MEAN TIME FOR THE DEVELOPMENT OF LARGE WORKLOADS AND LARGE QUEUE LENGTHS IN THE GI/G/1 QUEUE

by

Charles Knessl and Charles Tier

Department of Mathematics, Statistics and Computer Science

University of Illinois at Chicago

851 South Morgan Street

Chicago, IL 60607-7045

Keywords: queueing systems, asymptotics, singular perturbations

AMS Classification: 60K25, 34E15

Abstract

We consider the GI/G/1 queue described by either the workload $U(t)$ (unfinished work) or the number of customers $N(t)$ in the system. We compute the mean time until $U(t)$ reaches or exceeds the level K, and also the mean time until $N(t)$ reaches N_0. For the M/G/1 and GI/M/1 models, we obtain exact contour integral representations for these mean first passage times. We then compute the mean times asymptotically, as K and $N_0 \to \infty$, by evaluating these contour integrals. For the general GI/G/1 model, we obtain asymptotic results by a singular perturbation analysis of the appropriate backward Kolmogorov equation(s). Numerical comparisons show that the asymptotic formulas are very accurate even for moderate values of K and N_0.

1. Introduction

Queueing models arise in a wide variety of applications such as computer systems and communications networks. The mathematical analysis of these models typically involves the computation of certain performance measures, such as the steady-state queue length or workload distribution, the length of a busy period, etc. Of particular importance is the total number of customers ("jobs") or the size of the workload in the queueing system. If this total size becomes very large, the system performance may deteriorate and jobs may be lost or suffer long delays. For example, in the design of high speed communications systems, the buffer size at a switch is of crucial importance. If the buffer capacity is too small, arriving jobs may be frequently lost. Even if the buffer has a large capacity, a large buffer size below capacity may develop, resulting in unacceptably long delays. It is therefore natural to ask how long it will take before a large buffer size (as measured by either number of jobs or total workload) develops in a particular queueing system.

This research was supported in part by NSF Grant DMS-93-00136 and DOE Grant DE-FG02-93ER25168

In this paper we consider the GI/G/1 queue and compute the mean time needed for either the workload (total unfinished work), or the number of customers in the system, to reach or exceed some specified level. We denote the workload at time t by $U(t)$ and by $N(t)$ the number of customers. These stochastic processes are not Markovian, but may be imbedded in higher dimensional Markovian processess by using the supplementary variable technique. If we let $Z(t)$ be the elapsed time since the last arrival and $Y(t)$ be the elapsed service time of the customer presently being served, then $(U(t), Z(t))$ and $(N(t), Y(t), Z(t))$ are both Markov processes.

It is generally undesirable to have large queue lengths or large workloads. In a stable GI/G/1 queue the occurence of these events is very rare. However, having accurate measures of the probabilities of such rare events may be an important measure of performance and reliability. We therefore analyze the time for $U(t)$ to reach or exceed some large level K, and also the time for $N(t)$ to reach some large number N_0. For a model with a finite capacity of either workload or queue length, computing this time is the same as computing the time until the next customer is lost.

For the GI/G/1 model we denote the interarrival time density by $a(\cdot)$ and the service time density by $b(\cdot)$. Throughout the paper we assume that the Laplace transforms

$$\widehat{a}(\theta) = \int_0^\infty e^{-\theta z} a(z) dz, \qquad \widehat{b}(\theta) = \int_0^\infty e^{-\theta y} b(y) dy$$

are analytic for Re(θ)<0. Thus, all the moments are finite, and we denote the respective moments by

$$m_k = \int_0^\infty y^k b(y) dy, \quad M_k = \int_0^\infty z^k a(z) dz, \quad k \geq 1.$$

The traffic intensity is $\rho = m_1/M_1$ and we assume that $\rho < 1$, which guarantees that the queue is stable, and the processes $N(t)$ and $U(t)$ have steady-state distributions. For exponential arrivals we set $M_1 = 1/\lambda$ and for exponential service we set $m_1 = 1/\mu$.

Let $p(w)$ be the steady-state workload density, and let $p(n)$ be the steady-state probability that $N(t) = n$. For these quantities there are exact integral representations. A good summary of known exact results for the GI/G/1 model can be found in the book of Cohen [4]. From these integrals and our assumptions about the analyticity of the Laplace transforms, it is easy to obtain the asymptotic tail probabilities

$$p(w) \sim \alpha_1 e^{-cw}, \quad w \to \infty \tag{1.1}$$

$$p(n) \sim \alpha_2 \left[\int_0^\infty e^{-cz} a(z) dz \right]^n, \quad n \to \infty. \tag{1.2}$$

Here c is the unique positive solution of

$$\left[\int_0^\infty e^{-cz} a(z) dz \right] \left[\int_0^\infty e^{cy} b(y) dy \right] = 1, \quad c > 0. \tag{1.3}$$

Thus, the tail exponent in the workload is precisely $-c$, and the tail exponent in the queue length distribution is $\log[\widehat{a}(c)]$. The constant c corresponds to a simple pole in the integral representations of the distribution functions. The constants $\alpha_1,\ \alpha_2$ correspond to residues at this pole, and these are also readily computed. In sections 6 and 7 we explicitly identify α_1 and α_2.

Now let N be the mean time until the workload reaches or exceeds the level K. For the GI/G/1 model, N will generally depend upon the initial workload $U(0) = w$, and also on the initial value $Z(0) = z$. Thus, $N = N(w, z)$. We let T be the mean time until $N(t)$ reaches N_0. This function depends on the initial values $(N(0), Y(0), Z(0)) = (n, y, z)$, i.e. $T = T(n, y, z)$. We will show, however, that asymptotically as $K, N_0 \to \infty$ (and for initial values of $U(0), N(0)$ not close to K, N_0) the mean first passage times are independent of w, n, y, z; and we have

$$N(w, z) \sim \beta_1 e^{cK}, \quad K \to \infty \tag{1.4}$$

$$T(n, y, z) \sim \beta_2 \left[\int_0^\infty e^{cy} b(y) dy\right]^{N_0}, \quad N_0 \to \infty. \tag{1.5}$$

Thus, N and T grow exponentially at precisely the same rates as the corresponding steady-state probabilities decay. While this seems to be well-known and can be argued, for example, by using (1.1)-(1.2) and renewal theory, the computation of the constants β_1, β_2 appears to be a much more difficult task. The purpose of this paper is to give explicit formulas for these constants. The computation of β_1 and β_2 is essential if one is to obtain accurate numerical approximations to N and T; this is further discussed in section 8.

Previous work on tail exponents and tail probabilities includes Cohen [2], Iglehart [8], Neuts and Takahashi [16] and Sadowsky and Szpankowski [17]. These authors consider the m-server GI/G/m queue and/or various special cases of this model. Using the tail behavior as an exponential approximation is also discussed by Fredericks [5], Gaver and Shedler [6], and in the book of Tijms [18]. This type of approximation seems to be superior to the standard (exponential) heavy traffic approximation, and reduces to the latter in the heavy traffic limit (for the GI/G/1 model this is defined as $\rho \uparrow 1$). The difficulty in using this approximation is the computation of the constants α_1, α_2.

The asymptotic approach employed here makes use of singular perturbation methods such as boundary layer theory and asymptotic matching; general references for these techniques are Kevorkian and Cole [9] and Bender and Orszag [1]. Matkowsky and Schuss [15] developed a singular perturbation method for computing asymptotically first passage times for diffusion processes with small diffusion coefficients. We have extended this method to discrete random walks and Markov jump processes in [11-13]. In particular, in [12,13] we computed the first passage times N and T for a state-dependent M/G/1 model, where the arrival and service processes are allowed to depend on the present workload or queue length. Results for the standard M/G/1 model may be obtained simply by omitting the various

state-dependence. Recently [10], we have computed the mean time for large queue lengths to develop in tandem Jackson networks. Numerical comparisons in [10] show that the asymptotic results are in excellent agreement with exact (numerical) solutions. For this agreement to occur, it is essential to compute the constants in the asymptotic expansions (i.e. the analogs of β_2 in (1.5)).

For the M/M/1 model, it is trivial to compute exactly the first passage times N and T. For the more general M/G/1 and GI/M/1 models, we give exact integral representations for N and T (cf. Results 2–5). It is easy to then evaluate these integrals asymptotically and hence obtain (1.4) and (1.5). For these special models, we also obtain the asymptotics directly, by using perturbation techniques to analyze the appropriate backward Kolmogorov equation(s). Of course, the two methods yield identical results. For the general GI/G/1 model, we have not been able to obtain exact expressions for N and T, so that we use the perturbation method to obtain asymptotic formulas for $K, N_0 \to \infty$. The main results for the GI/G/1 model are summarized in Results 6 and 7, and these appear in sections 6 and 7, respectively.

While we only compute mean first passage times, similar techniques may be used for higher moments. However, in the asymptotic limit considered here, the first passage times tend to be exponentially distributed, so that the mean is sufficient to characterize the entire distribution. This was shown explicitly for singularly perturbed diffusion processes by Williams [19], and this calculation can be easily adapted to discrete random walks and jump processes.

The assumption on the analyticity of the Laplace transforms $\widehat{a}(\cdot)$ and $\widehat{b}(\cdot)$ is essential to our analysis. If, say, the service time density had only an algebraic tail, then it is likely that the mean first passage times N and T would have only algebraic growth in K and N_0, and also be more sensitive to the initial conditions w and n.

2. Queue length in the M/M/1 queue

We compute the mean time until $N(t)$, the number of customers in the system, reaches some large number N_0. Thus we define

$$\tau = \min[t : N(t) \geq N_0] = \min[t : N(t) = N_0] \tag{2.1}$$

and

$$T(n) = T(n; N_0) = E(\tau | N(0) = n), \quad 0 \leq n \leq N_0. \tag{2.2}$$

Note that since customers arrive individually, the time to reach or exceed N_0 is the same as the time to reach N_0.

The function $T(n)$ satisfies the backward equation

$$\lambda T(n+1) + \mu T(n-1) - (\lambda + \mu) T(n) = -1, \quad 1 \leq n \leq N_0 - 1 \tag{2.3}$$

with the boundary conditions

$$\lambda T(1) - \lambda T(0) = -1 \tag{2.4}$$

$$T(N_0) = 0. \tag{2.5}$$

Here λ and μ are the arrival and service rates, respectively. This difference equation is easily solved, and then the result may be asymptotically evaluated in the limit $N_0 \to \infty$. We set $\rho = \lambda/\mu$ and give the results below.

RESULT 1:

The mean time for $N(t)$ to reach N_0 in the M/M/1 queue is given by

$$T(n) = \frac{1}{\mu(1-\rho)^2}\left[\left(\frac{\mu}{\lambda}\right)^{N_0} - \left(\frac{\mu}{\lambda}\right)^{n}\right] + \frac{n - N_0}{\mu - \lambda}.$$

For $N_0 \to \infty$, asymptotic expansions for $T(n)$ are

(a) $N_0 \to \infty, \quad N_0 - n \to \infty$

$$T(n) \sim \frac{1}{\mu(1-\rho)^2}\left(\frac{1}{\rho}\right)^{N_0}$$

(b) $N_0 \to \infty, \quad N_0 - n = m = O(1), \quad m \geq 1$

$$T(n) \sim \frac{1}{\mu(1-\rho)^2}\left(\frac{1}{\rho}\right)^{N_0}[1 - \rho^m].$$

Even this very simple model reveals some important insights into the structure of $T(n)$. First, we note that $T(n)$ grows exponentially as $N_0 \to \infty$. Also, $T(n)$ is independent of the initial value $N(0) = n$, as long as n is not close to the "exit boundary" N_0. In [10], we have shown how to obtain the asymptotic results in (a) and (b) directly from the difference equation, by using singular perturbation techniques. This method was then extended to compute the time needed for large queue lengths to build up in tandem Jackson networks. For these problems exact results are not available, so that the direct asymptotic approach is needed.

We note that the last term in the exact expression for $T(n)$ (the term linear in n) is uniformly smaller than the first term(s), in the limit $N_0 \to \infty$, $0 \leq n < N_0$. It is in fact exponentially smaller. This last term is a particular solution to the difference equation (2.3), which has -1 as an inhomogeneous term. The parts of the exact expression for $T(n)$ that grow exponentially as n, $N_0 \to \infty$ satisfy the homogeneous form of (2.3). These observations are useful for developing the perturbation method.

3. Workload in the M/G/1 queue

We consider an M/G/1 model with arrival rate λ and service time density $b(\cdot)$, with finite moments of all order. The workload $U(t)$ is clearly a Markov process. We define the first passage time

$$\tau_* = \min[t : U(t) \geq K] \tag{3.1}$$

and its conditional mean

$$N(w) = E(\tau_* | U(0) = w), \quad 0 \leq w < K. \tag{3.2}$$

By definition, $N(w) = 0$ for $w \geq K$. Note that $U(t)$ can jump across the level K without actually hitting it. In this respect the jump process is different from a diffusion process or the discrete queue length process.

The backward equation satisfied by $N(w)$ is

$$-N'(w) - \lambda N(w) + \lambda \int_0^{K-w} N(w+z)b(z)dz = -1, \quad 0 < w < K \tag{3.3}$$

$$-\lambda N(0) + \lambda \int_0^K N(w+z)b(z)dz = -1. \tag{3.4}$$

The last equation may be replaced by the equivalent condition $N'(0^+) = 0$, which is obtained by setting $w = 0^+$ in (3.3) and subtracting equation (3.4).

For exponential service with $b(z) = \mu e^{-\mu z}$ and $\rho = \lambda/\mu$, we can easily convert (3.3)-(3.4) to the differential equation

$$-N''(w) + (\mu - \lambda)N'(w) = \mu \tag{3.5}$$

$$N'(0) = 0; \quad N'(K) + \lambda N(K) = 1. \tag{3.6}$$

Here $N(K)$ is understood to mean $N(K^-)$, as $N(K^+) = 0$. The solution to (3.5) and (3.6) is

$$N(w) = \frac{1}{\lambda(1-\rho)^2} e^{(\mu-\lambda)K} \left[1 - \rho e^{-(\mu-\lambda)(K-w)}\right] + \frac{\mu(w-K) - 1}{\mu - \lambda}. \tag{3.7}$$

As was the case for $T(n)$, we see that $N(w)$ is exponentially large in K and nearly constant, expect when $K - w = O(1)$, which corresponds to initial workloads close to the "capacity" K.

