THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 3, March 1982
Printed in U.S.A.

A Class of Approximations for the Waiting Time Distribution in a *GI/G/1* Queueing System

By A. A. FREDERICKS

(Manuscript received August 11, 1981)

Single server queueing models are important in the study of a wide variety of stochastic systems. A particularly important example is the study of task schedules for computer systems with real-time applications. In this paper, we present a class of approximations for the waiting time distribution in single server queueing systems with general independent (renewal) input and general (independent) service time distributions. These approximations allow the analyst to use as much (or as little) of the structure of the input and service processes as desired. Moreover, they also allow him to concentrate on specific quantities associated with the delay distribution that may represent the most relevant performance criteria. Examples include probability of delay, mean delay, and tails of the delay distribution. The methods given in this paper have been used to analyze the performance of task schedules in a variety of processor-based systems.

I. INTRODUCTION

Single server queueing models arise quite naturally in the study of a wide variety of stochastic server systems. A particularly important class is single processor computer systems, such as stored program control switching systems or nodes in a data communication network. In many such applications, it is important to keep as much of the structure of the interarrival and service time processes as possible in order to obtain realistic results; that is, the simplifying assumptions of exponential distributions cannot be made. While the resulting *GI/G/1* queue may be extremely difficult to analyze, one is often content with reasonable approximations that incorporate the main features of the problem. In addition, one often desires results that are reasonably simple analytically, since the behavior of the *GI/G/1* queueing model may be the input to the analysis of a more complex system.

Our main purpose here is to present a class of approximations for the waiting time distribution, $W(x)$, for such $GI/G/1$ systems which allows the analyst to use as much (or little) of the structure of the relevant input and service processes as desired. The resulting approximations can be extremely simple in form (e.g., a single exponential) or, with additional effort, more complex. In particular, we give relatively simple expressions for constants C and a, such that $W_A(x) = 1 - Ce^{-ax}$ provides a good fit to the probability of delay, $P_D = [1 - W(0)]$ and the mean delay, $\bar{w}$. Similar approximations are developed which provide a good fit to the tails of the delay distribution. Thus, these results can be used to study systems with a wide variety of performance criteria. Our results also lead to some new heavy traffic approximations, as well as a simple approximation for light traffic.

While we feel that the approximations presented here are of use in themselves, we hope that the results obtained will stimulate more active research into the general method presented. Of particular interest is the problem of obtaining more quantitative error bounds to guide the user in the application of these techniques.

We note that the methods given here have been used to analyze some rather complex $GI/G/1$ queueing systems which have arisen in the study of a certain class of computer systems with real-time applications.[1,2]

Some key results are summarized in the next section to give a general idea of the nature of this work. Our basic approach is introduced in Section III where several (single) exponential approximations are derived. In Section IV, we look briefly at the implications of the approximations for heavy traffic. In Section V, we look at some special cases which lead to analytic statements about the accuracy of our various approximations, and in Section VI, their accuracy is assessed via several numerical examples. Extensions to more general functional forms are considered in Section VII and illustrated via another numerical example. Some final remarks are given in Section VIII. A summary of our notation, key formulas, and approximations are given in the Appendix for easy reference.

II. SUMMARY OF SOME KEY RESULTS

The main idea of this work is to assume a functional form for an approximation $W_A(x)$ to the true waiting time distribution $W(x)$ for a $GI/G/1$ queue. The Lindley integral equation for $W(x)$ is then used to obtain the unknown coefficients. We obtain several approximations by assuming $W(x) \approx W_A(x) = 1 - Ce^{-ax}$. A key approximation of this type, referred to as approximation A^* has C and a determined by eqs. (7) and (9) with $a = a_E$ (also see pages 299 and 300):

$$C = C^* = \frac{1 - K(0)}{1 - \hat{K}_-(a_E)}$$

and

$$\frac{C}{a} = \frac{C^*}{a^*} = \frac{u_+}{K(0) - \hat{K}_-(a_E)},$$

where $K(u)$ is the distribution function for the difference, $\tilde{u}$,[†] between the service time and interarrival time and

$$\hat{K}_-(s) = \int_{-\infty}^{0} e^{su} dK(u)$$

$$\bar{u}_+ = \int_0^{\infty} u dK(u)$$

and a_E satisfies the characteristic equation

$$\int_{-\infty}^{\infty} e^{a_E u} dK(u) = 1.$$

(The equation numbers of this section correspond to those of later sections.) Limiting properties and special cases for this approximation, as well as others are discussed. Several numerical examples are also given to illustrate the accuracy of these approximations. For an example $H_2/E_3/1$ system, the maximum relative error in the approximation to P_D given by A* is found to be 5.0 percent, while the maximum relative error in the approximation to $\bar{w}$ given by A* is found to be 1.5 percent. Perhaps the simplest approximation considered (using only the first two moments of $\tilde{u}$) is the one referred to as approximation $A_{G,0}$, which is also of the form $W_A(x) = 1 - Ce^{-ax}$, but where now the constants C and a are determined by

$$a = a_H = \frac{-2\bar{u}}{\sigma_u^2}, \tag{14}$$

i.e., the heavy traffic a, and

$$C = C_{G,0} = [1 - \mathrm{Erf}(|\bar{u}|/\sqrt{2}\sigma u)]/[1 + \mathrm{Erf}(|\bar{u}|/\sqrt{2}\sigma u)], \tag{15}$$

where Erf(x) is the error function. Note that (15) has the limiting behavior for small ($\bar{u}$) (assuming σ_u tends to a finite limit)

† The symbol ~ will be used to denote a random variable (see the Appendix for a summary of notations used).

$$C_{G,0} \underset{\substack{\bar{u}\to 0\\(\rho\to 1)}}{\sim} 1 - 2\sqrt{\frac{2}{\pi}}\frac{|\bar{u}|}{\sigma_u}, \tag{16}$$

which then results in the following limit for the approximate mean delay, $\bar{w}_{G,0}$

$$\bar{w}_{G,0} = \frac{C_{G,0}}{a_H} \underset{\substack{\bar{u}\to 0\\(\rho\to 1)}}{\sim} \frac{1}{a_H} - \sigma_u\sqrt{\frac{2}{\pi}} \tag{17}$$

the right-hand side of eq. (17) is, in fact, a well-known lower bound for $\bar{w}$ (as $\bar{u} \to 0$, e.g., see Ref. 4).

The accuracy of these approximations, as well as several others we develop, are studied in considerably more detail in what follows. We also show how the approach taken leads to a class of approximations which can be made increasingly more accurate (with additional effort).

III. DEVELOPMENT OF THE EXPONENTIAL APPROXIMATION

We wish to obtain an approximate solution for the waiting time distribution, $W(x)$, for a $GI/G/1$ queueing system. More specifically, we have a single server queue where the interarrival times, $\tilde{t}$, between customers are independent and identically distributed, taken from a general distribution, $A(t)$ and the service times, $\tilde{\tau}$, are independently drawn from an arbitrary distribution $B(\tau)$. A first in-first out discipline is assumed. If we let $\tilde{u} = \tilde{\tau} - \tilde{t}$ and denote the distribution of $\tilde{u}$ by $K(u)$, then if an (equilibrium) waiting time distribution exists,† it satisfies the well-known Lindley integral equation[3]

$$W(x) = \int_{-\infty}^{x} W(x-y)dK(y), \qquad x \geq 0 \tag{1}$$

($W(x)$ is equal to 0 for $x < 0$).

In this section, we consider approximations $W_A(x)$ to $W(x)$ of the form

$$W_A(x) = 1 - Ce^{-ax}. \tag{2}$$

Our objective is to determine suitable values for the constants C and a. For this purpose, we will use the relation (1).

3.1 Pointwise matching

As is well known (see Ref. 5, pages 376 and 410), if the equation

$$\int_{-\infty}^{\infty} e^{au}dK(u) = 1, \tag{3}$$

† Throughout we assume stability, i.e., that $(1/\alpha) = \bar{\tau} < \bar{t} = (1/\lambda)$; $\rho = (\lambda/\alpha) < 1$.

possesses a real nonzero root, a_E, then

$$1 - W(x) \underset{x\to\infty}{\sim} C_\infty e^{-a_E x}. \tag{4}$$

Throughout this paper, we will tacitly assume that for all systems considered, eq. (3) does, in fact, possess such a root. Thus, if an exponential form is to be used, a_E is a reasonable choice for the parameter a in (2). On the other hand, substitution of (2) into (1) yields the relation

$$1 - Ce^{-ax} = K(x) - Ce^{-ax}\int_{-\infty}^{x} e^{ay}dK(y). \tag{5}$$

Hence, we see that if we choose $a = a_E$, (5) will be valid to order $0(e^{-a_E x})$† as x tends to infinity. We can now determine C by requiring that (5) be valid at another value of x. For x such that $[1 - K(x)] > 0$ we can solve (5) for C to obtain

$$C(x) = \frac{[1 - K(x)]e^{ax}}{1 - \displaystyle\int_{-\infty}^{x} e^{ay}dK(y)}. \tag{6}$$

Now if we are interested in approximating the behavior of $W(x)$ near $x = 0$, a reasonable choice might be to solve for C from (6) with $x = 0$; that is, choose

$$C = C_0(a) = \frac{1 - K(0)}{1 - \hat{K}_-(a)} = \frac{1 - \hat{K}_-(0)}{1 - \hat{K}_-(a)}, \tag{7}$$

where

$$\hat{K}_-(s) = \int_{-\infty}^{0} e^{su}dK(u).$$

[Note that clearly $0 \le C_0(a) \le 1$.]

