-state waiting-time distribution in the single-server queue with unlimited waiting room and the first-come first-served (FCFS) service discipline were first obtained by Borovkov [9, section 22], Cohen [13] and Pakes [26]. For recent extensions, see [6,22 and references therein].

The purpose of this paper is to obtain some corresponding partial results for the challenging multi-server queue, also with unlimited waiting room and the FCFS discipline. In particular, we derive a lower bound on the waiting-time tail probability in the M/GI/s model that allows us to conclude that the waiting-time distribution has a heavy-tailed distribution (with appropriate definition) whenever the service-time distribution does. This conclusion deduced from a lower bound parallels the conclusion deduced for general single-server fluid models in [12]. The bounding arguments have the appeal of quite simply determining the main qualitative behavior, even though they do not determine the full asymptotic behavior.

We also establish a second, very different, lower bound for the waiting-time tail probability in a heavy-traffic regime. Since the second lower bound leads to a much larger asymptote, it is natural to suspect that there must be a large gap between the first lower-bound asymptote and the true asymptote. However, we show that is not always the case, by showing that the first lower bound is asymptotically correct in light traffic. We also establish an upper bound that shows that the second lower bound is at least nearly correct in heavy traffic. Together, these results demonstrate that the actual asymptotic behavior must be quite complicated. We make a conjecture about the actual asymptotic behavior.

We state and discuss the main results in section 2 and provide proofs and additional details in the following sections. We conclude with a discussion of the transient waiting-time distribution. It turns out that an infinite-server lower bound also applies to the transient waiting-time distribution. That is important because steady state is approached very slowly with heavy-tailed service-time distributions. Indeed, steady state is approached so slowly that it may be preferable to use transient distributions in applications. The infinite-server lower bound is useful because it provides both a simple quantitative description of the rate of convergence to steady state and an approximation for the transient waiting-time distribution. Since these descriptions are for the infinite-server lower bound, they do not yield exact descriptions for $\mathrm{M}/\mathrm{G}/s$ models, but they can serve as useful approximations when s is not too small.

We conclude this introduction by mentioning papers by Boxma et al. [10] and Korshunov [24] that came to our attention (and were evidently done) after this paper was completed. They contain asymptotic results that tend to support the conjectures here.

2. Main results

To specify the model, let G be the service-time cumulative distribution function (cdf) and let $G^{\mathrm{c}}(t) \equiv 1 - G(t)$ be the associated complementary cdf (ccdf). We assume that G has finite mean m_1. Let G_{e} be the stationary-excess (or equilibrium residual-lifetime) cdf associated with G, defined by

$$G_{\mathrm{e}}(t) = m_1^{-1} \int_0^t G^{\mathrm{c}}(u)\,\mathrm{d}u, \quad t > 0, \tag{2.1}$$

and let $G_{\mathrm{e}}^{\mathrm{c}}$ be its ccdf. Let the interarrival time have finite mean λ^{-1} and let the traffic intensity be $\rho \equiv \lambda m_1/s$. To ensure model stability, we assume that $\rho < 1$. Let W be the steady-state waiting time until beginning service and let W^{c} be its ccdf; i.e., $W^{\mathrm{c}}(t) \equiv P(W > t)$, $t \geqslant 0$.

The lower bound for $W^{\mathrm{c}}(t)$ here is obtained by approximating the $\mathrm{M}/\mathrm{GI}/s$ model by the associated infinite-server $\mathrm{M}/\mathrm{GI}/\infty$ model. Such infinite-server models have proven very useful for approximating $\mathrm{M}_t/\mathrm{GI}/s/r$ models with time-dependent Poisson arrival processes; see section 8 below and [15,25 and references therein]. It is well known that the steady-state number of customers in the $\mathrm{M/GI}/s$ model is bounded below by the steady-state number of customers in the associated $\mathrm{M/GI}/\infty$ model. Of course there is no waiting before beginning service in the $\mathrm{M/GI}/\infty$ model, but it is easy to see that the steady-state waiting time ccdf $W^{\mathrm{c}}(t)$ in the $\mathrm{M/G}/s$ model is bounded below by the ccdf $F^{\mathrm{c}}(t)$ of the first passage time, starting in steady state, to a level with $s-1$ or fewer busy servers in the $\mathrm{M/G}/\infty$ model, assuming that all future arrivals are neglected. Somewhat surprisingly, perhaps, this first-passage-time ccdf F^{c} has a relatively simple form, being the ccdf of a Poisson distribution with mean $\rho s G^{\mathrm{c}}_{\mathrm{e}}(t)$, where $G^{\mathrm{c}}_{\mathrm{e}}$ is the ccdf associated with the service-time stationary-excess cdf G_{e} in (2.1).

To state the result, let $\Pi^{\mathrm{c}}(k;\lambda)$ be the ccdf of a Poisson distribution with mean λ; i.e., if X has the Poisson distribution, then $\Pi^{\mathrm{c}}(k;\lambda) \equiv P(X > k)$. We say that f is asymptotically equivalent to g as $t \to \infty$, and write $f(t) \sim g(t)$, if $f(t)/g(t) \to 1$ as $t \to \infty$.

Theorem 1. In the $\mathrm{M/GI}/s$ model, for all ρ, $0 < \rho < 1$, and $t > 0$,

$$W^{\mathrm{c}}(t) \geqslant F^{\mathrm{c}}(t) \equiv \Pi^{\mathrm{c}}\bigl(s-1; \rho s G^{\mathrm{c}}_{\mathrm{e}}(t)\bigr) \equiv \mathrm{e}^{-\rho s G^{\mathrm{c}}_{\mathrm{e}}(t)} \sum_{m=s}^{\infty} \frac{(\rho s G^{\mathrm{c}}_{\mathrm{e}}(t))^m}{m!}, \tag{2.2}$$

where

$$F^{\mathrm{c}}(t) \sim \frac{(\rho s)^s}{s!} G^{\mathrm{c}}_{\mathrm{e}}(t)^s \qquad \text{as } t \to \infty, \tag{2.3}$$

$$F^{\mathrm{c}}(t) \sim \frac{\mathrm{e}^{\rho s G^{\mathrm{e}}_{\mathrm{c}}(t)} (\mathrm{e}\rho G^{\mathrm{c}}_{\mathrm{e}}(t))^s}{(1 - \rho G^{\mathrm{c}}_{\mathrm{e}}(t))\sqrt{2\pi s}} \quad \text{as } s \to \infty, \tag{2.4}$$

and

$$F^{\mathrm{c}}(t) \sim \frac{(\mathrm{e}\rho G^{\mathrm{c}}_{\mathrm{e}}(t))^s}{\sqrt{2\pi s}} \quad \text{as } s \to \infty \text{ and then } t \to \infty. \tag{2.5}$$

