the exact analysis of G/G/m has been given by Kiefer and Wolfowitz [KIEF 55] involving the (usually impossible) task of solving an integral equation (which reduces to Lindley's Integral Equation for G/G/1). More recently, de Smit [DESM 73] has extended the theory due to Pollaczek [POLL 61] for G/G/m and has elaborated upon the G/M/m and G/H_R/m queues.

This completes our very rapid summary of the elements of queueing theory. We will need much of this material in the following chapters. It should be clear that a number of important behavioral properties for these queueing systems remain as yet unsolved. Nevertheless we are faced in the real world with applying the tools from queueing theory to solve immediate problems. The balance of this textbook discusses such problems and methods for applying the theory developed. Consequently, we begin with a rather advanced chapter in queueing systems where the goal is *not* to extend the rigorous theory as summarized here but rather to find *effective approximation methods* that permit one to use the theory in a true engineering sense.

REFERENCES

BROC 48 Brockmeyer, E., H. L. Halstrøm, and A. Jensen, "The Life and Works of A. K. Erlang," *Transactions of the Danish Academy of Technology and Science*, **2,** (1948).

DESM 73 de Smit, J. H. A., "Some General Results for Many Server Queues," pp. 153–169 and, "On the Many Server Queue with Exponential Service Times," *Advances in Applied Probability*, **5,** No. 1 (April 1973).

KIEF 55 Kiefer, J., and J. Wolfowitz, "On the Theory of Queues with Many Servers," *Transactions of the American Mathematics Society*, **78,** 1–18 (1955).

KLEI 75 Kleinrock, L., *Queueing Systems, Volume I: Theory*, Wiley-Interscience (New York), 1975.

POLL 61 Pollaczek, F., *Théorie Analytique de Problèms Stochastiques Relatifs à un Groupe de Lignes Telèphonique avec Dispositif d'attente*, Gauthiers-Villars (Paris), 1961.

TELE 70 *Telephone Traffic Theory Tables and Charts, Part 1*, Siemens Altiengesellschaft, Telephone and Switching Division (Munich), 1970.

2

Bounds, Inequalities, and Approximations

An exciting "new" branch of queueing theory is emerging that deals with methods for finding approximate or bounding behaviour for queues.* It is not hard to convince oneself that queueing theory is rather difficult and that exact results are hard to obtain; in fact, *many* of the interesting queueing phenomena have not as yet yielded to exact analysis (and perhaps never will!). Moreover, in those simpler systems where exact results can be obtained, their form is sometimes so complex as to render them ineffectual for practical applications.

If one examines why we study queueing theory in the first place, one readily admits that it is to answer questions regarding real queues in the real world. The mathematical structures we have created in attempting to describe these real situations are merely idealized fictions, and one must not become enamoured with them for their own sake if one is really interested in practical answers. We must face the fact that authentic queueing problems seldom satisfy the assumptions made throughout most of the literature available on queueing theory: stationarity is rare, independence occurs only occasionally, and ergodicity is not only unlikely but is also impossible to establish with measurements over a finite time! Therefore if our mathematical models are so crude, we should be willing to accept much less than an exact solution to the systems of equations they give rise to; rather, we should be happy to accept approximate solutions to these "approximate" mathematical models and hope that such solutions provide information about the behavior of real-world queues. Even more important is the search for "robust" qualitative behavior of queues which provides "rules of thumb" for estimating the

*Perhaps the first approximations used in queueing theory date back to Erlang himself through his introduction of the method of stages (see [KLEI 75], Section 4.2); he tried to approximate the underlying distributions of a queueing system with tractable analytic functions. The reader is referred to [SYSK 62] and to the elucidation of Erlang's work and era in [BROC 48] for some of the historical flavor.

behavior of complex systems. An excellent example of successful robust models is given in Chapter 4, where we apply Markovian queueing networks to multiaccess computer system modelling; Buzen [BUZE 74] discusses the way in which system structure can be used to generate simple robust models. A second example is the use of diffusion approximations in a variety of applications (see Sections 2.9, 4.13, and 4.14). A third example is described in Chapter 5 in which a robust model is developed for computer network delay. All three of these examples demonstrate the success in the use of simple (often Markovian) models to predict behavior of rather complicated real-world systems. Issues such as these are addressed in this chapter and we emphasize that this approach to the study of queues is relatively new and potentially highly rewarding.

The chapter is organized as follows. We begin by establishing a robust approximation for the distribution of queueing time in the G/G/1 heavy-traffic case ($\rho \to 1$ from below). This approximation lurks just beneath the surface of many of the results we have already seen. A tight upper bound on the average wait W is then established from first principles (good for $0 \leq \rho < 1$); lower bounds for W are more difficult to come by and certain available results are presented. (It is sad but true that even W cannot be expressed exactly in terms of the simple system parameters for the G/G/1 queue!) We also give a bound on the tail of the waiting time distribution in Section 2.4. Most of the results in the first four sections were inspired by the work of Kingman [KING 61, 62a, 62b, 64, 70] and pursued by Marshall [MARS 68c], Brummelle [BRUM 71, 73], and others [SUZU 70, MARC 74]. A simple discrete approximation for the G/G/1 system is then presented in Section 2.5 using techniques from elementary queueing theory. Next we make a few remarks concerning bounds on W for G/G/m. At this point in the chapter we abandon our former approach of attempting to find approximate solutions to the given system equations and take the point of view that we will approximate the stochastic processes themselves (that is, the arrival and departure processes). We begin with a "first-order" approximation whereby stochastic processes are replaced simply by their average values (perhaps time-dependent) and this leads us to the *fluid approximation* for queues. Next, we study a "second-order" approximation in which a stochastic process is represented both by its mean and its variance, and this gives us the *diffusion approximation* for queues. This diffusion approximation refines the fluid approximation by describing the time-dependent processes with means given by the fluid approximation but with a normal (Gaussian) distribution describing the fluctuations about that (possibly time-varying) mean. In the case of stable queues ($\rho < 1$) the limit of this diffusion approximation as $t \to \infty$ is in fact the heavy-traffic approximation of Section 2.1! These are related because the diffusion approximation assumes that the queue never empties, and this is just the kind of approximation made in the heavy-traffic case. Following this, a careful discussion of the diffusion approximation for the M/G/1 queue is given. These methods are then applied to give approximate solutions to the "rush-hour" behavior so common in practical life. The diffusion approximation methods have been studied by various workers including Newell [NEWE 65, 68, 71], Gaver [GAVE 68], Iglehart and Whitt [IGLE 69], McNeil [McNE 73] and Kobayashi [KOBA 74a, 74b].

2.1. THE HEAVY-TRAFFIC APPROXIMATION

In this section we study the behavior of the system G/G/1 in the *heavy-traffic case* [KING 62a]. This is the case where $\rho \cong 1$ (but remains strictly less than one, preserving stability). We establish the central result for heavy-traffic theory, which states that the waiting time distribution is, as an approximation, exponentially distributed with the mean wait given by $(\sigma_a^2 + \sigma_b^2)/2(1-\rho)\bar{t}$. This is a remarkable result and it pervades most of our approximation methods. (It is valid when the denominator is small compared to the square root of the numerator.)

Our point of departure in establishing the central result is Eq. (1.128) repeated below:

$$A^*(-s)B^*(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)} \tag{2.1}$$

We will examine this expression in the case $\rho \cong 1$. Let us begin by considering the Taylor series expansion for $B^*(s)$ and $A^*(-s)$ as follows:

$$B^*(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} B^{*(k)}(0) \tag{2.2}$$

However, from Eq. (1.17) we know that $B^{*(k)}(0) = (-1)^k \overline{x^k}$; using this and considering $B^*(s)$ near the origin ($s \to 0$), we have†

$$B^*(s) = 1 - \bar{x}s + \frac{\overline{x^2}s^2}{2!} + o(s^2) \tag{2.3}$$

Similarly, we have

$$A^*(-s) = 1 + \bar{t}s + \frac{\overline{t^2}s^2}{2!} + o(s^2) \tag{2.4}$$

† As usual, the notation $o(x)$ denotes any function which goes to zero faster than x, that is, $\lim_{x \to 0} [o(x)/x] = 0$.