We observe that $N(K^-) > 0 = N(K^+)$, so that the first passage time has a discontinuity at $w = K$. This is because for initial workloads just below K, the system's unfinished work cannot exceed K until the next customer arrives, and by the time this occurs the server has decreased the workload, possibly by a significant amount. Asymptotically, as $K \to \infty$, we have $N(K^-) \sim (1-\rho)N(0)$ so that $N(K^-)$ is in fact exponentially large in K, of the same order of magnitude as $N(0)$, the mean first passage time starting with zero workload. The factor $1 - \rho$ suggests that roughly a fraction $1 - \rho$ of the sample paths that start at $U(0) = K^-$ will have the workload decreased to some $O(1)$ value, before finally

undergoing the large deviation to cross $U(t) = K$. The remaining fraction ρ of sample paths that start at $U(0) = K^-$ will cross $U(t) = K$ in a short time period, before the server has had the chance to empty the system of the large workload. Since the first set of paths weigh the mean escape time by an exponentially large amount, $N(K^-)$ will also be asymptotically exponentially large.

Now consider (3.3) for general service time densities. By setting $u = K - w$ and using a Laplace transform over u, we obtain from (3.3) the contour integral representation

$$N(w) = \frac{1}{2\pi i} \int_{Br} \frac{e^{(K-w)s}[N(K) - 1/s]}{s - \lambda + \lambda \widehat{b}(s)} ds \tag{3.8}$$

where $\widehat{b}(s) = \int_0^\infty e^{-sz} b(z) dz$ is the Laplace transform of service time density, $N(K) = N(K^-)$, and the integration contour is a Bromwich contour on which Re(s)>0. Note that the integrand has a double pole at $s = 0$. The constant $N(K)$ is determined from (3.4), or, equivalently, from the condition $N'(0) = 0$; hence

$$N(K) = \frac{\dfrac{1}{2\pi i} \displaystyle\int_{Br} \frac{e^{Ks}}{s - \lambda + \lambda \widehat{b}(s)} ds}{\dfrac{1}{2\pi i} \displaystyle\int_{Br} \frac{s e^{Ks}}{s - \lambda + \lambda \widehat{b}(s)} ds} \equiv \frac{\text{NUM}(K)}{\text{DEN}(K)}. \tag{3.9}$$

When $b(z) = \mu e^{-\mu z}$, $\widehat{b}(s) = \mu/(\mu + s)$ and it is then easy to show that (3.8) and (3.9) reduces to (3.7).

We next examine the asymptotics of $N(w)$ as $K \to \infty$. We use two independent approaches. First, we expand the integrals in (3.8) and (3.9) as $K \to \infty$. Then we obtain the identical results by asymptotically analyzing (3.3) and (3.4) by perturbation methods.

The asymptotic behavior of the integrals in (3.9) is determined by the singularities of the integrands with the largest real part. The integrand in DEN(K) is analytic at $s = 0$, but has a pole at $s = -a$, where a satisfies the (real) transcendental equation

$$\lambda \int_0^\infty e^{aw} b(w) dw = a + \lambda, \quad a > 0. \tag{3.10}$$

The existence of the solution follows from our assumption that the moment generating function $M_b(\theta) = \int_0^\infty e^{\theta w} b(w) dw$ is analytic for Re(θ)>0, and the stability condition λm_1<1. The numerator in (3.9) has a pole at $s = 0$. Thus, computing the corresponding residues in (3.9) we obtain

$$\text{NUM}(K) \sim \frac{1}{1 - \rho}, \quad \text{DEN}(K) \sim \frac{a e^{-Ka}}{\lambda I_1(a) - 1}; \quad K \to \infty \tag{3.11}$$

where

$$I_1(a) = \int_0^\infty w e^{aw} b(w) dw. \tag{3.12}$$

It follows that $N(K^-)$ is again exponentially large as $K \to \infty$. Now consider the integral in (3.8). Since $N(K)$ is exponentially large, we can approximate $N(K) - 1/s \sim N(K)$ in the integrand, with

an error that is exponentially small. For $K - w = u = O(1)$, we cannot simplify (3.8) any further. For $K - w \to \infty$, the asymptotic behavior of $N(w)$ is determined by the simple pole in (3.8) at $s = 0$. Using (3.8) and (3.11) thus yields

$$N(w) \sim \frac{N(K^-)}{1-\rho} \sim \frac{\lambda I_1(a) - 1}{(1-\rho)^2 a} e^{Ka}; \quad K \to \infty, \quad K - w \to \infty. \tag{3.13}$$

An alternate approach to the asymptotics is as follows. We introduce the scaled variables

$$X = \frac{w}{K}, \quad \varepsilon = \frac{1}{K}, \quad N(w) = n(X) \tag{3.14}$$

and note that $\varepsilon \to 0$ when $K \to \infty$. In terms of these new variables, (3.3) becomes

$$-\varepsilon n'(X) - \lambda n(X) + \lambda \int_0^{(1-X)/\varepsilon} n(X + \varepsilon z) b(z) dz = -1 \tag{3.15}$$

and we use the boundary condition $n'(0) = 0$. For $1 - X \gg \varepsilon$, we expand $n(X)$ as

$$n(X) = C(\varepsilon) \left[n_0(X) + \varepsilon n_1(X) + \cdots \right] \tag{3.16}$$

and extend the upper limit on the integral in (3.15) to $+\infty$. Assuming further that $\varepsilon C(\varepsilon) \to \infty$ as $\varepsilon \to 0$ (which is a very mild assumption since we eventually find that $C(\varepsilon)$ grows exponentially as $\varepsilon \to 0$), we obtain from (3.15) to leading order the problem

$$(\lambda m_1 - 1) n_0'(X) = 0; \quad n_0'(0) = 0. \tag{3.17}$$

It follows that $n_0(X)$ is constant, and without loss of generality we set $n_0(X) = 1$. The expansion (3.16) breaks down near the exit boundary, where $X \approx 1$ (to be more precise, $1 - X = O(\varepsilon)$). Setting $u = (1 - X)/\varepsilon = K - w$, with $n(X) = N(w) = C(\varepsilon) \left[\mathcal{N}_0(u) + \varepsilon \mathcal{N}_1(u) + \cdots \right]$, we obtain from (3.15)

$$\mathcal{N}_0'(u) - \lambda \mathcal{N}_0(u) + \lambda \int_0^u \mathcal{N}_0(u - z) b(z) dz = 0, \quad u > 0. \tag{3.18}$$

The function $\mathcal{N}_0(u)$ is referred to as a "boundary-layer" approximation, since it is to be valid in the thin region $1 - X = O(\varepsilon)$, where $N(w)$ varies appreciably. By asymptotically matching the two expansions, we obtain

$$\mathcal{N}_0(u) \to 1 \quad \text{as} \quad u \to \infty. \tag{3.19}$$

Then (3.18) is easily solved using Laplace transforms, and we get the contour integral representation

$$\mathcal{N}_0(u) = \frac{1-\rho}{2\pi i} \int_{Br} \frac{e^{us}}{s - \lambda + \lambda \widehat{b}(s)} ds. \tag{3.20}$$

It also follows that the approximation $n(X) \sim C(\varepsilon) \mathcal{N}_0((1 - X)/\varepsilon)$ is valid uniformly, for $X \in [0, 1]$ (i.e. $w \in [0, K]$). It remains only to determine the constant $C(\varepsilon)$.

We denote the steady-state unfinished work distribution by

$$\begin{aligned} p(w)dw &= \lim_{t\to\infty} \Pr\left[U(t) \in (w, w+dw)\right], \quad w > 0 \\ A &= \lim_{t\to\infty} \Pr[U(t) = 0]. \end{aligned} \tag{3.21}$$

It satisfies the Takacs integro-differential equation

$$p'(w) - \lambda p(w) + \lambda \int_0^w p(w-z)b(z)dz + \lambda A b(w) = 0 \tag{3.22}$$

with

$$p(0) = \lambda A; \quad A + \int_0^\infty p(w)dw = 1. \tag{3.23}$$

The solution is well-known to be $A = 1 - \rho$ and

$$p(w) = \frac{\lambda(1-\rho)}{2\pi i} \int_{Br} \frac{1 - \widehat{b}(s)}{s - \lambda + \lambda\widehat{b}(s)} e^{ws} ds. \tag{3.24}$$

Now we multiply (3.3) by $p(w)$, integrate from $w = 0$ to $w = K$, and use (3.4), (3.22) and (3.23). After some simplification we are led to

$$p(K)N(K) = A + \int_0^K p(w)dw = 1 - \int_K^\infty p(w)dw. \tag{3.25}$$

This identity is exact for all $K > 0$. As $K \to \infty$, (3.25) can be replaced by the asymptotic relation

$$N(K) = N(K^-) \sim 1/p(K). \tag{3.26}$$

Now, from the perturbation expansion, $N(K) \sim C(\varepsilon)\mathcal{N}_0(0) = C(\varepsilon)(1-\rho)$. The asymptotics of $p(K)$ as $K \to \infty$ are easily obtained from (3.24), as the singularity with the largest real part is at $s = -a$ (cf. (3.10)). Computing the residue at this pole yields

$$p(K) \sim \lambda(1-\rho)\frac{1 - \widehat{b}(-a)}{1 + \lambda\widehat{b}'(-a)} e^{-aK} = \frac{(1-\rho)a}{\lambda I_1(a) - 1} e^{-aK}. \tag{3.27}$$

From (3.26) and (3.27) we find that

$$C(\varepsilon) \sim \frac{1}{(1-\rho)p(K)} \sim \frac{\lambda I_1(a) - 1}{(1-\rho)^2 a} e^{Ka}. \tag{3.28}$$

This result of course agrees with (3.13), which was obtained by asymptotically expanding the exact solution. We summarize our results below.

RESULT 2:

The mean time for $U(t)$ to reach or exceed K in the M/G/1 queue is given by

$$N(w) = \frac{1}{2\pi i}\int_{Br} \frac{e^{(K-w)s}}{s-\lambda+\lambda\widehat{b}(s)}\left[N(K^-) - \frac{1}{s}\right] ds,$$

$$N(K^-) = \left[\frac{1}{2\pi i}\int_{Br} \frac{e^{Ks}}{s-\lambda+\lambda\widehat{b}(s)} ds\right] / \left[\frac{1}{2\pi i}\int_{Br} \frac{se^{Ks}}{s-\lambda+\lambda\widehat{b}(s)} ds\right].$$

For $K \to \infty$, asymptotic expansions for $N(w)$ are

(a) $K \to \infty, \quad K - w \to \infty$

$$N(w) \sim \frac{\lambda I_1(a) - 1}{(1-\rho)^2 a} e^{Ka} \equiv C$$

(b) $K \to \infty, \quad K - w = u = O(1), \quad u > 0$

$$N(w) \sim C\frac{1-\rho}{2\pi i}\int_{Br} \frac{e^{us}}{s-\lambda+\lambda\widehat{b}(s)} ds.$$

Here $a > 0$ is the solution to (3.10), $I_1(a)$ is defined by (3.12), and $\widehat{b}(\cdot)$ is the Laplace transform of the service time density.

We again see that $N(w)$ is asymptotically constant, except near the exit boundary $w = K$.

4. Queue length in the M/G/1 queue

For the M/G/1 model, $N(t)$ is no longer a Markov process. Thus we let $Y(t)$ be the elapsed service time of the customer presently being served and consider the joint process $(N(t), Y(t))$, which is Markov. We denote the steady-state probabilities by

$$\begin{aligned} p(n,y)dy &= \lim_{t\to\infty} Pr\left[N(t) = n, Y(t) \in (y, y+dy)\right], \quad n \geq 1 \\ p(0) &= \lim_{t\to\infty} Pr\left[N(t) = 0\right] \\ p(n) &= \int_0^\infty p(n,y)dy, \quad n \geq 1. \end{aligned} \tag{4.1}$$

These satisfy the balance equations

$$\begin{aligned} p_y(n,y) &= \lambda p(n-1,y) - \left[\lambda + \mu(y)\right] p(n,y), \quad n \geq 2 \\ p_y(1,y) &= -\left[\lambda + \mu(y)\right] p(1,y), \\ p(1,0) &= \lambda p(0) + \int_0^\infty p(2,y)\mu(y)dy \\ \lambda p(0) &= \int_0^\infty p(1,y)\mu(y)dy \\ p(n,0) &= \int_0^\infty p(n+1,y)\mu(y)dy, \quad n \geq 2 \\ p(0) &+ \sum_{n=1}^\infty \int_0^\infty p(n,y)dy = 1. \end{aligned} \tag{4.2}$$

Subscripts are used to denote partial derivatives. Here $\mu(y)$ is the conditional departure rate given $Y(t) = y$, and is related to the service time density $b(\cdot)$ via

$$\mu(y) = \frac{b(y)}{\int_y^\infty b(t)dt}; \quad b(y) = \mu(y)\text{exp}\left[-\int_0^y \mu(t)dt\right]. \tag{4.3}$$

Assuming the stability condition $\lambda m_1 < 1$, (4.2) is easily solved using generating functions to give $p(0) = 1 - \rho$ and

$$p(n, y) = \exp\left[-\int_0^y \mu(t)dt\right] \frac{\lambda(1-\rho)}{2\pi i} \int_C \frac{(s-1)e^{\lambda(s-1)y}}{s^n[s - \widehat{b}(\lambda - \lambda s)]} ds, \quad n \geq 1 \tag{4.4}$$

where C is a small loop about $s = 0$. The integrand is analytic for $|s| < 1$ and at $s = 1$. As $n \to \infty$, the asymptotic behavior of the integral is determined by the simple pole at $s = 1 + a/\lambda$ (cf. (3.10)). Thus,

$$p(n, y) \sim \exp\left[-\int_0^y \mu(t)dt\right] \frac{(1-\rho)ae^{ay}}{\lambda I_1(a) - 1} \left(\frac{\lambda}{a+\lambda}\right)^n, \quad n \to \infty \tag{4.5}$$

and

$$p(n) \sim \frac{a(1-\rho)}{\lambda[\lambda I_1(a) - 1]} \left(\frac{\lambda}{a+\lambda}\right)^n, \quad n \to \infty. \tag{4.6}$$

Note that for the M/M/1 model, $a = \mu - \lambda$ and then $I_1 = \mu\lambda^{-2}$ and (4.6) reduces to the well known (exact) geometric distribution for this model.