Remark. Note that we can also establish the same lower bound asymptote in (2.3) by keeping only the term corresponding to $m = s$ in (2.2). Then

$$W^{\mathrm{c}}(t) \geqslant F^{\mathrm{c}}(t) \geqslant H^{\mathrm{c}}_1(t) \equiv \mathrm{e}^{-\rho s G^{\mathrm{c}}_{\mathrm{e}}(t)} \frac{(\rho s G^{\mathrm{c}}_{\mathrm{e}}(t))^s}{s!} \sim \frac{(\rho s)^s}{s!} G^{\mathrm{c}}_{\mathrm{e}}(t)^s \tag{2.6}$$

as $t \to \infty$. Also note that

$$W^{\mathrm{c}}(t) \geqslant H^{\mathrm{c}}_1(t) \geqslant A G^{\mathrm{c}}_{\mathrm{e}}(t)^s \quad \text{for } A = \frac{\mathrm{e}^{-\rho s}(\rho s)^s}{s!}. \tag{2.7}$$

Our primary goal is to apply this lower bound F^{c} in theorem 1 to deduce that in the $\mathrm{M/G}/s$ model the waiting-time ccdf W^{c} inherits a heavy-tail property from the service-time ccdf G^{c}, with appropriate definitions. However, we also point out that the lower bound may serve as a useful reference point and sometimes even a useful approximation, because it is remarkably tractable, because it applies for all t

and because the $\mathrm{M}/\mathrm{G}/s$ model with a heavy-tailed service-time distribution is very difficult to analyze. A heavy-tailed service-time distribution also makes it difficult to accurately estimate the waiting-time distribution with simulation; see [4].

One approach to carrying out numerical calculations with a heavy-tailed service-time distribution is to first approximate the heavy-tailed distribution by a hyperexponential distribution as in [17], or a more general phase-type distribution as in [5], and then solve the resulting $\mathrm{M}/\mathrm{H}_k/s$ or $\mathrm{M}/\mathrm{PH}/s$ queue, using algorithms such as in [7,28,29]. Of course, those approximations fail to capture the asymptotics as $t \to \infty$, but the approximations can be good for arbitrarily large t. For the $\mathrm{GI}/\mathrm{GI}/1$ queue, the calculations can be done directly as in [1].

Since F^{c} in (2.2) has a Poisson distribution, it can be computed. The normal approximation

$$\Pi^{\mathrm{c}}\bigl(s-1, \rho s G_{\mathrm{e}}^{\mathrm{c}}(t)\bigr) \approx \Phi^{\mathrm{c}}\left(\frac{s-0.5-\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)}{\sqrt{\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)}}\right) \tag{2.8}$$

should be a good approximation for $F^{\mathrm{c}}(t)$ if the argument

$$\alpha(s,t) \equiv \frac{s-0.5-\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)}{\sqrt{\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)}} \tag{2.9}$$

is not too large.

It is significant that both the lower-bound ccdfs F^{c} in (2.2) and H_1^{c} in (2.6) are asymptotically exact in light traffic.

Theorem 2. In the setting of theorem 1,

$$\lim_{\lambda \to 0} \frac{W^{\mathrm{c}}(t)}{H_1^{\mathrm{c}}(t)} = \lim_{\lambda \to 0} \frac{W^{\mathrm{c}}(t)}{F^{\mathrm{c}}(t)} = 1 \quad \text{for each } t > 0$$

for H_1^{c} in (2.6) and F^{c} in (2.3).

The basis for theorem 2 is a light-traffic result proved by Burman and Smith [11], which was reviewed Whitt in [35, section 2].

The tractable infinite-server lower bound can be a useful reference point, but it can grossly underestimate the true tail probabilities at higher traffic intensities, as we demonstrate by establishing a second lower bound for higher traffic intensities.

Theorem 3. In the $\mathrm{M}/\mathrm{GI}/s$ model, if $1 - s^{-1} < \rho < 1$, then

$$W^{\mathrm{c}}(t) \geqslant H_2^{\mathrm{c}}(t) \equiv p G_{\mathrm{e}}^{\mathrm{c}}(t) p_1(t) p_2(t), \tag{2.10}$$

where p, $p_1(t)$ and $p_2(t)$ are positive probabilities with $p_i(t) \to 1$ as $t \to \infty$ for $i = 1, 2$, so that

$$H_2^{\mathrm{c}}(t) \sim p G_{\mathrm{e}}^{\mathrm{c}}(t) \quad \text{as } t \to \infty. \tag{2.11}$$

Notice that (2.11) differs significantly from (2.3) and (2.7) when $s > 1$, because $G_e^c(t)$ appears on the right in (2.11) instead of $G_e^c(t)^s$.

We also apply an increasing-convex-order upper bound by Wolff [36] to show that this second lower-bound asymptote in theorem 3 is close to the true asymptote in this heavy-traffic regime. To state the result, we make a definition. We say that a cdf G is *subexponential* and belongs to class $\mathcal{S}$ if

$$(G * G)^c(t) \sim 2G^c(t) \quad \text{as } t \to \infty, \tag{2.12}$$

where $(G * G)^c$ is the ccdf of the two-fold convolution of G with itself. The subexponential distributions are a large subclass of the heavy-tailed distributions; see [19].

Theorem 4. In an $\mathrm{M/GI}/s$ model with $\rho < 1$ and service-time cdf G, if $G_e \in \mathcal{S}$, then

$$\liminf_{t\to\infty} \frac{P(W > t)}{G_e^c(t)} \leqslant \frac{\rho}{1-\rho}. \tag{2.13}$$

We now focus on the asymptotic behavior as $t \to \infty$. For the special case $s = 1$, the lower-bound asymptote in (2.3) and (2.6) underestimates the known limit in the heavy-tailed case (when $G_e^c \in \mathcal{S}$) only by the constant factor $(1-\rho)^{-1}$. Thus, for $s = 1$ the bound is sharp in the sense that is asymptotically correct in light traffic.

We are able to apply theorem 1 to deduce that the waiting-time ccdf is heavy tailed whenever the service-time ccdf is, with appropriate definitions. For that purpose, we introduce a new definition of a heavy-tailed distribution. Recall that a ccdf H^c is stochastically greater than or equal to another ccdf G^c if $H^c(t) \geqslant G^c(t)$ for all t. A ccdf G^c on $[0,\infty)$ is said to belong to the class $\mathcal{L}$ if $G^c(t) > 0$ for all $t > 0$ and

$$G^c(t-u) \sim G^c(t) \quad \text{as } t \to \infty \quad \text{for all } u > 0. \tag{2.14}$$

A ccdf H^c is *heavy tailed*, and we write $H^c \in \mathcal{H}$, if there is a ccdf $G^c \in \mathcal{L}$ such that

$$H^c(t) \geqslant G^c(t) \quad \text{for all } t.$$

It is known that $\mathcal{S} \subseteq \mathcal{L}$, so that $\mathcal{S} \subseteq \mathcal{H}$. An important subset of $\mathcal{S}$ is the set $\mathcal{P}$ of power-tail ccdfs; $G^c \in \mathcal{P}$ if

$$G^c(t) \sim \beta t^{-\alpha}, \quad \text{as } t \to \infty, \tag{2.15}$$

where α and β are positive constants.