We will refer to the exponential approximation of $W(x)$ obtained by choosing a to satisfy (3) ["(5) at $x = \infty$"] and C to satisfy (7) [(5) at $x = 0$] as approximation $A_{E,0}$, i.e., $W_{A_{E,0}} = 1 - C_0(a_E)e^{-a_E x}$. (The subscript E refers to the use of the exact dominant root—we will consider other alternatives shortly.)

Note that if the true delay distribution were given by a simple exponential, i.e., $W(x) = 1 - Ce^{-ax}$, then the value of $C(x)$ determined from (6) would be identically a constant and $W_{A_{E,0}}(x) \equiv W(x)$. Thus, the *variation* of $C(x)$ with x, as determined by (6), gives some indica-

† That is, the difference between the left-hand side and right-hand side will tend to zero even when multiplied by $Ce^{a_E x}$.

tion of *how* exponential the waiting time delay distribution is. A more "sophisticated" approximation might be obtained by minimizing the variation of $C(x)$ with respect to some suitable norm, however, we will concentrate here mostly on simpler techniques.

Note also that $C(x)$ as given by (6) appears in various (exponential) bounds for $W(x)$. For example, Kingman[6] and Rossberg and Siegel[7] prove the following inequality for the complementary waiting time distribution, $W^c(x) = 1 - W(x)$.

$$C_L e^{-a_E x} \le W^c(x) \le e^{-a_E x},$$

where a_E is given by (3) and

$$C_L = \inf_{\substack{x \ge 0 \\ [1-K(x)]>0}} C(x).$$

Ross[8] gives the bound

$$W^c(x) \le C_U e^{-a_E x},$$

where

$$C_U = \sup_{\substack{x \ge 0 \\ [1-K(x)]>0}} C(x).$$

(See also Ref. 9.) Hence, it seems quite natural to choose a specific value for x in $C(x)$ to obtain an exponential approximation to $W(x)$, with $C_L < C < C_U$.

3.2 An alternate method for determining C

As we shall see shortly, $A_{E,0}$ does provide a reasonable approximation to P_D, but the resulting approximation for $\bar{w}$ is not as good. One might consider choosing C from (6) with another choice of x; however, here we consider another alternative. If we compute the mean waiting time, $\bar{w}$, using eq. (1) we obtain

$$\bar{w} = \int_0^\infty x dW(x) = \int_0^\infty x \left[W(0)dK(x) + \int_{-\infty}^x d_x W(x-y)dK(y) \right]. \quad (8)$$

Again, assuming the exponential form (2) for $W_A(x)$ [and, hence, that $\bar{w}_A = (C/a)$], (8) yields

$$\bar{w}_A = \frac{C}{a} = \frac{\bar{u}_+}{K(0) - \hat{K}_-(a)} = \frac{\hat{K}'_+(0)}{\hat{K}_-(0) - \hat{K}_-(a)}, \quad (9)$$

where $\bar{u}_+ = \int_0^\infty u dK(u)$, $\hat{K}_+(s) = \int_0^\infty e^{su} dK(u)$ and $\hat{K}_-(s)$ is as in (7). If a is chosen from (3) and C then from (9), we will refer to the resulting approximation as $A_{E,1}$, i.e., $W_{A_{E,1}}(x) = 1 - C_1(a_E)e^{-a_E x}$. As we shall see, this will result in a good approximation for the mean delay via (9);

however, it is not clear from this relation that the resulting C need be less than 1. (It is clearly ≥ 0.) A correct probabilistic exponential form for $W_A(x)$ which incorporates both (7) (and, hence, results in a reasonable approximation for the probability of delay) and (9) (and hence results in a good approximation for the mean delay) can be obtained in the following manner. First, find a_E from (3), then $C = C^* = C_0(a_E)$ from (7), and then the approximate mean delay, $\bar{w} = C_1(a_E)/a_E$ from (9). The constant a is then chosen to be

$$a = a^* = \frac{C_0(a_E)}{C_1(a_E)}\, a_E.$$

We will refer to this approximation as A^*. We now look at a more organized way of obtaining such approximations.

3.3 Moment matching

Equation (9) was obtained by matching the first moment of the Lindley integral equation, i.e., from (8). We note that if we match the 0th moment in a similar way we are, in fact, also led to (7). That is, if we seek a solution to (1) of the form $1 - Ce^{-ax}$, and match the first two moments (0th and 1st) of (1) to determine *both* unknown constants, we are led to eqs. (7) and (9), i.e.,

$$C = \frac{1 - K(0)}{1 - \hat{K}_-(a)}$$

$$\frac{C}{a} = \frac{\bar{u}_+}{K(0) - \hat{K}_-(a)}$$

as a pair of equations in the two unknowns C and a. These can be combined and rewritten as

$$a = \frac{[K(0) - \hat{K}_-(a)][1 - K(0)]}{\bar{u}_+[1 - \hat{K}_-(a)]} \qquad (10)$$

$$C = \frac{1 - K(0)}{1 - \hat{K}_-(a)}. \qquad (11)$$

Thus, one can first determine a from (10) and then C from (11). Clearly, if such an a exists, $a \in [0, \infty]$, $C \in [0, 1]$.

Denoting the right-hand side of (10) by $f(a)$, we note that

$$f(a) \underset{a \to 0}{\longrightarrow} 0$$

$$f(a) \underset{a \to \infty}{\longrightarrow} \frac{K(0)[1 - K(0)]}{\bar{u}_+} < \infty$$

and also

$$f'(a) = \frac{-\hat{K}'_-(a)[1 - K(0)]^2}{\overline{u_+}[1 - \hat{K}_-(a)]^2}$$

has the properties

$$f'(a) > 0, \qquad a \geq 0$$

$$f'(0) = \frac{\bar{u}_+ - \bar{u}}{\bar{u}_+} > 1$$

$$f'(a) \underset{a \to \infty}{\longrightarrow} 0.$$

Moreover,

$$f''(a) = \frac{-[1 - K(0)]^2}{\bar{u}_+} \frac{(\hat{K}_-(a)[1 - \hat{K}_-(a)] + 2(\hat{K}_-(a))^2}{[1 - \hat{K}_-(a)]^3} < 0, \qquad a \geq 0.$$

Thus, we see that (10) possesses a unique positive root, which can be obtained via the iteration scheme

$$a_{n+1} = f(a_n)$$

starting from any positive a_0.

Thus, we can always determine the desired root $a_{0,1}$ from (10) and then the resulting $C = C_{0,1}$ from (11). This approximation

$$W_{A_{0,1}}(x) = 1 - C_{0,1}e^{-a_{0,1}x}$$

is referred to as $A_{0,1}$.

3.4 Light traffic

Under quite general conditions, the function $\hat{K}_-(a)$ defined in (7) and with a satisfying (3) has the limit

$$\hat{K}_-(a) \underset{\rho \to 0}{\longrightarrow} 0,$$

e.g., fix $B(\tau)$ and the "shape" of $A(t)$ and let $1/\lambda \to \infty$. Hence, for ρ small, (7) and (9) lead to the following intuitively appealing "light" traffic approximations (A_{LT})

$$C_{LT} = 1 - K(0) \tag{12}$$

$$\bar{w}_{LT} = \frac{C_{LT}}{a_{LT}} = \frac{\bar{u}_+}{K(0)}. \tag{13}$$

IV. HEAVY TRAFFIC

In the case of heavy traffic, an obvious possibility for the exponent a is the heavy traffic limit (e.g., see Ref. 4):

$$a_H = \frac{-2\bar{u}}{\sigma_u^2}, \tag{14}$$

where $\bar{u}$, σ_u^2 are the mean and variance, respectively, of $\tilde{u} = \tilde{\tau} - \tilde{t}$ (service minus interarrival time). We can use this value for a in place of a_E in the approximations of Section III. In particular, if we use $a = a_H$ in (7) we will refer to the resulting approximation as $A_{H,0}$. Note that the resulting C is strictly less than one unless $a_H = 0$ ($\rho = 1$). If we use $a = a_H$ in (9), we refer to the resulting approximation as $A_{H,1}$. These two approximations (as well as those we discuss next) offer potential improvements to the more standard heavy traffic approximation that can extend its applicability to lower load levels.

