View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Springer - Publisher Connector

Queueing Syst (2008) 58: 77–104
DOI 10.1007/s11134-008-9060-2

OpenAccess

The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process

David Alan Goldberg · Ward Whitt

Received: 4 July 2006 / Revised: 14 December 2007 / Published online: 23 January 2008
© The Author(s) 2008

Abstract This paper studies the last departure time from a queue with a terminating arrival process. This problem is motivated by a model of two-stage inspection in which finitely many items come to a first stage for screening. Items failing first-stage inspection go to a second stage to be examined further. Assuming that arrivals at the second stage can be regarded as an independent thinning of the departures from the first stage, the arrival process at the second stage is approximately a terminating Poisson process. If the failure probabilities are not constant, then this Poisson process will be nonhomogeneous. The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process serves as a remarkably tractable approximation, which is appropriate when there are ample inspection resources at the second stage. For this model, the last departure time is a Poisson random maximum, so that it is possible to give exact expressions and develop useful approximations based on extreme-value theory.

Keywords Queues with terminating arrival processes · Last departure time · Infinite-server queues · Non-stationary queues · Congestion caused by inspection · Two-stage inspection · Extreme-value theory

Mathematics Subject Classification (2000) 60K25 · 60G70 · 90B22

D.A. Goldberg
Operations Research Center, MIT, Cambridge, MA, USA
e-mail: dag3141@mit.edu

W. Whitt (✉)
IEOR Department, Columbia University, New York, NY, USA
e-mail: ww2040@columbia.edu

1 Introduction

In this paper we introduce what we believe is a new class of queueing problems: determining the probability distribution for the last departure time from a queue with a terminating arrival process. This problem arises from a model of two-stage inspection. It is assumed that there are n items to be inspected, and that these items begin arriving to be inspected at time 0. There are two stages of inspection, with preliminary screening done at the first stage and a more careful inspection done at the second stage for items "failing" inspection at the first stage. Outcomes of successive first-stage inspections (passing or failing) are regarded as independent and identically distributed (i.i.d.) Bernoulli random variables with probability p of failure. There may be multiple inspectors (servers) at each stage, working in parallel. The inspection times are i.i.d. random variables at each stage, with distributions depending on the stage. Thus the two-stage inspection process can be directly modelled as an acyclic open network of two multi-server queues, with Markovian routing for departures from the first queue, some exiting while others move on to the second queue.

For the model to be developed here, we assume that the arrival process to the first stage is sufficiently rapid relative to the service rate there that, from the perspective of the second stage, we can assume that the entire batch of n items arrives at the first stage at time 0. We also assume that the first-stage inspection process is a relatively routine process, nearly deterministic in nature. Let the first-stage inspection time have mean d and low variability. Let there be m servers working in parallel at stage 1. Then the time that all items finish first-stage inspection is approximately the deterministic time $\tau \equiv nd/m$. Under those assumptions, the total arrival process for second-stage inspection can thus be regarded as being a *terminating arrival process*, operating over the finite time interval $[0, \tau]$. We are interested in the last departure time from the second stage, because that is when the second stage of inspection will be finished.

If p is relatively small and n is relatively large, then the arrival process at the second stage can be regarded as approximately a Poisson process with constant arrival rate $\lambda = pm/d$, operating over the finite time interval $[0, nd/m]$. With s second-stage servers operating in parallel, we can regard the second-stage inspection as approximately a standard $M/G/s$ queue with arrival rate $\lambda = pm/d$, where the arrival process is turned off at time $\tau = nd/m$.

We were motivated by a mathematical model for inspecting shipping containers developed by Wein et al. [13]. Their model is quite elaborate, addressing a wide range of important issues. As a consequence, each specific issue—such as the congestion caused by the inspection scheme—had to depend on relatively simple sub-models. In particular, in [13] the extra delay at the second stage of congestion was modelled approximately as the steady-state sojourn time in the $M/G/s$ model. The idea is that the system could be regarded as being in steady state when the last container comes to the second stage of inspection. The remaining time for that container to finish the second stage of inspection after the first-stage congestion is complete should thus be conservatively approximated by the steady-state sojourn time at the second stage. Our analysis here stems from the observation that, with multiple second-stage inspection devices, inspections need not be completed in order of arrival, causing the expected remaining time until the last completed inspection actually to be larger than

the expected steady-state sojourn time of the last arrival. Our analysis focuses on that phenomenon. We analyze the shipping-container application further in [4].

So far, the arrival process at the second stage is homogeneous, but there are two compelling reasons for letting the arrival process be nonhomogeneous. First, we may be able to classify the items prior to inspection in a way that produces approximately independent failures, but with varying failure probabilities. If so, it should be possible to reduce the expected last departure time from second-stage inspection by inspecting the items that are more likely to fail first-stage inspection earlier in the inspection process at the first stage. To specify the arrival-rate function at the second stage, suppose that the k^{th} item to be inspected at stage 1 has failure probability p_k. Assuming that these failure probabilities are all relatively small, the arrival process to the second station is approximately a nonhomogeneous Poisson process with arrival-rate function

$$\lambda(t) = \frac{m}{d} p_{\lfloor (mt/d)+1 \rfloor} = \frac{p_k m}{d} \quad \text{for } \frac{(k-1)d}{m} \le t < \frac{kd}{m}, \quad 1 \le k \le n, \tag{1.1}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Given (1.1), we have total arrival rate

$$\int_0^\tau \lambda(t)\,dt = \sum_{k=1}^{n} \left(\frac{p_k m}{d}\right)\left(\frac{d}{m}\right) = \sum_{k=1}^{n} p_k. \tag{1.2}$$

Another way to get a nonhomogeneous arrival process at the second stage is to have multiple groups of items undergoing inspection, at different times. For example, suppose that two sets of items arrive to go through this two-stage inspection process. Let n_1 items arrive at one first-stage inspection station with m_1 servers at time t_1, while n_2 items arrive at a second first-stage inspection station with m_2 servers at time t_2. Let the first-stage inspection times at the two stations have means d_1 and d_2 and low variability. Let the first-stage failure probabilities be constant for each group, being p_1 and p_2, respectively. Let all items failing congestion from both groups proceed to a common second-stage inspection. Group i produces a Poisson arrival process at the second stage with rate $\lambda_i = p_i m_i / d_i$, operating over the finite interval $[t_1, t_1 + n_1 d_1]$, but at rate 0 for all other times. Then the second stage can be regarded as an $M_t/G/s$ queue with nonhomogeneous Poisson arrival process, having a piecewise-constant arrival-rate function equal to the sum of the two component arrival-rate functions.

Unfortunately, the distribution of the last departure time in an $M_t/G/s$ queue is quite complicated, even with a homogeneous Poisson arrival process. However, we observe that great simplification occurs if we can consider either of the two extremes: $s = 1$ or $s = \infty$. First, when $s = 1$, the remaining time T_τ after τ until the last departure time coincides with the workload (or virtual waiting time) at time τ, provided that it is positive, which is the case we are concerned with. The transient workload distribution is still somewhat complicated, but there are established methods for calculating its distribution [2]. Since the algorithm in [2] is for a piecewise-constant arrival-rate function, it is ideally suited for the multiple-group inspection just mentioned.

This paper is devoted to the other idealized case: $s = \infty$. The $M_t/G/\infty$ queue with a terminating arrival process is appropriate when there are ample inspecting

resources at the second station. Equivalently, the approximation is appropriate when the dominant portion of the time items spend at the second stage is the inspection time itself, rather than any waiting required before inspection can begin.

Here is how this paper is organized: In Sect. 2 we show that the remaining time T_τ until the last departure after τ can be represented as the maximum of a Poisson random number of i.i.d. random variables, and so is remarkably tractable, allowing us to apply results from extreme-value theory [6]. We give explicit expressions for the distribution of T_τ and its quantiles and moments. The distribution simplifies when the Poisson arrival process is homogeneous and when the service-time distribution is exponential or a finite mixture of exponentials.

Thereafter we focus on approximations, again drawing heavily on extreme-value theory. In Sect. 3 we develop approximations for transient distributions for two classes of service-time distributions: (i) those with a pure-exponential tail and (ii) those with a power tail. In Sect. 4 we develop associated approximations for the case in which τ is sufficiently large that the queue can be regarded as being approximately in steady state at time τ. We show that the transient behavior is often well approximated by the steady-state behavior in the exponential-tail case, provided that τ is not too small, but we identify difficulties in the power-tail case. These difficulties arise in the power-tail case because the order of two iterated limits matters. We propose ways to resolve this power-tail problem.

In Sect. 5 we evaluate the approximations by comparing with exact values of the mean, variance and several quantiles of the distribution of T_τ for several service-time distributions, obtained from numerical calculations based on Sect. 2. Finally, in Sect. 6 we draw conclusions. Additional numerical comparisons are contained in [8].

2 The remaining time until the last departure

Henceforth we consider an $M_t/G/\infty$ queue with a terminating arrival process. The service times are i.i.d. random variables distributed according to a random variable S with cumulative distribution function (cdf) G. The arrival-rate function $\lambda \equiv \{\lambda(t) : 0 \le t \le \tau\}$ is integrable. Let D be the last departure time. We want to determine the distribution of $T \equiv T_\tau \equiv (D-\tau)^+$, the remaining time after τ until the last departure.

As reviewed in [5], the number in system at the arrival-process terminating time τ has a Poisson distribution with mean

$$\nu_\tau = \int_0^\tau \lambda(u) G^c(\tau - u)\, du, \tag{2.1}$$

where $G^c(t) \equiv 1 - G(t)$. Moreover, there is a Poisson-random-measure representation discussed in the proof of Theorem 1 of [5] that enables us to calculate the conditional distribution of the remaining service times at time τ, given any number of customers still in service. By applying theorems about Poisson random measures; e.g., see Theorems 2.3–2.5 of [11], we obtain the following generalization of the well-known property for the stationary $M/G/\infty$ system; see p. 161 of Takács [12]:

Theorem 2.1 (Remaining service times) *Conditional on there being n customers in service at time τ, the remaining service times of those customers are i.i.d., each distributed as a random variable X_τ with ccdf*

$$G_\tau^c(x) \equiv P(X_\tau > x) \equiv \frac{1}{\nu_\tau} \int_0^\tau \lambda(u) G^c(\tau + x - u)\, du, \tag{2.2}$$

where ν_τ is the mean in (2.1).