The standard heavy-tail notions embody regularity properties for asymptotics as well as relatively large probabilities of large values. Hence, if G has a heavy tail by one of the standard definitions and if H is stochastically larger than G, i.e., if $H^c(t) \geqslant G^c(t)$ for all t, then it does *not* follow that H too has a heavy tail by the same definition. Hence, there is motivation for our introduction of the class $\mathcal{H}$.

Here is our main result.

Theorem 5. For an $\mathrm{M}/\mathrm{GI}/s$ model with $\rho < 1$ and service-time ccdf G^{c}, if $G^{\mathrm{c}} \in \mathcal{H}$, then $W^{\mathrm{c}} \in \mathcal{H}$.

It is intuitively clear that having more servers reduces the impact of a heavy-tailed service-time distribution, but that is not shown in theorem 5. We now obtain a result for a smaller class of heavy-tailed distributions that reveals this property. For that purpose, we say that a ccdf G^{c} on $[0, \infty)$ is regularly varying of index α, and write $G^{\mathrm{c}} \in \mathcal{R}(\alpha)$, if $G^{\mathrm{c}}(t) > 0$ for all $t > 0$ and

$$G^{\mathrm{c}}(ut) \sim u^{\alpha} G^{\mathrm{c}}(t) \quad \text{as } t \to \infty \tag{2.16}$$

for all $u > 0$; see [8]. Paralleling the class $\mathcal{H}$ defined above, we define a class $\mathcal{T}(-\alpha)$. A ccdf H^{c} is said to be dominated by a regularly varying tail of index-α, and we write $H^{\mathrm{c}} \in \mathcal{T}(-\alpha)$, if there exists a ccdf G^{c} in $\mathcal{R}(-\alpha)$ such that

$$H^{\mathrm{c}}(t) \geqslant G^{\mathrm{c}}(t) \quad \text{for all } t.$$

Theorem 6. For an $\mathrm{M}/\mathrm{GI}/s$ models with $\rho < 1$ and service-time ccdf G^{c}, if $G^{\mathrm{c}} \in \mathcal{T}(-\alpha)$ for $\alpha > 1$, then $W^{\mathrm{c}} \in \mathcal{T}(-s(\alpha - 1))$.

Unfortunately, we have been unable to determine the full asymptotic behavior of $W^{\mathrm{c}}(t)$ as $t \to \infty$, but theorems 1, 3 and 4 lead us to make a conjecture about the true asymptotic form.

Conjecture. In the stable $\mathrm{M}/\mathrm{GI}/s$ model with service-time cdf G, if $G_{\mathrm{e}} \in \mathcal{S}$, where $G_{\mathrm{e}}^{\mathrm{c}}$ is the service-time stationary-excess ccdf in (2.1), and if

$$s - k < \rho s < s - k + 1 \quad \text{for some } k,\ 1 \leqslant k \leqslant s, \tag{2.17}$$

then

$$W^{\mathrm{c}}(t) \sim \gamma G_{\mathrm{e}}^{\mathrm{c}}(\eta t)^{k} \quad \text{as } t \to \infty, \tag{2.18}$$

where γ and η are positive constants (as functions of t).

As indicated above, for $s = 1$, (2.18) is known to hold with $\gamma = \rho/(1-\rho)$, $\eta = 1$ and $k = 1$. For $s > 1$, conjecture (2.18) is consistent with theorems 1, 3 and 4. Here is some additional intuition behind (2.18) and (2.17) for $s > 1$: First, by (2.17), k is the minimum number of servers that can be removed (occupied by exceptionally long service times), so that the temporary traffic intensity without these servers exceeds 1, causing the workload to grow at positive rate. Second, the completed service times of the customers at k servers at an arbitrary time in steady state should be approximately distributed as k i.i.d. random variables with cdf G_{e}. (For example, that would be the case if the s servers were continuously busy; assuming that the cdf G is nonlattice.) Thus, the probability that there are k customers present who have all been in service for at least time ηt should be of order $\gamma G_{\mathrm{e}}^{\mathrm{c}}(\eta t)^{k}$ for constants γ and η. The scaling constant η in (2.18) is introduced because the workload grows at rate $\delta \equiv \rho s - (s-k)$ during this period of length ηt, so that the waiting time should be of order t for appropriate choice of the constant η.

If the conjecture is correct, then it seems evident that the asymptotic behavior will be quite complicated when the traffic intensity is at a boundary point, i.e., when $\rho s = s - k$ for some k, $1 \leqslant k \leqslant s-1$. When (2.17) does hold, we also make the stronger conjecture that (2.18) holds for the more general GI/GI/s model with a nonlattice interarrival-time distribution. (That is consistent with the known GI/GI/1 result.)

Here is how the rest of this paper is organized: We establish the two lower bounds in theorems 1 and 3 in sections 3 and 4. We then discuss heavy-tailed distributions and prove theorem 5 in section 5. We prove theorem 4 establishing the upper bound in section 6. We discuss the existence of finite moments in section 7. In particular, the lower-bound asymptote in (2.3) leads us to conjecture that, for $s > 1$, unlike for $s = 1$, the existence of the mean EW or other higher waiting-time moment EW^r depends upon more than the traffic intensity and the existence of service-time moments. Finally, we discuss a lower bound for the transient waiting-time distribution in section 8.

3. The infinite-server lower bound

In this section we prove theorem 1. We start by considering a general A/A/s model with unlimited waiting room, the FCFS discipline and arbitrary arrival and service processes. It is easy to see that, for each sample path of arrival times and service times, the remaining service times of all customers in the system at any time are bounded below by the remaining service times in the corresponding A/A/∞ system with the same arrival times and service times but infinitely many servers. Thus the waiting time for any arrival (or potential arrival) at any time t in the A/A/s system is bounded below by the first passage time to the state of $s-1$ or fewer busy servers in the A/A/∞ system starting at time t, with future arrivals shut off.

As a consequence, if steady-state distributions exist, then the steady-state waiting-time distribution in the A/A/s model is bounded below, in the sense of stochastic order, by the steady-state first passage time to level $s-1$ or below in the A/A/∞ model; i.e.,

$$W^{\mathrm{c}}(t) \geqslant F^{\mathrm{c}}(t) \quad \text{for all } t, \tag{3.1}$$

where F^{c} is the ccdf of the steady-state first passage time to $s-1$ or fewer busy servers in the A/A/∞ model, assuming that future arrivals are neglected.

We are able to usefully exploit the ordering (3.1) for the M/GI/s model, because there is a convenient exact characterization of the cdf F in the associated M/GI/∞ model. In particular, for the M/GI/∞ model, it is known that the steady-state number of busy servers has a Poisson distribution with mean ρs, independent of the service-time cdf G beyond its mean. Moreover, conditional on there being n busy servers in steady state, the n residual service times are distributed as n i.i.d. random variables with the service-time stationary-excess cdf G_{e}; see [31, p. 161]. (This $\mathrm{M/GI/\infty}$ property is also applied in [14,21].)