In many cases, where heavy traffic approximations are applied, the structure of $K(u)$ is either not known or too complex to be used in further analysis. Standard heavy traffic approximations generally only make use of the first two moments of $K(u)$, i.e., $\bar{u}$ and σ_u^2. We can also obtain results which use only $\bar{u}$, σ_u^2 by making the standard heavy traffic assumption that $K(u)$ is approximately Gaussian. For example, if we not only assume that $a = a_H$ in (7), but further that $K(u)$ is Gaussian, we obtain the following approximation for C using (7):

$$C_{G,0} = \frac{1 - \mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)}{1 + \mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)}. \tag{15}$$

We refer to this approximation as $A_{G,0}$. Note that (15) implies the limiting behavior for $|\bar{u}|$ small

$$C_{G,0} \underset{\substack{\bar{u}\to 0 \\ (\rho\to 1)}}{\sim} \frac{1 - \sqrt{\frac{2}{\pi}}\frac{|\bar{u}|}{\sigma_u}}{1 + \sqrt{\frac{2}{\pi}}\frac{|\bar{u}|}{\sigma_u}} \underset{\substack{\bar{u}\to 0 \\ (\rho\to 1)}}{\sim} 1 - 2\sqrt{\frac{2}{\pi}}\frac{|\bar{u}|}{\sigma_u}. \tag{16}$$

Now combining (16) with (14), we obtain the following (limiting) approximation to the mean delay for $|\bar{u}|$ small

$$\bar{w}_{G,0} = \frac{C_{G,0}}{a_H} \underset{\substack{\bar{u}\to 0 \\ (\rho\to 1)}}{\sim} \frac{1}{a_H} - \sigma_u\sqrt{\frac{2}{\pi}}. \tag{17}$$

In a similar manner, (9) yields the approximation $A_{G,1}$

$$\bar{w}_{G,1} = \frac{C_{G,1}}{a_H} = \frac{\frac{\sigma_u}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{u}}{\sigma_u}\right)^2} + \frac{\bar{u}}{2}\left[1 - \mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)\right]}{\mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)}, \tag{18}$$

which has the limit

$$\bar{w}_{G,1} = \frac{C_{G,1}}{a_H} \underset{\substack{\bar{u}\to 0 \\ (\rho\to 1)}}{\sim} \frac{\sigma_u^2}{2|\bar{u}|} - \frac{\sigma_u}{2}\sqrt{\frac{\pi}{2}} \underset{\substack{\bar{u}\to 0 \\ (\rho\to 1)}}{\sim} \frac{1}{a_H} - \frac{\sigma_u}{2}\sqrt{\frac{\pi}{2}}. \tag{19}$$

Now if $K(u)$ were, indeed, Gaussian, then the resulting (heavy traffic) mean delay, $\bar{w}_G$, is known to have the limiting bounds (e.g., see Ref. 4).

$$\frac{1}{a_H} - \sigma_u\sqrt{\frac{2}{\pi}} \le \bar{w}_G \le \frac{1}{a_H}. \tag{20}$$

Compare (17), (19), and (20).

V. SOME SPECIAL CASES

In this section, we look at the general behavior of these approximations for some special cases where exact results are readily obtainable. We begin by looking at the $GI/M/1$ case where the approximations of Section III are exact. This will serve as a simple illustrative example, as well as provide some additional insight into the behavior of the heavy traffic approximations of Section IV.

5.1 Case: GI/M/1

For a $GI/M/1$ system, we can write (for $u > 0$)

$$K(u) = \int_0^\infty [1 - e^{-\alpha(u+t)}]\mathrm{dA}(t) = 1 - pe^{-\alpha u}, \qquad u > 0, \tag{21}$$

where $1/\alpha$ is the mean of the (exponential) service time and

$$p = \int_0^\infty e^{-\alpha t}\mathrm{dA}(t) = \hat{A}(\alpha).$$

We, thus, have

$$\bar{u}_+ = \alpha p \int_0^\infty x e^{-\alpha x}\mathrm{dx} = p/\alpha \tag{22}$$

$$K(0) = 1 - p. \tag{23}$$

Now if a satisfies (3),

$$\hat{K}_-(a) = \int_{-\infty}^0 e^{au}\mathrm{dK}(u) = 1 - \int_0^\infty e^{au}\mathrm{dK}(u).$$

Hence,

$$\hat{K}_-(a) = 1 - \alpha p \int_0^\infty e^{au}e^{-\alpha u}\mathrm{du} = 1 + \frac{\alpha p}{a - \alpha}. \tag{24}$$

Using (23), (24) in (7) yields

$$C_0(a) = \frac{\alpha - a}{\alpha}, \tag{25}$$

while (22), (23), (24) in (9) yield

$$\frac{C_1(a)}{a} = \frac{\alpha - a}{a\alpha}. \tag{26}$$

If a is chosen to be the exact root of (3), then these both lead to the exact waiting time distribution. (Recall, for a $GI/M/1$ system

$$W(x) = 1 - \frac{\alpha - a_E}{\alpha} e^{-a_E x},$$

e.g., see Ref. 10.)

If we use the heavy traffic approximation for a, i.e., (14) then (25) or (26) can be used to obtain the following approximation for C

$$C_{H,0} = \frac{\alpha + \dfrac{2\bar{u}}{\sigma_u^2}}{\alpha} = 1 - \frac{2}{\alpha^2\sigma_u^2}\frac{(1-\rho)}{\rho}; \tag{27}$$

that is, a possible improvement to $C = 1$ for heavy traffic. Thus, the resulting mean delay is found to be

$$\bar{w}_{H,0} = \frac{C_{H,0}}{a_H} = \frac{1}{a_H} - \frac{1}{\alpha} = \frac{\sigma_\mu^2}{2|\bar{u}|} - \frac{1}{\alpha}. \tag{28}$$

Note that this correction term to $1/a_H$ as an approximation to the mean delay is typical. For example, for an $M/G/1$ system it is easy to show that the true mean delay, $\bar{w}$, satisfies

$$\bar{w} = \frac{1}{a_H} - \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{\alpha}\right).$$

5.2 Case: M/G/1

For this case, we have

$$K(u) = \int_0^\infty e^{\lambda(u-\tau)} dB(\tau) = qe^{\lambda u}, \qquad u < 0, \tag{29}$$

where $1/\lambda$ is the mean of the (exponential) interarrival time and

$$q = \int_0^\infty e^{-\lambda\tau} dB(\tau) = \hat{B}(\lambda).$$

We, thus, have

$$\bar{u}_+ = \int_0^\infty u dK(u) = \int_{-\infty}^\infty u dK(u) - \int_{-\infty}^0 u dK(u)$$

$$= \bar{u} - q\lambda \int_{-\infty}^0 u e^{\lambda u} du = \left(\frac{1}{\alpha} - \frac{1}{\lambda} + \frac{q}{\lambda}\right) \quad (30)$$

$$K(0) = q \quad (31)$$

and

$$\hat{K}_-(a) = q\lambda \int_{-\infty}^0 e^{au} e^{\lambda u} du = \frac{q\lambda}{a + \lambda}. \quad (32)$$

Using (7), we obtain the following approximation ($A_{E,0}$) for the probability of delay, P_D

$$C_0(a_E) = \frac{1 - q}{1 - \dfrac{q\lambda}{a_E + \lambda}}, \quad (33)$$

while from (9) we obtain the following approximation ($A_{E,1}$) to the mean delay, $\bar{w}$

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} = \frac{(\lambda - \alpha + q\alpha)(a_E + \lambda)}{q\alpha\lambda a_E}. \quad (34)$$

(Since we know that for an $M/G/1$ queue, the true value of P_D is ρ, we could use this fact, together with (33) to obtain a simple approximation for a_E; we discuss this possibility shortly.)

Using the known limits for the true $a = a_E$ as $\rho \to 1$, (33) readily yields the following:

$$C_0(a_E) \underset{\substack{\lambda \to 0 \\ (\rho \to 0)}}{\sim} \rho(1 - \rho\overline{\tau^2}\alpha^2) \underset{\substack{\lambda \to 0 \\ (\rho \to 0)}}{\sim} \rho \quad (35)$$

$$C_0(a_E) \underset{\substack{\lambda \to \alpha \\ (\rho \to 1)}}{\sim} 1 - \frac{q}{1-q}\frac{a}{\alpha} \underset{\substack{\lambda \to \alpha \\ (\rho \to 1)}}{\to} 1, \quad (36)$$

where we have used the fact that for $\lambda \to 0$

$$q = \int_0^\infty e^{-\lambda\tau} \mathrm{d}B(\tau) \underset{\lambda \to 0}{\sim} 1 - \frac{\lambda}{\alpha} + \frac{\lambda^2\overline{\tau^2}}{2}. \quad (37)$$

We similarly obtain from (34)

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} \underset{\substack{\lambda \to 0 \\ (\rho \to 0)}}{\sim} \frac{\lambda\overline{\tau^2}}{2(1-\rho)}\left(1 + \frac{\lambda}{a_E}\right) \underset{\substack{\lambda \to 0 \\ (\rho \to 0)}}{\sim} \frac{\lambda\overline{\tau^2}}{2(1-\rho)} \quad (38)$$

and for $\lambda \to \alpha$, clearly

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} \underset{\substack{\lambda\to\alpha \\ (\rho\to 1)}}{\sim} \frac{1}{a_E}. \tag{39}$$

Equation (38) shows that as $\lambda \to 0$, we obtain the familiar $P - K$ (Pollecek-Khinchin) formula for $\bar{w}$ in an $M/G/1$ system, while (39) shows that with

$$a = a_E, \qquad \bar{w}_{E,1} \underset{\rho\to 1}{\sim} \frac{1}{a_E} \underset{\rho\to 1}{\sim} \frac{1}{a_H}.$$