Now, since we are considering the heavy-traffic case, we recognize that our interest must lie in the distribution of *large* waiting times. Recall that the waiting time distribution, $W(y)$, has a Laplace transform $\Phi_+(s)$, whereas the density, $w(y)$, has a transform $W^*(s)=s\Phi_+(s)$. It can be seen that the behavior of $W(y)$ for large values of y is governed by that pole (singularity) of $\Phi_+(s)$ which has the smallest Re (s) in absolute value; this follows since the decay rate of each exponential term in $w(y)$ or $W(y)$ is inversely related to the (negative) real part of the pole associated with that term. The expression in Eq. (2.1) has a zero at $s=0$ and in fact, has an additional zero near $s=0$ for the heavy-traffic case; as we shall see, this additional zero forms the pole of $\Phi_+(s)$ that governs the behavior of large waiting times (this is merely the final-value theorem for Laplace transforms).† Let us find this nearby zero (which has a small but negative real part). Using the expansions in Eqs. (2.3) and (2.4) we have

$$\begin{aligned} A^*(-s)B^*(s)-1 &= \left(1-\bar{x}s+\frac{\overline{x^2}s^2}{2}\right)\left(1+\bar{t}s+\frac{\overline{t^2}s^2}{2}\right)-1+o(s^2) \\ &= 1+s(\bar{t}-\bar{x})+s^2\left(\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}\right)-1+o(s^2) \\ &= s\left[\bar{t}-\bar{x}+s\left(\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}\right)\right]+o(s^2) \qquad (2.5) \end{aligned}$$

From this last we clearly see the root at $s=0$. Solving for the second root in the vicinity of $s=0$ we first note that

$$\frac{\overline{x^2}}{2}+\frac{\overline{t^2}}{2}-\bar{x}\bar{t}=\frac{\sigma_b^2+\sigma_a^2}{2}+\frac{(\bar{x}-\bar{t})^2}{2} \qquad (2.6)$$

Since $\rho\cong 1$, we choose to assume at this point that the last term on the right-hand side of Eq. (2.6) (the squared difference of the first moments) is negligible compared to the first term in that equation (the sum of the variances). Using Eq. (2.6) under this approximation and dropping $o(s^2)$ (since we are examining the vicinity of the origin), we may then solve Eq. (2.5) for our second root (which we denote by s_0) as

$$\bar{t}-\bar{x}+s_0\frac{\sigma_b^2+\sigma_a^2}{2}\cong 0$$

which yields

$$s_0\cong -\frac{2\bar{t}(1-\rho)}{\sigma_a^2+\sigma_b^2} \qquad (2.7)$$

† One can already see the exponential approximation emerging from this single critical pole.

Clearly $s_0<0$. Note from this and Eq. (2.5) that s_0 is (approximately) the inverse of the mean residual life (see footnote on p. 16) of the random variable $\tilde{u}=\tilde{x}-\tilde{t}$. Thus, as an approximation near the origin, we have

$$A^*(-s)B^*(s)-1\cong s(s-s_0)\frac{(\sigma_a^2+\sigma_b^2)}{2}$$

Returning to our direct argument now, when s is near the origin we may then use the expression in Eq. (2.1) and arrive at the approximation

$$\Psi_+(s)\cong s(s-s_0)C \qquad (2.8)$$

where $C=\Psi_-(0)[\sigma_a^2+\sigma_b^2]/2$. In order to proceed to our solution for $\Phi_+(s)$, we see from Eq. (1.129) that we must evaluate the constant K; this we do by using Eq. (1.130) as follows:

$$K=\lim_{s\to 0}(s-s_0)C=-s_0C$$

which then gives from Eq. (1.129)

$$\Phi_+(s)\cong\frac{-s_0}{s(s-s_0)}$$

(The unknown constant C cancels!) Making a partial fraction expansion we have

$$\Phi_+(s)\cong\frac{1}{s}-\frac{1}{s-s_0}$$

Finally, using the expression for s_0, this inverts to give

$$W(y)\cong 1-\exp\left(-\frac{2\bar{t}(1-\rho)}{\sigma_a^2+\sigma_b^2}y\right) \qquad \blacksquare \quad (2.9)$$

This last gives us an approximation for the distribution of waiting time in the vicinity of large waiting times for $\rho\cong 1$. The factor s_0 is given specifically through Eq. (2.7). We note that the average wait W is given by $(-1/s_0)$ and so

$$W\cong\frac{(\sigma_a^2+\sigma_b^2)}{2(1-\rho)\bar{t}} \qquad \blacksquare \quad (2.10)$$

Equations (2.9) and (2.10) form the *central results* for heavy-traffic theory as applied to G/G/1. These results are extremely robust and give the general behavior of queues with long waiting times. From Eq. (2.10) we see that the numerator contribution to the average waiting time is due to fluctuations in the arrival and service processes, whereas the denominator (which dominates in the heavy-traffic case) depends only on first moments (in particular, on ρ). The exponential character of these large waiting times is in some sense a central limit theorem for queueing theory and we shall see it appear again in our diffusion approximations below *

2.2. AN UPPER BOUND FOR THE AVERAGE WAIT

The heavy-traffic approximation studied in the previous section leads to an exponential distribution of large waiting times whose mean is given by Eq. (2.10). In this section we are interested not in an approximation, but in a firm upper bound on the average wait W in the system G/G/1.

The following development is simple and is due again to Kingman [KING 62b]. We recall from Section 1.10 that the limiting random variable $\tilde{w}$ must have the same distribution as the random variable $(\tilde{w}+\tilde{u})^+$. Therefore, assuming the following moments exist, we must have

$$E[(\tilde{w})^k] = E\{[(\tilde{w}+\tilde{u})^+]^k\} \tag{2.11}$$

Let us introduce the definition

$$(X)^- \triangleq -\min[0, X] \tag{2.12}$$

Then, recalling that $(X)^+ \triangleq \max[0, X]$ we have the simple relationship

$$X = (X)^+ - (X)^- \tag{2.13}$$

and it must also be true from their definitions that

$$(X)^+(X)^- = 0 \tag{2.14}$$

Squaring Eq. (2.13) and using Eq. (2.14) we then see that

$$X^2 = [(X)^+]^2 + [(X)^-]^2 \tag{2.15}$$

Taking X to be a random variable, we may form expectations in Eq. (2.13) to yield

$$\bar{X} = \overline{(X)^+} - \overline{(X)^-} \tag{2.16}$$

And likewise, from Eq. (2.15) we have

$$\overline{X^2} = \overline{[(X)^+]^2} + \overline{[(X)^-]^2}$$

* Queues in series have also been studied by means of the heavy-traffic approximation [HARR 73]. Again it is shown that the total waiting time is asymptotically distributed in a way depending only on the mean and variance of the interarrival and service time distributions. When all variances are identical, then it is shown that the waiting time distribution is an exponential function of these moments.

Since $\sigma_X{}^2 = \overline{X^2} - (\bar{X})^2$, we may use the above relationships to yield

$$\sigma_X{}^2 = \sigma^2_{(X)^+} + \sigma^2_{(X)^-} + 2\overline{(X)^+}\;\overline{(X)^-} \tag{2.17}$$

This last result is true for any random variable X.