We analyze the time for $N(t)$ to reach N_0. We again define τ by (2.1) and set

$$\begin{gathered} T(n, y) = T(n, y; N_0) = E(\tau | N(0) = n, Y(0) = y), \quad n \geq 1 \\ T(0) = E(\tau | N(0) = 0). \end{gathered} \tag{4.7}$$

The function T satisfies the backward equation

$$\lambda T(n+1, y) + \mu(y)T(n-1, 0) - [\lambda + \mu(y)]T(n, y) + T_y(n, y) = -1, \quad 2 \leq n \leq N_0 - 2 \tag{4.8}$$

and the boundary conditions are

$$\begin{gathered} \lambda T(2, y) + \mu(y)T(0) - [\lambda + \mu(y)]T(1, y) + T_y(1, y) = -1 \\ \lambda T(1, 0) - \lambda T(0) = -1 \end{gathered} \tag{4.9}$$

and

$$\mu(y)T(N_0 - 2, 0) - [\lambda + \mu(y)]T(N_0 - 1, y) + T_y(N_0 - 1, y) = -1. \tag{4.10}$$

Note that, by definition, $T(N_0, y) = 0$ for all y.

For exponential service, $\mu(y) = \mu =$ constant and then (4.8)-(4.10) admits a solution independent of y, which leads to the expression in Result 1.

We solve (4.8)-(4.10) by using generating functions. The final result is

$$T(n,y) = \frac{N_0 - n}{\lambda} + \frac{1}{2\pi i}\int_C \frac{\int_y^\infty b(t)e^{\lambda(s-1)(t-y)}dt}{s^{N_0-n}\left[\widehat{b}(\lambda-\lambda s)-s\right]\int_y^\infty b(t)dt}\left[T(N_0-1,0)+\frac{1}{\lambda(s-1)}\right]ds \tag{4.11}$$

where

$$T(N_0-1,0) = \frac{\frac{1}{\lambda}\frac{1}{2\pi i}\int_C s^{-N_0}\frac{\widehat{b}(\lambda-\lambda s)}{\widehat{b}(\lambda-\lambda s)-s}ds}{\frac{1}{2\pi i}\int_C s^{-N_0}\frac{\widehat{b}(\lambda-\lambda s)(1-s)}{\widehat{b}(\lambda-\lambda s)-s}ds} \equiv \frac{\mathrm{NUM}_1(N_0)}{\mathrm{DEN}_1(N_0)}. \tag{4.12}$$

We examine $T(n,y)$ asymptotically as $N_0 \to \infty$. In this limit, the expansions of the integrals in (4.11) and (4.12) are determined by the singularities of the integrands whose distance from the orgin is minimal. For $\mathrm{DEN}_1(N_0)$ the dominant singularity is $s = 1 + a/\lambda$, and for $\mathrm{NUM}_1(N_0)$ the dominant singularity is $s = 1$. These integrals are asymptotic to (-1) times the corresponding residues; hence

$$\mathrm{NUM}_1(N_0) \sim \frac{1}{\lambda(1-\rho)}, \quad \mathrm{DEN}_1(N_0) \sim \frac{a}{\lambda}\frac{1}{\lambda I_1(a)-1}\left(\frac{\lambda}{\lambda+a}\right)^{N_0-1}. \tag{4.13}$$

For $N_0 \to \infty$, we can write $T(N_0-1,0) + 1/(\lambda(s-1)) \sim T(N_0-1,0)$, with an exponentially small error, in (4.11). Also, the term $(N_0-n)/\lambda$ is asymptotically negligible. If $N_0 \to \infty$ with $N_0 - n = m = O(1)$, no further simplification is possible. But if $N_0 - n \to \infty$, then the dominant singularity in the integral is the simple pole at $s = 1$, and we get

$$T(n,y) \sim \frac{T(N_0-1,0)}{1-\rho} \sim \frac{\lambda I_1(a)-1}{(1-\rho)^2 a}(1+\frac{a}{\lambda})^{N_0-1}; \quad N_0 \to \infty, \quad N_0 - n \to \infty. \tag{4.14}$$

We next analyze (4.8)-(4.10) by a perturbation method similar to that in section 3. We set

$$x = \frac{n}{N_0}, \quad \delta = \frac{1}{N_0}, \quad T(n,y) = t(x,y). \tag{4.15}$$

Then (4.8) becomes

$$t_y(x,y) + \lambda t(x+\delta,y) + \mu(y)t(x-\delta,0) - [\lambda+\mu(y)]t(x,y) = -1 \tag{4.16}$$

for $x = \delta, 2\delta, \cdots, 1-2\delta$, where we identify $t(0,0)$ with $t(0) = T(0)$. Using the expansion

$$t(x,y) = D(\delta)\left[t_0(x,y) + \delta t_1(x,y) + \cdots\right] \tag{4.17}$$

in (4.16), and assuming that $\delta D(\delta) \to \infty$ as $\delta \to 0$, we obtain at the first two orders in δ the equations

$$t_{0,y}(x,y) + \mu(y)\left[t_0(x,0) - t_0(x,y)\right] \equiv \mathcal{L}t_0 = 0 \tag{4.18}$$

$$\mathcal{L}t_1 = \partial_x \left[\mu(y)t_0(x,0) - \lambda t_0(x,y)\right]. \tag{4.19}$$

The only solution to (4.18) that doesn't grow exponentially in y is given by $t_0(x,y) = t_0(x)$, which is independent of y. Then (4.19) has, in general, no solution unless a solvability condition is satisfied. The solvability condition is obtained by multiplying the equation by

$$\exp\left[-\int_0^y \mu(\tau)d\tau\right] = \int_y^\infty b(\tau)d\tau$$

and integrating from $y = 0$ to $y = \infty$. Using $t_0(x,y) = t_0(x)$, (4.19) then becomes

$$(1 - \lambda m_1)t_0'(x) = 0$$

so that $t_0(x)$ is a constant, and we set $t_0(x) = 1$, as this constant can be incorporated into $D(\delta)$ in (4.17). Note that with (4.17), $t_0(x) = 1$ also satisfies the boundary conditions in (4.9) asymptotically.

We next construct a boundary layer correction to the expansion (4.17), which takes into account the boundary condition (4.10). We set

$$\frac{1-x}{\delta} = m = 1, 2, 3, \cdots \quad \text{with } t(x,y) = T(n,y) = \mathcal{T}(m,y)$$

and expand $\mathcal{T}$ as

$$\mathcal{T}(m,y) = D(\delta)\left[\mathcal{T}_0(m,y) + \delta\mathcal{T}_1(m,y) + \cdots\right]. \tag{4.20}$$

Then from (4.8) and (4.10) we obtain the problem

$$\mathcal{T}_{0,y}(m,y) + \lambda\left[\mathcal{T}_0(m-1,y) - \mathcal{T}_0(m,y)\right] + \mu(y)\left[\mathcal{T}_0(m+1,0) - \mathcal{T}_0(m,y)\right] = 0, \quad m \geq 1 \tag{4.21}$$

and the boundary condition (4.10) can be replaced by $\mathcal{T}_0(0,y) = 0$ for all y. By introducing the generating function

$$\mathcal{G}(s,y) = \sum_{m=1}^{\infty} s^m \mathcal{T}_0(m,y) \tag{4.22}$$

we obtain from (4.21)

$$\mathcal{G}_y(s,y) + \left[\lambda(s-1) - \mu(y)\right]\mathcal{G}(s,y) = \mu(y)\left[\mathcal{T}_0(1,0) - \frac{1}{s}\mathcal{G}(s,0)\right]. \tag{4.23}$$

Multiplying (4.23) by $e^{\lambda(s-1)y}\exp\left[-\int_0^y \mu(\tau)d\tau\right]$ and integrating from $y = 0$ to $y = \infty$ yields

$$\mathcal{G}(s,0) = \left[\frac{1}{s}\mathcal{G}(s,0) - \mathcal{T}_0(1,0)\right]\widehat{b}(\lambda - \lambda s). \tag{4.24}$$

Solving (4.24) for $\mathcal{G}(s,0)$, using the result in (4.23) and then integrating the resulting ODE in y, we obtain $\mathcal{G}(s,y)$ in term of $\mathcal{T}_0(1,0)$. Then we invert the generating function in (4.22) and obtain

$$\mathcal{T}_0(m,y) = \frac{\mathcal{T}_0(1,0)}{2\pi i}\int_C \frac{\displaystyle\int_y^\infty b(t)e^{\lambda(s-1)(t-y)}dt}{s^m\left[\widehat{b}(\lambda - \lambda s) - s\right]\displaystyle\int_y^\infty b(t)dt}ds. \tag{4.25}$$

In order for expansions (4.17) and (4.20) to be consistent, we must have $\mathcal{T}_0(m,y) \to 1$ as $m \to \infty$ for each y. As $m \to \infty$, the expansion of (4.25) is determined by the simple pole at $s = 1$, which is the singularity of the integrand that is closest to the origin. Noting that the residue is independent of y, the matching condition implies that

$$\mathcal{T}_0(1,0) = 1 - \rho = 1 - \lambda m_1. \tag{4.26}$$

It remains to determine the constant $D(\delta)$. To this end we use the steady-state balance equations (4.2). We multiply (4.8) by the steady-state probability $p(n,y)$, integrate from $y = 0$ to $y = \infty$, and sum from $n = 1$ to $n = N_0 - 1$. Integrating some of the integrals by parts, shifting indices in the summations, and using the fact that $p(n,y)$ satisfies the system of equations (4.2), we eventually obtain

$$\begin{aligned} p(N_0-1,0)T(N_0-1,0) &= p(0) + \sum_{n=1}^{N_0-1} \int_0^\infty p(n,y)dy \\ &= 1 - \sum_{n=N_0}^{\infty} \int_0^\infty p(n,y)dy. \end{aligned} \tag{4.27}$$

We evaluate this identity as $N_0 \to \infty$. The right side of (4.27) is asymptotically equal to 1, with an error that is exponentially small. From the perturbation method, we found that $T(N_0 - 1, 0) \sim D(\delta)\mathcal{T}_0(1,0) = D(\delta)(1-\rho)$ and $p(N_0 - 1, 0)$ can be evaluated using (4.5). Thus, solving (4.27) for $D(\delta)$ we obtain

$$D(\delta) \sim \frac{\lambda I_1(a) - 1}{(1-\rho)^2 a} \left(1 + \frac{a}{\lambda}\right)^{N_0 - 1}, \tag{4.28}$$

which agrees with (4.14). We summarize the main results.

RESULT 3:

The mean time for $N(t)$ to reach N_0 in the M/G/1 queue is given by

$$T(n,y) = \frac{N_0 - n}{\lambda} + \frac{1}{2\pi i} \int_C \frac{\int_y^\infty b(t) e^{\lambda(s-1)(t-y)} dt}{s^{N_0 - n} \left[\widehat{b}(\lambda - \lambda s) - s\right] \int_y^\infty b(t)dt} \left[T(N_0 - 1, 0) + \frac{1}{\lambda(s-1)}\right] ds,$$

$$T(N_0 - 1, 0) = \frac{1}{\lambda} \left[\frac{1}{2\pi i} \int_C s^{-N_0} \frac{\widehat{b}(\lambda - \lambda s)}{\widehat{b}(\lambda - \lambda s) - s} ds\right] \Big/ \left[\frac{1}{2\pi i} \int_C s^{-N_0} \frac{\widehat{b}(\lambda - \lambda s)(1-s)}{\widehat{b}(\lambda - \lambda s) - s} ds\right].$$

For $N_0 \to \infty$, asymptotic expansions for $T(n,y)$ are

(a) $N_0 \to \infty \quad N_0 - n \to \infty$

$$T(n,y) \sim \frac{\lambda I_1(a) - 1}{(1-\rho)^2 a} \left(1 + \frac{a}{\lambda}\right)^{N_0 - 1} \equiv D$$

(b) $N_0 \to \infty, \quad N_0 - n = m = O(1), \quad m \geq 1$

$$T(n,y) \sim D\frac{1-\rho}{2\pi i}\int_C \frac{\int_y^\infty b(t)e^{\lambda(s-1)(t-y)}dt}{s^m\left[\widehat{b}(\lambda-\lambda s)-s\right]\int_y^\infty b(t)dt}ds.$$

Here a, $I_1(a)$ and $\widehat{b}(\cdot)$ are as in Result 2.

When $b(y) = \mu e^{-\mu y}$, $\widehat{b}(\lambda-\lambda s) = \mu/[\mu+\lambda-\lambda s]$ and then T is independent of y. By evaluating the contour integrals we can easily show that Result 3 reduces to the corresponding parts in Result 1. We again observe that T is exponentially large and approximately constant, except near the exit boundary. Near the exit boundary T is still exponentially large, but now depends upon y and $m = N_0 - n$. The expression for $T(n,y)$ in Result 3 applies for all $n \geq 0$. Note that when $n \geq N_0$, the contour integral vanishes for all $y \geq 0$, as now the integrand is analytic at $s = 0$.