Now recall for the approximation A^*, we take $C^* = C_0(a_E)$ and

$$a^* = \frac{C_0(a_E)a_E}{C_1(a_E)}.$$

Hence, for A^*, the resulting exponent for the exponential approximation is given by

$$a^* = \frac{q\alpha\lambda a_E(1-q)}{(\lambda - \alpha + q\alpha)(a_E + \lambda - q\lambda)}. \tag{40}$$

From (40) we obtain

$$a^* \underset{\substack{\lambda\to 0 \\ (\rho\to 0)}}{\to} \frac{2}{\alpha\overline{\tau^2}} = \frac{1}{\bar{R}} \tag{41}$$

$$a^* \underset{\substack{\lambda\to\alpha \\ (\rho\to 1)}}{\sim} a_E\left\{1 - \left[\frac{a_E(1-q)}{\alpha}\right]\right\} \underset{\substack{\lambda\to\alpha \\ (\rho\to 1)}}{\sim} a_E, \tag{42}$$

where $\bar{R}$ is mean forward recurrence time of the service time distribution. Note we have used the fact that

$$a_E \underset{\rho\to 1}{\sim} 0.$$

The first of these is a very interesting relation. It is easy to show that for an $M/G/1$ system, that as $\rho \to 0$ the true dominant root a_E satisfies

$$\hat{B}(-a_E) = \int_0^\infty e^{a_E\tau}\mathrm{dB}(\tau) \underset{\rho\to 0}{\sim} \infty. \tag{43}$$

Hence, for example, if $\hat{B}(s)$ is rational, a_E tends to the smallest (in magnitude) pole of $\hat{B}(s)$, which is not consistent with (41). However, it is easy to show that for an $M/G/1$ system, the *mean* delay conditioned on being delayed, is just the mean forward recurrence time, $\bar{R}$—the dominant contribution is from the case where the arrival finds one customer in the system. That is, for $\rho \to 0$, the dominant root really may not be the best quantity to use for an approximation, indeed, the limit (41) is most likely preferable. We, thus, have via A^* an exponential approximation to the true delay distribution for $\rho \to 0$ which can be decidedly *nonexponential* but with the *correct* limiting probability of delay *and* mean delay.

The approximation $A_{0,1}$ also has some very interesting properties for the $M/G/1$ case. Considering (7) and (9) as defining two equations for the two unknowns C and a, we obtain the two relations

$$(a + \lambda) = \frac{Cq\lambda}{C + q - 1} \tag{44}$$

and

$$(a + \lambda) = \frac{Cq\alpha\lambda}{\lambda - \alpha + q\alpha}. \tag{45}$$

Combining (44) and (45) we find that

$$C = \frac{\lambda}{\alpha} = \rho. \tag{46}$$

That is, for the $M/G/1$ case, $A_{0,1}$ yields an *exact* expression for the probability of delay, P_D. Using (44) or (45), we obtain the following approximation for the dominant root:

$$a = \frac{\alpha\rho^2 q}{(\rho - (1 - q))} - \alpha\rho = \frac{\rho\alpha(1 - q)(1 - \rho)}{\rho - (1 - q)}. \tag{47}$$

Note that

$$a \underset{\rho \to 1}{\sim} \frac{\alpha(1 - q)(1 - \rho)}{q}, \tag{48}$$

but

$$a \underset{\rho \to 0}{\to} \frac{2}{\alpha\overline{\tau^2}}$$

as with A^*.

Note that since for an $M/G/1$ system, we know that $P_D = \rho$, we could have used this fact directly in either of the two approximations [(33), (34)] to obtain an approximation for the true dominant root, a_E. This, of course, would result in

$$a_E \cong \frac{\rho\alpha(1 - q)(1 - \rho)}{\rho - (1 - q)},$$

i.e., just (47).

For the mean delay, $\bar{w}$, $A_{0,1}$ yields the approximation

$$\bar{w}_{0,1} = \frac{\rho - (1-q)}{\alpha(1-q)(1-\rho)}, \tag{49}$$

which has the asymptotic properties

$$\bar{w}_{0,1} \underset{\substack{\lambda\to 0\\(\rho\to 0)}}{\sim} \frac{\lambda\overline{\tau^2}}{2\left(1-\rho-\dfrac{\lambda\alpha\overline{\tau^2}}{2}\right)} \tag{50}$$

$$\bar{w}_{0,1} \underset{\substack{\lambda\to \alpha\\(\rho\to 1)}}{\sim} \frac{q}{\alpha(1-q)(1-\rho)}. \tag{51}$$

Comparing the asymptotic behavior of $A_{0,1}$ and A^*, we see that while the former has a more desirable property for the approximation to P_D, the behavior of the approximations to $\bar{w}$ and a_E are more desirable for A^* than for $A_{0,1}$. This leads us to consider another approximation, one that combines the best properties of each. The simplest method is to define approximation A^{**} via $w_{A^{**}}(x) = 1 - C^{**}e^{-a^{**}x}$, where

$$C^{**} = C_{0,1}$$

and

$$a^{**} = \frac{C_{0,1}}{\bar{w}_{E,1}} = \frac{C_{0,1}a_E}{C_{E,1}}.$$

The use of the heavy traffic a_H, of course, yields similar results to those for $A_{E,0}$, $A_{E,1}$, and A^* as $\rho \to 1$. Again, there is no simplification for the Gaussian case.

VI. SOME NUMERICAL EXAMPLES

We consider several numerical examples of varying complexity to illustrate the accuracy of the various approximations. Because of the large number of possible combinations of approximations and quantities of interest, we will not cover all possibilities for every example.

6.1 Case: M/D/1

For the $M/D/1$ case,

$$K(u) = P(\tilde{u} = \bar{\tau} - \tilde{t} \le u) = \begin{cases} 1, & u > \bar{\tau} \\ e^{-\lambda(\bar{\tau}-u)}, & u \le \bar{\tau} \end{cases}.$$

Hence, we can readily obtain eq. (3) for the dominant root, a_E

$$\int_{-\infty}^{\infty} e^{au}\mathrm{dK}(u) = \frac{\lambda}{a+\lambda}\, e^{a\bar{\tau}} = 1. \tag{52}$$

Rewriting this equation as

$$\lambda\bar{\tau}(e^{a\bar{\tau}} - 1) = a\bar{\tau}, \tag{53}$$

it is clear that in addition to $a = 0$, there is always one positive real root, a_E, for any finite λ. Moreover,

$$a_E \underset{\lambda\bar{\tau}=\rho\to 1}{\longrightarrow} 0, \qquad a_E \underset{\substack{\lambda\to 0\\(\rho\to 0)}}{\longrightarrow} \infty.$$

The other quantities needed for the various approximations are

$$q = \hat{B}(\lambda) = e^{-\rho} \tag{54}$$

$$K(0) = q = e^{-\rho} \tag{55}$$

$$\hat{K}_{-}(a_E) = \frac{q\lambda}{a_E + \lambda} = \frac{e^{-\rho}\lambda}{a_E + \lambda} \tag{56}$$

$$\bar{u}_{+} = \left(\frac{1}{\alpha} - \frac{1}{\lambda} + \frac{q}{\lambda}\right) = \left(\bar{\tau} - \frac{1}{\lambda} + \frac{e^{-\rho}}{\lambda}\right) \tag{57}$$

$$\bar{u} = \left(\frac{1}{\alpha} - \frac{1}{\lambda}\right) = \bar{\tau} - \frac{1}{\lambda} \tag{58}$$

$$\sigma_u^2 = \frac{1}{\lambda^2} \tag{59}$$

and, hence,

$$a_H = \frac{-2\bar{u}}{\alpha_u^2} = \frac{2\rho(1-\rho)}{\bar{\tau}}. \tag{60}$$

With a_E from (53) and (54) to (60), we can compute the various quantities predicted by our approximations. For simplicity, we will assume $\bar{\tau} = 1$ in the following so that $\lambda = \rho$.

6.1.1 Probability of delay, P_D

Table I shows the value of P_D as predicted by the various approximations. Recall that P_D from $A_{0,1}$ is exact for this case. We see, however, that $A_{E,0}(A^*)$ produces a good approximation to this quantity. The value of P_D predicted by $A_{E,1}$ is seen to be significantly poorer, particularly for ρ small. Actually, our heavy traffic approximation $A_{H,0}$ does significantly better than $A_{E,1}$. Note that $A_{G,0}$ provides a reasonable

Table I—Probability of Delay, P_D, case $M/D/1$

ρ	True	$A_{0,1}(A^{**})$	$A_{E,0}(A^*)$	$A_{H,0}$	$A_{G,0}$	$A_{E,1}$	A_{LT}
0.01	0.0100	0.0100	0.0100	0.0149	0.1920	0.0327	0.0100
0.1	0.1000	0.1000	0.0975	0.1406	0.2256	0.1986	0.0952
0.5	0.5000	0.5000	0.4756	0.5647	0.4462	0.6170	0.3935
0.9	0.9000	0.9000	0.8864	0.8975	0.8524	0.9276	0.5934
0.99	0.9900	0.9900	0.9884	0.9885	0.9842	0.9928	0.6284

Table II—Mean delay, $\bar{w}$, case $M/D/1$

ρ	True	$A_{E,1}(A^*, A^{**})$	$A_{0,1}$	$A_{E,0}$	$A_{H,1}$	$A_{G,1}$	A_{HT}	A_{LT}
0.01	0.00505	0.00505	0.00507	0.00154	0.00759	12.53	50.51	0.00503
0.1	0.0556	0.0550	0.0565	0.0270	0.0832	1.589	5.556	0.0536
0.5	0.5000	0.4911	0.5415	0.3785	0.7026	1.033	2.000	0.3513
0.9	4.500	4.478	5.166	4.279	5.027	4.895	5.556	0.8379
0.99	49.50	49.56	57.55	49.35	50.13	49.88	50.51	0.9829

approximation in the heavy traffic region and that A_{LT} provides a good approximation in the light traffic region.