Now taking $X = \tilde{w} + \tilde{u}$, we see from Eq. (2.16) that $\bar{X} = \bar{w} + \bar{u}$ is given by

$$\bar{w} + \bar{u} = \overline{(\tilde{w}+\tilde{u})^+} - \overline{(\tilde{w}+\tilde{u})^-} \tag{2.18}$$

However, from Eq. (2.11) (with $k = 1$) we have $\bar{w} = \overline{(\tilde{w}+\tilde{u})^+}$, and so Eq. (2.18) may be rewritten as*

$$\bar{u} = -\overline{(\tilde{w}+\tilde{u})^-}$$

Furthermore, from Eq. (2.11) we have that

$$\sigma_{\tilde{w}}{}^2 = \sigma^2_{(\tilde{w}+\tilde{u})^+} \tag{2.19}$$

Once again, taking $X = \tilde{w} + \tilde{u}$ we see that the term $\sigma^2_{(X)^+}$ from Eq. (2.17) may be set equal to $\sigma_{\tilde{w}}{}^2$ due to the relationship in Eq. (2.19). Furthermore, since w_n and u_n are independent, it must be that $\sigma^2_{(\tilde{w}+\tilde{u})} = \sigma_{\tilde{w}}{}^2 + \sigma_{\tilde{u}}{}^2$, and so Eq. (2.17) finally takes the form

$$\sigma_{\tilde{w}}{}^2 + \sigma_{\tilde{u}}{}^2 = \sigma_{\tilde{w}}{}^2 + \sigma^2_{(X)^-} + 2\overline{(\tilde{w}+\tilde{u})^+}\;\overline{(\tilde{w}+\tilde{u})^-} \tag{2.20}$$

Regarding the last term in this equation, we have already established that $\overline{(\tilde{w}+\tilde{u})^+} = \bar{w}$ and $\overline{(\tilde{w}+\tilde{u})^-} = -\bar{u}$; using these and canceling the variance of $\tilde{w}$ from both sides of our last equation, we have

$$\sigma_{\tilde{u}}{}^2 = \sigma^2_{(X)^-} - 2\bar{w}\bar{u} \tag{2.21}$$

By definition $\tilde{u} = \tilde{x} - \tilde{t}$ and so, as we have seen many times before, $\bar{u} = \bar{t}(\rho - 1)$; similarly, since $\tilde{x}$ and $\tilde{t}$ are independent, it must be that $\sigma_{\tilde{u}}{}^2 = \sigma_{\tilde{t}}{}^2 + \sigma_{\tilde{x}}{}^2$. However, we already have notation for the variance of interarrival time and variance of service time, namely, $\sigma_a{}^2$ and $\sigma_b{}^2$, respectively. With these observations, and solving for $\bar{w}$ (which, in the past, we have written simply as W) we may rewrite Eq. (2.21) as follows:

$$W = \frac{\sigma_a{}^2 + \sigma_b{}^2}{2\bar{t}(1-\rho)} - \frac{\sigma^2_{(X)^-}}{2\bar{t}(1-\rho)}$$

Since variances are always non-negative, we may drop the last term in this equation and thereby create our final upper bound on the average

* We point out that the limiting random variable $\tilde{y} = \lim_{n\to\infty} y_n$ must have the same distribution as $(\tilde{w}+\tilde{u})^-$, as may be seen from Eq. (1.124).

waiting time:

$$W \leq \frac{\sigma_a^2 + \sigma_b^2}{2\bar{t}(1-\rho)} \quad ■ \quad (2.22)$$

This result is correct for $0 \leq \rho < 1$ and improves (is asymptotically sharp) as $\rho \rightarrow 1$.* This result is familiar! It is, in fact, the mean waiting time that we obtained in the previous section for the heavy-traffic approximation. What we now see is that the heavy-traffic approximation to the mean wait forms a strict upper bound for the mean wait in any G/G/1 queue. In Section 1.4 we boldly stated that the behavior of the mean waiting time for the queue M/M/1 was typical of most queueing systems in that the dominant behavior is due to a simple pole at $\rho = 1$; we have now confirmed that statement by our basic results in this and the previous section.

Our upper bound is essentially distribution-free in that it depends only on the first two moments of the service and interarrival time; this simplicity is a key virtue since often we are willing to specify only some gross properties of the input (e.g., mean and variance). Unfortunately, this simplicity does not extend to the lower bound, which we discuss next.

2.3. LOWER BOUNDS FOR THE AVERAGE WAIT

The simple upper bound obtained in the last section may also be derived easily (see Exercise 2.6) from Eq. (1.132), which we may express as follows:

$$W = W_U + \frac{1}{2}\bar{t}(1-\rho) - \frac{\overline{I^2}}{2\bar{I}} \qquad (2.23)$$

* We note that the upper bound exceeds the known exact mean wait for M/G/1 [as given by the P-K mean value formula in Eq. (1.82)] by $(\bar{x}+\bar{t})/2$, which is less than one interarrival time. Marchal has proposed that the upper bound in Eq. (2.22) be scaled down so that it is *exact* for M/G/1; thus his approximation is

$$W \cong \frac{1+C_b^2}{(1/\rho)^2 + C_b^2}\left[\frac{\sigma_a^2 + \sigma_b^2}{2\bar{t}(1-\rho)}\right]$$

where C_b, the service time coefficient of variation, is defined as $C_b = \sigma_b/\bar{x}$. Both he [MARC 74] and Gross [GROS 73] consider the effectiveness of this (and other) approximiations to W. Their numerical studies show that the fit to G/M/1 is good, so far as percentage error is concerned; for G/G/1 it is fair, degrading with an increase in the coefficient of variation of either the interarrival times or the service times, and improving as ρ increases.

where we have defined W_U to be our upper bound

$$W_U \triangleq \frac{\sigma_a^2 + \sigma_b^2}{2\bar{t}(1-\rho)} \qquad (2.24)$$

We also had an alternative expression for the mean wait equivalent to that given in Eq. (2.23) and expressed it as Eq. (1.131), which we repeat here:

$$W = -\frac{\overline{u^2}}{2\bar{u}} - \frac{\overline{y^2}}{2\bar{y}} \qquad (2.25)$$

These last two expressions for W form our point of departure in establishing lower bounds on the mean wait in G/G/1. It is clear that if we are to obtain such lower bounds, then we must place an upper bound on $\overline{I^2}/2\bar{I}$, which is the mean residual life of the idle time period I. We have already introduced the random variable $\tilde{y} = (\tilde{w} + \tilde{u})^-$ [see Section 1.10, Eqs. (1.124), (2.12) and the footnote on p. 33] whose mean residual life is shown in Exercise 2.6 to be equal to that of the idle time, that is,

$$\frac{\overline{I^2}}{2\bar{I}} = \frac{\overline{y^2}}{2\bar{y}}$$

and since $\overline{y^2} = \sigma_{\tilde{y}}^2 + (\bar{y})^2$, we see that our main task is to place an upper bound on the variance of $\tilde{y}$; in this endeavor we follow the approach of Kingman [KING 62b]. First recall from Eqs. (1.116) and (1.126) that

$$\bar{t}(1-\rho) = -\bar{u} = \bar{y} \qquad (2.26)$$

Now, since $\tilde{y}$ has the same distribution as $(\tilde{w} + \tilde{u})^-$, and furthermore, since $\tilde{w} \geq 0$, then from a stochastic point of view,* $\tilde{w} + \tilde{u} \geq \tilde{u}$ and $(\tilde{w} + \tilde{u})^- \leq (\tilde{u})^-$; thus we may write

$$(\tilde{y})^2 = [(\tilde{w} + \tilde{u})^-]^2 \leq [(\tilde{u})^-]^2$$

Finally,

$$\overline{y^2} \leq \overline{[(\tilde{u})^-]^2}$$

Using this last and taking advantage of Eq. (2.15) (with $X = \tilde{u}$) we may also write

$$\overline{y^2} \leq \overline{(\tilde{u})^2} - \overline{[(\tilde{u})^+]^2}$$

* To say that a random variable X_1 is stochastically smaller than X_2 means that $P[X_1 \leq x] \geq P[X_2 \leq x]$.

which upon application of Eq. (2.26) yields

$$\frac{\overline{[(\tilde{u})^+]^2}}{2\bar{y}} \leq -\frac{\overline{u^2}}{2\bar{u}} - \frac{\overline{y^2}}{2\bar{y}} \tag{2.27}$$

So, finally, we substitute back into Eq. (2.25) and establish the following lower bound on the average waiting time:*

$$W_K \triangleq \frac{\overline{[(\tilde{u})^+]^2}}{2\bar{t}(1-\rho)} \leq W \qquad \blacksquare \tag{2.28}$$

This is the first of our lower bounds. We note that it depends on much more than just the first two moments of our input process. This is not an especially tight bound, and in order to do better, we must place conditions on our arrival process, as we shall see later.