5. Queue length in the GI/M/1 queue

We consider the GI/M/1 model with service rate μ and interarrival time density $a(\cdot)$. The process $N(t)$ is not Markovian, but the joint process $(N(t), Z(t))$, where $Z(t)$ denotes the elapsed time since the last arrival, is Markov. Assuming the stability condition $\rho = (\mu M_1)^{-1} < 1$, we define the steady-state probabilities by

$$p(n,z)dz = \lim_{t\to\infty}\Pr\left[N(t) = n, Z(t) \in (z, z+dz)\right], \quad n \geq 0 \tag{5.1}$$

$$p(n) = \int_0^\infty p(n,z)dz, \quad n \geq 0. \tag{5.2}$$

These satisfy the balance equations

$$\begin{gathered} p_z(n,z) = \mu\left[p(n+1,z) - p(n,z)\right] - \lambda(z)p(n,z), \quad n \geq 1 \\ p_z(0,z) = \mu p(1,z) - \lambda(z)p(0,z) \\ p(n+1,0) = \int_0^\infty \lambda(z)p(n,z)dz, \quad n \geq 0 \\ p(0,0) = 0 \\ \sum_{n=0}^\infty \int_0^\infty p(n,z)dz = 1. \end{gathered} \tag{5.3}$$

Here $\lambda(z)$ is the conditional arrival rate given $Z(t) = z$; it is related to the interarrival time density $a(\cdot)$ via

$$\lambda(z) = \frac{a(z)}{\int_z^\infty a(\tau)d\tau}; \quad a(z) = \lambda(z)\text{exp}\left[-\int_0^z \lambda(\tau)d\tau\right]. \tag{5.4}$$

The solution to (5.3) is

$$\begin{aligned} p(n,z) &= \frac{1-b_*}{M_1}(b_*)^{n-1}e^{\mu(b_*-1)z}\exp\left[-\int_0^z \lambda(\tau)d\tau\right], \quad n \geq 1 \\ p(0,z) &= \frac{1}{M_1}\left[1-e^{\mu(b_*-1)z}\right]\exp\left[-\int_0^z \lambda(\tau)d\tau\right], \end{aligned} \tag{5.5}$$

where b_* is the unique solution to

$$b_* = \int_0^\infty e^{\mu(b_*-1)z}a(z)dz, \quad 0 < b_* < 1. \tag{5.6}$$

The marginal queue length probabilities are

$$p(n) = \frac{1-b_*}{\mu M_1}(b_*)^{n-1}, \quad n \geq 1; \quad p(0) = 1-\rho = 1-\frac{1}{\mu M_1}. \tag{5.7}$$

To compute the time until $N(t)$ reaches N_0, we again define the stopping time τ by (2.1) and set

$$T(n,z) = E(\tau | N(0) = n, Z(0) = z). \tag{5.8}$$

By definition, $T(N_0,z) = 0$ for $z \geq 0$. For $n < N_0$, we easily obtain the following backward equation for T

$$T_z(n,z) + \lambda(z)T(n+1,0) + \mu T(n-1,z) - [\lambda(z)+\mu]\,T(n,z) = -1, \quad 1 \leq n \leq N_0 - 2 \tag{5.9}$$

$$T_z(0,z) + \lambda(z)T(1,0) - \lambda(z)T(0,z) = -1 \tag{5.10}$$

$$T_z(N_0-1,z) + \mu T(N_0-2,z) - [\lambda(z)+\mu]\,T(N_0-1,z) = -1. \tag{5.11}$$

The last equation may be replaced by the condition $T(N_0,0) = 0$ and the requirement that (5.9) also hold for $n = N_0 - 1$. We solve (5.9)-(5.11) using generating functions and obtain

$$T(n,z) = T(1,0) + \frac{\int_z^\infty (t-z)a(t)dt}{\int_z^\infty a(t)dt} + \frac{M_1}{2\pi i}\int_C \frac{1}{(s-1)s^n[\widehat{a}(\mu-\mu s)-s]}\frac{\int_z^\infty a(t)e^{\mu(s-1)(t-z)}dt}{\int_z^\infty a(t)dt}ds. \tag{5.12}$$

The constant $T(1,0)$ is determined from the boundary condition $T(N_0,0) = 0$; hence

$$0 = T(1,0) + M_1 + \frac{M_1}{2\pi i}\int_C \frac{1}{(s-1)s^{N_0}}\frac{\widehat{a}(\mu-\mu s)}{\widehat{a}(\mu-\mu s)-s}ds. \tag{5.13}$$

Here $\widehat{a}(\cdot)$ is the Laplace transform of the interarrival time density. Expression (5.12) gives $T(n,z)$ for $0 \leq n \leq N_0 - 1$ and $z \geq 0$, and also for $n = N_0$ and $z = 0$. For $n > N_0$ or $n = N_0$ and $z > 0$, we obviously have $T(n,z) = 0$. It is easy to show that if $a(z) = \lambda e^{-\lambda z}$, then $T(n,z)$ is independent of z and we recover the first formula in Result 1.

We next evaluate T in the limit $N_0 \to \infty$. The integrand in (5.13) now has a unique pole inside the unit circle $|s| < 1$, at $s = b_*$ (cf. (5.6)). For large N_0 this pole determines the asymptotic behavior of $T(1,0)$, and we obtain

$$T(1,0) \sim \frac{M_1}{1-\mu J_1(b)} \frac{1}{1-b_*} \left(\frac{1}{b_*}\right)^{N_0-1} = \frac{\mu M_1}{b[1-\mu J_1(b)]} \left(\frac{\mu}{\mu - b}\right)^{N_0-1} \tag{5.14}$$

where $b = \mu(1-b_*) \in (0,\mu)$ and

$$J_1(b) = \int_0^\infty z e^{-bz} a(z) dz. \tag{5.15}$$

For $n \to \infty$, the contour integral in (5.12) also has a pole at $s = b_*$, and grows exponentially. For $N_0 - n \to \infty$, this integral is smaller asymptotically than $T(1,0)$ and we obtain $T(n,z) \sim T(1,0)$ for $N_0,\ N_0 - n \to \infty$. For $N_0 - n = m = O(1)$, the integral in (5.12) is of the same order of magnitude as $T(1,0)$. Evaluating the residue from $s = b_*$, we thus obtain

$$T(n,z) \sim T(1,0) \left[1 - \left(1 - \frac{b}{\mu}\right)^{m-1} \frac{\int_z^\infty e^{-b(t-z)} a(t) dt}{\int_z^\infty a(t) dt}\right], \quad m \geq 1$$

for $N_0 \to \infty$, $N_0 - n = m = O(1)$.

An alternate approach to the asymptotics is to analyze (5.9)-(5.11) by the perturbation method. We set $n = x/\delta$, $\delta = N_0^{-1}$ with $T(n,z) = t(x,z) = D(\delta)\left[t_0(x,z) + \delta t_1(x,z) + \cdots\right]$. Using these scaled variables in (5.9) and expanding for $\delta \to 0$, we obtain

$$t_{0,z}(x,z) + \lambda(z)\left[t_0(x,0) - t_0(x,z)\right] \equiv \mathcal{L}_1 t_0 = 0 \tag{5.16}$$

$$\mathcal{L}_1 t_1 = \partial_x\left[\mu t_0(x,z) - \lambda(z) t_0(x,0)\right]. \tag{5.17}$$

Solving these equations in a manner similar to that used to analyze (4.18) and (4.19), we eventually find that $t_0(x,z) = t_0(x) = 1$. Then the boundary condition (5.10) is also asymptotically satisfied.

The approximation $T(n,z) \sim D(\delta)$ breaks down for $x \approx 1$ ($n \approx N_0$) and does not satisfy the boundary condition (5.11). In the boundary layer we set $(1-x)/\delta = N_0 - n = m = O(1)$, with $T(n,z) = t(x,z) = \mathcal{T}(m,z) = D(\delta)\left[\mathcal{T}_0(m,z) + \delta \mathcal{T}_1(m,z) + \cdots\right]$. The leading term $\mathcal{T}_0$ satisfies the problem (cf. (5.10) and (5.11))

$$\mathcal{T}_{0,z}(m,z) + \mu\left[\mathcal{T}_0(m+1,z) - \mathcal{T}_0(m,z)\right] + \lambda(z)\left[\mathcal{T}_0(m-1,0) - \mathcal{T}_0(m,z)\right] = 0, \quad m \geq 1 \tag{5.18}$$

and we use the boundary condition $\mathcal{T}_0(0,0) = 0$. We also have the matching condition

$$\mathcal{T}_0(m,z) \to 1 \quad \text{as} \quad m \to \infty, \quad \text{for each } z. \tag{5.19}$$

Since (5.18) is "constant-coefficient" in m, we look for solutions in the form $\mathcal{T}_0 = \beta^m \alpha(z)$. Then (5.18) yields $\alpha'(z) + \mu(\beta - 1)\alpha(z) - \lambda(z)\alpha(z) + \lambda(z)\alpha(0)/\beta = 0$, or

$$\frac{d}{dz}\left\{\exp\left[-\int_0^z \lambda(t)dt\right] e^{\mu(\beta-1)z}\alpha(z)\right\} = -\frac{1}{\beta}a(z)\alpha(0)e^{\mu(\beta-1)z}. \tag{5.20}$$

Integrating (5.20) from $z = 0$ to $z = \infty$ yields

$$-\alpha(0) = -\frac{1}{\beta}\alpha(0)\int_0^\infty a(z)e^{\mu(\beta-1)z}dz.$$

For $\alpha(0) \neq 0$, this equation has two solutions: $\beta = 1$ and $\beta = b_*$ (cf. (5.6)). If $\beta = 1$ then (5.20) implies that $\alpha(z) = \alpha(0) =$ constant. If $\beta = b_*$, we set $b_* = 1 - b/\mu$ and then (5.20) has the solution

$$\alpha(z) = \frac{\alpha(0)}{1 - b/\mu}\frac{\int_z^\infty e^{-b(t-z)}a(t)dt}{\int_z^\infty a(t)dt}. \tag{5.21}$$

We superimpose these two solutions and write

$$\mathcal{T}_0(m, z) = C_1 + C_2\left(1 - \frac{b}{\mu}\right)^{m-1}\frac{\int_z^\infty e^{-b(t-z)}a(t)dt}{\int_z^\infty a(t)dt}. \tag{5.22}$$

The remaining constants are determined from the boundary condition $\mathcal{T}_0(0, 0) = 0$ and the matching condition (5.19); this yields $C_1 = 1$, $C_2 = -1$. It follows that a uniform approximation to T is $T(n, z) \sim D(\delta)\mathcal{T}_0(m, z)$.

We determine $D(\delta)$ by multiplying (5.9) by $p(n, z)$, integrating from $z = 0$ to $z = \infty$, and summing from $n = 1$ to $n = N_0 - 1$ (using $T(N_0, 0) = 0$). After some integration by parts, shifting of indices on the sums, and use of the fact that $p(n, z)$ satisfies (5.3), we obtain

$$\mu\int_0^\infty T(N_0 - 1, z)p(N_0, z)dz = \sum_{n=0}^{N_0-1}\int_0^\infty p(n, z)dz = 1 - \sum_{n=N_0}^{\infty}\int_0^\infty p(n, z)dz. \tag{5.23}$$

This identity may be viewed as a generalized Green's theorem (or Lagrange identity) for the linear operators in (5.3) and (5.9)-(5.11). As $N_0 \to \infty$, the right side of (5.23) is asymptotically equal to 1. To evaluate the left side we use $T(N_0 - 1, z) \sim D(\delta)\mathcal{T}_0(1, z)$, and $p(N_0, z)$ is given by (5.5). Now,

$$\exp\left[-\int_0^z \lambda(t)dt\right] = \int_z^\infty a(t)dt$$

and integration by parts yields

$$\int_0^\infty \exp\left[-\int_0^z \lambda(\tau)d\tau\right] e^{-bz}dz = \frac{1}{b}\left[1 - \int_0^\infty e^{-bz}a(z)dz\right] = \frac{1}{\mu},$$

$$\int_0^\infty \left[\int_z^\infty e^{-bt}a(t)dt\right] dz = \int_0^\infty ze^{-bz}a(z)dz = J_1(b).$$

Using these identities in (5.23) leads to

$$D(\delta)\frac{b}{M_1}\left(1-\frac{b}{\mu}\right)^{N_0-1}\left[\frac{1}{\mu}-J_1(b)\right]\sim 1.$$

Solving for $D(\delta)$ regains the expression in (5.14). This again establishes the equivalence of the two asymptotic approaches. Below we summarize our results.

RESULT 4:

The mean time for $N(t)$ to reach N_0 in the GI/M/1 queue is given by

$$\begin{aligned}T(n,z) &= \frac{M_1}{2\pi i}\int_C \frac{-1}{(s-1)s^{N_0}}\frac{\widehat{a}(\mu-\mu s)}{\widehat{a}(\mu-\mu s)-s}ds+\frac{\int_z^\infty (t-z)a(t)dt}{\int_z^\infty a(t)dt}-M_1\\ &\quad+\frac{M_1}{2\pi i}\int_C \frac{1}{(s-1)s^n[\widehat{a}(\mu-\mu s)-s]}\frac{\int_z^\infty a(t)e^{\mu(s-1)(t-z)}dt}{\int_z^\infty a(t)dt}ds.\end{aligned}$$

For $N_0\to\infty$, asymptotic expansions for $T(n,z)$ are

(a) $N_0\to\infty,\quad N_0-n\to\infty$

$$T(n,z)\sim\frac{\mu M_1}{b[1-\mu J_1(b)]}\left(\frac{\mu}{\mu-b}\right)^{N_0-1}\equiv D$$

(b) $N_0\to\infty,\quad N_0-n=m=O(1),\quad m\geq 1$

$$T(n,z)\sim D\left\{1-\left(1-\frac{b}{\mu}\right)^{m-1}\frac{\int_z^\infty e^{-b(t-z)}a(t)dt}{\int_z^\infty a(t)dt}\right\}.$$

Here $b=\mu(1-b_*)$ where b_* is given by (5.6), $J_1(b)$ is defined in (5.15), and $\widehat{a}(\cdot)$ is the Laplace transform of the interarrival time density.

The overall asymptotic structure is similar to that for the M/M/1 and M/G/1 models. For the GI/M/1 model the structure of the solution in the boundary layer (result in part (b)) is simpler than the corresponding M/G/1 result. We also note that if $N_0=1$, the exact result becomes

$$T(0,z)=\frac{\int_z^\infty (t-z)a(t)dt}{\int_z^\infty a(t)dt},$$

which is just the mean residual life on the renewal process that governs the arrivals. Now the first passage time measures the first arrival time to an empty system.

6. Workload in the GI/G/1 queue

We consider $U(t)$ in the GI/G/1 model. We again let $Z(t)$ be the elapsed time since the last arrival and analyze the Markov process $(U(t), Z(t))$. The steady-state distribution of this process is well-known to be equivalent to solving a Wiener-Hopf problem. Since our approach, which uses the supplementary variable technique, is different from that used in most books on queueing theory, we briefly give the details of the computation of the steady-state distribution. Our main goal is the computation of the time for $U(t)$ to reach or exceed K. We can compute this exactly for the GI/M/1 model, but not for the general GI/G/1 case. For the latter case we derive asymptotic results by using the perturbation method that we outlined in the previous sections. This method requires that (i) we know the tail behavior of the steady-state distribution for the joint process $(U(t), Z(t))$ and (ii) we use the balance equations satisfied by this distribution.