To characterize the distribution of T, let $\{X_n : n \geq 1\}$ be a sequence of i.i.d. random variables, each distributed as a random variable $X \equiv X_\tau$ having cdf G_τ. Let $M_n \equiv \max\{X_1, \ldots, X_n\}$, $n \geq 1$; it has cdf $P(M_n \leq x) = G_\tau(x)^n$. Let $N \equiv N_\tau$ be a random variable independent of $\{X_n : n \geq 1\}$ with a Poisson distribution having mean ν_τ. Then

$$T \stackrel{\mathrm{d}}{=} M_N. \tag{2.3}$$

That structure allows us to obtain a simple explicit expression for the distribution of T in terms of ν_τ in (2.1) and G_τ in (2.2). See Embrechts et al. [6] for background on extreme-value theory; see Sect. 4.3 and Example 5.3.5 there for the case of random indices.

When $\lambda(t) = \lambda$, it is convenient to work with the classical *stationary-excess cdf*

$$G_e(x) \equiv \frac{1}{E[S]} \int_0^x G^c(u)\, du; \tag{2.4}$$

e.g., p. 432 of [10]. Let S_e be a random variable with cdf G_e. It is well known that $E[S_e^k] = E[S^{k+1}]/(k+1)E[S]$.

For any random variable Y with a continuous cdf, let its quantile function be $q_Y \equiv q_Y(x)$ such that $P(Y \leq q_Y(x)) = x$. We write $f(x) \sim g(x)$ as $x \to \infty$ when $f(x)/g(x) \to 1$ as $x \to \infty$.

Theorem 2.2 (The cdf of T) (a) *For any $x > 0$,*

$$P(T \leq x) = e^{-\nu_\tau G_\tau^c(x)}, \tag{2.5}$$

where ν_τ and G_τ^c are given in (2.1) and (2.2). As a consequence, $P(T > x) \sim \nu_\tau G_\tau^c(x)$ as $x \to \infty$ and

$$q_T(x) = q_{X_\tau}\left(1 - \frac{\log(1/x)}{\nu_\tau}\right), \quad e^{-\nu_\tau} < x < 1. \tag{2.6}$$

(b) *If, in addition, $\lambda(t) = \lambda$ for $t \geq 0$, then $\nu_\tau = \lambda E[S] G_e(\tau)$, $G_\tau^c(x) = [G_e(\tau + x) - G_e(x)]/G_e(\tau)$ and*

$$P(T \leq x) = e^{-\lambda E[S](G_e^c(x) - G_e^c(\tau + x))}. \tag{2.7}$$

Proof Conditioning and unconditioning on N, we obtain

$$P(T \le x) = \sum_{n=0}^{\infty} \frac{e^{-\nu_\tau} \nu_\tau^n}{n!} G_\tau(x)^n = e^{-\nu_\tau G_\tau^c(x)}. \tag{2.8}$$

The limit for $P(T > x)$ uses $e^{-x} = 1 - x + x^2/2 + o(x^2)$ as $x \to 0$. □

We can calculate moments by combining (2.5) or (2.7) with the classical formula $E[T^j] = \int_0^\infty j x^{j-1} P(T > x)\, dx$. The distribution of T has a remarkably simple form when the arrival rate is constant and G is exponential. Indeed, then T has exactly a Gumbel distribution, restricted to the positive halfline. Let W be a random variable with the (standard) *Gumbel distribution*, i.e.,

$$P(W \le x) \equiv \exp\{-e^{-x}\}, \quad -\infty < x < +\infty, \tag{2.9}$$

with mode 0, $E[W] = 0.5772$ (Euler's constant), $\mathrm{Var}(W) = \pi^2/6 = 1.644$ and quantile function $q_W(x) = -\log\log(1/x)$ [9]. Let $\overset{\mathrm{d}}{=}$ mean equality in distribution. Let $(x)^+ \equiv \max\{x, 0\}$.

Corollary 2.1 (Exponential service times) *If $\lambda(t) = \lambda$ for $t \ge 0$ and G is exponential with mean $1/\eta$, then $G_\tau = G_e = G$, $\nu_\tau = \lambda(1 - e^{-\eta\tau})/\eta$ and*

$$P(T \le x) = e^{-(\lambda/\eta)(1-e^{-\eta\tau})e^{-\eta x}}, \quad x \ge 0, \tag{2.10}$$

or, equivalently,

$$T \overset{\mathrm{d}}{=} \frac{1}{\eta}\left(\log\left(\frac{\lambda(1-e^{-\eta\tau})}{\eta}\right) + W\right)^+, \tag{2.11}$$

where W is the standard Gumbel random variable in (2.9). Consequently,

$$q_T(x) = \frac{1}{\eta}\left(\log\left(\frac{\lambda(1-e^{-\eta\tau})}{\eta}\right) - \log\log(1/x)\right)^+. \tag{2.12}$$

When $P(T = 0) = e^{-\nu_\tau}$ is negligible, the approximation obtained from (2.11) by ignoring the positive part function will be excellent, yielding

$$E[T] \approx \frac{1}{\eta}\left(\log\left(\frac{\lambda(1-e^{-\eta\tau})}{\eta}\right) + 0.5772\right) \quad \text{and} \quad \mathrm{Var}(T) \approx \frac{\pi^2}{6\eta^2} = \frac{1.644}{\eta^2}.$$

There also is a convenient explicit expression for the distribution of T when G is hyperexponential (a mixture of exponential distributions, denoted by H_k). To get it, we simply combine Corollary 2.1 with the following mixture result. Let $T(\lambda, G)$ denote the random variable T as a function of the arrival-rate function $\lambda(t)$ and the service-time cdf G.

Corollary 2.2 (Mixtures) *If*

$$G(x) = \sum_{i=1}^{n} p_i G_i(x), \quad x \geq 0, \tag{2.13}$$

where G_i is a cdf for each i, $p_i > 0$ for each i and $p_1 + \cdots + p_n = 1$, then

$$P(T(\lambda, G) \leq x) = \prod_{i=1}^{n} P(T(p_i\lambda, G_i) \leq x), \tag{2.14}$$

where $P(T(\lambda, G) \leq x)$ is given in (2.5).

Proof This is easily verified from (2.5). It also can be proved by interpreting the service time as depending on the "customer type," where the customer is type i with probability p_i. Then the arrival processes of the different types are independent Poisson processes, so $T(\lambda, G)$ is the maximum of the last-departure times for the n types, which are independent, leading to (2.14). □

When the arrival rate is constant and τ is large, we can regard the infinite-server queue as being in steady state. Moreover, the convergence as $\tau \to \infty$, is monotone in the sense of stochastic order.

Corollary 2.3 (Steady state) *If $\lambda(t) = \lambda$ for $t \geq 0$ and $\tau \to \infty$, then*

$$\nu_\tau = \lambda E[S] G_e(\tau) \uparrow \lambda E[S] \equiv \nu_\infty \equiv \nu, \tag{2.15}$$

$$G_\tau^c(x) = \frac{G_e(\tau + x) - G_e(x)}{G_e(\tau)} \to G_e^c(x) \quad \textit{for all } x, \tag{2.16}$$

$$\nu_\tau G_\tau^c(x) = \lambda E[S](G_e^c(x) - G_e^c(\tau + x)) \uparrow \lambda E[S] G_e^c(x) \quad \textit{for all } x, \tag{2.17}$$

$$P(T_\tau > x) = 1 - e^{-\nu_\tau G_\tau^c(x)} \uparrow 1 - e^{-\nu_\infty G_e^c(x)} \equiv P(T_\infty > x) \quad \textit{for all } x, \tag{2.18}$$

and

$$q_{T_\tau}(x) \uparrow q_{T_\infty}(x) = q_{S_e}\left(1 - \frac{\log(1/x)}{\lambda E[S]}\right), \qquad e^{-\lambda E[S]} < x < 1. \tag{2.19}$$

Since ν_τ, G_τ and $P(T \leq x)$ can be computed, we can directly determine when the steady-state approximation is reasonable. The stochastic monotonicity implies that the steady-state values serve as upper bounds. From (2.15), we see that $(\nu_\infty - \nu_\tau)/\nu_\infty = G_e^c(\tau)$, where G_e is the stationary-excess cdf in (2.4), as in formula (23) of [5].

3 Approximations

In this section we develop approximations, using two approaches: (i) direct asymptotics and (ii) extreme-value theory as in Chap. 3 of Embrechts et al. [6] and Crow

et al. [3]. Both depend on the tail-probability asymptotics for G^c. We consider two common cases: (i) a pure-exponential tail and (ii) a power tail. We show that both extreme-value approximations are asymptotically correct in heavy traffic, i.e., as the arrival rate increases.

3.1 A pure-exponential tail

Let the service-time cdf G have a pure-exponential tail; i.e., $G^c(x) \sim \gamma e^{-\eta x}$ as $x \to \infty$ for $\eta > 0$ and $\gamma > 0$. We first show that the pure-exponential-tail property is inherited by the cdf G_τ in (2.2), leaving the asymptotic decay rate η unchanged. (For the following limit, we apply the dominated convergence theorem using the assumed integrability of the arrival-rate function.)

Theorem 3.1 (Inheritance by G_τ) *If $G^c(x) \sim \gamma e^{-\eta x}$ as $x \to \infty$, then $G_\tau^c(x) \sim \gamma_\tau e^{-\eta x}$ as $x \to \infty$ for cdf G_τ in (2.2), where*

$$\gamma_\tau = (\gamma/\nu_\tau) \int_0^\tau \lambda(u) e^{-\eta(\tau-u)}\, du, \tag{3.1}$$

with ν_τ in (2.1). If, in addition, $\lambda(t) = \lambda$ for $t \geq 0$, then $\nu_\tau = \lambda E[S] G_e(\tau)$ and

$$\gamma_\tau = \frac{\gamma(1 - e^{-\eta\tau})}{\eta E[S] G_e(\tau)}. \tag{3.2}$$

By Theorem 3.1, $\nu_\tau G_\tau^c(x) \sim \xi e^{-\eta x}$ as $x \to \infty$, where

$$\xi \equiv \gamma \int_0^\tau \lambda(u) e^{-\eta(\tau-u)}\, du = \nu_\tau \gamma_\tau. \tag{3.3}$$

Based on this, we propose the direct asymptotic approximation

$$P(T \leq x) = e^{-\nu_\tau G_\tau^c(x)} \approx e^{-\xi e^{-\eta x}}, \quad x \geq 0, \tag{3.4}$$

or, equivalently,

$$T \approx \frac{1}{\eta} (\log(\xi) + W)^+ \approx \frac{1}{\eta} (\log(\xi) + W), \tag{3.5}$$

where W is the Gumbel random variable in (2.9) and ξ is the constant in (3.3). When $\lambda(t) = \lambda$ for $t \geq 0$,

$$\xi = \frac{\lambda\gamma(1 - e^{-\eta\tau})}{\eta}. \tag{3.6}$$

Note that approximation (3.5) with (3.6) coincides with the exact formula (2.11) when G is exponential (where $\gamma = 1$). From (3.5), we obtain

$$E[T] \approx \eta^{-1}[\log(\nu_\tau \gamma_\tau) + 0.5772] \tag{3.7}$$

and, assuming that ν_τ is suitably large,

$$\mathrm{Var}(T) \approx \frac{1.644}{\eta^2}. \tag{3.8}$$

We now consider the second approach based on extreme-value theory. We will show that a simple form of this new approximation coincides with approximation (3.5) above. This second approach uses extreme-value theory to approximate the distribution of M_n. Let $\Rightarrow$ denote convergence in distribution. As reviewed in [6], under the exponential-tail assumption in Theorem 3.1, $M_n - (\log(n\gamma_\tau)/\eta) \Rightarrow W/\eta$ as $n \to \infty$, which supports the approximation

$$M_n \approx \eta^{-1}[\log(n\gamma_\tau) + W], \tag{3.9}$$

for W in (2.9). From (3.9), we obtain the approximation

$$q_{M_n}(x) \approx \eta^{-1}[\log(n\gamma_\tau) - \log\log(1/x)]. \tag{3.10}$$

Notice that M_n, as measured by $q_{M_n}(x)$, grows like $\log(n\gamma_\tau)$ as $n \to \infty$, but the spread, as measured by $q_{M_n}(x_2) - q_{M_n}(x_1)$, is asymptotically constant, so that the distribution of M_n concentrates (relatively) as n increases.