6.1.2 Mean delays, $\bar{w}$

Table II shows the mean delays resulting from the various approximations. We see that $A_{E,1}(A^*)$ provides the best approximation to the true $\bar{w}$. The approximations $A_{H,1}$ and $A_{0,1}$ are somewhat similar, with $A_{H,1}$ being better at higher loads, while $A_{0,1}$ appears better in the midrange. The approximation $A_{E,0}$ appears quite poor for low loads but improves with increasing load. While not comparable with $A_{H,1}$, $A_{G,1}$ does provide some improvement over A_{HT}. In the light load region, A_{LT} is seen to provide a good approximation. It is interesting to note that while $A_{H,1}$, $A_{G,1}$, and A_{HT} all use the heavy traffic a_H from (60), which has the poor behavior

$$a_H \underset{\rho \to 0}{\longrightarrow} 0,$$

this only results in the anomalous behavior

$$\bar{w}_A \underset{\rho \to 0}{\longrightarrow} \infty$$

for $A_{G,1}$ and A_{HT}.

6.1.3 Dominant root, a_E

Table III shows the exponent that would be used in the (exponential) approximation for $W(x)$ corresponding to each of the approximations. Recall that for $A_{E,0}$ and $A_{E,1}$ the exponent is exact, i.e., $a = a_E$. We see that A^* is somewhat better than $A_{0,1}$ for heavier loads and both tend to 2 as $\rho \to 0$—the inverse of the mean forward recurrence time of the service time. As should be expected, the heavy traffic a, a_H, is quite poor for light traffic.

Table III—Dominant root, a, $M/D/1$

ρ	True	A^*	$A_{0,1}$	A_H
0.01	6.475	1.974	1.974	0.0198
0.1	3.615	1.775	1.770	0.1800
0.5	1.256	0.9685	0.9234	0.5000
0.9	0.2071	0.1979	0.1742	0.1800
0.99	0.0200	0.0199	0.0172	0.0198

Note that the differences between P_D and $\bar{w}$ as predicted by $A_{E,0}$ and $A_{E,1}$, particularly for light loads, indicates that the true delay distribution is essentially nonexponential. Yet, we see that assuming an exponential form can result in excellent approximations for P_D and $\bar{w}$. Moreover, the reasonable results predicted by $A_{H,0}$ and $A_{H,1}$ for light loads where a_H is far from any reasonable choice of a dominant root indicates that these techniques are somewhat robust with respect to the choice of the exponent, a.

6.2 Case: M/E₂/1

We now look at several of the resulting approximations for the waiting time distribution, $W_A(x)$, and compare these with the exact results for an $M/E_2/1$ system. Specifically, we assume $A(t) = 1 - e^{-\lambda t}$ and that $B(t)$ is the convolution of two exponentials with unit mean. Hence, we have for the respective Laplace Stieltes transforms

$$\hat{A}(s) = \frac{\lambda}{\lambda + s}$$

$$\hat{B}(s) = \frac{1}{(1 + s)^2}. \tag{61}$$

For this system, the exact delay distribution, $W(x)$ can readily be obtained via standard technique (e.g., see Ref. 10). The result is

$$W(x) = 1 - C_1 e^{-a_1 x} - C_2 e^{-a_2 x}, \tag{62}$$

where

$$a_1 = \frac{(2 - \lambda) - [\lambda(4 + \lambda)]^{1/2}}{2}$$

$$a_2 = \frac{(2 - \lambda) + [\lambda(4 + \lambda)]^{1/2}}{2}$$

$$C_1 = \frac{(1 - \rho)(1 - a_1)^2}{a_1(a_2 - a_1)}$$

$$C_2 = \frac{-(1 - \rho)(1 - a_2)^2}{a_1(a_2 - a_1)}.$$

From (61) and (62), we can readily compute all of the quantities we need. We have

$$q = \hat{B}(\lambda) = \frac{1}{(1 + \lambda)^2}; \qquad \bar{\tau} = \frac{1}{\alpha} = 2; \qquad \bar{u} = 2 - \frac{1}{\lambda},$$

and

$$\sigma_u^2 = 2 + \frac{1}{\lambda^2},$$

yielding expressions analogous to (54) to (60).

Tables IV, V, and VI show comparisons of the exact delay distribution, as well as the mean delay, with some of the approximations we developed for $\rho = \lambda\bar{\tau} = 2\lambda = 0.1, 0.5$, and 0.9. The approximations for the probability of delay, P_D, and mean delay, $\bar{w}$, show similar behavior to that for the $M/D/1$ case. In addition, we see that the approximation $A_{E,1}$ seems to have the best tail behavior (followed closely by A^{**}).

Thus, from these two examples, we might conclude that A^{**} is the best overall approximation, while $A_{E,1}$ has slightly better tail behavior.

6.3 Case: $D/D_2/1$ ($D/G/1$)

We briefly consider here the $D/G/1$ case and a simple numerical example which illustrates some interesting properties of our approximations.

For the $D/G/1$ case, we have

$$K(u) = \begin{cases} B(u+d); & u \geq -d \\ 0; & u < -d' \end{cases}$$

Table IV—$P(\text{Delay} \leq x)$ for case $M/E_2/1$

x	True	$A_{E,0}$	$A_{E,1}$	$A_{0,1}$	A^*	A^{**}
0	0.9000	0.9014	0.8760	0.9000	0.9014	0.9000
0.5	0.9220	0.9323	0.9148	0.9257	0.9268	0.9261
1.0	0.9414	0.9534	0.9414	0.9448	0.9460	0.9454
5.0	0.9962	0.9977	0.9971	0.9950	0.9950	0.9951
10.0	0.9999	0.9999	0.9999	0.9997	0.9997	0.9998
$\bar{w}$	0.1667	0.1314	0.1654	0.1680	0.1654	0.1654

Note: $\rho = 0.1$, True $W(x) = 1 - 0.1667e^{-0.7500x} + 0.06667e^{-1.200x}$.

Table V—$P(\text{Delay} \leq x)$ for case $M/E_2/1$

x	True	$A_{E,0}$	$A_{E,1}$	$A_{0,1}$	A^*	A^{**}
0	0.5000	0.5119	0.4666	0.5000	0.5119	0.5000
0.5	0.5644	0.5922	0.5547	0.5742	0.5859	0.5776
1.0	0.6272	0.6593	0.6277	0.6374	0.6488	0.6431
5.0	0.9084	0.9192	0.9117	0.8998	0.9058	0.9073
10.0	0.9848	0.9866	0.9854	0.9799	0.9818	0.9828
$\bar{w}$	1.500	1.357	1.483	1.556	1.483	1.483

Note: $\rho = 0.5$, True $W(x) = 1 - 0.5532e^{-0.3596x} + 0.05317e^{-1.390x}$.

Table VI—$P(\text{Delay} \leq x)$ for case $M/E_2/1$

x	True	$A_{E,0}$	$A_{E,1}$	$A_{0,1}$	A^*	A^{**}
0	0.1000	0.1057	0.0919	0.1000	0.1057	0.1000
0.5	0.1244	0.1354	0.1220	0.1278	0.1349	0.1296
1.0	0.1509	0.1641	0.1511	0.1548	0.1632	0.1582
5.0	0.3497	0.3617	0.3518	0.3426	0.3584	0.3557
10.0	0.5359	0.5446	0.5374	0.5198	0.5397	0.5388
$\bar{w}$	13.50	13.26	13.46	14.33	13.46	13.46

Note: $\rho = 0.9$, True $W(x) = 1 - 0.911e^{-0.06745x} + 0.01110e^{-1.483x}$.

where d is the constant interarrival time. Thus,

$$K(0) = B(d)$$

$$\hat{K}_-(a) = \int_{-\infty}^{0} e^{ay} dK(y) = 1 - \int_{0}^{\infty} e^{ay} dK(y) = 1 - \hat{K}_+(a)$$

$$\overline{u_+} = \int_{0}^{\infty} x dK(x) = \hat{K}'_+(0) ,$$

where a satisfies

$$\int_{-\infty}^{\infty} e^{ay} dK(y) = \hat{K}_-(0) + \hat{K}_0(0) = 1 .$$

This class of systems $(D/G/1)$ is particularly important in the analysis of a certain class of schedules for computer systems with real time applications. For example, see Refs. 1 and 2 where the approximations given here are applied to the study of these schedules. The service time distribution for these systems are generally discrete in nature. Here we look at a simple special case with

$$A(t) = U(t - d)$$

$$B(\tau) = p_1 U(\tau - s_1) + p_2 U(\tau - s_2) ,$$

where $U(x)$ is the unit step function, $p_1 + p_2 = 1$ and $s_1 < d < s_2$. From these, we readily obtain

$$K(0) = p_1$$

$$\hat{K}_-(a) = p_1 e^{a(s_1 - d)}$$

$$\hat{K}_+(a) = p_2 e^{a(s_2 - d)}$$

$$\overline{u_+} = \hat{K}'_+(0) = p_2(s_2 - d) ,$$

and a_E satisfies

$$\hat{K}_-(a) + \hat{K}_+(a) = \bar{e}^{ad}(p_1 e^{as_1} + p_2 e^{as_2}) = 1 .$$