We thus define

$$p(w,z)dwdz = \lim_{t\to\infty} \Pr\left[U(t) \in (w, w+dw),\ Z(t) \in (z, z+dz)\right] \tag{6.1}$$

$$A(z)dz = \lim_{t\to\infty} \Pr\left[U(t) = 0,\ Z(t) \in (z, z+dz)\right] \tag{6.2}$$

and assume the stability condition $\rho = m_1/M_1 < 1$. The balance equations are given by

$$-p_z(w,z) + p_w(w,z) - \lambda(z)p(w,z) = 0; \qquad z, w > 0 \tag{6.3}$$

$$-A'(z) + p(0,z) - \lambda(z)A(z) = 0; \qquad z > 0 \tag{6.4}$$

$$p(w,0) = \int_0^w b(y) \int_0^\infty \lambda(u)p(w-y,u)dudz + b(w)\int_0^\infty A(u)\lambda(u)du \tag{6.5}$$

$$A(0) = 0 \tag{6.6}$$

$$\int_0^\infty A(z)dz + \int_0^\infty \int_0^\infty p(w,z)dwdz = 1. \tag{6.7}$$

Here $\lambda(z)$ is as in section 5.

From (6.3), (6.4) and (6.6) it follows that

$$p(w,z) = R(w+z)\text{exp}\left[-\int_0^z \lambda(\tau)d\tau\right] \tag{6.8}$$

$$A(z) = \left[\int_0^z R(u)du\right]\text{exp}\left[-\int_0^z \lambda(\tau)d\tau\right]. \tag{6.9}$$

Then (6.5) yields an integral equation for $R(\cdot)$. By setting

$$R(w) = \frac{1}{M_1}\frac{d}{dw}\int_0^w b(y)W(w-y)dy \tag{6.10}$$

this integral equation can be reduced to the Wiener-Hopf (Lindley) equation

$$W(w) = \int_0^\infty c(w-u)W(u)du, \quad W(\infty) = 1. \tag{6.11}$$

The kernel in (6.11) is

$$c(w) = \int_{\max\{0,-w\}}^\infty a(u)b(u+w)du \tag{6.12}$$

and we have used the normalization condition (6.7) to infer the value of $W(\infty)$. Using standard Wiener-Hopf techniques, we write the solution to (6.11) as

$$W(w) = \frac{M_1 - m_1}{2\pi i}\int_{Br} \frac{e^{sw}}{\widehat{a}(-s)\widehat{b}(s) - 1}\exp\left[\Gamma_*(s)\right] ds \tag{6.13}$$

where

$$\begin{aligned}
\Gamma_*(s) &= \frac{1}{2\pi i}\int_{Br_+}\left(\frac{1}{\omega - s} - \frac{1}{\omega}\right)\log\left[\widehat{a}(-\omega)\widehat{b}(\omega) - 1\right] d\omega \\
\widetilde{\Gamma}(s) &= \frac{1}{2\pi i}\int_{Br_-}\left(\frac{1}{\omega - s} - \frac{1}{\omega}\right)\log\left[\widehat{a}(-\omega)\widehat{b}(\omega) - 1\right] d\omega \\
\exp\left[\Gamma_*(s)\right] &= \left[\widehat{a}(-s)\widehat{b}(s) - 1\right]\exp\left[\widetilde{\Gamma}(s)\right].
\end{aligned} \tag{6.14}$$

Given our assumptions that the Laplace transform $\widehat{a}(-s)$ (resp. $\widehat{b}(s)$) is analytic for Re$(s) > 0$ (resp. Re$(s)<0$), we have $\widehat{a}(-s)\widehat{b}(s) - 1$ analytic and nonzero in some strip $0<$Im$(s)<\varepsilon_0$. The Bromwich contour in (6.13) lies within this strip, and on Br_+ (resp. Br_-) we have Re$(s)<$Re$(\omega)<\varepsilon_0$ (resp. $0<$Re$(\omega)<$Re(s)). The various contours are sketched in Figure 1. Note that $\Gamma_*(s)$ is analytic at $s = 0$ (with $\Gamma_*(0) = 0$), but $\exp[\widetilde{\Gamma}(s)]$ has a simple pole at $s = 0$. In view of (6.8)-(6.10) and (6.13), we have the integral representations

$$p(w,z) = \exp\left[-\int_0^z \lambda(\tau)d\tau\right]\frac{1-\rho}{2\pi i}\int_{Br}\frac{e^{s(w+z)}s\widehat{b}(s)}{\widehat{a}(-s)\widehat{b}(s) - 1}e^{\Gamma_*(s)}ds \tag{6.15}$$

$$A(z) = \exp\left[-\int_0^z \lambda(\tau)d\tau\right]\frac{1-\rho}{2\pi i}\int_{Br}\frac{(e^{sz}-1)\widehat{b}(s)}{\widehat{a}(-s)\widehat{b}(s) - 1}e^{\Gamma_*(s)}ds. \tag{6.16}$$

Note that $\Gamma_*(s)$ is analytic in the left half-plane Re$(s) < \varepsilon_0$. From (6.16), it is easy to show that $A = \int_0^\infty A(z)dz = 1 - \rho$, as is well-known.

We evaluate the tail behavior of $p(w,z)$. Let c be the unique positive solution of the (real) transcendental equation

$$\left[\int_0^\infty e^{-cz}a(z)dz\right]\left[\int_0^\infty e^{cy}b(y)dy\right] = 1, \quad c > 0. \tag{6.17}$$

Then the integrand in (6.15) has a pole at $s = -c$, and this pole determines the asymptotic behavior of $p(w,z)$ as $w \to \infty$. We have

$$p(w,z) \sim e^{-cz}\exp\left[-\int_0^z \lambda(\tau)d\tau\right]\frac{(1-\rho)cI_0(c)\Gamma_0(c)}{I_1(c)J_0(c) - I_0(c)J_1(c)}e^{-cw}, \quad w \to \infty \tag{6.18}$$

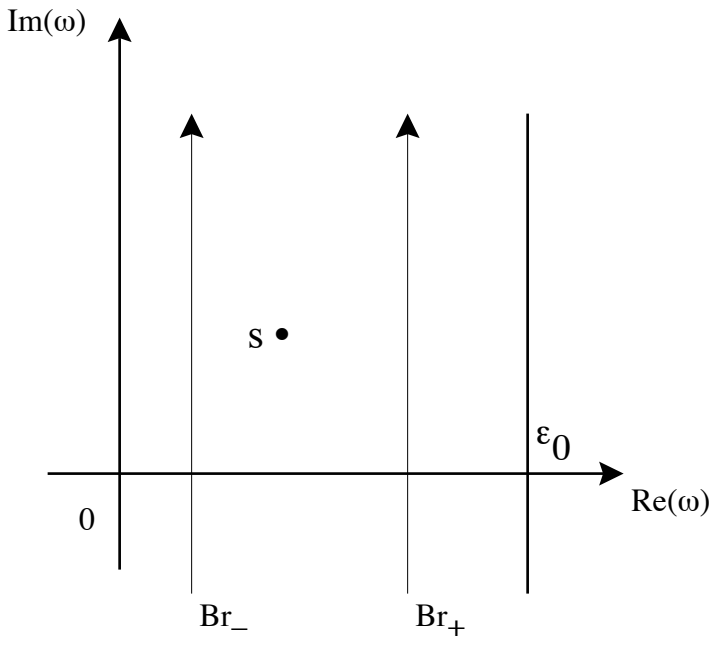

Figure 1: A sketch of the integration contours in the complex ω-plane.

where

$$\begin{aligned} I_j(c) &= \int_0^\infty y^j e^{cy} b(y) dy; \quad j = 0, 1 \\ J_j(c) &= \int_0^\infty y^j e^{-cy} a(y) dy; \quad j = 0, 1 \end{aligned} \tag{6.19}$$

and

$$\Gamma_0(c) = \exp\left[\Gamma_*(-c)\right] = \exp\left\{ \frac{1}{2\pi i} \int_{Br_+} \left(\frac{1}{\omega + c} - \frac{1}{\omega} \right) \log\left[\widehat{a}(-\omega)\widehat{b}(\omega) - 1 \right] d\omega \right\}. \tag{6.20}$$

The contour in (6.20) may be taken along the imaginary axis, with an indentation about $\omega = 0$ in the right half-plane. Note that equation (6.17) may be compactly written as $J_0(c)I_0(c) = 1$. Since

$$\int_0^\infty e^{-cz} \exp\left[-\int_0^z \lambda(\tau) d\tau \right] dz = \frac{1 - J_0(c)}{c} = \frac{I_0(c) - 1}{cI_0(c)},$$

the marginal density of $U(t)$ has the tail

$$p(w) = \int_0^\infty p(w, z) dz \sim \frac{(1-\rho)(I_0 - 1)\Gamma_0}{I_1 J_0 - I_0 J_1} e^{-cw}, \quad w \to \infty. \tag{6.21}$$

In special cases, the contour integrals in (6.14) may be explicitly evaluated and the results then become more explicit. For example, for the M/G/1 model we easily evaluate (6.14) to get $\exp[\Gamma_*(s)] = \lambda/(\lambda - s)$ and the expression for $p(w) = \int_0^\infty p(w, z) dz$ is equivalent to that in (3.24). In this case

$c = a$ (cf. (3.10) and (6.17)), $\Gamma_0 = \lambda/(\lambda + a)$, $J_0 = \lambda/(\lambda + a)$, $I_0 = 1 + a/\lambda$, and $J_1 = \lambda/(\lambda + a)^2$. Then (3.27) agrees with (6.21) (with $w = K$).

For the GI/M/1 model, we have $c = b$ (cf. (5.6) and (6.17)) and evaluating (6.14) yields

$$\exp[\Gamma_*(s)] = \frac{b(\mu\widehat{a}(-s) - \mu - s)}{s(s+b)(\mu M_1 - 1)}$$

and then integrating (6.15) over z gives

$$p(w) = \frac{b}{\mu M_1} e^{-bw} \quad \text{(GI/M/1 queue)}. \tag{6.22}$$

Also, we now have $\Gamma_0 = \exp[\Gamma_*(-b)] = [1 - \mu J_1(b)] / [\mu M_1 - 1]$, $I_0 = \mu/(\mu - b)$, $J_0 = 1 - b/\mu$, and $I_1 = \mu/(\mu - b)^2$. Then the asymptotic result (6.21) reduces to the exact expression in (6.22).

We compute the mean time for $U(t)$ to reach or exceed K. Let τ_* be as in (3.1) and set

$$N(w, z) = E(\tau_* | U(0) = w, Z(0) = z). \tag{6.23}$$

The backward equation is easily obtained as

$$-N_w(w,z) + N_z(w,z) - \lambda(z)N(w,z) + \lambda(z)\int_0^{K-w} N(w+y, 0)b(y)dy = -1; \quad z < 0,\ 0 < w < K \tag{6.24}$$

with the boundary condition

$$N_z(0,z) - \lambda(z)N(0,z) + \lambda(z)\int_0^K N(y,0)b(y)dy = -1, \quad z > 0. \tag{6.25}$$

The latter may be replaced by the equivalent "reflecting" condition $N_w(0^+, z) = 0,\ z > 0$. When $a(z) = \lambda e^{-\lambda z}$, we have $\lambda(z) = \lambda =$ constant and then $N(w, z) = N(w)$ so that (6.24)-(6.25) reduces to (3.3)-(3.4). Note that an arrival causes the clock on the renewal process $Z(t)$ to be reset to zero, and thus the second argument in $N(\cdot, \cdot)$ in (6.24) inside the integral is zero. Also, from the definition (3.1), we see that $N(w, z)$ for $w \geq K$ for all z. From now on, we will use $N(K, z)$ to mean $N(K^-, z)$.

We can construct the (exact) solution to (6.24) for the GI/M/1 model with $b(y) = \mu e^{-\mu y}$. It is then easy to evaluate this expression asymptotically as $K \to \infty$. Below we give only the final results.

RESULT 5:

The mean time for $U(t)$ to reach or exceed K in the GI/M/1 queue is given by

$$\begin{aligned} N(w,z) \;&= \frac{\mu M_1}{2\pi i}\int_{Br} \frac{\mu}{\mu - s}\frac{\widehat{a}(s)e^{sK}}{s(s - \mu + \mu\widehat{a}(s))}ds + \frac{\int_z^\infty (t-z)a(t)dt}{\int_z^\infty a(t)dt} \\ &\quad -M_1 - \frac{\mu M_1}{2\pi i}\int_{Br}\frac{e^{sw}}{s(s-\mu+\mu\widehat{a}(s))}\frac{\int_z^\infty e^{-s(t-z)}a(t)dt}{\int_z^\infty a(t)dt}ds. \end{aligned}$$

For $K \to \infty$, asymptotic expansions for $N(w, z)$ are

(a) $K \to \infty, \quad K - w \to \infty$

$$N(w,z) \sim \frac{\mu M_1}{b[1-\mu J_1(b)]} e^{Kb} \equiv C$$

(b) $K \to \infty, \quad K - w = u = O(1), \quad u > 0$

$$N(w,z) \sim C\left[1 - e^{-bu}\frac{\int_z^\infty e^{-b(t-z)}a(t)dt}{\int_z^\infty a(t)dt}\right].$$

Here b, $J_1(b)$ and $\widehat{a}(\cdot)$ are as in Result 4. On the contour Br we have $b < \mathrm{Re}(s) < \mu$.

Note that the first formula in Result 5 applies for $0 \le w < K$ and all z. For other values of (w,z), we clearly have $N(w,z) = 0$ by definition.