Combining (2.3) and (3.9), we obtain the approximation

$$T \stackrel{\mathrm{d}}{=} M_N \approx \eta^{-1}[\log(N\gamma_\tau) + W], \tag{3.11}$$

where N and W are independent random variables with the Poisson and Gumbel distributions, respectively. Assuming that the Poisson mean ν_τ is not too small, N is approximately normally distributed. (We assume that the normal random variable has negligible probability of being negative.) Further, we can use the more elementary approximation corresponding to $N \approx E[N]$, yielding approximation (3.5) for ξ in (3.3) (which becomes (3.6) when $\lambda(t)$ is constant).

Under (3.5), we can approximate the quantiles of T by

$$q_T(x) \approx \eta^{-1}[\log(\xi) - \log\log(1/x)], \tag{3.12}$$

for ξ in (3.3).

The analysis above shows that these approximations are asymptotically correct as the arrival-rate function increases. That is the standard heavy-traffic limiting regime for infinite-server models; see Sect. 10.3 of [14]. To formulate the heavy-traffic limit, we consider a sequence of models indexed by n, letting the service-time cdf G and the time horizon τ remain unchanged. We let the arrival-rate function in model n be $\lambda_n(t) = n\lambda(t)$ for some initial arrival-rate function λ and then let $n \to \infty$.

Theorem 3.2 (Heavy-traffic limit) *If $n \to \infty$ in the sequence of models specified above with $\lambda_n(t) = n\lambda(t)$, $0 \le t \le \tau$, and fixed cdf G having an exponential tail as in Theorem 3.1, then $\nu_{\tau,n} = n\nu_{\tau,1} \to \infty$ while $\gamma_{\tau,n} = \gamma_{\tau,1}$ in (3.1) for all n,*

$$\eta\left[T_n - \log(n\nu_{\tau,1}\gamma_{\tau,1})\right] \Rightarrow W, \tag{3.13}$$

where T_n denotes T_τ as a function of n and W has the Gumbel distribution in (2.9), and the approximations above are asymptotically correct.

Proof Let N_n be the number of customers in the system at time τ in model n, which we have observed has a Poisson distribution. By the scaling, $n^{-1}N_n \Rightarrow \nu_{\tau,1}$ as $n \to \infty$. Use the Skorohod representation theorem to replace that convergence in distribution by convergence with probability 1 (w.p.1), see Theorem 3.2.2 of [14], and condition on one sample point, yielding $n^{-1}N_n \to \nu_{\tau,1}$ w.p.1 as $n \to \infty$. Next apply the classic extreme-value theorem to get $\eta[T_n - \log(N_n\gamma_{\tau,1})] \Rightarrow W$, where W has the Gumbel distribution in (2.9). Finally, we get the desired (3.13) from this limit and the convergence-together theorem, Theorem 11.4.7 of [14], by noting that

$$\log(N_n\gamma_{\tau,1}) - \log(n\nu_{\tau,1}\gamma_{\tau,1}) = \log\left(\frac{N_n\gamma_{\tau,1}}{n\nu_{\tau,1}\gamma_{\tau,1}}\right) \to 0.$$

Since we get that convergence in distribution in the w.p.1 representation, we get the same convergence in distribution in general. □

3.2 A power tail

Motivated by the possibility that the second-stage inspection-time distribution might have a heavy tail, in this subsection we assume that the service-time cdf G has a power tail; i.e.,

$$G^c(x) \sim \gamma x^{-\alpha} \quad \text{as } x \to \infty, \tag{3.14}$$

for $\alpha > 0$ and $\gamma > 0$. Paralleling Theorem 3.1, we see that the power-tail property is inherited by the cdf G_τ in (2.2), with the same exponent α.

Theorem 3.3 *If G^c satisfies (3.14), then $G_\tau^c(x) \sim \gamma_\tau x^{-\alpha}$ as $x \to \infty$ for G_τ^c in (2.2), where*

$$\gamma_\tau = (\gamma/\nu_\tau)\int_0^\tau \lambda(u)\,du, \tag{3.15}$$

with γ from (3.14) and ν_τ in (2.1). If, in addition, $\lambda(t) = \lambda$ for $t \ge 0$, then $\nu_\tau = \lambda E[S]G_e(\tau)$ and $\gamma_\tau = \gamma\tau/E[S]G_e(\tau)$.

Paralleling (3.4), we have the following direct asymptotic approximation based on Theorem 3.3:

$$P(T \le x) = e^{-\nu_\tau G_\tau^c(x)} \approx e^{-\xi x^{-\alpha}}, \quad \text{where } \xi = \nu_\tau\gamma_\tau. \tag{3.16}$$

Approximation (3.16) is equivalent to

$$T \approx \xi^{1/\alpha} Y_\alpha, \tag{3.17}$$

where Y_α is a random variable with the standard *Fréchet distribution* on $[0, \infty)$, i.e.,

$$P(Y_\alpha \le x) = e^{-x^{-\alpha}}, \quad x \ge 0; \tag{3.18}$$

see Chapter 3 of [6], especially Sect. 3.3.1. The standard Fréchet random variable Y_α can be related to the standard Gumbel random variable W in (2.9) by $Y_\alpha \stackrel{\mathrm{d}}{=} e^{W/\alpha}$.

Paralleling (3.9), we can again apply extreme-value theory to approximate the distribution of M_n. Under assumption (3.14), we have $M_n/(\gamma_\tau n)^{1/\alpha} \Rightarrow Y_\alpha$, which supports the approximation

$$M_n \approx (\gamma_\tau n)^{1/\alpha}\, Y_\alpha. \tag{3.19}$$

It is easy to see that the quantiles of Y_α are

$$q_{Y_\alpha}(x) = \left(\frac{1}{\log(1/x)}\right)^{1/\alpha}. \tag{3.20}$$

If x is close to 1, then $\log(1/x) \approx (1/x) - 1 \approx 1 - x$. Combining (3.19) and (3.20), we obtain

$$q_{M_n}(x) \approx \left(\frac{\gamma_\tau n}{\log(1/x)}\right)^{1/\alpha}. \tag{3.21}$$

In contrast to (3.10), (3.21) shows that, under (3.14), M_n grows like $n^{1/\alpha}$ instead of $\log(n)$. (But for finite n, we may well have $n^{1/\alpha} < \log(n)$.) Moreover, the spread, as measured by $q_{M_n}(x_2) - q_{M_n}(x_1)$, grows like $n^{1/\alpha}$ instead of being asymptotically constant.

Now, returning to T, we can combine (2.3) and (3.19) to get

$$T \stackrel{\mathrm{d}}{=} M_N \approx (\gamma_\tau N(\nu_\tau, \nu_\tau))^{1/\alpha}\, Y_\alpha = N(\nu_\tau\gamma_\tau, \nu_\tau\gamma_\tau^2)^{1/\alpha} Y_\alpha, \tag{3.22}$$

for α in (3.14), ν_τ in (2.1), and γ_τ in (3.15) . The mean thus has the approximation

$$E[T] = E[M_N] \approx E[(N(\nu_\tau\gamma_\tau, \nu_\tau\gamma_\tau^2))^{1/\alpha} Y_\alpha] = E[N(\nu_\tau\gamma_\tau, \nu_\tau\gamma_{\tau^2})^{1/\alpha}]E[Y_\alpha]. \tag{3.23}$$

If, in addition, the normal distribution can be approximated by its mean, then we have again (3.17), which is equivalent to the direct asymptotic approximation in (3.16).

If, in addition, $\lambda(t) = \lambda$ for $t \geq 0$, then $\xi = \lambda\gamma\tau$ and we obtain associated approximations for the mean and variance, paralleling those given in (3.7) and (3.8) (assuming $\alpha > 1$ for the mean formula and $\alpha > 2$ for the variance formula):

$$T \approx (\lambda\gamma\tau)^{1/\alpha} Y_\alpha, \tag{3.24}$$

$$E[T] \approx (\lambda\gamma\tau)^{1/\alpha}\Gamma(1 - (1/\alpha)), \quad \alpha > 1, \tag{3.25}$$

and

$$\mathrm{Var}[T] \approx (\lambda\gamma\tau)^{2/\alpha}(\Gamma(1 - (2/\alpha)) - \Gamma(1 - (1/\alpha))^2), \quad \alpha > 2. \tag{3.26}$$

Then we can combine (3.17) and (3.20) to obtain

$$q_T(x) \approx \left(\frac{\xi}{\log(1/x)}\right)^{1/\alpha}, \tag{3.27}$$

for ξ in (3.16), which becomes $\lambda\gamma\tau$ when $\lambda(t) = \lambda$.

We conclude by stating an analog of Theorem 3.2—the heavy-traffic limit that follows from the analysis above. We consider the same sequence of models indexed by n, but where the fixed cdf G now has a power tail instead of a pure-exponential tail. We omit the proof, which is essentially the same as before.

Theorem 3.4 *If $n \to \infty$ in the sequence of models specified above with $\lambda_n(t) = n\lambda(t)$, $0 \le t \le \tau < \infty$, and fixed cdf G satisfying (3.14), then $\nu_{\tau,n} = n\nu_{\tau,1} \to \infty$ while $\gamma_{\tau,n} = \gamma_{\tau,1}$ in (3.15) for all n,*

$$\frac{T_n}{(n\nu_{\tau,1}\gamma_{\tau,1})^{1/\alpha}} \Rightarrow Y_\alpha, \tag{3.28}$$

where T_n denotes T_τ as a function of n and Y has the Fréchet distribution in (3.18), and the approximations in (3.22)–(3.27) are asymptotically correct.