Moreover, it is elementary to see that the ccdf F^{c} in (3.1) is the tail of a Poisson distribution, because each of the Poisson random number of customers initially in the system at time 0 is still there t time units later with probability $G_{\mathrm{e}}^{\mathrm{c}}(t)$, independent of the other customers. Hence, we have an independent thinning of the original Poisson population, which is again Poisson. The associated analytic derivation is

$$\begin{aligned} F^{\mathrm{c}}(t) &= \sum_{n=s}^{\infty} \frac{\mathrm{e}^{-\rho s}(\rho s)^n}{n!} \sum_{k=s}^{n} \binom{n}{k} G_{\mathrm{e}}^{\mathrm{c}}(t)^k G_{\mathrm{e}}(t)^{n-k} \\ &= \frac{\mathrm{e}^{-\rho s}(\rho s)^s}{s!} \sum_{n=0}^{\infty} \frac{(\rho s)^n}{(s+n)!} \sum_{l=0}^{n} \binom{s+n}{s+l} G_{\mathrm{e}}^{\mathrm{c}}(t)^{s+l} G_{\mathrm{e}}(t)^{n-l} \\ &= \frac{\mathrm{e}^{-\rho s}(\rho s)^s}{s!} \sum_{l=0}^{\infty} \frac{(\rho s G_{\mathrm{e}}^{\mathrm{c}}(t))^l}{(s+l)!} \sum_{n=l}^{\infty} \frac{(\rho s G_{\mathrm{e}}(t))^{n-l}}{(n-l)!} \\ &= \mathrm{e}^{-s\rho G_{\mathrm{e}}^{\mathrm{c}}(t)} \sum_{n=s}^{\infty} \frac{(\rho s G_{\mathrm{e}}^{\mathrm{c}}(t))^n}{n!} = \Pi^{\mathrm{c}}\bigl(s-1; \rho s G_{\mathrm{e}}^{\mathrm{c}}(t)\bigr). \end{aligned} \tag{3.2}$$

It thus remains to establish the asymptotic relations for this special Poisson distribution, which of course is classic. First, recall that the ccdf of a Poisson distribution can always be expressed as the cdf of an Erlang distribution, using the familiar inverse relation between partial sums of nonnegative random variables and counting processes. Equivalently, we can express F^{c} in terms of special functions. In particular,

$$\begin{aligned} F^{\mathrm{c}}(t) &= \frac{\gamma(s, \rho s G_{\mathrm{e}}^{\mathrm{c}}(t))}{\Gamma(s)} = \mathrm{e}^{-\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)} \frac{(\rho s G_{\mathrm{e}}^{\mathrm{c}}(t))^s}{s!} M\bigl(1, s+1, \rho s G_{\mathrm{e}}^{\mathrm{c}}(t)\bigr) \\ &\sim \frac{(\rho s G_{\mathrm{e}}^{\mathrm{c}}(t))^s}{s!} \quad \text{as } t \to \infty, \end{aligned} \tag{3.3}$$

where γ is the incomplete gamma function, Γ is the gamma function and M is the confluent hypergeometric function, which satisfies $M(1, s+1, 0) = 1$; see [3, 6.5.2, 6.5.12, 13.1.2 and 13.5.5]. As noted in the remark, we can get the same asymptote as $t \to \infty$ by keeping only the terms with $G_{\mathrm{e}}^{\mathrm{c}}(t)^s$ in (3.2), i.e.,

$$\begin{aligned} F^{\mathrm{c}}(t) \geqslant H_1^{\mathrm{c}}(t) &\equiv \sum_{n=s}^{\infty} \frac{\mathrm{e}^{-\rho s}(\rho s)^n}{n!} \binom{n}{s} G_{\mathrm{e}}^{\mathrm{c}}(t)^s G_{\mathrm{e}}(t)^{n-s} \\ &= \frac{(\rho s)^s\, \mathrm{e}^{-\rho s} G_{\mathrm{e}}^{\mathrm{c}}(t)^s}{s!} \sum_{n=s}^{\infty} \frac{(\rho s)^{n-s}}{(n-s)!} G_{\mathrm{e}}(t)^{n-s} \\ &= \frac{(\rho s)^s}{s!} \mathrm{e}^{-\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)} G_{\mathrm{e}}^{\mathrm{c}}(t)^s \sim \frac{(\rho s)^s}{s!} G_{\mathrm{e}}^{\mathrm{c}}(t)^s \quad \text{as } t \to \infty. \end{aligned} \tag{3.4}$$

Moreover, from [3, 13.1.2], it follows that for any x

$$M(1, 1+s, \rho s x) \to 1 + \rho x + (\rho x)^2 + \cdots = (1-\rho x)^{-1}, \tag{3.5}$$

so that

$$F^{\mathrm{c}}(t) \sim \frac{\mathrm{e}^{-\rho s G_{\mathrm{e}}^{\mathrm{c}}(t)}}{(1-\rho G_{\mathrm{e}}^{\mathrm{c}}(t))} \frac{(\rho s G_{\mathrm{e}}^{\mathrm{c}}(t))^s}{s!} \quad \text{as } s \to \infty. \tag{3.6}$$

Applying Stirling's formula to (3.6), we obtain (2.4).

4. The minimal-stability lower bound

In this section we prove theorem 3, which holds under the condition that (2.17) holds for $k = 1$, i.e., $1 - s^{-1} < \rho < 1$. Following Scheller-Wolf and Sigman [27], we refer to this case as the minimal-stability case.

In order to consider the virtual waiting time of a potential arrival at time 0 in steady state, we look at the system at time $-Bt$, which is also in steady state. By the Poisson-Arrivals-See-Time-Averages (PASTA) property, see [37], the waiting time of an actual arrival has the same distribution.

We assume that the servers are numbered and that new arrivals and waiting customers are always assigned to the lowest-numbered free server. We then focus on server number 1. Server 1 experiences an alternating series of idle and busy periods, with the busy periods made up of a succession of service times. (Since we have a Poisson arrival process, the empty state is a regeneration point for the model; e.g., see [32].) If we look at this server at an arbitrary time in steady-state, which we have set as $-Bt$, with probability p, $0 < p < 1$, this server is busy and, conditional on this server being busy, the remaining service time exceeds Bt with probability $G_{\mathrm{e}}^{\mathrm{c}}(Bt)$, where G_{e} is the service-time stationary excess cdf. (See [20] for more on this property.)

Let $A(t, u)$ count the Poisson number of arrivals in the interval $(t, u]$ and let V_k denote the kth service time. By the Poisson, i.i.d. and finite-mean assumptions, we have the laws of large numbers (LLNs)

$$u^{-1} A(t, t+u) \to \lambda \quad \text{w.p. } 1 \quad \text{as } u \to \infty \tag{4.1}$$

and

$$n^{-1}(V_1 + \cdots + V_n) \to m_1 \equiv EV_1 \quad \text{w.p. } 1 \quad \text{as } n \to \infty. \tag{4.2}$$

Because the arrival process is assumed to be Poisson, the arrival process after time $-Bt$ is independent of the remaining service time observed for the customer being served by server 1 at time $-Bt$. By (4.1), for any $\varepsilon > 0$,

$$P\bigl(A(-Bt, 0) > \lambda B\bigl((1-\varepsilon)t\bigr)\bigr) \to 1 \quad \text{as } t \to \infty. \tag{4.3}$$

Given n arrivals in $(-Bt, 0]$, the waiting time of an arrival at time 0 is bounded below by the case in which all of these n arrivals occur together at time $-Bt$. Since server 1 is occupied during $(-Bt, 0]$, work is cleared by at most rate $s-1$ during $(-Bt, 0]$ and then afterwards at rate s. The waiting time of the potential customer at time 0 will exceed t if the sum of the $(n-s+1)$ smallest of the n service times exceed $(s-1)Bt+t$. (All but $s-1$ of the original customers in the system at time $-Bt$ must have been served.) This sum in turn is greater than or equal to the overall sum, say S_n, minus $(s-1)$ times the maximum M_n.