We now consider the general GI/G/1 model. While we have not been able to solve (6.24)-(6.25) exactly in this case, the perturbation method in the previous sections can be readily extended to the general model. We begin by establishing an identity between the solution of (6.3)-(6.7) and $N(w,z)$. As before, we multiply (6.24) by $p(w,z)$ and integrate the resulting expression over the strip $0 < w < K$, $0 < z < \infty$. To this we add $A(z)$ times equation (6.25), integrated from $z = 0$ to $z = \infty$. After some integration by parts and use of equations (6.3)-(6.6), we obtain

$$\int_0^\infty N(K^-,z)p(K,z)dz = \int_0^\infty A(z)dz + \int_0^\infty \int_0^K p(w,z)dwdz = 1 - \int_K^\infty \int_0^\infty p(w,z)dzdw. \quad (6.26)$$

This identity generalizes (3.25). The right side is asymptotically equal to 1. To evaluate the left side as $K \to \infty$, we use (6.18) with $w = K$. Then we must know the asymptotic expansion of $N(K^-,z)$, which corresponds to computing the mean first passage time for an initial workload just below the exit boundary. We also note that for the GI/M/1 model, $p(w,z)$ has the simple form

$$p(w,z) = \frac{b}{M_1} e^{-bz} \exp\left[-\int_0^z \lambda(\tau)d\tau\right] e^{-bw} \quad (6.27)$$

while $N(w,z)$ is quite complicated (cf. Result 5). By using (6.27) and Result 5 to evaluate the first integral in (6.26), we have verified that (6.26) is indeed satisfied.

To evaluate $N(w,z)$ as $K \to \infty$ we again use the perturbation method with $w = X/\varepsilon$, $\varepsilon = 1/K$. Away from the exit boundary we find that N is a constant, i.e. $N(w,z) \sim C(\varepsilon)$. In the boundary layer we set $(1-X)/\varepsilon = K - w = u = O(1)$ and expand N as $N(w,z) \sim C(\varepsilon)\mathcal{N}_0(u,z)$. From (6.24) we obtain, for u and $z > 0$,

$$\mathcal{N}_{0,u}(u,z) + \mathcal{N}_{0,z}(u,z) - \lambda(z)\mathcal{N}_0(u,z) + \lambda(z)\int_0^u \mathcal{N}_0(u-y,0)b(y)dy = 0 \quad (6.28)$$

and the matching condition is

$$\mathcal{N}_0(u,z) \to 1 \quad \text{as} \quad u \to \infty, \quad \text{for all } z. \quad (6.29)$$

Setting

$$\mathcal{N}_0(u,z) = \exp\left[\int_0^z \lambda(\tau)d\tau\right]\bar{\mathcal{N}}(u,z) = \left[\int_z^\infty a(\tau)d\tau\right]^{-1}\bar{\mathcal{N}}(u,z), \tag{6.30}$$

we obtain from (6.28)

$$\bar{\mathcal{N}}_u(u,z) + \bar{\mathcal{N}}_z(u,z) + a(z)\int_0^u \bar{\mathcal{N}}(u-y,0)b(y)dy = 0. \tag{6.31}$$

Furthermore, we write the solution of (6.31) in the form

$$\bar{\mathcal{N}}(u,z) = \int_z^\infty a(t)G(u+t-z)dt. \tag{6.32}$$

In view of (6.29) and (6.30), we have $G(\infty) = 1$. Using (6.32) in (6.31) we find that

$$G(u) = \int_0^u \bar{\mathcal{N}}(u-y,0)b(y)dy = \bar{\mathcal{N}}(u,0) * b(u). \tag{6.33}$$

Using (6.33) back in (6.32) yields

$$\bar{\mathcal{N}}(u,z) = \int_z^\infty a(t)\left\{\int_0^{u+t-z} \bar{\mathcal{N}}(u+t-z-y,0)b(y)dy\right\}dt, \tag{6.34}$$

which expresses $\bar{\mathcal{N}}(u,z)$ in terms of $\bar{\mathcal{N}}(u,0)$. By setting $z = 0$ in (6.34), we obtain an integral equation for $\bar{\mathcal{N}}(u,0)$. Since $G(\infty) = 1$, it also follows from (6.33) that $\bar{\mathcal{N}}(\infty,0) = 1$. After some elementary manipulation, this integral equation becomes

$$\bar{\mathcal{N}}(u,0) = \int_0^\infty c(u-\xi)\bar{\mathcal{N}}(\xi,0)d\xi, \quad \bar{\mathcal{N}}(\infty,0) = 1 \tag{6.35}$$

where $c(u)$ is as in (6.12). But (6.35) is precisely the Lindley integral equation, which is satisfied by the steady-state unfinished work distribution at arrival instants (cf. (6.11)-(6.13)). Hence,

$$\bar{\mathcal{N}}(u,0) = \frac{M_1 - m_1}{2\pi i}\int_{Br} \frac{e^{su}}{\widehat{a}(-s)\widehat{b}(s) - 1}\exp\left[\Gamma_*(s)\right]ds \tag{6.36}$$

where $\mathrm{Re}(s) > 0$ on Br. We thus obtain the boundary layer approximation from (6.36), (6.34), and (6.30). Also,

$$N(K^-,z) \sim C(\varepsilon)\mathcal{N}_0(0^+,z) = C(\varepsilon)\exp\left[\int_0^z \lambda(\tau)d\tau\right]\bar{\mathcal{N}}(0^+,z),$$

which when used in (6.26) along with (6.18) leads to

$$C(\varepsilon)\frac{(1-\rho)c\Gamma_0 I_0}{I_1 J_0 - I_0 J_1}e^{-cK}\int_0^\infty e^{-cz}\bar{\mathcal{N}}(0,z)dz \sim 1, \quad K \to \infty. \tag{6.37}$$

We evaluate explicitly the integral in (6.37). From (6.36) and (6.34) we obtain

$$\bar{\mathcal{N}}(u,z) = \frac{M_1 - m_1}{2\pi i}\int_{Br} \frac{\widehat{b}(s)e^{\Gamma_*(s)}}{\widehat{a}(-s)\widehat{b}(s) - 1}\left\{\int_z^\infty a(t)e^{s(u+t-z)}dt\right\}ds. \tag{6.38}$$

Setting $u = 0$ and using

$$\int_0^\infty e^{-cz}\left[\int_z^\infty e^{s(t-z)}a(t)dt\right]dz = \frac{\widehat{a}(-s)-\widehat{a}(c)}{s+c}$$

we obtain from (6.38)

$$\int_0^\infty e^{-cz}\bar{\mathcal{N}}(0,z)dz = \frac{M_1-m_1}{2\pi i}\int_{Br}\frac{\widehat{b}(s)(\widehat{a}(-s)-\widehat{a}(c))}{(s+c)(\widehat{a}(-s)\widehat{b}(s)-1)}e^{\Gamma_*(s)}ds. \tag{6.39}$$

Using (6.14), the last integral may be written as

$$\frac{M_1-m_1}{2\pi i}\left\{\int_{Br}\frac{1}{s+c}e^{\Gamma_*(s)}ds + \int_{Br}\frac{1-\widehat{b}(s)\widehat{a}(c)}{s+c}e^{\widetilde{\Gamma}(s)}ds\right\}. \tag{6.40}$$

Now $\Gamma_*(s)$ (resp. $\widetilde{\Gamma}(s)$) is analytic in the left (resp. right) half-plane Re$(s)<\varepsilon_0$ (resp. Re$(s)>0$). Also, it is easy to show that as $|s| \to \infty$ in the corresponding half-planes, we have $\exp[\Gamma_*(s)] \sim (\text{const.})/s$ and $\exp[\widetilde{\Gamma}(s)] \sim (\text{const.})/s$. It follows that the second integral in (6.40) vanishes, as can seen by closing the Bromwich contour in the right half-plane, where the integrand is analytic. The first integral in (6.40) can be evaluated by closing the contour in the left half-plane. In this region the only singularity is the simple pole at $s = -c$. Hence we have

$$\int_0^\infty e^{-cz}\bar{\mathcal{N}}(0,z)dz = (M_1-m_1)\exp\left[\Gamma_*(-c)\right] = (M_1-m_1)\Gamma_0(c). \tag{6.41}$$

Using (6.41) in (6.37) we solve for $C(\varepsilon)$, and then the asymptotic expansion of $N(w,z)$ is completely determined. We summarize the final results below.

RESULT 6:

Asymptotic expansions for the mean time for $U(t)$ to reach or exceed K in the GI/G/1 queue are given by

(a) $K \to \infty, \quad K-w \to \infty$

$$N(w,z) \sim \frac{I_1(c)J_0(c)-I_0(c)J_1(c)}{(1-\rho)^2cM_1I_0(c)\Gamma_0^2(c)}e^{cK} \equiv C$$

(b) $K \to \infty, \quad K-w = u = O(1)$

$$N(w,z) \sim C\frac{M_1-m_1}{2\pi i}\int_{Br}\frac{\widehat{b}(s)e^{\Gamma_*(s)}}{\widehat{a}(-s)\widehat{b}(s)-1}\frac{\int_z^\infty a(t)e^{s(u+t-z)}dt}{\int_z^\infty a(t)dt}ds.$$

Here

$$I_j(c) = \int_0^\infty y^j e^{cy}b(y)dy; \quad J_j(c) = \int_0^\infty y^j e^{-cy}a(y)dy \quad (j=0,1),$$

c is the unique positive solution of $I_0(c)J_0(c)=1$, and

$$\Gamma_0(c)=\exp\left[\Gamma_*(-c)\right]=\exp\left\{\frac{1}{2\pi i}\int_{Br_+}\left(\frac{1}{\omega+c}-\frac{1}{\omega}\right)\log\left[\widehat{a}(-\omega)\widehat{b}(\omega)-1\right]d\omega\right\}.$$

On the contour Br we have $0<\mathrm{Re}(s)<\varepsilon_0$, and Br_+ may be taken as the imaginary axis with an indentation about $\omega=0$ in the right half-plane.

The overall structure of N for the GI/G/1 model is similar to that we obtained for the special cases. The numerical evaluation of the contant C requires that we evaluate (numerically) the contour integral in $\Gamma_0(c)$. This can be done analytically, in closed form, if $a(\cdot)$ and $b(\cdot)$ have rational Laplace transforms.

For the M/G/1 model we have $c=a$, $a(z)=\lambda e^{-\lambda z}$, $M_1=1/\lambda$, $\exp[\Gamma_*(s)]=\lambda/(\lambda-s)$, $\Gamma_0=J_0=I_0^{-1}=\lambda/(a+\lambda)$, $J_1=\lambda/(a+\lambda)^2$ and then Result 6 reduces to parts (a) and (b) in Result 2. The dependence on z in Result 6(b) disappears.

For the GI/M/1 model we have $c=b$, $b(y)=\mu e^{-\mu y}$,

$$\frac{e^{\Gamma_*(s)}}{\widehat{a}(-s)\widehat{b}(s)-1}=e^{\tilde{\Gamma}(s)}=\frac{b(s+\mu)}{s(s+b)}\frac{1}{\mu M_1-1},$$

$\Gamma_0=\rho[1-\mu J_1(b)]/(1-\rho)$ and $(I_1J_0-I_0J_1)/(bI_0\Gamma_0)=(1-\rho)M_1/b$. Then Result 6 reduces to parts (a) and (b) in Result 5. The contour integral in part (b) of Result 6 may now be explicitly evaluated, as the integrand has but two poles in the region $\mathrm{Re}(s)<\varepsilon_0$, at $s=0$ and $s=-b$. The residues at these poles correspond to the two terms in Result 5, part (b).

7. Queue length in the GI/G/1 queue

We compute asymptotically the mean time for $N(t)$ to reach N_0 in the GI/G/1 model. We consider the Markov process $(N(t),Y(t),Z(t))$, where $Y(t)$ is the elapsed service time and $Z(t)$ is the age on the renewal process that governs the arrivals. The perturbation method requires that we know the tail behavior of the joint steady-state distribution of the 3-dimensional Markov process. The method also uses the balance equations satisfied by this distribution function.

We hence define

$$p(n,y,z)dydz=\lim_{t\to\infty}\Pr\left[N(t)=n,Y(t)\in(y,y+dy),Z(t)\in(z,z+dz)\right],\quad n\geq 2 \tag{7.1}$$

$$P(1,y)dy=\lim_{t\to\infty}\Pr\left[N(t)=1,Y(t)=Z(t)\in(y,y+dy)\right] \tag{7.2}$$

$$p(1,y,z)dydz=\lim_{t\to\infty}\Pr\left[N(t)=1,Y(t)\in(y,y+dy),Z(t)\in(z,z+dz)\right] \tag{7.3}$$

$$p(0,z)dz=\lim_{t\to\infty}\Pr\left[N(t)=0,Z(t)\in(z,z+dz)\right]. \tag{7.4}$$

Note that we have decomposed the probability that $N(t)=1$ into the two pieces (7.2) and (7.3). There are two ways that there can be a single customer present in the system. If a customer arrives to an

empty system, then $N(t) = 1$ and $Y(t) = Z(t)$, until the next arrival or departure. This accounts for the probability "mass" in (7.2). We can also have $N(t) = 1$ if the system has two customers and the one being served finishes service, and then we have $N(t) = 1$ and $Z(t) > Y(t)$, until another arrival or departure occurs. It follows that the support of $p(1, y, z)$ is $z \geq y$. We can also combine (7.2) and (7.3) and write

$$\widetilde{p}(1, y, z) = P(1, y)\delta(y - z) + H(z - y)p(1, y, z) \tag{7.5}$$

where $\delta(\cdot)$ is the Dirac delta function and $H(\cdot)$ is the Heaviside step function.

The balance equations are now

$$-p_y(n, y, z) - p_z(n, y, z) - [\mu(y) + \lambda(z)]\, p(n, y, z) = 0; \quad n \geq 2,\ y > 0,\ z > 0 \tag{7.6}$$

$$p(n - 1, 0, z) = \int_0^\infty \mu(y)p(n, y, z)dy; \quad n \geq 2,\ z > 0 \tag{7.7}$$

$$p(n + 1, y, 0) = \int_0^\infty \lambda(z)p(n, y, z)dz; \quad n \geq 2,\ y > 0. \tag{7.8}$$

When $n = 1$, (7.6) holds in the region $z > y > 0$. A careful consideration of the various transitions when $N(t) = 1$ or $N(t) = 0$ leads to the boundary conditions

$$-P_y(1, y) - [\mu(y) + \lambda(y)]\, P(1, y) = 0, \quad y > 0 \tag{7.9}$$

$$p(2, y, 0) = \int_y^\infty \lambda(z)p(1, y, z)dz + \lambda(y)P(1, y), \quad y > 0 \tag{7.10}$$

$$-p_z(0, z) - \lambda(z)p(0, z) + \int_0^z \mu(y)p(1, y, z)dy + \mu(z)P(1, z) = 0, \quad z > 0 \tag{7.11}$$

$$p(0, 0) = 0 \tag{7.12}$$

$$\int_0^\infty \lambda(z)p(0, z)dz = P(1, 0) \tag{7.13}$$

$$\sum_{n=2}^{\infty} \int_0^\infty \int_0^\infty p(n, y, z)dydz + \iint_{z>y>0} p(1, y, z)dydz + \int_0^\infty P(1, y)dy + \int_0^\infty p(0, z)dz = 1. \tag{7.14}$$

Here $\lambda(z)$ and $\mu(y)$ were defined in the previous sections; they represent the arrival and departure rates, conditioned on $Z(t) = z$ and $Y(t) = y$.