4 The steady-state approximation

In this section we suppose that the arrival-rate function λ is constant and $E[S] < \infty$, so that the resulting stationary $M/G/\infty$ model will approach steady state as time evolves. If the termination time τ is relatively large, then the $M/G/\infty$ system can be regarded as approximately in steady state at the termination time τ, as indicated in Corollary 2.3.

In this case we can allow the termination time τ to be random, provided that it is independent of the history of the queueing system. We want to determine the distribution of T_∞, the remaining time after the arrival process terminates until the last departure.

We will first describe the steady-state behavior for the case of a pure-exponential tail, and show that steady-state approximations are often reasonable. Afterwards, we will consider the more problematic case of a power tail.

4.1 The case of a pure-exponential tail

Suppose that the service-time cdf G has a pure-exponential tail as in Theorem 3.1. Then, just as in that theorem, the stationary-excess cdf G_e also has a pure-exponential tail with the same decay rate η. To show that, we can apply the following basic lemma from p. 17 of Erdélyi [7].

Lemma 4.1 *If $f(x) \sim g(x)$ as $x \to \infty$, then*

$$\int_x^\infty f(u)\,du \sim \int_x^\infty g(u)\,du \quad \text{as } x \to \infty.$$

In particular, we can apply Lemma 4.1 to establish

Theorem 4.1 *If* $G^c(x) \sim \gamma e^{-\eta x}$ *as* $x \to \infty$, *then*

$$G_e^c(x) \sim \gamma_e e^{-\eta x} \quad \text{as } x \to \infty, \tag{4.1}$$

for $\gamma_e = \gamma/(\eta E[S])$.

As a consequence, we have

$$T_\infty \stackrel{\mathrm{d}}{=} M_N \approx \eta^{-1}[\log(N(\nu\gamma_e, \nu\gamma_e^2)) + W]. \tag{4.2}$$

Reasoning as in (3.7) and (3.8), for reasonably large $\lambda E[S]$, we get

$$E[T_\infty] \approx \eta^{-1}\left[\log(\lambda\gamma/\eta) + 0.5772\right] \tag{4.3}$$

and, assuming that $\nu = \lambda E[S]$ is suitably large,

$$\mathrm{Var}(T_\infty) \approx \frac{1.644 + (1/\nu)}{\eta^2} \approx \frac{1.644}{\eta^2}. \tag{4.4}$$

Further, we can approximate the normal random variable by its mean and get

$$T_\infty \approx \eta^{-1}\left[\log(\lambda\gamma/\eta) + W\right], \tag{4.5}$$

$$q_{T_\infty}(x) \approx \eta^{-1}\left[\log(\lambda\gamma/\eta) - \log\log(1/x)\right]. \tag{4.6}$$

We can state another heavy-traffic limit, which follows from the analysis above. The proof is essentially the same as for Theorem 3.2.

Theorem 4.2 *Consider a family of* $M/G/\infty$ *models in steady state, indexed by the constant arrival rate* λ. *Let* T_λ *denote* T_∞ *as a function of* λ. *If* $\lambda \to \infty$ *with* $G^c(x) \sim \gamma e^{-\eta x}$ *as* $x \to \infty$, *then* $\nu_\lambda = \lambda E[S] \to \infty$ *while* γ_e *in Theorem 4.1 remains unchanged*,

$$\eta\left[T_\lambda - \log(\lambda\gamma/\eta)\right] \Rightarrow W, \tag{4.7}$$

and the approximations in (4.5)–(4.6) *are asymptotically correct*.

4.2 Two-moment approximations

For the case of a pure-exponential tail, we can go further, drawing upon [3], and approximate the two asymptotic parameters η and γ by appropriate functions of the first two moments of the cdf G or, equivalently, of the mean $E[S]$ and the SCV c^2. For this purpose, we apply approximation (1.9)–(1.12) of [3], which is based on the hyperexponential distribution (H_2, mixture of two exponentials) for $c^2 \geq 1$ and the shifted-exponential distribution when $c^2 \leq 1$. Letting $c \equiv \sqrt{c^2}$, the approximation is

$$\frac{1}{\eta} \approx \begin{cases} E[S]c^2, & c^2 \geq 1, \\ E[S]c, & c^2 \leq 1, \end{cases} \tag{4.8}$$

and

$$\gamma \approx \psi(c^2) \equiv \begin{cases} \frac{c^2+1}{2(c^2)^2} \approx \frac{1}{c^2}, & c^2 \geq 1, \\ e^{\{(1-\sqrt{c^2})/\sqrt{c^2}\}} \approx \frac{1}{c}, & c^2 \leq 1. \end{cases} \tag{4.9}$$

Note that this approximation makes $\gamma < 1$ when $c^2 > 1$, but $\gamma > 1$ when $c^2 < 1$.

Applying the simple rough approximations for η and γ in (4.8) and (4.9), we obtain simple rough approximations for the distribution of T_∞ and its first moments in (4.2)–(4.5) that depend on only three parameters: λ, $E[S]$ and c^2. For example, for $c^2 \geq 1$, we can combine (4.3)–(4.6), (4.8) and (4.9) to get the following approximations

$$\begin{aligned} T_\infty &\approx E[S]c^2\left[\log(\lambda E[S]) + W\right], \\ E[T_\infty] &\approx E[S]c^2\left[\log(\lambda E[S]) + .5772\right], \\ \mathrm{Var}[T_\infty] &\approx 1.644(E[S]c^2)^2, \\ q_{T_\infty}(x) &\approx E[S]c^2\left[\log(\lambda E[S]) - \log\log(1/x)\right]. \end{aligned} \tag{4.10}$$

Corresponding approximations hold for $c^2 \leq 1$.

4.3 Transition to steady state with a pure-exponential tail

Given the convergence in distribution of T_τ to T_∞ established in Corollary 2.3, it would be natural to expect the approximations for the distribution of T_∞ to be the limit as $\tau \to \infty$ of the associated approximations for the distribution of T_τ. Indeed, that is the case with a pure-exponential tail, since $(\lambda\gamma/\eta)(1 - e^{-\eta\tau}) \to \lambda\gamma/\eta$ as $\tau \to \infty$. That shows that the approximation for the distribution of T_τ, as given in (3.5) and (3.6), transitions in the limit to the approximation for the distribution of T_∞, as given in (4.5). This fortunate state of affairs occurs because the two iterated limits $\lim_{\tau\to\infty}\lim_{x\to\infty}$ and $\lim_{x\to\infty}\lim_{\tau\to\infty}$ coincide when the service-time distribution has a pure-exponential tail. We will soon see that this property does not hold with a power tail.

4.4 The case of a power tail

We now turn to the case of a power tail in steady state, as in (3.14), which requires $\alpha > 1$ in order to have $E[S] < \infty$. Unlike for the cdf G_τ in Theorem 3.3, the stationary-excess cdf G_e has a power tail with a different exponent. Even though $G_\tau \to G_e$ as $\tau \to \infty$, the asymptotics are different. We again apply Lemma 4.1.

Theorem 4.3 *If the cdf G satisfies (3.14) with $\alpha > 1$, then*

$$G_e^c(x) \sim \gamma_e x^{-(\alpha-1)} \quad \text{as } x \to \infty, \tag{4.11}$$

for the stationary-excess cdf G_e in (2.4), where

$$\gamma_e = \frac{\gamma}{E[S](\alpha-1)}. \tag{4.12}$$

Henceforth assume that $\alpha > 1$. Now in steady state with a power-tail service-time distribution, paralleling (3.19), we can apply Theorem 4.3 to obtain the approximation $M_n \approx (\gamma_e n)^{1/(\alpha-1)} Y_{\alpha-1}$. Then, paralleling (3.22), we get approximation

$$T_\infty \stackrel{\mathrm{d}}{=} M_N \approx (\gamma_e N(\nu,\nu))^{1/(\alpha-1)}\, Y_{\alpha-1} \stackrel{\mathrm{d}}{=} N(\nu\gamma_e, \nu\gamma_e^2)^{1/(\alpha-1)} Y_{\alpha-1}. \tag{4.13}$$

Again reasoning as in (4.3) and (4.4), if the normal distribution can be approximated by its mean and the arrival rate is a constant λ (and $\alpha > 2$ for the mean formula and $\alpha > 3$ for the variance formula)

$$T_\infty \approx \left(\frac{\lambda\gamma}{\alpha-1}\right)^{\frac{1}{\alpha-1}} Y_{\alpha-1}, \tag{4.14}$$

$$E[T_\infty] \approx \left(\frac{\lambda\gamma}{\alpha-1}\right)^{\frac{1}{\alpha-1}} \Gamma\left(1-\frac{1}{\alpha-1}\right), \quad \alpha > 2, \tag{4.15}$$

and

$$\mathrm{Var}[T_\infty] \approx \left(\frac{\lambda\gamma}{\alpha-1}\right)^{\frac{1}{\alpha-1}} \left(\Gamma\left(1-\frac{2}{\alpha-1}\right) - \Gamma\left(1-\frac{1}{\alpha-1}\right)^2\right), \quad \alpha > 3. \tag{4.16}$$

If $\alpha \le 2$, then the steady-state mean is infinite; if $\alpha \le 3$, then the steady-state variance is infinite. Paralleling (3.27) and using the above assumptions, here we get the following approximation for the quantiles of the distribution of T:

$$q_{T_\infty}(x) \approx \left(\frac{\lambda\gamma}{(\alpha-1)\log(1/x)}\right)^{1/(\alpha-1)}. \tag{4.17}$$

We can state an analog of Theorems 3.4 and 4.2, which follows from the analysis above.

Theorem 4.4 *Consider a family of $M/G/\infty$ models in steady state, indexed by the constant arrival rate λ. Let T_λ denote T_∞ as a function of λ. If $\lambda \to \infty$ with fixed cdf G satisfying (3.14) with $\alpha > 1$, then $\nu_\lambda = \lambda E[S] \to \infty$ while γ_e in (4.12) remains unchanged,*

$$\frac{T_\lambda}{(\lambda\gamma/(\alpha-1))^{1/(\alpha-1)}} \Rightarrow Y_{\alpha-1}, \tag{4.18}$$

and the approximations in (4.13)–(4.17) are asymptotically correct (with the specified conditions on α).

Note that different scaling appears in Theorems 3.4 and 4.4. We thus should expect problems in the approximations for large τ.

4.5 The transition to steady state with a power tail

Unlike what we saw in Sect. 4.3, when we take the limit as $\tau \to \infty$ of the approximation for the transient distribution of T_τ as given in (3.24), we obtain $(\lambda\gamma\tau)^{1/\alpha} Y_\alpha \to$

∞ w.p.1. That clashes with the approximation for the steady-state distribution given in (4.14).