Marshall [MARS 68a, b] has established a lower bound on W different from that given in Eq. (2.28). This new bound is an improvement over the other (W_K) in the light-traffic case, and the converse is true in the heavy-traffic case. To establish this new bound, our point of departure is once again the basic relationship

$$w_{n+1} = \max[0, w_n + u_n]$$

This piecewise linear expression takes on the value zero whenever $u_n < -w_n$; therefore, if we condition on the event $w_n = y \geq 0$, then any calculation for the expected value of w_{n+1} need only consider the range for which $u_n \geq -y$, and in this range it must be true that $w_{n+1} = y + u_n$. We may therefore form the conditional expectation on w_{n+1} as

$$E[w_{n+1} \mid w_n = y] = \int_{u=-y}^{\infty} (y+u)\, dP[u_n \leq u] \tag{2.29}$$

Recall that $P[u_n \leq u] = C(u)$. Integrating by parts, it is easy to show that the integral in the following equation is identical to that in Eq. (2.29), namely,

$$E[w_{n+1} \mid w_n = y] = \int_{u=-y}^{\infty} [1 - C(u)]\, du \tag{2.30}$$

this being good for all $y \geq 0$. It is convenient to define $g(y)$ as the integral above, that is,

$$g(y) \triangleq \int_{-y}^{\infty} [1 - C(u)]\, du$$

* We use the notation W_K for this lower bound since it is due to Kingman.

Now let us show that $g(y)$ is convex.* We observe that the PDF $C(u)$ is nondecreasing with u (for all u in the range $-\infty \leq u \leq \infty$), and so $C(-u)$ is nonincreasing with u; therefore, $1 - C(-u)$ is nondecreasing with u. We also have

$$\frac{dg(y)}{dy} = 1 - C(-y) \tag{2.31}$$

Due to the property for $1 - C(-u)$, we see that $dg(y)/dy$ is nondecreasing with y; thus $g(y)$ is convex.

Let us now proceed with the calculation of W. We define $W_n(y) \triangleq P[w_n \leq y]$. Unconditioning Eq. (2.30), we then have

$$\begin{aligned} E[w_{n+1}] &= \int_0^{\infty} E[w_{n+1} \mid w_n = y]\, dW_n(y) \\ &= \int_0^{\infty} \int_{-y}^{\infty} [1 - C(u)]\, du\, dW_n(y) \\ &= \int_0^{\infty} g(y)\, dW_n(y) \end{aligned}$$

Thus

$$E[w_{n+1}] = E[g(w_n)] \tag{2.32}$$

where the expectation on the right-hand side of this equation is with respect to the distribution of the random variable w_n. However, we have already shown that $g(y)$ is a convex function of its argument. Thus we may apply Jensen's inequality, which states, for any convex function g of a random variable X, that we must have

$$E[g(X)] \geq g(E[X]) \tag{2.33}$$

From Eqs. (2.32) and (2.33) we therefore have

$$E[w_{n+1}] \geq g(E[w_n])$$

If we allow $n \to \infty$ we obtain

$$W \geq g(W) \tag{2.34}$$

Let us now consider the equation $y = g(y)$, that is,

$$y = \int_{-y}^{\infty} [1 - C(u)]\, du \tag{2.35}$$

* That is, for $y_1 \leq y_2$ and $0 \leq \alpha \leq 1$, $g(y)$ will be shown to have the following property:

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha) g(x_2)$$

This is equivalent to requiring that $dg(y)/dy$ be nondecreasing.

where $y \geq 0$. We are interested in the value of y that satisfies this equation since, as we shall see, this value will be our lower bound, which we denote by W_M (the subscript reminds us that it is due to Marshall). We may rewrite Eq. (2.35) as

$$y = \int_{-y}^{0} [1 - C(u)]\,du + g(0) \tag{2.36}$$

for $y \geq 0$. In Figure 2.1 we show y and $g(y)$ versus y. Note that $g^{(1)}(0) = 1 - C(0^-) = P[u_n \geq 0] \geq 0$. We see that a solution to Eq. (2.35) will be obtained if and only if the two curves shown in Figure 2.1 intersect; of course this point of intersection is W_M. Let us next show that these curves cross *exactly once* (for $y \geq 0$) and therefore W_M is unique. We note from Eq. (2.36) that for $g(0) = 0$, $y = 0 = W_M$ will be a solution (and if W_M is to be our lower bound on W, then this value is useless). Moreover, if $g(0) > 0$, then the two curves will cross if and only if for sufficiently large y we have

$$\begin{aligned} y &> g(y) \\ &= g(0) + \int_{-y}^{0} [1 - C(u)]\,du \\ &= g(0) + y - \int_{-y}^{0} C(u)\,du \end{aligned}$$

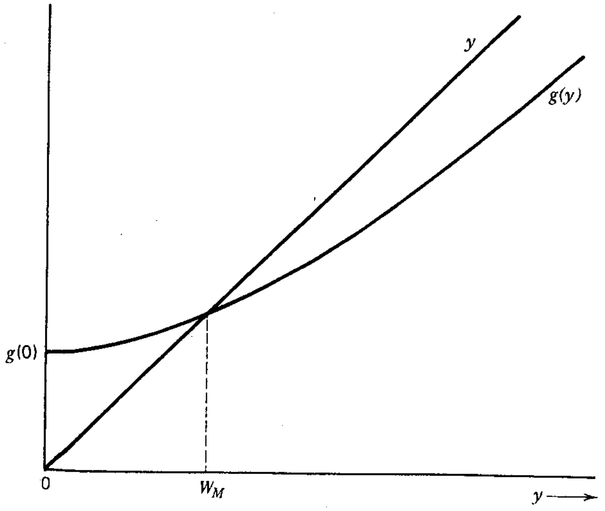

Figure 2.1 Location of the lower bound W_M.

This last condition reduces to

$$g(0) < \int_{-y}^{0} C(u)\,du \tag{2.37}$$

Now, as $y \to \infty$, this last integral is simply

$$\int_{-\infty}^{0} C(u)\,du = -E[\min(0, u_n)] = E[(u_n)^-]$$

Furthermore, $g(0)$ may be written as

$$\begin{aligned} g(0) &= \int_{0}^{\infty} [1 - C(u)]\,du \\ &= E[\max(0, u_n)] = E[(u_n)^+] \end{aligned}$$

Thus

$$\begin{aligned} E[\tilde{u}] &= E[(u_n)^+] - E[(u_n)^-] \\ &= g(0) - \int_{-\infty}^{0} C(u)\,du \end{aligned}$$

But $E[\tilde{u}] = -\bar{t}(1-\rho)$, and so as $y \to \infty$, the condition in Eq. (2.37) is equivalent to the condition

$$\bar{t}(\rho - 1) < 0$$

or

$$\rho < 1 \tag{2.38}$$

The condition expressed in inequality (2.38) is the condition that guarantees that both curves cross and thereby guarantees a (nontrivial) solution to Eq. (2.35); however, inequality (2.38) is our usual condition for stable queueing systems! Moreover, since we have just shown (for $\rho < 1$) that

$$\lim_{y \to \infty} (y - g(y)) > 0$$

and since $g(y)$ is a convex function, these curves will cross exactly once [if they crossed more than once, then by the convexity of $g(y)$ they would cross exactly twice, and the last inequality above would have to be reversed—a contradiction]. However, one might inquire whether these two curves can coincide over some interval $(a \leq y < b)$, say. This would require that $dg(y)/dy = 1$ over this interval. However, due to Eq. (2.31) our assumptions would then also require that $1 - C(-y) = 1$ over this region. But the function $1 - C(-y) \leq 1$ and is nondecreasing with y, thereby requiring that $dg(y)/dy = 1$ over the *entire* range $(a \leq y)$. Thus we come to the conclusion that if the curves coincide over any finite interval, then they must coincide over a semi-infinite interval and will never separate from one another; however, condition (2.38) guarantees that they will separate. We have thus arrived at a contradiction, thereby removing the possibility of the two curves coinciding over any finite range.