Combining these features, we obtain

$$\begin{aligned} P(W>t) \geqslant H_2^{\mathrm{c}}(t) \equiv pG_{\mathrm{e}}^{\mathrm{c}}(Bt)P\big(A(-Bt,0)>\lambda B(1-\varepsilon)t\big) \\ \times P\big(S_n-(s-1)M_n>(s-1)Bt+st\big), \end{aligned} \tag{4.4}$$

where $n=\lfloor \lambda B(1-\varepsilon)t\rfloor$.

However, $n^{-1}M_n \to 0$ as $n\to\infty$ by lemma 7 below and $n^{-1}S_n\to m_1$ as $n\to\infty$ by (4.2), so that, for any $\varepsilon>0$,

$$P\big(S_n-(s-1)M_n>m_1(1-\varepsilon)n\big)\to 1 \quad \text{as } n\to\infty. \tag{4.5}$$

Hence, in order to show that

$$H_2^{\mathrm{c}}(t)\sim pG_{\mathrm{e}}^{\mathrm{c}}(Bt) \quad \text{as } t\to\infty \tag{4.6}$$

for $H_2^{\mathrm{c}}(t)$ in (4.4), it suffices to choose B suitably large and ε suitably small so that

$$\lambda B(1-\varepsilon)^2 m_1 > (s-1)B+s \tag{4.7}$$

or, equivalently,

$$B\big(\rho s(1-\varepsilon)^2-(s-1)\big)>s. \tag{4.8}$$

However, by (2.17), we can achieve (4.8); in particular, by choosing ε suitably small, B can be any number larger than s/δ, where $\delta\equiv\rho s-(s-1)>0$.

One would expect that this argument also extends to cover the more general GI/GI/s, G/GI/s and G/G/s models, exploiting assumed LLNs as in (4.1) and (4.2), but there is some dependence, conditioning on the state of server 1 at time $-Bt$, that needs to be controlled.

We need the following to complete the proof above.

Lemma 7. If $EV_1=m_1<\infty$, then $n^{-1}M_n\Rightarrow 0$ as $n\to\infty$.

Proof. Note that

$$P(M_n\leqslant nx)=G(nx)^n=\left(1-\frac{nG^{\mathrm{c}}(nx)}{n}\right)^n.$$

However, since $EV_1<\infty$, $G^{\mathrm{c}}(x)$ is integrable, which implies that $nG^{\mathrm{c}}(nx)\to 0$ as $n\to\infty$. Since $(1-n^{-1}c_n)^n\to \mathrm{e}^{-c}$ as $n\to\infty$ if $c_n\to c$, $P(M_n\leqslant nx)\to \mathrm{e}^0=1$. □

5. Properties of heavy-tailed distributions

We now elaborate on the heavy-tail property and prove theorems 5 and 6. In addition to the classes of cdfs G^{c} on $[0,\infty)$ with $G^{\mathrm{c}}(t)>0$ for all t defined in section 2, we define two more. We say that $G^{\mathrm{c}}\in\mathcal{S}^*$ if G has finite mean m_1 and

$$\int_0^t G^{\mathrm{c}}(t-u)G^{\mathrm{c}}(u)\,\mathrm{d}u\sim 2m_1G^{\mathrm{c}}(t)\quad\text{as } t\to\infty. \tag{5.1}$$

We say that $G^{\mathrm{c}}\in\mathcal{C}_3$, i.e., class III according to [1], if for all $\varepsilon>0$

$$\int_0^\infty \mathrm{e}^{\varepsilon t}\,\mathrm{d}G(t)=\infty. \tag{5.2}$$

The following orderings are known or easy to establish.

Proposition 8. The classes are ordered by

$$\mathcal{S}^*\subseteq\mathcal{S}\subseteq\mathcal{L}\subseteq\mathcal{H}\subseteq\mathcal{C}_3\quad\text{and}\quad\mathcal{R}(-\alpha)\subseteq\mathcal{T}(-\alpha)\subseteq\mathcal{H}.$$

For $\alpha>1$, $\mathcal{R}(-\alpha)\subseteq\mathcal{S}^*$. For $\alpha>0$, $\mathcal{R}(-\alpha)\subseteq\mathcal{S}$.

We now consider how properties of G induce corresponding properties in its stationary-excess cdf G_{e}.

Proposition 9. Let G have finite mean.

(i) If $G\in\mathcal{R}(-\alpha)$ for $\alpha\geqslant 1$, then $G_{\mathrm{e}}\in\mathcal{R}(-(\alpha-1))$.

(ii) If $G\in\mathcal{S}^*$, then $G_{\mathrm{e}}\in\mathcal{S}$.

(iii) If $G\in\mathcal{L}$, then $G_{\mathrm{e}}\in\mathcal{L}$.

(iv) If $G\in\mathcal{H}$, then $G_{\mathrm{e}}\in\mathcal{H}$.

(v) If $G^{\mathrm{c}}\in\mathcal{T}(-\alpha)$ for $\alpha\geqslant 1$, then $G_{\mathrm{e}}^{\mathrm{c}}\in\mathcal{T}(-(\alpha-1))$.

(vi) If $G\in\mathcal{C}_3$, then $G_{\mathrm{e}}\in\mathcal{C}_3$.

Proof. (i) See [18, VIII.9]. The case $\alpha=1$ follows from the fact that if $L(t)$ is slowly varying and $L(t)/t$ is integrable, then $\int_x^\infty(L(u)/u)\,\mathrm{d}u$ is slowly varying.

(ii) See [23].

(iii) Write

$$G_{\mathrm{e}}^{\mathrm{c}}(t-u)=m_1^{-1}\int_t^\infty G^{\mathrm{c}}(v-u)\,\mathrm{d}v$$

and recall that integration preserves tail equivalence; see [16, p. 17].

(iv) Suppose that $G^{\mathrm{c}}(t)\geqslant H^{\mathrm{c}}(t)$ for all $t\geqslant 0$ where $H\in\mathcal{L}$. By (iii), $H_{\mathrm{e}}^{\mathrm{c}}\in\mathcal{L}$, but

$$G_{\mathrm{e}}^{\mathrm{c}}(t)=\frac{1}{m_1(G)}\int_t^\infty G^{\mathrm{c}}(u)\,\mathrm{d}u\geqslant\frac{1}{m_1(G)}\int_t^\infty H^{\mathrm{c}}(u)\,\mathrm{d}u\geqslant\frac{m_1(H)}{m_1(G)}H_{\mathrm{e}}^{\mathrm{c}}(t), \tag{5.3}$$

where $m_1(G) \geqslant m_1(H)$, so that the right side of (5.3) can itself be regarded as a lower bound ccdf in $\mathcal{L}$.