To better understand this phenomenon, we examine the two iterated limits $\lim_{\tau\to\infty}\lim_{x\to\infty}$ and $\lim_{x\to\infty}\lim_{\tau\to\infty}$ for the distribution of T_τ. We will assume that $\lambda(t)=\lambda$, a constant, and that the distribution of G is Pareto with shift parameter θ and exponent parameter α; specifically, let the service-time ccdf be the Pareto ccdf

$$G^c(x)\equiv\gamma(x+\theta)^{-\alpha},\quad x\geq 0, \tag{4.19}$$

where $\gamma=\theta^\alpha$. To have mean $E[S]=1$, we let $\theta=\alpha-1$. The associated SCV is $c^2=\alpha/(\alpha-2)$, for $\alpha>2$.

Using the form of T_τ given in (2.7), we will examine the two iterated limits with respect to

$$-\log\left(P(T_\tau\leq x)\right)=\lambda E[S]\left(G_e^c(x)-G_e^c(\tau+x)\right). \tag{4.20}$$

For the Pareto ccdf in (4.19),

$$G_e^c(x)=\left(\frac{1}{E[S]}\right)\left(\frac{\alpha-1}{x+\theta}\right)^{\alpha-1},\quad x\geq 0. \tag{4.21}$$

Hence,

$$\begin{aligned}&\lambda E[S]\left(G_e^c(x)-G_e^c(\tau+x)\right)\\&\quad=\lambda(\alpha-1)^{(\alpha-1)}\left(\left(\frac{1}{x+\theta}\right)^{\alpha-1}-\left(\frac{1}{x+\theta+\tau}\right)^{\alpha-1}\right) \qquad (4.22)\\&\quad=\lambda(\alpha-1)^{(\alpha-1)}\left(\frac{(x+\theta+\tau)^{\alpha-1}\ -\ (x+\theta)^{\alpha-1}}{(x+\theta)^{\alpha-1}(x+\theta+\tau)^{\alpha-1}}\right). \qquad (4.23)\end{aligned}$$

Then, letting $r\equiv r(x,\tau,\theta)\equiv(x+\theta+\tau)/(x+\theta)$, we get

$$\lambda E[S]\left(G_e^c(x)-G_e^c(\tau+x)\right)=\lambda\left(\frac{\alpha-1}{x+\theta}\right)^{\alpha-1}\left(1-r^{-(\alpha-1)}\right). \tag{4.24}$$

There are now three cases to consider:

First, if $x\gg\tau\gg 0$, which corresponds to the transient case and the iterated limit $\lim_{\tau\to\infty}\lim_{x\to\infty}$, then

$$1-r^{-(\alpha-1)}=1-\left(1+\frac{\tau}{x+\theta}\right)^{-(\alpha-1)}\approx 1-e^{-\frac{\tau(\alpha-1)}{x+\theta}}\approx\frac{\tau(\alpha-1)}{x+\theta} \tag{4.25}$$

and

$$\lambda E[S](G_e^c(x)-G_e^c(\tau+x))\approx\lambda\tau\left(\frac{\alpha-1}{x+\theta}\right)^{\alpha}\approx\lambda\tau\left(\frac{\alpha-1}{x}\right)^{\alpha}, \tag{4.26}$$

implying that $T_\tau\approx(\alpha-1)(\lambda\tau)^{1/\alpha}Y_\alpha$, which coincides with the transient approximation given in (3.24).

Second, if $\tau \gg x \gg 0$, which corresponds to the steady-state case and the iterated limit $\lim_{x\to\infty}\lim_{\tau\to\infty}$, then $1 - r^{-(\alpha-1)} \approx 1$ and

$$\lambda E[S](G_e^c(x) - G_e^c(\tau + x)) \approx \lambda \left(\frac{\alpha - 1}{x + \theta}\right)^{\alpha-1} \approx \lambda \left(\frac{\alpha - 1}{x}\right)^{\alpha-1}, \tag{4.27}$$

implying that

$$T_\tau \approx (\alpha - 1)\lambda^{1/(\alpha-1)} Y_{\alpha-1}, \tag{4.28}$$

which coincides with the steady-state approximation given in (4.14). Notice that formulas (4.26) and (4.27) differ in two ways: Formula (4.26) has an extra factor τ and a larger exponent on the $(\alpha - 1)/x$ term. The τ factor causes T_τ to explode as $\tau \to \infty$.

Indeed, when we actually are in steady state, we have an exact relation for the distribution of T_∞ associated with this Pareto distribution, paralleling Corollary 2.1. Combining Corollary 2.3 and (4.21), we obtain the steady-state formulas

$$P(T_\infty \le x) = e^{-\lambda((\alpha-1)/(x+\theta))^{(\alpha-1)}}, \quad x \ge 0, \tag{4.29}$$

so that

$$P(T_\infty + \theta \le x) = P(T_\infty \le x - \theta) = e^{-\lambda((\alpha-1)/x)^{(\alpha-1)}}, \quad x \ge 0,\ x \ge \theta, \tag{4.30}$$

and

$$P\left((\alpha - 1)\lambda^{1/(\alpha-1)} T_\infty + \theta \le x\right) = P(Y_{\alpha-1} \le x) \equiv e^{-x^{-(\alpha-1)}}, \quad x \ge \theta. \tag{4.31}$$

By (4.30), we see that the asymptotic steady-state approximation for higher quantiles of T will exceed the exact values by exactly θ, provided that the system is indeed in steady state. (That is borne out by the numerical examples; e.g., see Table 3.)

Finally, there is the third case in which $\tau \approx x \gg 0$, where neither of the above approximations would be appropriate. In other words, the transient extreme value approximations are only appropriate when x is very large relative to τ, and the steady-state approximations are only appropriate when τ is very large relative to x. The third case represents an in-between zone, where we would expect both our transient and steady-state approximations to fail.

As reasonable overall approximations for the moments and quantiles of T_τ, we can take the *minimum* of the transient and steady-state approximations. This modified approximation is supported by Corollary 2.3, which shows that the steady-state time T_∞ is always a stochastic upper bound for the transient time T_τ. Moreover, we know that the transient approximations diverge to infinity, so that they inevitably become inaccurate. So we should disregard the transient approximation when it exceeds the steady-state approximation. From (3.27) and (4.17), we see that the two approximations will coincide for the quantiles for the value of τ satisfying the equation

$$q_{T_\tau}(x) \approx \left(\frac{\lambda\gamma\tau}{\log(1/x)}\right)^{1/\alpha} = \left(\frac{\lambda\gamma}{(\alpha - 1)\log(1/x)}\right)^{1/(\alpha-1)} \approx q_{T_\infty}(x), \tag{4.32}$$

where $\gamma = (\alpha - 1)^{\alpha}$, yielding the *matching time*

$$\tau^* = \left(\frac{\lambda}{\log(1/q)}\right)^{1/(\alpha-1)}. \tag{4.33}$$

This matching time τ^* provides a rough approximation for the time that the system can be considered to have reached steady state with respect to a given quantile. The simple formula clearly shows the dependence upon λ, x and α.

We now proceed to examine the third case more carefully. To do so, assume that $y = b\tau$ for some positive constant b, and then let $\tau \to \infty$. Then we have

$$r(y, \tau, \theta) \equiv r(\tau, \theta) = \frac{(b+1)\tau + \theta}{b\tau + \theta} \to \frac{b+1}{b} \quad \text{as } \tau \to \infty. \tag{4.34}$$

Thus, from (4.24), we see that

$$\lambda E[S](G_e^c(y) - G_e^c(\tau + y)) \to \lambda \left(\frac{\alpha - 1}{y + \theta}\right)^{(\alpha-1)} (1 - (b/(b+1))^{\alpha-1}) \quad \text{as } \tau \to \infty. \tag{4.35}$$

Combining (4.20) and (4.35), we obtain an approximation for $P(T \le y)$. Reasoning as in (4.27), we obtain

$$\begin{aligned} \lambda E[S](G_e^c(y) - G_e^c(\tau + y)) &\approx \lambda \left(1 - \left(\frac{b}{b+1}\right)^{(\alpha-1)}\right) \left(\frac{\alpha - 1}{y + \theta}\right)^{\alpha-1}, \\ &\approx \lambda \left(1 - \left(\frac{b}{b+1}\right)^{(\alpha-1)}\right) \left(\frac{\alpha - 1}{y}\right)^{\alpha-1}, \end{aligned} \tag{4.36}$$

implying that

$$T \approx (\alpha - 1)\lambda^{1/(\alpha-1)} \left(1 - \left(\frac{b}{b+1}\right)^{(\alpha-1)}\right)^{\frac{1}{\alpha-1}} Y_{\alpha-1}. \tag{4.37}$$

Paralleling (4.17), we get the following quantile approximation

$$\begin{aligned} q_T(x) &\approx (\alpha - 1) \left(\frac{\lambda}{\log(1/x)}\right)^{1/(\alpha-1)} \left(1 - (b/(b+1))^{(\alpha-1)}\right)^{1/(\alpha-1)} \\ &= \left(1 - (b/(b+1))^{(\alpha-1)}\right)^{1/(\alpha-1)} q_{T_\infty}(x). \end{aligned} \tag{4.38}$$

In this third limiting regime, the approximate quantile $q_T(x)$ is a fixed fraction of the steady-state quantile $q_{T_\infty}(x)$ depending on b and α. The steady-state quantile is approached as b decreases toward 0. We can expect this approximation depending upon b to perform well for quantiles $q_T(x) = b\tau$ when τ is suitably large.

5 Numerical comparisons

We now see how the various approximations perform by making comparisons with exact values from Sect. 2. We only consider the case of a homogeneous Poisson arrival process. We consider values of τ for which the queue can be considered to be in steady state at time τ and values for which it cannot. We first consider the case of a pure-exponential tail and then a power tail. In all cases, we normalize so that $E[S] = 1$. We performed extensive calculations for several service-time distributions over a wide range of arrival rates λ and terminal times τ. Extensive results appear in [8]; we present the highlights here.

All exact computations were performed with *Mathematica* using the formulas in Sect. 2. We use binary search with the exact cdf in (2.5) in order to numerically calculate the exact quantiles of the distribution of T, just as in [3]. These values were also verified through simulation in many cases. (For large values of ν_τ, it became prohibitively expensive to perform high-accuracy simulations.)

For all distributions with a pure-exponential tail, we considered 7 values of λ and 9 values of τ, yielding $7 \times 9 = 63$ cases in all. The 7 values of λ are 2^{2n-1} for $n = 0, 1, \ldots, 6$, while the 9 values of τ are 2^{n-1} for $n = 0, 1, \ldots, 8$. For all distributions with a power tail, we considered 6 values of λ and 10 values of τ, yielding $6 \times 10 = 60$ cases in all. The 6 values of λ are 2^{4n-1} for $n = 0, 1, \ldots, 5$, while the 10 values of τ are 2^{3n-1} for $n = 0, 1, \ldots, 9$. We let both λ and τ vary over a wider range for the power-tailed distributions. All numerical values are given to four significant digits.