In view of (7.6), (7.9) and (7.11), we have

$$p(n, y, z) = \exp\left[-\int_0^y \mu(\tau)d\tau - \int_0^z \lambda(t)dt\right] R(n, z - y), \quad n \geq 1 \tag{7.15}$$

$$P(1, y) = \exp\left[-\int_0^y \mu(\tau)d\tau - \int_0^y \lambda(t)dt\right] R_1 \tag{7.16}$$

$$p(0, z) = \exp\left[-\int_0^z \lambda(t)dt\right] R_0(z). \tag{7.17}$$

Note that $R(1,z)=0$ for $z<0$. From (7.7), (7.8), (7.10)–(7.13) we find that

$$\int_0^\infty b(y)R(n,z-y)dy = R(n-1,z); \quad n\geq 2, \quad z>0 \tag{7.18}$$

$$\int_0^\infty a(z)R(n,z-y)dz = R(n+1,-y); \quad n\geq 2, \quad y>0 \tag{7.19}$$

$$\int_y^\infty a(z)R(1,z-y)dz + a(y)R_1 = R(2,-y), \quad y>0 \tag{7.20}$$

$$0 = -R_0'(z) + \int_0^z b(y)R(1,z-y)dy + b(z)R_1, \quad z>0 \tag{7.21}$$

$$R_0(0)=0 \tag{7.22}$$

$$\int_0^\infty a(z)R_0(z)dz = R_1. \tag{7.23}$$

These equations are easily solved by using the results in section 6. We first observe that the joint probability that the system is empty and $Z(t)<z$, is the same whether we are measuring the workload or the number of customers. Hence, $p(0,z)=A(z)$ and from (6.16) we obtain

$$p(0,z) = \exp\left[-\int_0^z \lambda(t)dt\right]\frac{1-\rho}{2\pi i}\int_{Br}\frac{e^{z\theta}\widehat{b}(\theta)}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}e^{\Gamma_*(\theta)}d\theta \tag{7.24}$$

where $\Gamma_*(\theta)$ is given by (6.14). Note that it is irrelevant whether the integrand contains the factor $e^{z\theta}$ or $e^{z\theta}-1$, since $p(0,0)=(1-\rho)(2\pi i)^{-1}\int_{Br}\widehat{b}(\theta)\exp\left[\widetilde{\Gamma}(\theta)\right]d\theta=0$, as this last integrand is analytic for Re(θ)>0.

In view of (7.5), we write

$$\widetilde{R}(1,z-y) = R_1\delta(y-z) + R(1,z-y), \quad z\geq y. \tag{7.25}$$

Then we use $R_0(z)$ (cf. (7.17) and (7.24)) in (7.21) and solve this equation using Laplace transforms. This leads to the integral representation

$$\widetilde{R}(1,z) = \frac{1-\rho}{2\pi i}\int_{Br}\frac{\theta e^{z\theta}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}e^{\Gamma_*(\theta)}d\theta, \quad z\geq 0, \tag{7.26}$$

from which one can compute R_1 and $R(1,z)$ (cf. (7.25)). We can then use (7.15) and (7.16) to compute (7.2) and (7.3). Using (7.25) and (7.26) in (7.20) to compute $R(2,z)$ for $z<0$, we get

$$R(2,z) = \frac{1-\rho}{2\pi i}\int_{Br}\frac{\theta\widehat{a}(-\theta)e^{z\theta}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}e^{\Gamma_*(\theta)}d\theta, \quad z<0. \tag{7.27}$$

On the contour Br we have 0<Re(θ)<ε_1, where $\widehat{a}(-\theta)$ is analytic for Re(θ)<ε_1. But by continuing (7.27) to positive values of z, we find that (7.18) is satisfied when $n=2$. Then we can compute $R(3,z)$ for $z<0$ from (7.19) with $n=2$. The resulting integral representation can be continued to

positive values of z, and then (7.18) will be satisfied for $n = 3$. Continuing in this recursive manner, we obtain

$$R(n,z) = \frac{1-\rho}{2\pi i}\int_{Br} \frac{\theta e^{\theta z}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1} e^{\Gamma_*(\theta)}\left[\widehat{a}(-\theta)\right]^{n-1} d\theta, \quad n \geq 2 \tag{7.28}$$

where $0 \leq \mathrm{Re}(\theta) < \varepsilon_1$ on Br. Expression (7.28) remains valid for $n = 1$, if we interpret the left side as $\widetilde{R}(1,z)$. Finally, the probability density $p(n,y,z)$ can be computed from (7.15) and (7.28). The marginal queue length probabilities are given by, for $n \geq 2$,

$$p(n) = \int_0^\infty \int_0^\infty p(n,y,z)dydz = \frac{1-\rho}{2\pi i}\int_{Br} \frac{(1-\widehat{b}(\theta))(\widehat{a}(-\theta)-1)}{\theta(\widehat{a}(-\theta)\widehat{b}(\theta)-1)} e^{\Gamma_*(\theta)}\left[\widehat{a}(-\theta)\right]^{n-1} d\theta. \tag{7.29}$$

This representation is equivalent to that given by Cohen in [1, pg. 307, eq. (5.140)]. Expression (7.29) also applies to $n = 1$, where

$$p(1) = \int_0^\infty P(1,y)dy + \int\int_{z>y>0} p(1,y,z)dydz$$

and, of course, $p(0) = 1-\rho$.

The tail behavior as $n \to \infty$ is easily obtained by shifting the Br contours in (7.28) and (7.29) to the left, past the pole at $\theta = -c$. By computing the residues at this pole we obtain

$$\begin{aligned} p(n,y,z) \sim \exp\left[-\int_0^z \lambda(t)dt - \int_0^y \mu(\tau)d\tau\right] e^{c(y-z)} \\ \times \frac{(1-\rho)c\Gamma_0(c)}{I_1(c)J_0(c) - I_0(c)J_1(c)}\left[\int_0^\infty e^{-ct}a(t)dt\right]^{n-1}, \quad n \to \infty. \end{aligned} \tag{7.30}$$

and

$$p(n) \sim \frac{(1-\rho)\Gamma_0(I_0-1)(1-J_0)}{c(I_1J_0 - I_0J_1)}\left[J_0\right]^{n-1}, \quad n \to \infty. \tag{7.31}$$

Here c, Γ_0, I_j, J_j are as in Result 6.

Next we analyze the time to reach N_0. We again define τ by (2.1) and set

$$\begin{aligned} T(n,y,z) = E(\tau | N(0) = n, Y(0) = y, Z(0) = z), \quad n \geq 1 \\ T(0,z) = E(\tau | N(0) = 0, Z(0) = z). \end{aligned} \tag{7.32}$$

The backward equation is now

$$\begin{aligned} T_y(n,y,z) + T_z(n,y,z) + \mu(y)T(n-1,0,z) + \lambda(z)T(n+1,y,0) \\ -\left[\lambda(z)+\mu(y)\right]T(n,y,z) = -1, \quad 2 \leq n \leq N_0 - 1 \end{aligned} \tag{7.33}$$

and we use the boundary conditions $T(N_0,y,0) = 0$ for $y > 0$, and

$$T_z(0,z) - \lambda(z)T(0,z) + \lambda(z)T(1,0,0) = -1. \tag{7.34}$$

Equation (7.33) remains valid when $n = 1$, if we identify $T(0,0,z)$ with $T(0,z)$.

We have not been able to solve exactly for T, so that we compute T asymptotically as $N_0 \to \infty$. The standard perturbation method shows that away from the exit boundary $T(n,y,z) \sim D(\delta)$, $\delta = N_0^{-1}$. In the boundary layer we set $m = N_0 - n$ and $T(n,y,z) = D(\delta)[\mathcal{T}_0(m,y,z) + \delta\mathcal{T}_1(m,y,z) + \cdots]$. Then (7.33) leads to

$$\begin{aligned}\mathcal{T}_{0,y}(m,y,z) + \mathcal{T}_{0,z}(m,y,z) + \mu(y)\mathcal{T}_0(m+1,0,z) + \lambda(z)\mathcal{T}_0(m-1,y,0) \\ -[\lambda(z)+\mu(y)]\mathcal{T}_0(m,y,z) = 0.\end{aligned} \tag{7.35}$$

The matching condition is

$$\mathcal{T}_0(m,y,z) \to 1 \text{ as } m \to \infty, \text{ for all } y \text{ and } z. \tag{7.36}$$

Also, $T(N_0,y,0) = 0$ implies that $\mathcal{T}_0(0,y,0) = 0$. Setting

$$\mathcal{T}_0(m,y,z) = \exp\left[\int_0^z \lambda(t)dt + \int_0^y \mu(\tau)d\tau\right]\bar{\mathcal{T}}(m,y,z) \tag{7.37}$$

leads to

$$(\partial_y + \partial_z)\bar{\mathcal{T}}(m,y,z) + b(y)\bar{\mathcal{T}}(m+1,0,z) + a(z)\bar{\mathcal{T}}(m-1,y,0) = 0. \tag{7.38}$$

We represent the solution to (7.38) in the form

$$\bar{\mathcal{T}}(m,y,z) = \int_z^\infty a(t)\int_y^\infty b(\tau)G(m,z-y-t+\tau)d\tau dt. \tag{7.39}$$

Then (7.38) is satisfied for all $y, z > 0$ provided that

$$\begin{aligned}G(m,z) &= \int_0^\infty a(t)G(m-1,z-t)dt, \quad z > 0 \\ G(m,-z) &= \int_0^\infty b(\tau)G(m+1,\tau-z)d\tau, \quad z > 0.\end{aligned} \tag{7.40}$$

In view of (7.18), (7.19) and (7.40), we see that $G(m,z)$ satisfies the same equations as $R(m,-z)$. It can be easily verified that the contour integral

$$G(m,z) = \frac{\gamma}{2\pi i}\int_{Br} \frac{e^{-z\theta}}{\hat{a}(-\theta)\hat{b}(\theta) - 1} e^{\Gamma_*(\theta)}\left[\hat{a}(-\theta)\right]^{m-1} d\theta \tag{7.41}$$

satisfies (7.40). The matching condition (7.36) implies (in view of (7.37) and (7.39)) that $G(\infty,z) = 1$. Now, as $m \to \infty$, the behavior of $G(m,z)$ is determined by the simple pole at $\theta = 0$. Recall that on Br, $0<\text{Re}(\theta)<\varepsilon_1$. Since $\exp[\Gamma_*(0)] = \hat{a}(0) = 1$, we have $G(\infty,z) = \gamma/(M_1 - m_1)$ so that we must choose $\gamma = M_1 - m_1 = M_1(1-\rho)$. We only need to verify that the boundary condition $\bar{\mathcal{T}}(0,y,0) = 0$

is satisfied. In view of (7.39) and (7.41), we have

$$\begin{aligned}
\bar{\mathcal{T}}(0,y,0) &= \int_0^\infty a(t)\int_y^\infty b(\tau)G(0,\tau-t-y)d\tau dt \\
&= \frac{M_1-m_1}{2\pi i}\int_{Br}\frac{e^{\Gamma_*(\theta)}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}\frac{1}{\widehat{a}(-\theta)}\left\{\int_0^\infty a(t)\int_y^\infty b(\tau)e^{(t+y-\tau)\theta}d\tau dt\right\}d\theta \\
&= \frac{M_1-m_1}{2\pi i}\int_{Br}\frac{e^{\Gamma_*(\theta)}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}\left\{\int_y^\infty b(\tau)e^{(y-\tau)\theta}d\tau\right\}d\theta \\
&= \frac{M_1-m_1}{2\pi i}\int_y^\infty b(\tau)\left\{\int_{Br}e^{\tilde{\Gamma}(\theta)}e^{(y-\tau)\theta}d\theta\right\}d\tau \\
&= 0,
\end{aligned}$$

since the integral over Br vanishes for all $y<\tau$, as the integrand is analytic for Re(θ)>0. We have thus solved for the boundary layer function $\mathcal{T}_0(m,y,z)$, and it remains only to determine the constant $D(\delta)$. We also note that $G(m,z)$ is related to the steady-state probabilities via $\partial_z G(m,-z)=M_1R(m,z)$ (for $m\geq 2$).

We next establish an identity between $T(n,y,z)$ and $p(n,y,z)$. We multiply (7.33) by $p(n,y,z)$, sum from $n=2$ to $n=N_0-1$, and integrate over y and z. After some integration by parts, shifts of indices on the summations, and use of equations (7.6)-(7.13), we obtain the identity

$$\begin{aligned}
\int_0^\infty T(N_0-1,0,z)p(N_0-1,0,z)dz &= 1-\sum_{n=N_0}^\infty\int_0^\infty\int_0^\infty p(n,y,z)dydz \\
&= 1-\sum_{n=N_0}^\infty p(n).
\end{aligned} \tag{7.42}$$

For the GI/M/1 model, $\mu p(N_0,z)=\int_0^\infty \mu(y)p(N_0,y,z)dy=p(N_0-1,0,z)$ so that (7.42) reduces to (5.23), as now $T(n,y,z)$ is independent of y. For the M/G/1 model, (7.42) reduces to (4.27), since $T(n,y,z)=T(n,y)$.