(v) Use part (i) and the proof of (iv).

(vi) See [2, lemma 3.2]. □

For our application to theorems 5 and 6, it is also important to consider how the properties apply to the ccdf $G_1^{\mathrm{c}}(t)G_2^{\mathrm{c}}(t)$ associated with the minimum of two independent random variables with ccdf's G_1^{c} and G_2^{c}. The following is immediate.

Proposition 10. Let G_1^{c} and G_2^{c} be two ccdf's on $[0, \infty)$ with $G_i^{\mathrm{c}}(t) > 0$ for all $t > 0$ and let $G^{\mathrm{c}}(t) = G_1^{\mathrm{c}}(t)G_2^{\mathrm{c}}(t)$.

(i) If $G_i \in \mathcal{R}(-\alpha_i)$ for $\alpha_i > 0$ and $i = 1, 2$, then $G \in \mathcal{R}(-(\alpha_1 + \alpha_2))$.

(ii) If $G_i^{\mathrm{c}} \in \mathcal{L}$ for $i = 1, 2$, then $G^{\mathrm{c}} \in \mathcal{L}$.

(iii) If $G_i^{\mathrm{c}} \in \mathcal{H}$ for $i = 1, 2$, then $G^{\mathrm{c}} \in \mathcal{H}$.

(iv) If $G_i^{\mathrm{c}} \in \mathcal{T}(-\alpha_i)$ for $\alpha_i \geqslant 0$ and $i = 1, 2$, then $G^{\mathrm{c}} \in \mathcal{T}(-(\alpha_1 + \alpha_2))$.

We are motivated to introduce the class $\mathcal{H}$ because the standard heavy-tail notions embody regularity properties for asymptotics as well as large tails.

Example. To illustrate, we give an example of a ccdf in $\mathcal{H}$ but not in $\mathcal{L}$. First, the ccdf

$$G^{\mathrm{c}}(t) = \mathrm{e}^{-\int_1^t z^{-1}\,\mathrm{d}z}, \quad t > 1, \tag{5.4}$$

is in $\mathcal{L}$. Let

$$\alpha(t) = 2^n, \quad 2^n \leqslant t < 2^{n+1}$$

for $n \geqslant 0$. Since $\alpha(t) \leqslant t$ for all $t > 0$, we have the ordering

$$H^{\mathrm{c}}(t) \equiv G^{\mathrm{c}}\bigl(\alpha(t)\bigr) \geqslant G^{\mathrm{c}}(t) \quad \text{for all } t > 0.$$

Hence, the ccdf H^{c} is in $\mathcal{H}$. However, H^{c} is not in $\mathcal{L}$, because

$$\begin{aligned} 1 = \liminf_{t\to\infty} \frac{H^{\mathrm{c}}(t-u)}{H^{\mathrm{c}}(t)} < \limsup_{t\to\infty} \frac{H^{\mathrm{c}}(t-u)}{H^{\mathrm{c}}(t)} &= \limsup_{n\to\infty} \mathrm{e}^{+\int_{2^n}^{2^{n+1}} z^{-1}\,\mathrm{d}z} \\ = \exp\bigl(\ln\bigl(2^{n+1}\bigr) - ln\bigl(2^n\bigr)\bigr) = 2. \end{aligned} \tag{5.5}$$

Proof of theorem 5. Now we prove theorem 5, using theorem 1. Suppose that $G^{\mathrm{c}} \in \mathcal{H}$. Then there exists $H^{\mathrm{c}} \in \mathcal{L}$ such that $G^{\mathrm{c}}(t) \geqslant H^{\mathrm{c}}(t)$ for all t. That ordering implies that a steady-state waiting time in the system with service-time cdf H, say $W(H)$, exists and, in addition, by making a sample path comparison, as in [34, theorem 4],

$$P\bigl(W(G) > t\bigr) \geqslant P\bigl(W(H) > t\bigr) \quad \text{for all } t. \tag{5.6}$$

Now, considering the $\mathrm{M}/\mathrm{GI}/s$ system with service-time cdf H, we first have $H_{\mathrm{e}} \in \mathcal{L}$ by proposition 9(iii). Then by (2.3) and proposition 10(ii), it follows that $F^{\mathrm{c}}(H) \in \mathcal{L}$ too, for F^{c} defined in (2.2). (If $G_1^{\mathrm{c}} \in \mathcal{L}$ and $G_1^{\mathrm{c}} \sim G_2^{\mathrm{c}}$, then $G_2^{\mathrm{c}} \in \mathcal{L}$.) Thus, by (2.2), $W^{\mathrm{c}}(H) \in \mathcal{H}$. Finally, by (5.6), $W^{\mathrm{c}}(G) \in \mathcal{H}$, as claimed. □

Proof of theorem 6. Let H^{c} be a ccdf in $\mathcal{R}(-\alpha)$ with $G^{\mathrm{c}}(t) \geqslant H^{\mathrm{c}}(t)$ for all $t > 0$. Just as in the proof of theorem 5, (5.6) holds. By proposition 9(i), $H_{\mathrm{e}} \in \mathcal{R}(-(\alpha-1))$. Then, by (2.3) and proposition 10(i), $F^{\mathrm{c}}(H) \in \mathcal{R}(-s(\alpha-1))$ for F^{c} in (2.2). (If $G_1^{\mathrm{c}} \in \mathcal{R}(-\alpha)$ and $G_1^{\mathrm{c}} \sim G_2^{\mathrm{c}}$, then $G_2^{\mathrm{c}} \in \mathcal{R}(-\alpha)$.) Hence, first, $W^{\mathrm{c}}(H) \in \mathcal{T}(-s(\alpha-1))$ by (2.2) and, second, $W^{\mathrm{c}}(G) \in \mathcal{T}(-s(\alpha-1))$ by (5.6). □

6. Upper bounds

In this section we prove theorem 4. The proof is based on a stochastic comparison between multi-server and single-server queues. Given a $\mathrm{GI}/\mathrm{GI}/s$ model, let W_1 be the steady-state waiting time in the associated single-server queue obtained by assigning successive arrivals cyclically to the s servers, i.e., the $\mathrm{GI}/\mathrm{GI}/1$ system in which the interarrival-time distribution is the s-fold convolution of the interarrival-time in the original s-server system. Wolff [36] showed that W is bounded above by W_1 in increasing convex stochastic order, i.e.,

$$Ef(W) \leqslant Ef(W_1) \tag{6.1}$$

for all nondecreasing convex real-valued functions f or, equivalently (see [30, p. 9]),

$$\int_t^\infty P(W > u)\,\mathrm{d}u \leqslant \int_t^\infty P(W_1 > u)\,\mathrm{d}u \quad \text{for all } t. \tag{6.2}$$

However, the ordering cannot be extended to ordinary stochastic order, i.e., (6.1) need not hold for all nondecreasing f or, equivalently, we need *not* have

$$P(W > t) \leqslant P(W_1 > t) \quad \text{for all } t; \tag{6.3}$$

see [33]. Nevertheless, we can apply (6.2) to deduce that

$$\liminf_{t\to\infty} \frac{P(W > t)}{P(W_1 > t)} \leqslant 1. \tag{6.4}$$

When G_{e} is subexponential, we have

$$\lim_{t\to\infty} \frac{P(W_1 > t)}{G_{\mathrm{e}}^{\mathrm{c}}(t)} = \frac{\rho}{1-\rho}. \tag{6.5}$$

Hence, we can combine (6.4) and (6.5) to obtain theorem 4.