5.1 A pure-exponential tail

We considered examples with $c^2 = 1$, $c^2 = 0.6 < 1$ and $c^2 = 4.0 > 1$. For $c^2 = 1$, we consider an exponential distribution, as in Sect. 2 of [3]. For $c^2 = 0.6$, we consider a hypoexponential distribution, i.e., the convolution of two exponential distributions with different means, as in Sect. 6 of [3]. Specifically, the service-time ccdf is

$$G^c(x) = 1.618e^{-1.38313x} - 0.621e^{-3.61011x}, \quad x \geq 0. \tag{5.1}$$

For $c^2 = 4$, we consider an H_2 distribution, which has ccdf

$$G^c(x) = pe^{-\eta x} + (1-p)e^{-\delta x}, \quad x \geq 0. \tag{5.2}$$

The three parameters η, δ and p in (5.2) are chosen so that the mean is 1, the SCV is $c^2 = 4.0$ and the proportion of the mean provided by the dominant exponential component (with rate η, where $\eta < \delta$) is

$$r = \frac{p/\eta}{(p/\eta) + ((1-p)/\delta)} = \frac{p}{\eta}. \tag{5.3}$$

Formulas relating different parameterizations are given in [3]. In this case, $G^c(x) \sim \gamma e^{-\eta x}$ as $x \to \infty$ with $\gamma = p$. We consider H_2 distributions with three possible values of r: $r = 0.25$, $r = 0.50$ and $r = 0.75$, as in Sect. 4 of [3].

5.1.1 Exponential service times

Consistent with Corollary 2.1, the results for the exponential distribution are spectacular; results are shown in [8]. For the transient approximation, the only error is due to the probability $P(T = 0) = e^{-\nu_\tau}$. When $\nu_\tau = 8$, $P(T = 0) = 0.0003$, so that for $\nu_\tau \geq 8$, $P(T = 0)$ is negligible and the transient approximation is accurate to four significant digits. All approximations break down when ν_τ, the mean number in queue at time τ, is too small. The breakdown point occurs approximately at $\nu_\tau = 1$. When $\nu_\tau < 1$, the approximations are often negative.

When $\tau \geq 8$, the $M/G/\infty$ queue can be regarded as being in steady state; as anticipated from (2.15). Then the steady-state approximation agrees with the transient approximation. For smaller values of τ, especially for $\tau \leq 2$, the transient approximation is significantly better than the steady-state approximation, as we would expect.

5.1.2 Hypoexponential service times

Results for the hypoexponential distribution with $c^2 = 0.6$ are displayed in Table 1. We show results for 4 values of λ: 2, 8, 32 and 2048, and 2 values of τ: 0.5 and 8.0. We display the mean and variance and three quantiles: $q_T(0.50)$, $q_T(0.95)$, and $q_T(0.9999)$.

For the two higher quantiles, $q_T(0.95)$, and $q_T(0.9999)$, the transient approximation performs well for all these cases. The only poor performance of the transient approximation for other characteristics occurs for $\lambda = 2$ and $\tau = 0.5$, where $\nu_\tau = 0.881 < 1$. For $\tau = 8.0$, the steady-state approximations agree with the transient approximations, with the exception of the variance when $\lambda = 2.0$. For almost all cases with $\tau = 8$, the approximations are yielding accuracy to all four significant digits displayed.

Overall, we conclude that, for service-time distributions with a pure-exponential tail and $c^2 \leq 1$, the transient approximation should perform well provided that $\nu_\tau \geq 1$. Moreover, the steady-state approximation ought to agree with the transient approximation for $\tau \geq 8$. As with the exponential distribution, the transient approximation should be significantly better than the steady-state approximation when $\tau \leq 2$.

Finally, we discuss the steady-state two-moment approximation presented in Sect. 4.2. The two-moment steady-state approximation also performs quite well in this case, consistently producing errors less than 10%. As in [3], we conclude that it performs quite well for $c^2 \leq 1$ when c^2 does not differ greatly from 1.

5.1.3 Hyperexponential service times

We display results for H_2 service times with $c^2 = 4.0$ and $r = 0.5$ in Table 2. We show results for 3 values of λ: 8, 128 and 2048, and 3 values of τ: 1.0, 8.0 and 32.0. We consider larger values of τ in Table 2 than in Table 1 because it takes the system longer to reach steady state with the H_2 service-time distribution. Referring to (2.15), we see that the "mean time to approach steady state," $E[S_e] = (c^2 + 1)/2 = 2.5$ here, compared to $E[S_e] = (c^2 + 1)/2 = 0.8$ for the hypoexponential service-time distribution considered above.

Table 1 A comparison of approximations with exact values for the characteristics of the distribution of T, the remaining time until the last departure. The service distribution G is hypoexponential, the convolution of two exponentials, having mean $E[S] = 1$ and $c^2 = 0.6$; the ccdf is in (5.1). The arrival rate is a constant λ. The arrival process is turned off at time τ. The key model parameters ν_τ and γ_τ are as given in formulas (2.1) and (3.2). The transient approximations for the mean, variance and quantiles are given in (3.7), (3.8) and (3.12). The steady-state approximations for the mean, variance and quantiles are given in (4.3), (4.4) and (4.6). The corresponding two-moment steady-state approximations for the mean, variance and quantiles are given in (4.3)–(4.6) with (4.8) and (4.9)

Hypoexponential service-time distribution with $c^2 = 0.6$								
Model parameters								
λ	2.0	2.0	8.0	8.0	32	32	2048	2048
τ	0.5	8.0	0.5	8.0	0.5	8.0	0.5	8.0
ν_τ	0.881	1.996	3.522	7.982	14.09	31.93	901.6	2043
γ_τ	1.326	1.172	1.326	1.172	1.326	1.172	1.326	1.172
$\nu_\tau \gamma_\tau$	1.168	2.340	4.672	9.358	18.69	37.43	1196	2396
$P(T = 0) = e^{-\nu_\tau}$	0.415	0.136	0.030	0.000	0.000	0.000	0.000	0.000
Performance measures								
$E[T]$	0.612	1.033	1.511	2.030	2.532	3.036	5.541	6.043
Transient approx.	0.530	1.032	1.532	2.034	2.534	3.036	5.541	6.043
Steady st. approx.	1.032	1.032	2.034	2.034	3.036	3.036	6.043	6.043
Steady st. 2-mt.	0.984	0.984	2.058	2.058	3.132	3.132	6.353	6.353
Var[T]	0.669	0.831	0.896	0.868	0.864	0.861	0.860	0.860
Trans. & steady st.	0.860	0.860	0.860	0.860	0.860	0.860	0.860	0.860
Steady st. 2-mt.	0.987	0.987	0.987	0.987	0.987	0.987	0.987	0.987
$q_T(0.50)$	0.273	0.864	1.371	1.880	2.381	2.884	5.389	5.891
Transient approx.	0.377	0.880	1.380	1.882	2.382	2.884	5.389	5.891
Steady st. approx.	0.880	0.880	1.882	1.882	2.884	2.884	5.891	5.891
Steady st. 2-mt.	0.821	0.821	1.895	1.895	2.968	2.968	6.190	6.190
$q_T(0.95)$	2.259	2.762	3.262	3.764	4.264	4.767	7.271	7.773
Transient approx.	2.260	2.762	3.262	3.764	4.264	4.767	7.271	7.773
Steady st. approx.	2.762	2.762	3.764	3.764	4.767	4.767	7.773	7.773
Steady st. 2-mt.	2.838	2.838	3.911	3.911	4.985	4.985	8.207	8.207
$q_T(0.9999)$	6.771	7.274	7.774	8.276	8.776	9.278	11.78	12.29
Transient approx.	6.771	7.274	7.774	8.276	8.776	9.278	11.78	12.29
Steady st. approx.	7.274	7.274	8.276	8.276	9.278	9.278	12.29	12.29
Steady st. 2-mt.	7.671	7.671	8.745	8.745	9.819	9.819	13.04	13.04

As before, we display the mean, the variance and the three quantiles: $q_T(0.50)$, $q_T(0.95)$, and $q_T(0.9999)$. As in the previous two examples, the transient approximation for the high quantiles is accurate in all cases to the full four significant digits displayed. The steady-state approximation differs significantly from the transient approximation for $\tau = 1$, differs only slightly for $\tau = 8.0$ and is identical (to the four digits displayed) for $\tau = 32.0$.

Table 2 A comparison of approximations with exact values for the characteristics of the distribution of T, the remaining time until the last departure. The service distribution G is hyperexponential, the mixture of two exponentials, having mean $E[S] = 1$, $c^2 = 4.0$ and $r = 0.5$; the ccdf is in (5.2). The arrival rate is a constant λ. The arrival process is turned off at time τ. The key model parameters ν_τ and γ_τ are as given in formulas (2.1) and (3.2). The transient approximations for the mean, variance and quantiles are given in (3.7), (3.8) and (3.12). The steady-state approximations for the mean, variance and quantiles are given in (4.3), (4.4) and (4.6). The corresponding two-moment steady-state approximations for the mean, variance and quantiles are given in (4.10)

Hyperexponential service-time distribution with $c^2 = 4.0$ and $r = 0.5$									
Model parameters									
λ	8.0	8.0	8.0	128.0	128	128	2048	2048	2048
τ	1.0	8.0	32.0	1.0	8.0	32.0	1.0	8.0	32.0
ν_τ	4.129	7.341	7.997	66.06	117.5	128.0	1057	1879	2047
γ_τ	0.196	0.455	0.500	0.196	0.455	0.500	0.196	0.455	0.500
$\nu_\tau \gamma_\tau$	0.807	3.341	3.997	12.92	53.46	63.95	206.6	855.3	1023
$P(T = 0) = e^{-\nu_\tau}$	0.016	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Performance measures									
$E[T]$	3.479	8.021	8.767	13.91	20.21	21.01	26.21	32.51	33.31
Transient approx.	1.611	7.912	8.708	13.91	20.21	21.01	26.21	32.51	33.31
Steady st. approx.	8.711	8.711	8.711	21.01	21.01	21.01	33.31	33.31	33.31
Steady st. 2-mt.	10.63	10.63	10.63	21.72	21.72	21.72	32.81	32.81	32.81
Var[T]	17.64	30.73	31.41	32.36	32.38	32.38	32.38	32.38	32.38
Trans. & steady st.	32.38	32.38	32.38	32.38	32.38	32.38	32.38	32.38	32.38
Steady st. 2-mt.	26.32	26.32	26.32	26.32	26.32	26.32	26.32	26.32	26.32
$q_T(0.50)$	1.749	6.978	7.773	12.98	19.28	20.07	25.28	31.58	32.37
Transient approx.	0.676	6.978	7.773	12.98	19.28	20.07	25.28	31.58	32.37
Steady st. approx.	7.776	7.776	7.776	20.08	20.08	20.08	32.38	32.38	32.38
Steady st. 2-mt.	9.784	9.784	9.784	20.87	20.87	20.87	31.96	31.96	31.96
$q_T(0.95)$	12.23	18.53	19.32	24.53	30.83	31.62	36.83	43.13	43.93
Transient approx.	12.23	18.53	19.32	24.53	30.83	31.62	36.83	43.13	43.93
Steady st. approx.	19.33	19.33	19.33	31.63	31.63	31.63	43.93	43.93	43.93
Steady st. 2-mt.	20.20	20.20	20.20	31.29	31.29	31.29	42.38	42.38	42.38
$q_T(0.9999)$	39.91	46.21	47.01	52.21	58.51	59.31	64.51	70.81	71.61
Transient approx.	39.91	46.21	47.01	52.21	58.51	59.31	64.51	70.81	71.61
Steady st. approx.	47.01	47.01	47.01	59.31	59.31	59.31	71.61	71.61	71.61
steady st. 2-mt.	45.16	45.16	45.16	56.25	56.25	56.25	67.34	67.34	67.34