Thus, (7) and (9) readily yield

$$P_D = C_0(a_E) = \frac{1 - p_1}{1 - p_1 e^{a_E(s_1 - d)}} \tag{63}$$

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} = \frac{p_2(s_2 - d)}{p_1 - p_1 e^{a_E(s_1 - d)}} . \tag{64}$$

For simplicity, we consider the special case $s_1 = 0$, $s_2 = 2d$. In this case, the exact solution satisfies the difference equation

$$P_i = p_1 P_{i+1} + p_2 P_{i-1} , \tag{65}$$

where $P_i = P$(work in the system = id at an arbitrary arrival epoch), i.e., $P_i = P$(arrival delayed id). Taking $d = 1$, the solution to (65) is readily found to be

$$P_i = P(\text{Delay} = i) = (1 - r)r^i; \qquad r = \frac{p_2}{p_1}. \tag{66}$$

Hence,

$$P_D = r = \frac{p_2}{p_1} \tag{67}$$

$$\bar{w} = \frac{r}{1 - r} = \frac{p_2}{p_1 - p_2} \tag{68}$$

and $a_E = \ln(r^{-1})$.

Now (63) and (64) yield

$$C_0(a_E) = \frac{1 - p_1}{1 - p_2} = \frac{p_2}{p_1} \tag{69}$$

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} = \frac{p_2}{p_1 - p_2}, \tag{70}$$

i.e., $A_{E,0}$ gives the *exact* value for P_D and $A_{E,1}$ gives the *exact* value $\bar{w}$. In particular, this implies that the approximation $A_{E,0}$ is *exact* at grid points—$W_{A_{E,0}}(x)$ is exact for $x = i$, an integer.

Table VII shows the resulting delay distributions for several of our approximations. We see here that A^* seems to be the best overall approximation. Note that $A_{0,1}$ (and hence A^{**}) result in somewhat poorer approximations for this case. This may indicate that A^* is potentially more robust.

As another comparison for the mean delays predicted by our approximations, we have included on the table (denoted by $\bar{w}[K - L]$) the results of using an approximation for the mean delay given by

Table VII—$P(\text{Delay} \leq i)$ for case $D/D_2/1$

i	True	$A_{E,0}$	$A_{E,1}$	A^*	$A_{0,1}$
0	0.1818	0.1818	0.0970	0.1818	0.1282
1	0.3306	0.3306	0.2612	0.3178	0.2331
2	0.4523	0.4523	0.3955	0.4312	0.3254
3	0.5519	0.5519	0.5054	0.5258	0.4066
4	0.6334	0.6334	0.5953	0.6046	0.4780
5	0.7000	0.7000	0.6689	0.6704	0.5408
6	0.7546	0.7546	0.7291	0.7252	0.5960
7	0.7992	0.7992	0.7784	0.7709	0.6446
8	0.8357	0.8357	0.8187	0.8089	0.6874
9	0.8655	0.8655	0.8516	0.8407	0.7250
10	0.8900	0.8900	0.8786	0.8672	0.7581
$\bar{w}$	4.50	4.08	4.50	4.50	6.80

Note: $\rho = 0.9$, True $P(\text{Delay} = i) = \left[1 - \left(\frac{9}{11}\right)\right]\left(\frac{9}{11}\right)^i$, $\bar{w}(K - L) = 6.43$.

Kramer and Largenbach-Belz.[11] This approximation is essentially a heuristic extension of an approximation originally obtained by Heyman[12] via diffusion techniques. The rather simple form of the approximation $\bar{w}(K-L)$ given below makes its use very appealing.

$$\bar{w}(K-L) = \frac{\bar{\tau}\rho e^{-\gamma}(C_\tau^2 + C_t^2)}{2(1-\rho)},$$

where

$$\gamma = \begin{cases} \dfrac{2(1-\rho)(1-C_t^2)^2}{3\rho(C_\tau^2 + C_t^2)} & C_t^2 \le 1 \\[2ex] \dfrac{(1-\rho)(C_t^2 - 1)}{4C_\tau^2 + C_t^2} & C_t^2 \ge 1 \end{cases}$$

and C_τ, C_t are, respectively, the coefficients of variation of the service time and the interarrival time. (Note that $\bar{w}(K-L)$ is exact for an $M/G/1$ system.)

6.4 Case: $H_2/E_3/1$

As a last numerical example, we consider an $H_2/E_3/1$ system. Specifically,

$$\hat{A}(s) = \frac{p_1\lambda_1}{\lambda_1 + s} + \frac{p_2\lambda_2}{\lambda_2 + s}$$

$$\hat{B}(s) = \frac{1}{(1+s)^3}.$$

Again, the exact waiting time is relatively easy to find via standard techniques (see Ref. 10), which yield

$$W(x) = 1 - C_1 e^{-a_1 x} - C_2 e^{-a_2 x} - \bar{C}_2 e^{-\bar{a}_2 x},$$

where the a_i are the roots of the equation

$$\hat{K}(s) = \hat{A}(-s)\hat{B}(s) = 1,$$

$\bar{a}$ denotes the complex conjugate of a, and the C_i's are the corresponding residues which are readily obtainable (see Ref. 10).

For the approximations of the preceding sections, we find that

$$\bar{u} = 3 - \frac{p_1}{\lambda_1} - \frac{p_2}{\lambda_2}$$

$$\sigma_u^2 = 3 + 2\left(\frac{p_1}{\lambda_1^2} + \frac{p_2}{\lambda_2^2}\right) - \left(\frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2}\right)^2$$

$$K(0) = p_1 q_1 + p_2 q_2$$

$$\hat{K}_-(a) = \frac{p_1 q_1 \lambda_1}{\lambda_1 + a_1} + \frac{p_2 q_2 \lambda_2}{\lambda_2 + a_1},$$

Table VIII—$P(\text{Delay} \le x)$ for case $H_2/E_3/1$

x	True	A^*	A^{**}	$A_{0,1}$	$A'_{0,1}$
0	0.7834	0.7897	0.7835	0.7835	0.7835
0.5	0.8121	0.8265	0.8225	0.8212	0.8080
1	0.8406	0.8569	0.8544	0.8523	0.8377
5	0.9731	0.9694	0.9702	0.9680	0.9740
10	0.9977	0.9955	0.9959	0.9953	0.9978
20	0.9999844	0.9999054	0.9999223	0.9998966	0.9999849
$\bar{w}$	0.5536	0.5458	0.5458	0.5662	0.5523

Note: $\rho = 0.2$, True $P(\text{Delay} \le x) = 1.0 - 0.3260e^{-0.4974x} + e^{-1.215x}[0.10944\cos(0.3129x) - 0.08673\sin(0.3129x)]$, $W_{A'_{0,1}}(x) = 1 - 0.3161e^{-0.4974x} + 0.0995e^{-1.215x}$, $\bar{w}(K-L) = 0.5476$.

Table IX—$P(\text{Delay} \le x)$ for case $H_2/E_3/1$

x	True	A^*	A^{**}	$A_{0,1}$	$A'_{0,1}$
0	0.5728	0.5867	0.5730	0.5730	0.5727
0.5	0.6145	0.6408	0.6305	0.6280	0.6104
1	0.6578	0.6877	0.6803	0.6752	0.6550
5	0.8997	0.8982	0.8996	0.8913	0.9008
10	0.9806	0.9749	0.9764	0.9723	0.9808
20	0.9993	0.9985	0.9987	0.9982	0.9993
40	0.9999990	0.99999997	0.9999998	0.9969756	0.9999990
$\bar{w}$	1.497	1.475	1.475	1.561	1.494

Note: $\rho = 0.4$, True $P(\text{Delay} \le x) = 1.0 - 0.5176e^{-0.3285x} + e^{-1.265x}[0.09038\cos(0.3793x) - 0.06644\sin(0.3793x)]$, $W_{A'_{0,1}}(x) = 1 - 0.5124e^{-0.3285x} + 0.0851e^{-1.265x}$, $\bar{w}(K-L) = 1.488$.