Scheller-Wolf and Sigman [27] have established conditions for W to have finite moments that also support the conjecture when $s = 2$ and $\rho < 1/2$. They show that (i) $EW < \infty$ if $EV^{3/2} < \infty$ and (ii) $EW^2 < \infty$ if $EV^2 < \infty$. We can apply

Chebychev's inequality to convert finite moments into tail-probability upper bounds; i.e., if $EW^r < \infty$, then

$$P(W > t) \leqslant \frac{EW^r}{t^r}. \tag{6.6}$$

To illustrate, suppose that $s = 2$ and $G^{\mathrm{c}}(t) \sim At^{-(\varepsilon+3/2)}$ for small positive ε. The infinite-server lower bound yields

$$W^{\mathrm{c}}(t) \geqslant F^{\mathrm{c}}(t) \sim A_1 G_{\mathrm{e}}^{\mathrm{c}}(t)^2 \sim A_2 t^{-(1+2\varepsilon)}, \tag{6.7}$$

while $EV^{3/2} < \infty$, so that $EW < \infty$ by Scheller-Wolf and Sigman [27], and

$$P(W > t) \leqslant \frac{EW}{t}. \tag{6.8}$$

Inequalities (6.7) and (6.8) show that the exponents in the two power-tail bounding ccdfs differ by only 2ε, where ε was arbitrary.

7. Moments

Bounds on the waiting-time tail probabilities allow us to determine whether moments are finite or not. For this purpose, recall that a nonnegative random variable V with cdf G satisfies $EV^r < \infty$ if and only if $t^{r-1}G^{\mathrm{c}}(t)$ is integrable over $(0, \infty)$; see [18, p. 151].

First notice that our minimal-stability lower bound in theorem 3 allows us to deduce that $EW^r = \infty$ whenever $EV^{r+1} = \infty$ and $k = 1$ in (2.17), which is one side of an equivalence established by Scheller-Wolf and Sigman [27].

Next we assume that the service-time ccdf G^{c} has a power tail, i.e., (2.15) holds. Then $EV^r = \infty$ for $r > 0$ if and only if $\alpha \leqslant r$, where α is the exponent of the power tail. Hence we can apply the infinite-server lower bound theorem 1 to deduce the following.

Theorem 11. Suppose that the service-time ccdf G^{c} has a power tail as in (2.15).

(i) If $EV^r = \infty$ for $r > 0$, then $EW^{s(r-1)} = \infty$.

(ii) When $s = 2$, $EV^{3/2} < \infty$ if and only if $EW < \infty$.

(iii) When $s = 2$, $EV^2 < \infty$ if and only if $EW^2 < \infty$.

Proof. (i) Since G^{c} has a power tail with exponent α, $G_{\mathrm{e}}^{\mathrm{c}}$ and the ccdf H_2^{c} in (2.10), which is a lower bound for $W^{\mathrm{c}}(t)$, have power tails with exponents $\alpha - 1$ and $s(\alpha - 1)$, respectively. If $EV^r = \infty$, then the exponent α of the power tail of G^{c} must satisfy $\alpha \leqslant r$. Since $\alpha \leqslant r$, $s(\alpha - 1) \leqslant s(r - 1)$, so that we must have $EW^{s(r-1)} = \infty$.

(ii) and (iii) Apply part (i) together with upper bounds of [27, proposition 4.1]. □

We now point out that our infinite-server lower bound does *not* support part (i) of theorem 11 in general, i.e., without the power-tail condition. Suppose that the service-time ccdf satisfies

$$G^{\mathrm{c}}(t) \sim \frac{A}{t^{\alpha}(\log t)^{\beta}} \quad \text{as } t \to \infty, \tag{7.1}$$

so that $EV^{\alpha} < \infty$ if and only if $\beta > 1$. Let $\beta = 3/4$, so that $EV^{\alpha} = \infty$. It is easy to see that

$$G_{\mathrm{e}}^{\mathrm{c}}(t) \sim \frac{A'}{t^{\alpha-1}(\log t)^{\beta}} \quad \text{as } t \to \infty, \tag{7.2}$$

so that

$$H_1^{\mathrm{c}}(t) \sim \frac{A''}{t^{s(\alpha-1)}(\log t)^{s\beta}} \quad \text{as } t \to \infty \tag{7.3}$$

for H_1^{c} in (2.6), which implies that

$$\int_0^{\infty} t^{s(\alpha-1)} \, \mathrm{d}H_1(t) < \infty \quad \text{for all } s \geqslant 2, \tag{7.4}$$

making it impossible to deduce that $EW^{s(\alpha-1)} = \infty$. We conjecture that $EW^{s(\alpha-1)} < \infty$ in this example, but that is yet to be proved because H_1 in (7.4) is only a lower bound.

8. The transient waiting-time distribution

An infinite-server lower bound also applies with time-dependent non-Poisson arrival processes. As a special case, it applies to the transient distribution in the $\mathrm{M}/\mathrm{GI}/s$ model.

Of course the transient waiting-time distribution depends on the initial conditions. To have a relatively simple description, we assume that the $\mathrm{M}/\mathrm{G}/s$ system starts off empty at time 0. We are then interested in the time-dependent waiting-time ccdf. Let $W(t)$ be the waiting time before beginning service for a customer arriving at time t. (Since the arrival process in Poisson, the distribution is unaffected by the presence of an arrival at time t; i.e., the virtual and actual waiting-time distributions coincide.) Let $W^{\mathrm{c}}(u;t) \equiv P(W(t) > u)$ be the transient ccdf at time t. Then, paralleling (2.2), we have

$$W^{\mathrm{c}}(u;t) \geqslant F^{\mathrm{c}}(u;t) \quad \text{for all } u > 0, \tag{8.1}$$

where $F^{\mathrm{c}}(\cdot;t)$ is the ccdf of the first passage time to $s-1$ or fewer busy servers in the infinite-server model, starting off with the transient $\mathrm{M}/\mathrm{G}/\infty$ distribution at time t and neglecting all subsequent arrivals. As a consequence, we have the following result.