Thus, for $\rho<1$, W_M is the *unique* solution to Eq. (2.35). Now, if $W_M=0=g(0)$, then clearly $W\geq W_M$. On the other hand, if $W_M>0$, then $g(0)>0$ as shown above, and therefore for all $0\leq y<W_M$ we have that

$$y<g(0)+\int_{-y}^{0}[1-C(u)]\,du \tag{2.39}$$

due to the uniqueness arguments given above (see also Figure 2.1). Suppose now that $W<W_M$; in this case we would then be able to write

$$W<g(W) \tag{2.40}$$

since W would fall in the range for which Eq. (2.39) holds. However, Eq. (2.40) directly contradicts Eq. (2.34), and therefore we conclude that

$$W_M\leq W \qquad \blacksquare \tag{2.41}$$

which finally establishes the lower bound we were seeking. The value for the lower bound W_M is given as the unique solution to Eq. (2.35). Comparing this calculation with that required for Kingman's lower bound W_K, we see that they both require nontrivial computations.

We comment here that in Exercise 2.7 we show by methods similar to those described above that upper and lower bounds on the variance of the waiting time may be given as

$$\sigma_b^{\,2}\leq\sigma_{\tilde{w}}^{\,2}\leq\sigma_a^{\,2}+\sigma_b^{\,2}-2W_M\bar{t}(1-\rho) \qquad \blacksquare \tag{2.42}$$

If we are willing to place some simple constraints on the interarrival time distribution $A(t)$, then we find that we can simplify the lower bound on the average waiting time considerably. These constraints require that we define certain properties of the mean residual life and of the failure rate of distribution functions; the mean residual life and failure rate were defined in the footnote on p. 16. The following definitions are commonly used in reliability theory [BARL 65].

DEFINITION OF γ-MRLA (AND γ-MRLB): A nondiscrete distribution function F has its mean residual life bounded above (below) by γ [and is then said to be γ-MRLA (γ-MRLB)] if and only if

$$\int_t^{\infty}\frac{1-F(u)}{1-F(t)}\,du\underset{(\geq)}{\leq}\gamma \tag{2.43}$$

for all t and $1-F(t)>0$.

DEFINITION OF DMRL (AND IMRL): A nondiscrete distribution F has decreasing (increasing) mean residual life DMRL (IMRL) if and only if

$$\int_t^{\infty}\frac{1-F(u)}{1-F(t)}\,du \text{ decreases (increases) with } t \tag{2.44}$$

for $t\geq 0$ and $1-F(t)>0$.

DEFINITION OF IFR (AND DFR): A nondiscrete distribution function F has increasing (decreasing) failure rate IFR (DFR) if and only if for any $\varepsilon>0$ we have that

$$\frac{F(t+\varepsilon)-F(t)}{1-F(t)} \text{ increases (decreases) with } t \tag{2.45}$$

for all $t>0$ and for $1-F(t)>0$.

The first definition merely describes distributions whose mean residual life may be bounded independent of the age of the random variable. The second definition describes distribution functions whose mean residual life behaves monotonically with the age of the random variable. The third definition describes distribution functions whose death rate (failure rate) behaves monotonically with age. It can be shown that

$$\text{IFR}\subset\text{DMRL}\subset\bar{X}-\text{MRLA} \tag{2.46}$$

where $\subset$ is read as "implies" and $\bar{X}$ is the mean value of the random variable under consideration.

We now wish to apply the notion of the mean residual life and the definitions for this quantity described above for the case of the interarrival distribution $A(t)$ in our queueing system G/G/1. For an interarrival time distribution that is γ-MRLA in the system G/G/1 we use the special notation γ-MRLA/G/1, whereas if $A(t)$ is IFR, then we write IFR/G/1. In Exercise 2.8 we show for the queueing system γ-MRLA/G/1 that

$$\frac{\overline{I^2}}{2\bar{I}}\leq\gamma \tag{2.47}$$

where I is the random variable describing the idle time as earlier. As we commented previously, as soon as we are able to place an upper bound on the mean residual idle time, as we have just done, then Eq. (2.23) will immediately provide for us a lower bound on the mean wait W. We judiciously choose $\gamma = \bar{t}$, in which case Eqs. (2.23) and (2.47) give

$$\begin{aligned} W &\ge W_U + \frac{1}{2}\,\bar{t}(1-\rho) - \bar{t} \\ &= W_U - \frac{1}{2}\,\bar{t}(1+\rho) \end{aligned}$$

Thus, for the queueing system $\bar{t}$-MRLA/G/1, we have the upper and lower bounds on the mean wait given by

$$W_U - \frac{1}{2}\,\bar{t}(1+\rho) \le W \le W_U \qquad \blacksquare \quad (2.48)$$

If we now apply Little's result to this last equation and recall that $\bar{N}_q$ denotes the average number of customers in the queue, then we may bound this quantity as

$$\lambda W_U - \frac{1+\rho}{2} \le \bar{N}_q \le \lambda W_U \qquad \blacksquare \quad (2.49)$$

where $\lambda = 1/\bar{t}$ is the average arrival rate of customers to this queue. This last equation gives upper and lower bounds on the expected queue size; note that the difference between these bounds is less than unity!

In Exercise 2.11 we show for the IFR/G/1 queue that

$$\frac{\overline{I^2}}{2\bar{I}} \le \frac{\overline{t^2}}{2\bar{t}} = \frac{1}{2}\,\bar{t}(1+C_a^2) \qquad (2.50)$$

where $\overline{t^2}$ and C_a $(=\sigma_a/\bar{t})$ are the second moment and the coefficient of variation, respectively, for the interarrival time. Thus again we have an upper bound on the mean idle time, and so we may apply this to Eq. (2.23) to yield the following lower bound on the mean wait:

$$\begin{aligned} W &\ge W_U + \frac{1}{2}\,\bar{t}(1-\rho) - \frac{1}{2}\,\bar{t}(1+C_a^2) \\ &= W_U - \frac{1}{2}\,\bar{t}(\rho + C_a^2) \end{aligned}$$

Combining this with our upper bound we have that the IFR/G/1 queue has a mean waiting time bounded as follows:

$$W_U - \frac{1}{2}\,\bar{t}(C_a^2 + \rho) \le W \le W_U \qquad \blacksquare \quad (2.51)$$

It can easily be shown that any distribution that is IFR must have a coefficient of variation less than unity; therefore, the lower bound in Eq. (2.51) is tighter than the lower bound in Eq. (2.48). This is a reflection of the relationship in Eq. (2.46) which states that the IFR constraint is the strongest among the three. Applying Little's result here we find

$$\lambda W_U - \frac{C_a^2 + \rho}{2} \le \bar{N}_q \le \lambda W_U \qquad \blacksquare \quad (2.52)$$

which again reduces the range of uncertainty for the average queue size to less than one customer. For example, the system D/G/1 which is IFR and for which $C_a^2 = 0$ results in an average queue size that is bounded to within one-half a customer.

Except for these last two cases of $\bar{t}$-MRLA/G/1 and IFR/G/1, the lower bounds we have found in this section are not simply expressed in terms of the first two moments of the interarrival and service time distributions (which was the happy situation with regard to the upper bound of Section 2.2). Marchal [MARC 74] has developed such a lower bound which we now present. Our approach, once again, is to place an upper bound on $\overline{I^2}/2\bar{I}$ in Eq. (2.23), and our point of departure is the expression for y_n given in Eq. (1.124). We have already noted that the limiting form of this random variable, $\tilde{y}$, may be expressed as

$$\tilde{y} = -\min\,[0, \tilde{w} + \tilde{u}]$$

We have already shown (in Exercise 2.6) that $\overline{I^2}/2\bar{I} = \overline{y^2}/2\bar{y}$, and so we will study the moments of $\tilde{y}$. We may rewrite $\tilde{y}$ as

$$\begin{aligned} \tilde{y} &= \max\,[0, -\tilde{w} - \tilde{u}] \\ &= \max\,[0, \tilde{t} - \tilde{x} - \tilde{w}] \end{aligned}$$