Overall, poor performance is only seen in the minimal case with $\lambda = 8.0$ and $\tau = 1.0$. That can be explained by the product $\nu_\tau \gamma_\tau = 0.807 < 1$. The important role of $\nu_\tau \gamma_\tau$ for the hyperexponential distribution is discussed in [3]. (From (2.16) we see that G_τ is indeed an H_2 cdf when G is H_2 and $\lambda(t) = \lambda$.) In the asymptotic extreme-value approximation for the maximum of hyperexponential random variables, all but the dominant term of the hyperexponential mixture (that component

exponential with greatest mean) are ignored. Viewing each hyperexponential service time as being drawn from each component distribution with some probability, $\nu_\tau \gamma_\tau$ represents the expected number of service times in the maxima computation that have the dominant distribution. If the dominant component occurred too infrequently, we would expect the asymptotic approximation to perform badly, since extreme value theory does not hold for maxima of very few random variables. In particular, experience indicates that significant difficulties occur if $\nu_\tau \gamma_\tau \leq 1$. (This same criterion applies to the exponential distribution, because then $\nu_\tau \gamma_\tau = \nu_\tau$, since $\gamma_\tau = 1$.)

Again the two-moment steady-state approximation performs quite well, but it is tuned to perform well for r near 0.5. Indeed, our other tables show that the performance for H_2 service times with $c^2 = 4.0$ depends strongly on the third parameter r, just as in [3]. The main transient and steady-state approximations apply for these other values of r, just as for $r = 0.5$, but the two-moment approximations remain unchanged, and thus perform quite badly for those other values of r. For example, for $\lambda = 8.0$ and $\tau = 32$, $q_T(0.9999) = 75.88$, 47.01 and 35.48 for $r = 0.25$, 0.50 and 0.75, respectively. The two-moment approximation $q_T(0.9999) \approx 45.2$ serves for all three. It should be recognized that the two-moment approximations are only rough approximations, comparable to what you would get if you used an H_2 distribution with $r = 0.5$ as an approximation for some other H_2 distribution.

5.2 A power tail

To illustrate a power tail, we consider the Pareto ccdf in (4.19). To have mean $E[S] = 1$, we let $\theta = \alpha - 1$. The associated SCV is $c^2 = \alpha/(\alpha - 2)$, for $\alpha > 2$. We considered 3 examples with this Pareto distribution having parameter triples $(\alpha, \theta, c^2) = (8, 7, 1.33)$, $(4, 3, 2.00)$ and $(2.5, 1.50, 5.0)$. For the case $(2.5, 1.50, 5.0)$, a finite steady-state variance, $\mathrm{Var}(T_\infty)$, does not exist since $\alpha = 2.5 < 3$. We display a subset of these results for the cases $(4, 3, 2.00)$ and $(2.5, 1.5, 5.0)$ in Tables 3 and 4.

Due to slower convergence to steady state and the necessity of larger λ for the approximations to be highly accurate, λ and τ must be allowed to take much larger values than in the exponential-tail setting. In Table 3 we consider 3 values of λ: $2^3 = 8$, $2^{11} = 2048$ and $2^{15} = 32,770$ and 3 values of τ; $2^2 = 4$, $2^{11} = 2048$ and $2^{17} \approx 1.31E5 \equiv 1.31 \times 10^5$. Table 4 takes a different view, considering a very high arrival rate, but 8 different values for the termination time τ, ranging from 0.5 to $8.389E6 \equiv 8.389 \times 10^6$.

As noted in Sect. 4.5, the approximations for $E[T]$ and the quantiles of T are actually approximations for $T + \theta$, which arises since the asymptotics approximate $1/(x + \theta)$ by $1/x$. As λ goes to infinity, the approximations are nevertheless asymptotically correct, since θ is just some constant. Thus the ratio of the approximation to the true value still goes to 1. Once we make this adjustment, we see greater accuracy for Pareto service times.

The main observation from Table 3 is that we do not get the consistent high accuracy across almost all cases that we saw with the distributions having a pure-exponential tail. The story is very different for a power tail. A principal problem is the inconsistency between the transient and steady-state approximations discussed

Table 3 A comparison of approximations with exact values for the characteristics of the distribution of T, the remaining time until the last departure. The service distribution G is Pareto with mean $\mathrm{E}S = 1$, $c^2 = 2.0$, $\alpha = 4.0$ and $\theta = 3.0$. The arrival rate is a constant λ. The arrival process is turned off at time τ. The key model parameters ν_τ and γ_τ are as given in Theorem 3.3. The transient approximations for the mean, variance and quantiles are given in (3.25), (3.26) and (3.27). The steady-state approximations for the mean, variance and quantiles are given in (4.15), (4.16) and (4.17). The approximation for the quantiles depending on the parameter b are given in (4.38)

Pareto service-time distribution with $\alpha = 4.0$ and $c^2 = 2.0$									
Model parameters									
λ	8.0	8.0	8.0	2048	2048	2048	32,770	32,770	32,770
τ	4.0	2048.0	1.311E5	4.0	2048	1.311E5	4.0	2048	1.311E5
ν_τ	7.370	8.000	8.000	1887	2048	2048	3.02E4	3.28E4	3.28E4
γ_τ	351.7	1.66E5	1.06E7	351.7	1.66E5	1.06E7	351.7	1.66E5	1.06E7
$\nu_\tau \gamma_\tau$	2592	1.33E6	8.49E7	6.64E5	3.40E8	2.17E10	1.06E7	5.44E9	3.48E11
$P(T=0) = e^{-\nu_\tau}$	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Performance measures									
E[T]	4.143	5.125	5.125	30.08	48.58	48.59	65.00	126.9	127.0
Trans. approx.	8.744	41.59	117.6	34.97	166.4	470.61	69.95	332.7	941.1
Stdy. st. approx.	8.125	8.125	8.125	51.59	51.59	51.59	130.0	130.0	130.0
Var[T]	13.14	30.33	30.43	219.9	1201	1226	881.7	7385	7784
Trans. approx.	13.79	312	2496	220.6	4992	3.99E4	882.4	2.00E4	1.60E5
Stdy. st. approx	30.43	30.43	30.43	1227	1227	1227	7790	7790	7790
$q_T(0.50)$	3.226	3.780	3.780	26.39	40.05	40.05	57.61	105.5	105.5
Trans. approx.	7.820	37.2	105.2	31.28	148.8	420.8	62.56	297.6	841.7
Stdy. st. approx.	6.780	6.78	6.78	43.05	43.05	43.05	108.5	108.5	108.5
$b = 0.1$	6.778	6.778	6.778	43.04	43.04	43.04	108.5	108.5	108.5
$b = 1.0$	6.485	6.485	6.485	41.18	41.18	41.18	103.8	103.8	103.8
$b = 10.0$	4.264	4.264	4.264	27.07	27.07	27.07	68.23	68.23	68.23
$q_T(0.95)$	10.21	13.15	13.15	55.03	99.53	99.54	115.0	255.3	255.4
Trans. approx.	14.99	71.32	201.7	59.97	285.3	806.9	119.9	570.6	1614
Stdy. st. approx.	16.15	16.15	16.15	102.5	102.5	102.5	258.4	258.4	258.4
$b = 0.1$	16.15	16.15	16.15	102.5	102.5	102.5	258.3	258.3	258.3
$b = 1.0$	15.45	15.45	15.45	98.04	98.04	98.04	247.2	247.2	247.2
$b = 10.0$	10.16	10.16	10.16	64.46	64.46	64.46	162.5	162.5	162.5
$q_T(0.9999)$	66.4	126.3	126.3	280.4	811.4	817.8	565.8	1980	2065
Trans. approx.	71.35	339.4	960	285.4	1358	3840	570.8	2715	7680
Stdy. st. approx.	129.3	129.3	129.3	820.8	820.8	820.8	2068	2068	2068
$b = 0.1$	129.3	129.3	129.3	820.6	820.6	820.6	2068	2068	2068
$b = 1.0$	123.7	123.7	123.7	785.1	785.1	785.1	1978	1978	1978
$b = 10.0$	81.31	81.31	81.31	516.2	516.2	516.2	1300	1300	1300

Table 4 A comparison of approximations with exact values for the characteristics of the distribution of T, the remaining time until the last departure. The service distribution G is Pareto with mean $ES = 1$, $c^2 = 5.0$, $\alpha = 2.5$ and $\theta = 1.5$. The arrival rate is held fixed at a constant $\lambda = 5.243 \times 10^5$. The arrival process is turned off at time τ for 8 values of τ over a broad range, from 0.5 to 8.389×10^6. The key model parameters ν_τ and γ_τ are as given in Theorem 3.3. The transient approximations for the mean, variance and quantiles are given in (3.25), (3.26) and (3.27). The steady-state approximations for the mean, variance and quantiles are given in (4.15), (4.16) and (4.17). For each performance measure, an asterisk marks the smallest transient approximation value that is less than the corresponding steady-state approximation value