Theorem 12. In the $\mathrm{M/G}/s$ model starting out empty at time 0, for all positive λ, u and t,

$$W^{\mathrm{c}}(u;t) \geqslant F^{\mathrm{c}}(u;t) = \Pi\bigl(s-1; \lambda m_1\bigl[G_{\mathrm{e}}^{\mathrm{c}}(u) - G_{\mathrm{e}}^{\mathrm{c}}(t+u)\bigr]\bigr), \tag{8.2}$$

where

$$F^{\mathrm{c}}(u;t) \sim \frac{(\lambda m_1)^s}{s!}\bigl[G_{\mathrm{e}}^{\mathrm{c}}(u) - G_3^{\mathrm{c}}(t+u)\bigr]^s \quad \text{as } u \to \infty. \tag{8.3}$$

Proof. The number $N(t)$ of customers in the $\mathrm{M/G}/\infty$ model at time t has a Poisson distribution with mean

$$EN(t) = \lambda m_1 G_{\mathrm{e}}(t), \tag{8.4}$$

where λ is the arrival rate, m_1 is the mean service time and G_{e} is the service-time stationary-excess cdf in (2.1); see [15, (20), p. 740]. Moreover, the number of customers arriving in $[0,t]$ that are still present at time u has a Poisson distribution with mean

$$EN(t,u) = \lambda \int_0^t G^{\mathrm{c}}(u+s)\,\mathrm{d}s = \lambda \int_u^{u+t} G^{\mathrm{c}}(s)\,\mathrm{d}s = \lambda m_1\bigl[G_{\mathrm{e}}^{\mathrm{c}}(u) - G_{\mathrm{e}}^{\mathrm{c}}(t+u)\bigr] \tag{8.5}$$

by the Poisson-random measure argument in the proof of Eick et al. [15, theorem 1]. □

From theorem 12, we can see the approach to steady state as $t \to \infty$. From (8.4) we see that $EN(t)$ approaches the steady-state mean $EN(\infty) = \lambda m_1$. Moreover, we obtain an explicit expression for the rate of convergence:

$$\frac{EN(\infty) - EN(t)}{EN(\infty)} = G_{\mathrm{e}}^{\mathrm{c}}(t), \quad t > 0. \tag{8.6}$$

So, for heavy-tailed ccdfs G^{c}, the convergence of $EN(t)$ to $EN(\infty)$ is relatively slow.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions, Queueing Systems 16 (1994) 311–338.

[2] J. Abate and W. Whitt, Asymptotics for $\mathrm{M/G/1}$ low-priority waiting-time tail probabilities, Queueing Systems 25 (1997) 173–233.

[3] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC, 1972).

[4] S. Asmussen, *Stochastic Simulation with a View towards Stochastic Processes*, Lecture Notes No. 2 (MaPhySto, Centre for Mathematical Physics and Stochastics, University of Aarhus, Denmark, 1999).

[5] S. Asmussen, O. Nerman and M. Olsson, Fitting phase type distributions via the EM algorithm, Scand. J. Statist. 23 (1996) 419–441.

[6] S. Asmussen, H. Schmidli and V. Schmidt, Tail probabilities for non-standard risk and queueing processes with subexponential jumps, Dept. Math. Stat., Lund University, Sweden (1997).

[7] D. Bertsimas, An exact FCFS waiting time analysis for a general class of $\mathrm{G/G}/s$ queueing systems, Queueing Systems 3 (1988) 305–320.

[8] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation* (Cambridge Univ. Press, Cambridge, 1989).

[9] A.A. Borovkov, *Stochastic Processes in Queueing Theory* (Springer, New York, 1976) (translation of 1972 Russian book).

[10] O.J. Boxma, Q. Deng and A.P. Zwart, Waiting-time asymptotics for the $\mathrm{M}/\mathrm{G}/2$ queue with heterogeneous servers, unpublished manuscript, CWI, Amsterdam.

[11] D.Y. Burman and D.R. Smith, A light-traffic theorem for multi-server queues, Math. Oper. Res. 8 (1983) 15–25.

[12] G.L. Choudhury and W. Whitt, Long-tail buffer-content distributions in broadband networks, Performance Evaluation 30 (1996) 177–190.

[13] J.W. Cohen, Some results on regular variation for distributions in queueing and fluctuations theory, J. Appl. Probab. 10 (1973) 343–353.

[14] N.G. Duffield and W. Whitt, Control and recovery from rare congestion events in a large multi-server system, Queueing Systems 26 (1997) 69–104.

[15] S.G. Eick, W.A. Massey and W. Whitt, The physics of the $\mathrm{M}_t/\mathrm{G}/\infty$ queue, Oper. Res. 41 (1993) 731–742.

[16] A. Erdélyi, *Asymptotic Expansions* (Dover, New York, 1956).

[17] A. Feldmann and W. Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, Performance Evaluation 31 (1998) 245–279.

[18] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, 2nd ed. (Wiley, New York, 1971).

[19] C.M. Goldie and C. Klüppelberg, Subexponential distributions, in: *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*, eds. R. Adler, R. Feldman and M.S. Taqqu (Birkhauser, Boston, 1998) pp. 435–459.

[20] L.V. Green, A limit theorem on subintervals of interrenewal times, Oper. Res. 30 (1982) 210–216.

[21] A.G. Greenberg, R. Srikant and W. Whitt, Resource sharing for book-ahead and intermediate-request calls, IEEE/ACM Trans. Networking 7 (1999) 10–22.

[22] P.R. Jelenković and A.A. Lazar, Subexponential asymptotics of a Markov-modulated random walk with queueing applications, J. Appl. Probab. 35 (1998) 325–347.

[23] C. Klüppelberg, Subexponential distributions and integrated tails, J. Appl. Probab. 25 (1988) 132–141.

[24] D.A. Korshunov, On waiting-time distribution in $\mathrm{GI}/\mathrm{G}/2$ queueing system with heavy tailed service times, unpublished manuscript.

[25] W.A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, Queueing Systems 25 (1997) 157–172.

[26] A.G. Pakes, On the tails of waiting-time distributions, J. Appl. Probab. 12 (1975) 555–564.

[27] A. Scheller-Wolf and K. Sigman, Delay moments for FIFO $\mathrm{GI}/\mathrm{GI}/s$ queues, Queueing Systems 25 (1997) 77–95.

[28] L.P. Seelen, An algorithm for $\mathrm{Ph}/\mathrm{Ph}/c$ queues, European J. Oper. Res. 23 (1986) 118–127.

[29] J.H.A. de Smit, A numerical solution for the multi-server queue with hyperexponential service times, Oper. Res. Lett. 2 (1983) 217–224.

[30] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, New York, 1983).

[31] L. Takács, *Introduction to the Theory of Queues* (Oxford Univ. Press, New York, 1962).

[32] W. Whitt, Embedded renewal processes in the $\mathrm{GI}/\mathrm{G}/s$ queue, J. Appl. Probab. 9 (1972) 650–658.

[33] W. Whitt, On stochastic bounds for the delay distribution in the $\mathrm{GI}/\mathrm{G}/s$ queue, Oper. Res. 29 (1981) 604–608.

[34] W. Whitt, Comparing counting processes and queues, Adv. in Appl. Probab. 13 (1981) 207–220.

[35] W. Whitt, Comparison conjectures about the $\mathrm{M}/\mathrm{G}/s$ queue, Oper. Res. Lett. 2 (1983) 203–210.

[36] R.W. Wolff, An upper bound for multi-channel queues, J. Appl. Probab. 14 (1977) 884–888.

[37] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 223–231.