Now, since $\tilde{x}$, $\tilde{t}$ and $\tilde{w}$ are all non-negative random variables, then from this last expression, it must be that $\tilde{y}$ is stochastically smaller than $\tilde{t}$. It then follows that $\overline{y^k} \le \overline{t^k}$. Now, since $\overline{t^2} = \sigma_a^2 + (1/\lambda)^2$ we have

$$\overline{y^2} \le \sigma_a^2 + \frac{1}{\lambda^2}$$

But $\bar{y} = (1-\rho)/\lambda$ and so

$$\frac{\overline{y^2}}{2\bar{y}} = \frac{\overline{I^2}}{2\bar{I}} \le \frac{\lambda(\sigma_a^2 + 1/\lambda^2)}{2(1-\rho)}$$

Substituting this upper bound into Eq. (2.23), we finally obtain

$$W_U - \frac{\rho(2-\rho) + C_a^2}{2\lambda(1-\rho)} \le W \qquad \blacksquare \quad (2.53)$$

This may also be expressed as

$$\frac{\rho^2 C_b^2 + \rho(\rho - 2)}{2\lambda(1-\rho)} \le W \qquad \blacksquare \quad (2.54)$$

This is the lower bound we were seeking. Note that it is not symmetrical in σ_a^2 and σ_b^2, as is W_U in Eq. (2.22). This bound will be non-negative only for service time coefficients of variation that satisfy $C_b^2 \geq (2-\rho)/\rho$. The exact value for W for the system M/G/1 exceeds this lower bound by an amount $\bar{x}/(1-\rho)$; therefore, the bound degrades as ρ increases (but we have seen that the upper bound improves as ρ increases). The main virtue of this bound seems to be its simplicity.

Let us now look for bounds on the waiting time distribution itself rather than on the mean wait.

2.4. BOUNDS ON THE TAIL OF THE WAITING TIME DISTRIBUTION

We recognize that a customer's waiting time is the sum of the service times for all those customers he finds in the queue upon his arrival plus the residual service time for the customer he finds in service. Of course, each of the queued customers' service times is independent and identically distributed, and so we might expect that a result similar to the Chernoff bound [KLEI 75] would perhaps provide an upper and lower bound on the tail of the waiting time distribution. This is indeed the case, and we follow Kingman's approach [KING 70] in establishing these bounds.

Once again we begin with the equation $w_{n+1} = \max[0, w_n + u_n]$. Therefore, for $y > 0$ we may write

$$P[w_{n+1} \geq y] = P[w_n + u_n \geq y]$$

Conditioning this on the value for u_n and recognizing that $P[w_n \geq 0] = 1$, we have

$$P[w_{n+1} \geq y] = \int_{-\infty}^{\infty} P[w_n \geq y-u]\, dC(u)$$

$$= \int_{-\infty}^{y} P[w_n \geq y-u]\, dC(u) + 1 - C(y) \qquad (2.55)$$

Now let us consider $C^*(-s) \triangleq E[e^{su_n}]$ where s is taken to be a real (rather than a complex) variable; we recognize that s must lie in a restricted range if this transform is to remain bounded. In particular, if there exists a real s' such that $B^*(-s') \triangleq E[e^{s'\tilde{x}}] < \infty$, then a permissible range for s is $0 \leq s \leq s'$. Furthermore, there will be a range in which $C^*(-s) \leq 1$ (for example, in this stable case, $C^*(0) = 1$ and for $s = 0$, $dC^*(-s)/ds = \bar{u} < 0$, thus identifying a neighborhood in this range), and we let s_0 denote the largest value for s such that this remains true. We may thus write the following inequality:

$$e^{-s_0 y} \geq e^{-s_0 y} C^*(-s_0)$$

$$= e^{-s_0 y} \int_{-\infty}^{\infty} e^{s_0 u}\, dC(u)$$

$$= \int_{-\infty}^{\infty} e^{-s_0(y-u)}\, dC(u) \qquad (2.56)$$

Now since $s_0 > 0$, for the range $u \geq y$ it must be that $e^{-s_0(y-u)} \geq 1$; thus inequality (2.56) may be extended to

$$e^{-s_0 y} \geq \int_{-\infty}^{y} e^{-s_0(y-u)}\, dC(u) + \int_{y}^{\infty} dC(u)$$

$$= \int_{-\infty}^{y} e^{-s_0(y-u)}\, dC(u) + 1 - C(y) \qquad (2.57)$$

so long as $y > 0$.

Let us now assume that w_0 (an initial customer's waiting time) is chosen so that $P[w_0 \geq y] \leq e^{-s_0 y}$; we wish to prove that this hypothesis carries over for all w_n. We prove this by induction, assuming that we have already established its truth up to the nth step, that is $P[w_n \geq y] \leq e^{-s_0 y}$. Then applying this to Eq. (2.55) we have

$$P[w_{n+1} \geq y] \leq \int_{-\infty}^{y} e^{-s_0(y-u)}\, dC(u) + 1 - C(y)$$

But this right-hand side is exactly the expression we bounded in Eq. (2.57), and so we conclude that $P[w_{n+1} \geq y] \leq e^{-s_0 y}$ also, completing the inductive proof. Thus we have established the following exponential bound on the tail of the equilibrium waiting time distribution (by letting $n \to \infty$):

$$P[\tilde{w} \geq y] \leq e^{-s_0 y} \qquad (2.58)$$

where, as we stated earlier, s_0 is found from

$$s_0 = \sup\{s > 0 : C^*(-s) \leq 1\}$$

The result given in Eq. (2.58) is, as we had predicted, similar to the form of the Chernoff bound. It is possible also to prove that this tail has a lower bound of a similar form [KING 70], which combines with Eq. (2.58) to give

$$\gamma e^{-s_0 y} \leq 1 - W(y) \leq e^{-s_0 y} \qquad \blacksquare \quad (2.59)$$

where we have used our usual notation $W(y) \triangleq P[\tilde{w} \le y]$ and where γ must satisfy the inequality

$$\gamma \le \frac{1 - C(y)}{\int_y^{\infty} e^{-s_0(y-u)}\, dC(u)} \tag{2.60}$$

for all values of $y > 0$; therefore, γ is the smallest value that the ratio in this last equation takes on. From these bounds on the distribution function itself it is trivial to show that the mean wait may also be bounded by

$$\frac{\gamma}{s_0} \le W \le \frac{1}{s_0} \tag{2.61}$$

These bounds on W are sometimes sharper than those we considered earlier.

Kobayashi [KOBA 74c] also derives the Kingman upper bound in Eq. (2.58) using Kolmogorov's inequality for submartingales; Ross [ROSS 74] improves on Kingman's upper bound and studies these results for some special cases.

2.5. SOME REMARKS FOR G/G/m

So little is known about the queue G/G/m that any results available for its approximate behavior are extremely worthwhile. Much of the work has been addressed at bounding the mean wait and it is this which we discuss below.

As we know, the appropriate definition for the utilization factor of this system is

$$\rho = \frac{\bar{x}}{m\bar{t}} \tag{2.62}$$

and it has been shown [KIEF 55] that the condition for stability in this case is still

$$\rho < 1$$

Now the most general multiple-server queue that we have so far seen is G/M/m, and from Eq. (1.115) we observed that the conditional pdf for waiting time is exponentially distributed with parameter $m\mu(1-\sigma)$ where $\bar{x} = 1/\mu$; in the heavy-traffic case we expect the unconditional waiting time density to approach this conditional density, and so in that case we may write

$$W \cong \frac{1}{m\mu(1-\sigma)} \qquad \rho \to 1 \tag{2.63}$$

We must solve for the value of σ, which is given as the appropriate root of Eq. (1.104) repeated here:

$$\sigma = A^*(m\mu - m\mu\sigma)$$

Making the change of variable $\alpha = m\mu(1-\sigma)$ the last equation becomes

$$1 - \frac{\alpha\bar{x}}{m} = A^*(\alpha)$$

If we now expand $A^*(\alpha)$ in a power series about the origin as in Eq. (2.4) we have

$$1 - \frac{\alpha\bar{x}}{m} = 1 - \bar{t}\alpha + \frac{\overline{t^2}\alpha^2}{2!} + o(\alpha^2)$$