Table X—$P(\text{Delay} \le x)$ for case $H_2/E_3/1$

x	True	A^*	A^{**}	$A_{0,1}$	$A'_{0,1}$
0	0.1785	0.1906	0.1791	0.1791	0.1785
0.5	0.2041	0.2253	0.2148	0.2122	0.2024
1.0	0.2325	0.2586	0.2489	0.2440	0.2314
5.0	0.4609	0.4779	0.4738	0.4562	0.4616
10.0	0.6588	0.6632	0.6627	0.6398	0.6592
50.0	0.9912	0.9899	0.9904	0.9866	0.9912
100	0.9999093	0.9998741	0.9998874	0.9997827	0.9999094
$\bar{w}$	9.285	9.230	9.230	9.967	9.928

Note: $\rho = 0.8$, True $P(\text{Delay} \le x) = 1.0 - 0.8517e^{-0.0915x} + e^{-1.321x}[0.03026\cos(0.4548x) - 0.02125\sin(0.4548x)]$, $W_{A'_{0,1}}(x)$;q $1 - 0.8517^{-0.0915x} + 0.0302e^{-1.321x}$, $\bar{w}(K-L) = 9.266$.

where

$$q_1 = \hat{B}(\lambda_i) = \frac{1}{(\lambda_i + 1)^3}.$$

Tables VIII, IX, and X compare our various approximations for this case with $p_1 = 0.25$, $\lambda_2 = 2\lambda_1$ and $\rho = 0.2$, 0.4, and 0.8.† (The approximation denoted by $A'_{0,1}$ will be discussed shortly.) Here we see again that A^* and A^{**} yield quite good approximations over a wide range.

Again, for comparison, we have included the approximation $\bar{w}(K -$

† The roots needed for the exact solution (and the approximations) were obtained by using a program developed by A. E. Eckberg.

L). Unlike the $D/D_2/1$ case, we see that $\bar{w}(K - L)$ provides a good approximation to the mean delay for these cases.

VII. EXTENSIONS

As we have seen, even with the assumption of only a single exponential for the form of $W(x)$, we are led to a wide variety of approximations by using the Lindley integral equation (1) to determine the two unknown coefficients. We consider here some extensions for the case where we wish to choose a more complicated form for $W(x)$.

7.1 Method of moments for hyperexponentials

In many cases, $W(x)$ admits the expansion

$$W(x) = 1 - \sum_{i=1}^{\infty} c_i e^{-a_i x}, \tag{71}$$

where the coefficients C_i, a_i are not necessarily real. In such cases, an approximation of the form

$$W_A(x) = 1 - \sum_{i=1}^{m} C_i e^{-a_i x} \tag{72}$$

seems most appropriate. In general, of course, we do not know the form of $W(x)$. However, if the equation

$$\hat{K}(-s) = \hat{A}(s)\hat{B}(-s) = 1 \tag{73}$$

can be solved more than one nonzero root, then it seems reasonable to attempt to incorporate additional roots in an approximation of the form (72). One can discretize (1) to obtain a set of (implicit) equations for the needed C_i; however, we show here that the method of moments introduced in Section III can readily be extended to obtain a set of linear equations for these quantities. For this purpose, it is convenient to consider the equation for $W^c(x) = 1 - W(x)$ corresponding to (1), i.e.,

$$L[W^c](x) = W^c(x) - 1 + K(x) - \int_{-\infty}^{x} W^c(x-y)dK(y) = 0. \tag{74}$$

We denote the Lth moment of (74) by μ_L:

$$\mu_L = \int_0^{\infty} x^L dL[W^c](x). \tag{75}$$

Using (74) in (75) we obtain

$$\mu_0 = \int_0^{\infty} dW^c(x) + \int_0^{\infty} dK(x) + \int_{-\infty}^{0} W^c(-y)dK(y)$$

$$\mu_L = \int_0^\infty x^L dW^c(x) + \int_0^\infty x^L dK(x)$$

$$+ L \int_0^\infty x^{L-1} \int_{-\infty}^x W^c(x-y)dK(y)dx, \qquad L > 0. \tag{76}$$

With the notation

$$\overline{w_L^c} = \int_0^\infty x^L dW^c(x), \qquad \overline{u_L} = \int_0^\infty x^L dK(x)$$

(76) can be written

$$\mu_0 = \overline{w_0^c} + \overline{u_0} + \int_{-\infty}^0 W^c(-y)dK(y)$$

$$\mu_L = \overline{w_L^c} + \overline{u_L}$$

$$+ L \int_0^\infty x^{L-1} \int_{-\infty}^x W^c(x-y)dK(y)dx, \qquad L > 0. \tag{77}$$

[Note $\bar{u}_0 = 1 - K(0)$ and $\bar{u}_1 = \bar{u}_+$ of the preceding sections.]

Now using the form

$$W_A^c(x) = \sum_i C_i \bar{e}^{a_i x} \tag{78}$$

$\mu_0 = 0$ leads to

$$\sum_i C_i(K_i - 1) = -\overline{u_0}, \tag{79}$$

where

$$K_i = \int_{-\infty}^0 e^{a_i y} dK(y).$$

After some algebra, $\mu_L = 0$ for $L > 0$ yields

$$\sum_i C_i \left(\frac{L!K_i}{a_i^L} + \frac{L\bar{u}_{L-1}}{a_i} + \sum_{k=1}^{L-1} \frac{L(L-1)\cdots(L-k)}{a_i^{k+1}} \bar{u}_{L-1-k} - \frac{L!}{a_i^L} \right)$$

$$= -\bar{u}_L. \tag{80}$$

The empty sum is taken to be zero.

Thus, given a set of roots a_i, $i = 1, \cdots, m$ of (73) eqs. (79) and (80) allow us to readily compute the desired coefficient C_i. (Note that this method is identical to the common techniques of "method of moments" frequently used for other classes of integral equations [e.g., see Ref. 13).] We refer to the resulting approximation as

$$A_{0,1,\cdots,m-1}^E.$$

Table XI—Dominant residue for case $H_2/E_1/1$

ρ	True	$C_{E,0}$	$C_{E,1}$	$C_{0,1}$	$C_1(A'_{0,1})$
0.2	0.3260	0.2103	0.2714	0.2164	0.3161
0.4	0.5176	0.4133	0.4841	0.4270	0.5124
0.8	0.8517	0.8094	0.8444	0.8209	0.8517

7.2 Numerical example ($H_2/E_3/1$)

Note that if we apply the results of (71) to the $H_2/E_3/1$ example of Section VI and include all three roots, a_1, a_2, a_3, then the resulting delay distribution will be exact [that is, if all roots are given, (80) is equivalent to the standard methods of determining the appropriate residues]. To see how we can use more structure to improve our approximations, we will use the true dominant root a_1 as one of our roots, but take $a'_2 = \mathrm{Re}(a_2)$ as another root, the resulting approximation is shown on Tables VIII, IX, and X, where it is denoted by $A'_{0,1}$. We see that the inclusion of this additional term results in an approximation that more closely captures the structure of the true delay distribution. This is perhaps best illustrated by Table XI. Here we have given the values for the dominant residue, i.e., the coefficient of the exponential with the dominant exponent, for the various approximations. We see that of the three single exponential approximations $A_{E,1}$ has a C value nearest the true value. This shows why $A_{E,1}$ tends to have better tail behavior than the other single exponential approximation. However, we see that $A'_{0,1}$ resulted in an excellent approximation to the true dominant residue, even though we did not use another exact root. Hence, the excellent agreement on Tables VIII to X for the extreme tails of the distribution. This behavior can be very important in studying computer systems with dedicated real-time applications where often criteria are specified in the 10^{-5} probability range, i.e., the probability of delay greater than T shall be less than 10^{-5}.

VIII. CONCLUSIONS

The basic idea we have exploited is to choose a functional form for an approximation $W_A(x)$ to a true delay distribution, say, $W(x)$, and use the well-known Lindley integral equation to find the undetermined coefficients. For the case where $W_A(x)$ is exponential—$W_A(x) = 1 - Ce^{-ax}$—we have used this technique to develop several approximations, some of which make use of the explicit structure of the relevant service and interarrival time distributions, while others require only moment information.

Although not always the best choice, A^* seems to provide the robustness that one would require of a good approximation. The resulting approximation for the mean delay is excellent and the resulting probability of delay quite good, although A^{**} provides a better value for the latter. For predicting tails of distributions, we see that $A_{E,1}$ is the best of the simple exponential approximations. We have also seen that increasing the complexity of the forms of the waiting time distribution assumed in the approximation, e.g., using more than one exponential, can result in extremely accurate predictions of the tails of the delay distribution.

IX. ACKNOWLEDGMENTS

I am deeply indebted to A. E. Eckberg and D. L. Jagerman for many helpful discussions and suggestions. I would also like to thank D. F. DeMaio for her helpful comments on this paper. I am also very appreciative of the many useful observations and suggestions of the reviewers.