As before, the right side of (7.42) $\to 1$ as $N_0\to\infty$. To evaluate the left side we use (7.30) (with $n=N_0-1$ and $y=0$) and $T(N_0-1,0,z)\sim D(\delta)\mathcal{T}_0(1,0,z)$. We thus obtain

$$\begin{aligned}
&\frac{1}{D}\int_0^\infty T(N_0-1,0,z)p(N_0-1,0,z)dz \\
&\quad\sim \frac{(1-\rho)c\Gamma_0}{I_1J_0-I_0J_1}\,[J_0]^{N_0-2}\int_0^\infty e^{-cz}\int_z^\infty a(t)\int_0^\infty b(\tau)G(1,z-t+\tau)d\tau dtdz \\
&\quad= \frac{(1-\rho)c\Gamma_0}{I_1J_0-I_0J_1}\,[J_0]^{N_0-2}\,\frac{M_1-m_1}{2\pi i}\int_{Br}\frac{\widehat{b}(\theta)e^{\Gamma_*(\theta)}}{\widehat{a}(-\theta)\widehat{b}(\theta)-1}\left\{\int_0^\infty e^{-cz}\int_z^\infty a(t)e^{\theta(t-z)}dtdz\right\}d\theta.
\end{aligned}$$

The last integral can be evaluated as in (6.38)-(6.40). We have thus determined $D(\delta)$ and we summarize our results below.

RESULT 7:

Asymptotic expansions for the mean time for $N(t)$ to reach N_0 in the GI/G/1 queue are given by

(a) $N_0 \to \infty, \quad N_0 - n \to \infty$

$$T(n,y,z) \sim \frac{I_1(c)J_0(c) - I_0(c)J_1(c)}{(1-\rho)^2 c M_1 \Gamma_0^2(c)} \left[\int_0^\infty e^{cy} b(y)dy\right]^{N_0-2} \equiv D$$

(b) $N_0 \to \infty, \quad N_0 - n = m = O(1), \quad m \geq 1$

$$T(n,y,z) \sim D \frac{\int_z^\infty a(t) \int_y^\infty b(\tau) G(m, z-y-t+\tau) d\tau dt}{\int_z^\infty a(t)dt \int_y^\infty b(\tau)d\tau}$$

$$G(m,z) = \frac{M_1 - m_1}{2\pi i} \int_{Br} \frac{e^{-z\theta}}{\widehat{a}(-\theta)\widehat{b}(\theta) - 1} e^{\Gamma_*(\theta)} \left[\widehat{a}(-\theta)\right]^{m-1} d\theta.$$

The various quantities are defined in Result 6. On the contour Br, $0 < \mathrm{Re}(\theta) < \varepsilon_1$, where $\widehat{a}(-\theta)$ is analytic for $\mathrm{Re}(\theta) < \varepsilon_1$.

It is easy to show that Result 7 reduces to the asymptotic results in Result 3 for the M/G/1 model, and to those in Result 4 for the GI/M/1 model. For the M/G/1 model, we have $\widehat{a}(-\theta) = \lambda/(\lambda - \theta)$ and then the contour integral in $G(m,z)$ may be deformed into a small loop about $\theta = \lambda$. For the GI/M/1 model, the integrand in $G(m,z)$ has only two poles (at $\theta = 0$ and $\theta = -c = -b$) in the half-plane $\mathrm{Re}(\theta) < \varepsilon_1$.

8. Discussion and numerical results

We have computed the mean time needed for large workloads and large queue lengths to develop in the GI/G/1 queue. If either the arrival times or service times are exponentially distributed, we have derived exact contour integral representations for the mean first passage times. From these the asymptotic results are easily obtained. For the general model we have computed the mean time asymptotically, by using singular perturbation methods to analyze the appropriate backward equation(s).

The asymptotic method requires that we know explicitly the tail behavior of the steady-state distribution of the queue length or workload. For the GI/G/1 model this may be easily obtained from the Wiener-Hopf theory. The method also requires that we carefully analyze the first passage time for initial conditions that are close to the exit boundary (i.e. $N(0) \simeq N_0$, $U(0) \simeq K$). Using asymptotic methods, we have shown that this involves solving a second Wiener-Hopf problem. However, the second problem is very similar in structure to the computation of the steady-state distribution.

In principle, it should be possible to extend the asymptotic method to other models, such as the m-server GI/G/m queue. For this model, the tail exponent in the steady-state distribution(s) is easily

computed, but it is very difficult to compute the constant in the asymptotic relation(s). The latter would seem to require that we have some exact representation of the steady-state distribution(s). This is known for the GI/M/m model but not for the M/G/m model. Indeed, the analysis of even the M/G/2 queue is quite difficult (see Hokstadt [7]; Cohen [3]; and Knessl, Matkowsky, Schuss and Tier [14]). Thus, while the asymptotic method reduces the first passage time problem to two simpler problems, the solution of the latter may itself be difficult.

We next discuss the numerical accuracy of the asymptotic results. We would like to get some idea as to how large N_0 and K must be before there is good agreement between the exact and asymptotic results. Let N and T be the exact values of the mean first passage times, and let N_{asy} and T_{asy} be the asymptotic formulas we give in Results 1–7, for initial conditions not close to the exit boundary. From Results 1–5, we can easily obtain the estimates

$$\begin{aligned} N(w,z)/N_{asy}(w,z) &= 1+O(EST), \quad K\to\infty \\ T(n,y,z)/T_{asy}(n,y,z) &= 1+O(EST), \quad N_0\to\infty \end{aligned} \tag{8.1}$$

where EST denotes terms that are exponentially small in the appropriate limits (and thus smaller than any power of K^{-1} or N_0^{-1}). The estimate (8.1) is obviously true for the M/M/1 model (cf. Result 1 and (3.7)). It can easily be shown to hold for the M/G/1 and GI/M/1 models by estimating the various contour integrals in Results 2–5. We conjecture that (8.1) is also true for the general GI/G/1 case. This suggests that the asymptotic results should be highly accurate, even for moderate values of K and N_0.

In Tables 1-3 we compare the asymptotic and exact formulas in Result 2 for the M/E_2/1 queue, which has service time density $b(y)=(2\mu)^2 y e^{-2\mu y}$ with $m_1=1/\mu$. We set $\lambda=1$ and in Tables 1, 2, and 3 we have μ = 4, 2, and 4/3; respectively. The respective traffic intensities are thus ρ = 0.25, 0.50, and 0.75. We also set $w=0$ and compare the exact result to the asymptotic result $N(0)\sim C$ in part (a) in Result 2. The contour integrals in the exact answer are easily evaluated since the various integrands have only two or three poles. Also, the solution to (3.10) is

$$a=\frac{1}{2}\left[4\mu-\lambda-\sqrt{\lambda^2+8\mu\lambda}\right]=\frac{\mu}{2}\left[4-\rho-\sqrt{\rho^2+8\rho}\right].$$

In Table 1 we consider values of K in the range $5\le K\le 10$. The exact and asymptotic answers agree to 6 decimal places. Since the traffic intensity is quite small, the mean first passage time is very large (about 10^{10}) even when $K=5$. In Table 2 we increase ρ to 0.5. There is still good agreement between exact and asymptotic answers, with the worst maximum relative error being less than 1% when $K=5$. When $K=10$, the two answers agree to 5 decimal places. In Table 3 we further increase ρ to 0.75. Now the error is unacceptable (about 40%) when $K=5$, but decreases to about 5% when $K=10$ and to under 1% when $K=15$. As $\rho\to 1$ the asymptotics are no longer valid. In Tables 1-3 the

relative error is always under 1% for values of K and ρ which have $K(1-\rho) \geq 4$. This shows that the asymptotic result is quite useful even for moderate values of K, provided ρ is not close to one.

In Tables 4-6 we compare the asymptotic and exact formulas in Result 3 for the M/D/1 queue, which has service time density $b(y) = \delta(y - 1/\mu)$. We consider initial conditions $(n, y) = (0, 0)$ and use the asymptotic result $T(0,0) \sim D$ in part (a) in Result 3. We set $\lambda = 1$ and Tables 4, 5, and 6 have $\mu = 4(\rho = 0.25)$, $\mu = 2(\rho = 0.50)$, and $\mu = 4/3(\rho = 0.75)$; respectively. The exact answer was obtained by using symbolic methods to compute the residues at $s = 0$ for the various contour integrals. To get dependable numerical answers when $\rho = 0.25$, it was necessary to do the calculation using 30 digits of precision. Now $a > 0$ satisfies $\lambda + a = \lambda \exp(a/\mu)$, which we solved numerically. In Tables 4-6 we consider values of N_0 in the range $5 \leq N_0 \leq 15$. When $\rho = 0.25$, Table 4 shows that we get agreement within 1% even when $N_0 = 5$. When $N_0 = 15$, the asymptotic and exact answers agree to 6 decimal places. In Table 5 we increase ρ to 0.5. The error is about 3% when $N_0 = 5$ but decreases to under 1% for $N_0 \geq 7$; when $N_0 = 15$ we get agreement to 6 decimal places. Table 6 has $\rho = 0.75$. When $N_0 = 5$, the error is an unacceptable 50%, but decreases to about 3% when $N_0 = 10$ and to about 0.3% when $N_0 = 15$. Throughout Tables 4-6 the error is always under 1% for $N_0(1-\rho) \geq 4$.

References

[1] **C. M. Bender** and **S. A. Orszag**, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[2] **J. W. Cohen**, *Some results on regular variation for distributions in queueing and fluctuation theory*, J. Appl. Probab. 10, 1970, pp. 343-353.

[3] **J. W. Cohen**, *On the M/G/2 queueing model*, Stoch. Proc. Appl. 12, 1982, pp. 231-248.

[4] **J. W. Cohen**, *The Single Server Queue*, North-Holland, Amsterdam, 1982.

[5] **A. A. Fredricks**, *A class of approximations for the waiting time distribution in a GI/G/1 queueing system*, Bell Syst. Tech. J. 61, 1982, pp. 295-325.

[6] **D. P. Gaver** and **G. S. Shedler**, *Approximate models for processor utilization in multiprogrammed computer systems*, SIAM J. Comput. 2, 1973, pp. 183-192.

[7] **P. Hokstad**, *On the steady-state solution of the M/G/2 queue*, Adv. Appl. Probab. 11, 1979, pp. 240-255.

[8] **D. L. Iglehart**, *Extreme values in the GI/G/1 queue*, Ann. Math. Statist. 43, 1972, pp. 627-635.

[9] **J. Kevorkian** and **J. D. Cole**, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[10] **C. Knessl** and **C. Tier**, *Asymptotic properties of first passage times for tandem Jackson networks I: buildup of large queue lengths*, Comm. Statist.–Stochastic Models 11, 1995, pp. 133-162.

[11] **C. Knessl**, **B. J. Matkowsky**, **Z. Schuss** and **C. Tier**, *An asymptotic theory of large deviations for Markov jump processes*, SIAM J. Appl. Math. 46, 1985, pp. 1006-1028.

[12] **C. Knessl**, **B. J. Matkowsky**, **Z. Schuss** and **C. Tier**, *Asymptotic analysis of a state-dependent M/G/1 queueing system*, SIAM J. Appl. Math. 46, 1986, pp. 483-505.

[13] **C. Knessl**, **B. J. Matkowsky**, **Z. Schuss** and **C. Tier**, *On the performance of state-dependent single server queues*, SIAM J. Appl. Math. 46, 1986, pp. 657-697.

[14] **C. Knessl**, **B. J. Matkowsky**, **Z. Schuss** and **C. Tier**, *An integral equation approach to the M/G/2 queue*, J. Oper. Res. 38, 1990, pp. 506-518.

[15] **B. J. Matkowsky** and **Z. Schuss**, *The exit problem for randomly perturbed dynamical systems*, SIAM J. Appl. Math. 33, 1977, pp. 365-382.

[16] **M. Neuts** and **Y. Takahashi**, *Asymptotic behavior of the stationary distribution in the GI/PH/c queue with heterogeneous servers*, Z. Wahr. 57, 1981, pp. 441-452.

[17] **J. S. Sadowsky** and **W. Szpankowski**, *Maximum queue length and waiting time revisited: G/G/c queue*, Probab. in Eng. & Inform. Sci. 6, 1992, pp. 157-170.

[18] **H. C. Tijms**, *Stochastic Modeling and Analysis: A Computational Approach*, Wiley, New York, 1986.

[19] **M. Williams**, *Asymptotic exit time distributions*, SIAM J. Appl. Math. 42, 1982, pp. 149-154.

Table 1: $\lambda = 1, \quad \mu = 4$

K	Exact	Asymptotic
5	.100519E11	.100519E11
6	.102811E13	.102811E13
7	.105156E15	.105156E15
8	.107554E17	.107554E17
9	.110007E19	.110007E19
10	.112515E21	.112515E21

Table 2: $\lambda = 1, \quad \mu = 2$

K	Exact	Asymptotic
5	3328.58	3340.61
6	14063.80	14077.83
7	59310.03	59326.06
8	249990.73	250008.76
9	.105355E7	.105357E7
10	.443989E7	.443991E7

Table 3: $\lambda = 1, \quad \mu = 4/3$

K	Exact	Asymptotic
5	80.49	111.17
6	141.22	175.91
7	239.65	278.33
8	397.71	440.39
9	650.12	696.81
10	1051.83	1102.52
11	1689.77	1744.45
12	2701.46	2760.14
13	4304.52	4367.21
14	6843.29	6909.98
15	10862.56	10933.24

Table 4: $\lambda = 1, \quad \mu = 4$

N_0	Exact	Asymptotic
5	3464.00	3458.64
6	35657.97	35785.35
7	369810.20	370258.57
8	.383086E7	.383093E7
9	.396417E8	.396373E8
10	.410127E9	.410114E9
11	.424330E10	.424330E10
12	.439038E11	439040E11
13	.454259E12	.454259E12
14	.470006E13	.470006E13
15	.486299E14	.486299E14

Table 5: $\lambda = 1, \quad \mu = 2$

N_0	Exact	Asymptotic
5	178.24	183.36
6	637.66	644.11
7	2255.39	2262.70
8	7940.53	7948.56
9	27913.28	27922.21
10	98076.86	98086.89
11	.344554E6	.344565E6
12	.121039E7	.121041E7
13	.425199E7	.425201E7
14	.149367E8	.149367E8
15	.524706E8	.524706E8

Table 6: $\lambda = 1, \quad \mu = 4/3$

N_0	Exact	Asymptotic
5	40.72	59.13
6	81.10	102.52
7	153.32	177.73
8	280.71	308.11
9	503.74	534.15
10	892.60	926.00
11	1568.92	1605.32
12	2743.58	2782.99
13	4782.19	4824.60
14	8318.53	8363.93
15	14451.32	14499.72