Since we are considering the heavy-traffic case, we see that $\alpha \ll 1$ [that is, $W \gg \bar{x}$ and $\sigma \cong 1$; see Eq. (2.63)], and so we may neglect the higher-order terms; neglecting $o(\alpha^2)$ and solving for α we have

$$\alpha \cong \frac{2\bar{t}(1-\rho)}{\sigma_a^2 + \bar{t}^2}$$

but since $m\bar{t} \cong \bar{x}$ and since for the exponential service time $\sigma_b^2 = \bar{x}^2$, we may rewrite this last expression as

$$\alpha \cong \frac{2\bar{t}(1-\rho)}{\sigma_a^2 + (1/m^2)\sigma_b^2}$$

We may finally use this result in Eq. (2.63) to give the following† as the approximate mean wait in G/M/m as $\rho \to 1$:

$$W \cong \frac{\sigma_a^2 + (1/m^2)\sigma_b^2}{2\bar{t}(1-\rho)} \qquad \blacksquare \tag{2.64}$$

This observation led Kingman [KING 64] to generalize from G/M/m to G/G/m and to suggest (conjecture) for the heavy-traffic approximation for G/G/m that the waiting time should be distributed exponentially with a mean wait given by Eq. (2.64). This conjecture has recently been established by Köllerström [KOLL 74]; thus the Kingman–Köllerström approximation to the waiting time distribution for G/G/m is

$$W(y) \cong 1 - \exp\left(-\frac{2\bar{t}(1-\rho)}{\sigma_a^2 + (\sigma_b^2/m^2)}\, y\right) \qquad \blacksquare \tag{2.65}$$

The proof of this result uses a G/G/1 approximation to G/G/m in heavy traffic with interarrival times t_n and service times x_n/m. (The Brumelle

† Note for $m = 1$, that this approximation for W reduces exactly to Kingman's G/G/1 approximation.

lower bound below also uses this approach.) This G/G/1 approximation was developed earlier by Kiefer and Wolfowitz [KIEF 55]. This heavy-traffic approximation for $W(y)$ also implies that the heavy-traffic approximation for W is as given in Eq. (2.64) for G/G/m.

Suzuki and Yoshida [SUZU 70] have shown that Kingman's conjecture is truly an upper bound for W for $\rho \le 1/m$. Kingman himself [KING 70] suggests that the approximation is an upper bound for $0 \le \rho < 1$, but does not prove it, and so far this remains only a conjecture. We state the known *bounds* on W without proof. Kingman [KING 70] derives the following upper bound for the mean wait:

$$W \le \frac{\sigma_a^2 + (1/m)\sigma_b^2 + [(m-1)/m^2]\bar{x}^2}{2\bar{t}(1-\rho)} \tag{2.66}$$

Brumelle [BRUM 71] also finds this upper bound for G/G/m.

As for the lower bounds on G/G/m, Kingman [KING 70] shows the following:

$$W \ge \frac{2W^*\bar{t} - (\sigma_b^2 + m\sigma_a^2) - [(m-1)/m]\bar{x}^2}{2\bar{x}} \triangleq K_L \tag{2.67}$$

where W^* is the average waiting time in a G/G/1 system with service times $\{x_n\}$ and interarrival times $\{mt_n\}$. Brumelle [BRUM 71] also gives a lower bound in the following form:

$$W \ge \hat{W} - \frac{[(m-1)/m]\overline{x^2}}{2\bar{x}} \triangleq B_L \tag{2.68}$$

where $\hat{W}$ is the average waiting time in a G/G/1 system with service times $\{x_n/m\}$ and interarrival times $\{t_n\}$. Let us compare these last two bounds. If one plots the unfinished work for Kingman's special single-server system (whose average wait is W^* and whose average unfinished work will be denoted by $\bar{U}^*$) and if one also plots the unfinished work for Brumelle's equivalent single-server system (whose average waiting time is given by $\hat{W}$ and whose average unfinished work will be denoted by $\bar{\hat{U}}$) then one readily finds that

$$\overline{U^*} = m\bar{\hat{U}} \tag{2.69}$$

This is easily seen by comparing the two unfinished work functions and recognizing that the average of the unfinished work on a scaled time axis is independent of the scaling. In Chapter 3 below, we show that [see Eq. (3.23)] the average unfinished work, $\bar{U}$, is simply

$$\bar{U} = \rho W + \frac{\overline{x^2}}{2\bar{t}}$$

for G/G/1. If we form $\bar{U}^*$ and $\bar{\hat{U}}$ it is clear from this last equation that

$$W^* = m\hat{W} \tag{2.70}$$

Now if we subtract Kingman's lower bound from Brumelle's lower bound and denote this by $B_L - K_L$ we have

$$B_L - K_L = \hat{W} - \frac{\bar{t}}{\bar{x}} W^* + \frac{m\sigma_a^2 + (1/m)\sigma_b^2}{2\bar{x}}$$

Using Eq. (2.70) we then see

$$B_L - K_L = \hat{W}\left(\frac{\rho - 1}{\rho}\right) + \frac{\sigma_a^2 + (1/m^2)\sigma_b^2}{2\bar{t}\rho}$$

Now the average wait $\hat{W}$ in Brumelle's single-server system clearly has an upper bound given by Eq. (2.22), where the mean service time is $\bar{x}/m$ and the service time variance is σ_b^2/m^2; throughout we maintain the definition for ρ as given in Eq. (2.62). Therefore we may write

$$\hat{W} \le \frac{\sigma_a^2 + (1/m^2)\sigma_b^2}{2\bar{t}(1-\rho)}$$

Using this inequality in the expression for $B_L - K_L$ we immediately have

$$B_L - K_L \ge 0$$

which clearly shows that Brumelle's lower bound is tighter (larger) than Kingman's lower bound.

In summary then the best published bounds for the average wait in G/G/m are[†]

$$\hat{W} - \frac{[(m-1)/m]\overline{x^2}}{2\bar{x}} \le W \le \frac{\sigma_a^2 + (1/m)\sigma_b^2 + [(m-1)/m^2]\bar{x}^2}{2\bar{t}(1-\rho)} \quad ■ \tag{2.71}$$

One sees that these bounds are consistent with the Kingman–Köllerström heavy-traffic approximation of an exponentially distributed waiting time [Eq. (2.65)] with mean given by Eq. (2.64). The term $\hat{W}$ is Brumelle's single-server system to which we may apply any of our earlier bounds; in particular if one is willing to assume more about the interarrival times such as we did in Section 2.3 (for example IFR) then a tighter lower bound may be obtained.

An improvement in the upper bound may be found for the special case G/M/m. Although G/M/m has been solved exactly, as we saw in Section 1.9, we observed there that the solution required the difficult calculation of J and R_k $(k = 0, 1, \ldots, m-2)$. Therefore an easily calculated bound

[†] For $\rho \le 1/m$, the upper bound can be tightened by the results of Suzuki and Yoshida mentioned above.

serves a useful purpose. The key result here (due to Brumelle [BRUM 73]) is once again to consider his single-server system G/M/1 with service times $\{x_n/m\}$ and interarrival times $\{t_n\}$; again we denote all variables for this system by a caret. Brumelle shows that $P[w_n > y] < P[\hat{w}_n > y]$, which yields, as a corollary, $W \le \hat{W}$. To calculate $\hat{W}$, we need deal only with a single-server system, which avoids the calculation of J and R_k; the mean wait $\hat{W}$ is in fact given in Eq. (1.103), which involves finding σ from Eq. (1.100)—a much simpler task. On the other hand, to make the G/M/1 calculation even easier, we may use our earlier result in Eq. (2.22), which is good for any G/G/1 system, to obtain finally for G/M/m

$$W \le \hat{W} \le \frac{\sigma_a{}^2 + (\sigma_b{}^2/m^2)}{2\bar{t}(1-\rho)} \qquad \blacksquare \quad (2.72)$$

which is an improvement over Eq. (2.71) and which shows that the Kingman–Köllerström heavy-traffic approximation is, in fact, an upper bound to W for G/M/m. In fact, the bounds are rather tight, since we now have shown* that for G/M/m,

$$\hat{W} - \left(\frac{m-1}{m}\right)\bar{x} \le W \le \hat{W} \qquad \text{(G/M/m)} \qquad \blacksquare \quad (2.73)$$

Using Little's result, we have

$$\bar{\hat{N}} - \rho(m-1) \le \bar{N} \le \bar{\hat{N}} \qquad \text{(G/M/m)} \qquad \blacksquare \quad (2.74)$$

and since $\rho < 1$, we have bounded the average number in system to within $m-1$ of its true value [and this true value happens also to be within $m-1$ of the average number in the equivalent G/M/1 system, i.e., $\bar{\hat{N}} - \bar{N} \le \rho(m-1)$].