APPENDIX

Summary of Notations and Formulas

$\tilde{t}$ = interarrival time

$A(t)$ = interarrival time distribution

L.S.T. = Laplace-Stieltes Transform

$\hat{A}(s)$ = L.S.T. of $A(t)$

E = Expected value

$\bar{t} = E(\tilde{t}) = 1/\lambda$

$\tilde{\tau}$ = service time

$B(\tau)$ = service time distribution

$\hat{B}(s)$ = L.S.T. of $B(\tau)$

$\bar{\tau} = E(\tilde{\tau}) = 1/\alpha$

$\rho = \lambda/\alpha$

$\tilde{u} = \tilde{\tau} - \tilde{t}$

$K(u)$ = distribution function of $\tilde{u}$

$\hat{K}(s)$ = L.S.T. of $K(u) = \hat{A}(-s)\hat{B}(s)$

$\bar{u} = E(\tilde{u}) = 1/\alpha - 1/\lambda$

σ_u^2 = variance of u

$$\hat{K}_-(s) = \int_{-\infty}^{0} e^{su} dK(u)$$

$$\hat{K}_+(s) = \int_{0}^{\infty} e^{su} dK(u)$$

$$\hat{K}(s) = \hat{K}_-(-s) + \hat{K}_+(-s)$$

$$a_E = \text{positive real root of } \int_{-\infty}^{\infty} e^{au} dK(u) = 1$$

a_i = ordered roots of characteristic equation,

$\hat{K}(-a) = A(s)B(-s) = 1(a_1 = a_E)$

$$\bar{u}_+ = \int_0^\infty x dK(x) = \hat{K}'_+(0)$$

$K(0) = \hat{K}_-(0)$

$a_H = -2\bar{u}/\sigma_u^2$

$p = \hat{A}(\alpha)$

$q = \hat{B}(\lambda)$

$W(x)$ = waiting time distribution

P_D = true probability delay greater than zero

$\bar{u}$ = true mean delay

$W^C(x) = 1 - W(x)$ = complimentary waiting time distribution

$L[W^C](x)$ = Lindley integral equation (complimentary)

$$L[W^C](x) = W^C(x) - 1 + K(x) - \int_{-\infty}^{x} W^C(x-y)dK(y)$$

$$\mu_L = \int_0^\infty x^L dL[W^C](x)$$

$K_i = \hat{K}_-(a_i)$

Approximate waiting time distribution (single exponential):

$$W_A(x) = 1 - Ce^{-ax}$$

$$C = \text{approximate probability of delay}$$

$$\bar{w}_A = C/a = \text{approximate mean delay}.$$

Approximate $A_{E,0}$:

$$a = a_E$$

$$C = C_0(a_E) = \frac{1 - K(0)}{1 - \hat{K}_-(a_E)} = \frac{1 - \hat{K}_-(0)}{1 - \hat{K}_-(a_E)}.$$

Approximation $A_{E,1}$:

$$a = a_E$$

$$\bar{w}_{E,1} = \frac{C_1(a_E)}{a_E} = \frac{\bar{u}_+}{K(0) - \hat{K}_-(a_E)} = \frac{\hat{K}'_+(0)}{\hat{K}_-(0) - \hat{K}_-(a_E)}.$$

Approximation $A_{0,1}$:

$a = a_{0,1}$ solution of

$$a = \frac{[(K(0) - \hat{K}_-(a)][1 - K(0)]}{\bar{u}_+[1 - \hat{K}_-(a)]} = \frac{[\hat{K}_-(0) - \hat{K}_-(a)][1 - \hat{K}_-(0)]}{\hat{K}'_+(0)[1 - \hat{K}_-(a)]}$$

$$C = C_{0,1} = \frac{1 - K(0)}{1 - \hat{K}-(a_{0,1})} = \frac{1 - \hat{K}_-(0)}{1 - \hat{K}_-(a_{0,1})}.$$

Approximation A^*:

$$C = C^* = C_0(a_E) \text{ (from } A_{E,0})$$

$$\bar{w}^* = \frac{C^*}{a^*} = \frac{C_1(a_E)}{a_E} \text{ (from } A_{E,1})$$

$$\text{determines } a^* = \frac{C_0(a_E)}{C_1(a_E)} a_E .$$

Approximation A^{**}:

$$C = C^{**} = C_{0,1} \text{ (from } A_{0,1})$$

$$\bar{w}^{**} = \frac{C^{**}}{a^{**}} = \frac{C_1(a_E)}{a_E} \text{ (from } A_{E,1})$$

$$\text{determines } a^{**} = \frac{C_{0,1}}{C_1(a_E)} a_E .$$

Approximation $A_{H,0}$:

$$a = a_H = \frac{-2\bar{u}}{\sigma_u^2}$$

$$C = C_{H,0} = C_0(a_H) = \frac{1 - K(0)}{1 - \hat{K}_-(a_H)} = \frac{1 - \hat{K}_-(0)}{1 - \hat{K}_-(a_H)}.$$

Approximation $A_{H,1}$:

$$a = a_H$$

$$\bar{w}_{H,1} = \frac{C_{H,1}}{a_H} = \frac{C_1(a_H)}{a_H} = \frac{\bar{u}_+}{K(0) - \hat{K}_-(a_H)} = \frac{K'_+(0)}{\hat{K}_-(0) - \hat{K}_-(a_H)}.$$

Approximation A_{HT}:

$$a = a_H$$

$$\bar{w}_{HT} = \frac{C_{HT}}{a_H} = \frac{1}{a_H}.$$

Approximation $A_{G,0}$:

$$a = a_H$$

$$C_{G,0} = \frac{1 - \mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)}{1 + \mathrm{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)} .$$

Approximation $A_{G,1}$:

$$a = a_H$$

$$\bar{w}_{G,1} = \frac{C_{G,1}}{a_H} = \frac{\frac{\sigma_u}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{u}}{\sigma_u}\right)^2} + \frac{\bar{u}}{2}\left[1 - \text{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)\right]}{\text{Erf}\left(\frac{-\bar{u}}{\sqrt{2}\sigma_u}\right)}.$$

Approximation A_{LT}:

$$C_{LT} = 1 - K(0) = 1 - \hat{K}_-(0)$$

$$\bar{w}_{LT} = \frac{C_{LT}}{a_{LT}} = \frac{\bar{u}_+}{K(0)} = \frac{\hat{K}'_+(0)}{\hat{K}_-(0)}.$$

Approximate waiting time distribution (multiple exponentials):

$$W_A(x) = 1 - \sum_{i=1}^{m} C_i e^{-a'_i x}.$$

Approximation $A^E_{0,1,\cdots,m}$:

$$a_i \text{ roots of } \hat{K}(-a) = 1$$

$$a'_i = a_i, \qquad i = 1, \cdots, m$$

C_i determined from

$$\sum_{i=1}^{m} C_i \left(\frac{L!K_i}{a_i^L} + \frac{L\bar{u}_{L-1}}{a_i} + \sum_{k=1}^{L-1} \frac{L(L-1)\cdots(L-k)}{a_i^{k+1}} \bar{u}_{L-1-k} - \frac{L!}{a_i^L}\right) = -\bar{u}_L,$$

where the empty sum is taken to be zero.

Approximation $A'_{0,1}$:

$$a_i \text{ roots of } \hat{K}(-s) = \hat{A}(s)\hat{B}(-s) = 1$$

$$a'_1 = a_1 = a_E$$

$$a'_2 = \text{Re}(a_2)$$

C_i determined from

$$(1 - K_1)C_1 + (1 - K_2)C_2 = \overline{u_0} = 1 - K(0)$$

$$\frac{[K(0) - K_1]}{a'_1} C_1 + \frac{[K(0) - K_2]}{a'_2} C_2 = \bar{u}_1 = \bar{u}_+.$$

REFERENCES

1. A. A. Fredericks, "Analysis of a Class of Schedule with Real Time Applications," *Performance of Computer Systems*, M. Arrato, A. Butrimenko, and E. Gelinbe, eds., Amsterdam, New York, and Oxford: North-Holland Publishing Company, pp. 201–16.
2. A. A. Fredericks, "Analysis and Design of Processor Schedules for Real Time

Applications," *Proceedings of the Conference on Applied Probability and Computer Science, The Interface,* R. Disney and T. J. Ött, eds., Boston: Birkhauser, March, 1982.
3. D. V. Lindley, "The Theory of Queues with a Single Server," Proc. Cambridge Philosophical Soc., *48* (1952), pp. 277–89.
4. J. F. C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues," *Congestion Theory*, W. L. Smith, and W. E. Wilkinson, eds., Chapel Hill: Univ. N. Carolina Press, 1964, pp. 137–69.
5. W. Feller, *An Introduction to the Theory of Probability and its Applications*, New York: John Wiley, 1966, 1971.
6. J. F. C. Kingman, "Inequalities in the Theory of Queues," J. Royal Statist. Soc. *B32* (1970), pp 102–10.
7. H. J. Rossberg, and C. Siegel, "The Importance of Kingman's Integral-Inequalities . . . ," Math. Operationsforsh. Statist. *5* (1974), pp. 687–99.
8. S. M. Ross, "Bonds on the Delay Distribution in $GI/G/1$ Queues," J. Appl. Prod. *11* (1974), pp. 417–21.
9. R. Bergmann, and D. Stoyan, "Exponential Bounds for the Waiting-Time Distribution Function in $GI/G/1$," J. Appl. Prob., *13* (1976), pp. 411–7.
10. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, New York: John Wiley, 1974.
11. W. Kramer and M. Lagenbach-Belz, "Approximate Formulae for the Delay in the Queueing System $GI/G/1$," Congressbook, 8th Int. Teletraffic Conf., Melbourne, 1976, pp. 235.1–235.8.
12. D. P. Heyman, "A Diffusion Approximation for the $GI/G/1$ Queues in Heavy Traffic," B.S.T.J., *54* (1975), pp. 1637–46.
13. A. Yu. Luchka, *The Method of Averaging Functional Corrections*, *Theory and Applications*, New York: Academic Press, 1965.