Pareto service-time distribution with $\alpha = 2.5$ and $c^2 = 5.0$								
Model parameters								
λ	5.243E5	5.243E5	5.243E5	5.243E5	5.243E5	5.243E5	5.243E5	5.243E5
τ	0.5	4.0	256.0	2048	1.638E4	1.311E5	1.049E6	8.389E6
ν_τ	1.838E5	4.496E5	5.241E5	5.243E5	5.243E5	5.243E5	5.243E5	5.243E5
γ_τ	3.931	12.85	705.8	5644	4.515E4	3.612E5	2.890E6	2.312E7
$\nu_\tau \gamma_\tau$	7.224E5	5.779E6	3.699E8	2.959E9	2.367E10	1.894E11	1.515E12	1.212E13
$P(T = 0)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Performance measures								
E[T]	326.7	751.1	3856	8211	1.502E4	2.115E4	2.427E4	2.546E4
Tran. ap.	328.4	754.6	3983	9150	2.102E4*	4.829E4	1.109E5	2.549E5
St. ap.	2.613E4	2.613E4	2.613E4	2.613E4	2.613E4	2.613E4	2.613E4	2.613E4
Var[T]	1.154E5	6.093E5	1.697E7	8.920E7	4.530E8	1.974E9	6.918E9	2.128E10
Tran. ap.	1.154E5	6.093E5	1.697E7	8.958E7	4.728E8	2.496E9	1.317E10	6.952E10
St. ap.	∞	∞	∞	∞	∞	∞	∞	∞
$q_T(0.50)$	253.6	583.2	2970	6174	1.036E4	1.224E4	1.244E4	1.245E4
Tran. ap.	255.4	586.7	3097	7114*	1.634E4	3.755E4	8.626E4	1.982E5
St. ap.	1.245E4	1.245E4	1.245E4	1.245E4	1.245E4	1.245E4	1.245E4	1.245E4
$b = 0.1$	1.222E4	1.222E4	1.222E4	1.222E4	1.222E4	1.222E4	1.222E4	1.222E4
$b = 1.0$	9308	9308	9308	9308	9308	9308	9308	9308
$b = 10.0$	3247	3247	3247	3247	3247	3247	3247	3247
$q_T(0.95)$	721.8	1659	8645	1.916E4	3.895E4	6.182E4	6.911E4	7.061E4
Tran. ap.	723.6	1662	8774	2.016E4	4.631E4*	1.064E5	2.444E5	5.615E5
St. ap.	7.065E4	7.065E4	7.065E4	7.065E4	7.065E4	7.065E4	7.065E4	7.065E4
$b = 0.1$	6.935E4	6.935E4	6.935E4	6.935E4	6.935E4	6.935E4	6.935E4	6.935E4
$b = 1.0$	5.282E4	5.282E4	5.282E4	5.282E4	5.282E4	5.282E4	5.282E4	5.282E4
$b = 10.0$	1.843E4	1.843E4	1.843E4	1.843E4	1.843E4	1.843E4	1.843E4	1.843E4
$q_T(0.9999)$	8778	2.017E4	1.063E5	2.436E5	5.538E5	1.227E6	2.495E6	3.961E6
Tran. ap.	8780	2.017E4	1.065E5	2.446E5	5.619E5	1.291E6	2.966E6*	6.814E6
St. ap.	4.527E6	4.527E6	4.527E6	4.527E6	4.527E6	4.527E6	4.527E6	4.527E6
$b = 0.1$	4.444E6	4.444E6	4.444E6	4.444E6	4.444E6	4.444E6	4.444E6	4.444E6
$b = 1.0$	3.385E6	3.385E6	3.385E6	3.385E6	3.385E6	3.385E6	3.385E6	3.385E6
$b = 10.0$	1.181E6	1.181E6	1.181E6	1.181E6	1.181E6	1.181E6	1.181E6	1.181E6

in Sect. 4.5. That section is important background for interpreting Table 3. As suggested there, we do indeed obtain a much better overall approximation for moments and quantiles if we use the minimum of the transient and steady-state approximations.

Even for the very large values of λ and τ considered in Table 3, we have only limited accuracy. But if we look closely, we see that the accuracy is not too bad. First, the steady-state approximation is consistently accurate for the two larger times $\tau = 2048$ and $\tau = 1.311E5$, consistent with (4.29). In most cases the error in the steady-state approximation is precisely the shift $\theta = 3.0$. After that adjustment, the steady-state approximations in these cases are mostly accurate to the four displayed significant digits.

In Table 3, the performance of the transient approximation is significantly worse than the steady-state approximation, even in its best regions. However, the transient approximation is actually reasonably accurate for the shorter time $\tau = 4.0$. For very small τ, such as $\tau = 0.5$, there is an error of almost exactly $\theta = 3.0$, just as for the steady-state approximation, but this error grows as τ increases. For $\tau = 4.0$, we see an absolute error of about 5 in all cases. For the extremely large values of λ we consider, this nearly constant error of about 5 at $\tau = 4.0$ becomes relatively negligible.

With respect to the time at which steady-state is reached, we can still use the relation (2.15). The cdf G_e will show that the rate of approach to equilibrium is indeed much slower now. As a rough estimate, we can use the time τ^* in (4.33) at which the transient approximation equals the steady-state approximation. We see the advantage of the special approximation for quantiles in (4.38) in one case: When $\tau = 2048$ and $b = 1$, we have the approximation $q_T(0.9999) \approx 1978 \approx b\tau = 2048$; consistent with that, the exact value there is 1980.

Table 4 considers the most variable Pareto distribution (of the four we consider) with $\alpha = 2.5$. Here the remaining time until the last departure in steady-state, T_∞, fails to have a finite variance. Accordingly, we see the transient variances steadily increasing without bound. In this case, the approach to steady state is very slow. We see that the lower quantile $q_T(0.5)$ has reached steady-state for the last three values of τ, but the highest quantile $q_T(0.9999)$ has not reached steady state even by the final termination time $\tau = 8.4 \times 10^6$.

In this view, including very large λ and smaller τ, we see that the transient approximation looks very good in many cases, while the steady-state approximation does not. For the smallest value of τ, $\tau = 0.5$, the error in the transient approximation is very close to the shift $\theta = 1.5$. As τ increases, the error in the transient approximation increases, but here the relative accuracy is consistently good.

6 Conclusions

Motivated by the two-stage inspection problem, we studied the remaining time T after the arrival-process termination time τ until the last departure in the $M_t/G/\infty$ queue. Formula (2.3) shows that T can be represented as the maximum of a Poisson random number of i.i.d. random variables each distributed as the cdf G_τ in (2.2), where the Poisson random number has mean ν_τ in (2.1). As a consequence, we were able to give explicit expressions for the cdf and the quantiles of the distribution of T

in Theorem 2.2. That result implies the important Corollaries 2.1, 2.2 and 2.3. They give convenient explicit expressions for the distribution of T for exponential and hyperexponential service-time cdf's and establish that the steady-state cases serve as limiting upper bounds for the corresponding transient cases in the stochastic-order sense.

Most of the paper went beyond the exact relations in Sect. 2 to study approximations. We applied asymptotic methods to develop approximations for the distribution and characteristics of T for service-time cdf's G that (i) have an exponential tail and (ii) have a power tail. In Sects. 3 and 4 we considered the transient and steady-state cases. In all four cases, we established heavy-traffic limits (allowing λ to increase) under which these approximations are asymptotically correct.

Consistent with previous experience, e.g., [1] and [3], we found that the power-tail cases presented many difficulties, including lower-quality approximations. We observed that the two iterated limits for $P(T > x)$ involving $\lim_{x\to\infty}\lim_{\tau\to\infty}$ and $\lim_{\tau\to\infty}\lim_{x\to\infty}$ agree for the case of a pure-exponential tail, but do not agree for the case of a power tail. That explains why the exponents in the heavy-traffic limits in Theorems 3.4 and 4.4 do not agree. To obtain insight and new approximations, in Sect. 4.5 we introduced a new double limit in which $\lim_{x\to\infty}$ and $\lim_{\tau\to\infty}$ with $x = b\tau$ for some constant b, and applied it to Pareto service-time cdf's. For the power-tail case, we suggested using the minimum of transient and steady-state approximations for moments and quantiles.

We evaluated the approximations by performing extensive numerical comparisons, most of which appear in [8]. Highlights were presented in four tables here. The numerical results show that the approximations are remarkably effective for the exponential-tail case, provided that certain conditions are satisfied, as detailed in Sect. 5. In particular, we must be sure that the mean ν_τ is not too small. The steady-state approximation coincides with the transient approximation when τ is large enough, but can greatly overestimate if it is not. The required value for τ is of order $E[S_e]$, as shown by (2.15). The two-moment approximations from [3] were introduced in Sect. 4.2 and shown to be useful for the exponential-tail case here too. Simple rough two-moment approximations for the case $c^2 \geq 1$ are given in (4.10). Formulas (4.8) and (4.9) can be used to obtain corresponding formulas for $c^2 \leq 1$.

The power-tail case is much more problematic. Tables 3 and 4 show that the approximations can yield good results here too, but care is required. The parameters λ and τ were much larger in these Pareto tables. Unlike the exponential-tail case, the transient approximation for quantiles of T gets arbitrarily bad if we increase τ enough. We suggested the *minimum* of the transient and steady-state approximations for moments and quantiles as overall approximations. So far, we conclude that the approximations in the power-tail case are less reliable, so that checking with exact calculations is more important. Nevertheless, we showed that it is possible to generate useful approximations in the power-tail case.

Acknowledgements This research was done at Columbia University, where David Goldberg was an undergraduate and Ward Whitt was supported by NSF grant DMI-0457095.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Abate, J., Whitt, W.: Asymptotics for M/G/1 low-priority waiting-time tail probabilities. Queueing Syst. **25**, 173–223 (1997)
2. Choudhury, G.L., Lucantoni, D.M., Whitt, W.: Numerical solution of $M_t/G_t/1$ queues. Oper. Res. **45**, 451–463 (1997)
3. Crow IV, C.S., Goldberg, D., Whitt, W.: Two-moment approximations for maxima. Oper. Res. **55**, 532–548 (2007)
4. Crow IV, C.S., Goldberg, D., Whitt, W.: Congestion caused by inspection (in preparation)
5. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_t/G/\infty$ queue. Oper. Res. **41**, 731–742 (1993)
6. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. Springer, New York (1997)
7. Erdélyi, A.: Asymptotic Expansions. Dover, New York (1956)
8. Goldberg, D.A., Whitt, W.: The last departure time from an $M_t/G/\infty$ queue with a terminating arrival process: additional numerical results. Working paper, Columbia University (2006). Available at http://columbia.edu/~ww2040
9. Johnson, N.L., Kotz, S.: Continuous Univariate Distributions—I. Wiley, New York (1970)
10. Ross, S.M.: Introduction to Probability Models, 8th edn. Academic, New York (2003)
11. Serfozo, R.F.: Point processes. In: Stochastic Models. Handbooks in Operations Research and Management Sciences, vol. 2, pp. 1–93. North-Holland, Amsterdam (1990)
12. Takács, L.: Introduction to the Theory of Queues. Oxford University Press, New York (1962)
13. Wein, L.M., Wilkins, A.H., Baveja, M., Flynn, S.E.: Preventing the importation of illicit nuclear materials in shipping containers. Risk Anal. **26**, 1377–1393 (2006)
14. Whitt, W.: Stochastic-Process Limits. Springer, New York