If we now take advantage of Marchal's lower bound for G/G/1 in Eq. (2.54) and use it with Brumelle's lower bound in Eq. (2.68), we arrive at a simple explicit lower bound G/G/m as follows. In particular, we bound $\hat{W}$ by

$$\frac{\rho^2 C_b{}^2 - \rho(2-\rho)}{2\lambda(1-\rho)} \le \hat{W}$$

Using this in Eq. (2.68), we get

$$\frac{\rho^2 C_b{}^2 - \rho(2-\rho)}{2\lambda(1-\rho)} - \frac{[(m-1)/m]\overline{x^2}}{2\bar{x}} \le W \qquad \blacksquare \quad (2.75)$$

This is a simpler explicit lower bound for G/G/m.

The results of this section only begin to provide some answers for G/G/m; much more work needs to be done in this area.

* We note that $\overline{x^2}/2\bar{x} = \bar{x}$ for the exponential service time.

2.6. A DISCRETE APPROXIMATION

So far in this chapter we have handled the complexity of the G/G/1 queue by finding approximations and bounds for the exact solution. Throughout most of the rest of this chapter we take a different point of view: rather than attempt an approximate solution for the original problem, we attempt an exact solution for an approximation of the original problem. That is, we purposefully distort the equations of motion for the given G/G/1 queue and reformulate them in a fashion that permits the system equations to be solved. In this section we discuss a rather crude discrete approximation.

The key to the approach in this section is to alter the input distributions [$A(t)$ and $B(x)$] in such a way that our basic recurrence relationship [given again in Eq. (2.76) below] permits a direct analytic solution for the distribution of waiting time,

$$w_{n+1} = \max[0, w_n + u_n] \qquad (2.76)$$

We observe that the iterative application of this equation is quite straightforward when the interarrival time and service time are both discrete random variables whose only nonzero values occur at the instants $k\tau$ $(k = 0, 1, 2, \ldots)$ where τ is the basic time unit. In such cases one may write down the limit of such recursions to yield a set of linear difference equations that may then be handled by the method of z-transforms [KLEI 75]. We see that this approach requires little more sophistication than that which one uses in elementary queueing theory. If our original random variables are of this discrete nature to begin with, then we have a simple method for giving the *exact* distribution of waiting time. On the other hand, if our given random variables are continuous, then we are faced with an approximation problem; that is, we must approximate the continuous random variables with discrete ones in a fashion that preserves the essence of the solution we seek. Just how one goes about choosing this approximation is as yet basically unstudied and the only recommendation we make at this point is that if one wishes to represent a continuous distribution with a finite set of discontinuities then one should use this approximation to match as many of the moments of the original distribution as possible, working from the first moment and proceeding upwards. We emphasize again, however, that the precision of this approximation has only begun to be studied.

Perhaps the best way to present this method is through an example. We avoid the question of how one should approximate a continuous random variable and assume we begin with discrete interarrival time and service time distributions. Thus we assume, by way of example,

$$A(t)=\begin{cases}0 & t<2\tau\\ 1 & t\geq 2\tau\end{cases}$$

$$B(x)=\begin{cases}0 & x<0\\ \frac{1}{2} & 0\leq x<3\tau\\ 1 & 3\tau\leq x\end{cases}$$

Since we deal only with discrete random variables, let us define $a(k)=P[t_n=k\tau]$ and $b(k)=P[x_n=k\tau]$; we may then display these discrete functions as in Figure 2.2. Of course these also could have been represented as pdf's with impulses at these same points.

If we are to apply Eq. (2.76) we must find the probability distribution for u_n. We define $c(k)=P[u_n=k\tau]$; since $u_n=x_n-t_{n+1}$, we see that $c(k)$ must in general be given by the following discrete convolution:

$$\begin{aligned}c(k)&=a(-k)\circledast b(k)\\ &=\sum_{i=-\infty}^{\infty}a(-k+i)b(i)\end{aligned}$$

So long as the representation $a(k)$ and $b(k)$ contain a small number of terms, then this convolution is easily carried out by hand; for our example it is trivial and leads to

$$c(k)=\begin{cases}\frac{1}{2} & k=-2\\ \frac{1}{2} & k=1\\ 0 & \text{otherwise}\end{cases}$$

which is shown in Figure 2.3.

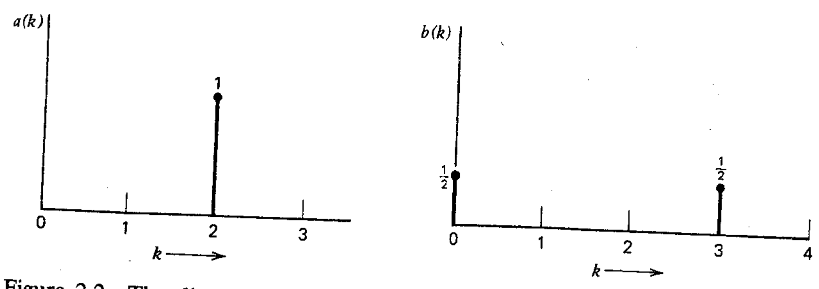

Figure 2.2 The discrete probabilities.

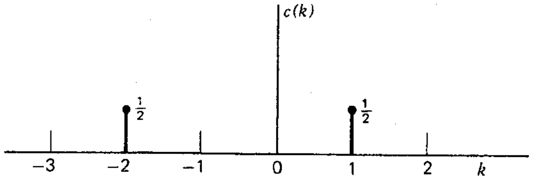

Figure 2.3 The probability $P[u_n=k\tau]$.

In order to apply the recursion in Eq. (2.76) we need a starting value, so let us assume for this example that $w_0=0$. Furthermore, we define the probability (mass) function as

$$p_n(k)=P[w_n=k\tau]$$

We are now ready to apply the recursion, which means that we must carry out the operations described in Eq. (1.121). Assuming that we have applied this recursion up to the calculation of $p_n(k)$, we may proceed as follows. First we find the probability distribution for the random variable w_n+u_n, which means that we must convolve $p_n(k)$ with $c(k)$; then the effect of the operator π means that we must sweep the probability in the negative half-line up to the origin (which in our discrete problems is simply a matter of addition) giving the next stage of the recursion, namely, $p_{n+1}(k)$. Carrying this operation out for our problem we generate the sequence shown in Figure 2.4. Starting in the upper left-hand corner of this figure we see the initial waiting time distribution; forming its convolution with $c(k)$ we get the distribution shown in the upper right-hand corner. Sweeping the probability in the negative half-line up to the origin we then easily find $p_1(k)$; this convolved with $c(k)$ gives the figure to its right, which when its negative half-line is swept up to the origin gives us $p_2(k)$, and so on, as we follow the arrows through the sequence of probability (mass) functions. Our object is to find the limiting probability function defined as

$$p(k)\triangleq\lim_{n\to\infty}p_n(k)$$

In order for this ergodic distribution to exist we require that $\rho<1$, which is equivalent to requiring that $E[u_n]<0$; for this example we have $\bar{x}=\frac{3}{2}$, $\bar{t}=2$ and so $\rho=\bar{x}/\bar{t}=\frac{3}{4}$ and $E[u_n]=\bar{x}-\bar{t}=-\frac{1}{2}$. The procedure for writing down the equations that describe the probabilities $p(k)$ is easily seen once we understand what is happening in Figure 2.4. We require that the stable distribution, after being convolved with $c(k)$ and then swept up to the origin, is exactly as it was before these two operations; that is, Eq. (1.122) must hold,. which in our notation becomes $p(k)=\pi(p(k)\circledast c(k))$. Furthermore we note that the π operator only affects the term $p(0)$, and so the form for $c(k)$ tells us exactly how a given term $p(k)$ is related to its

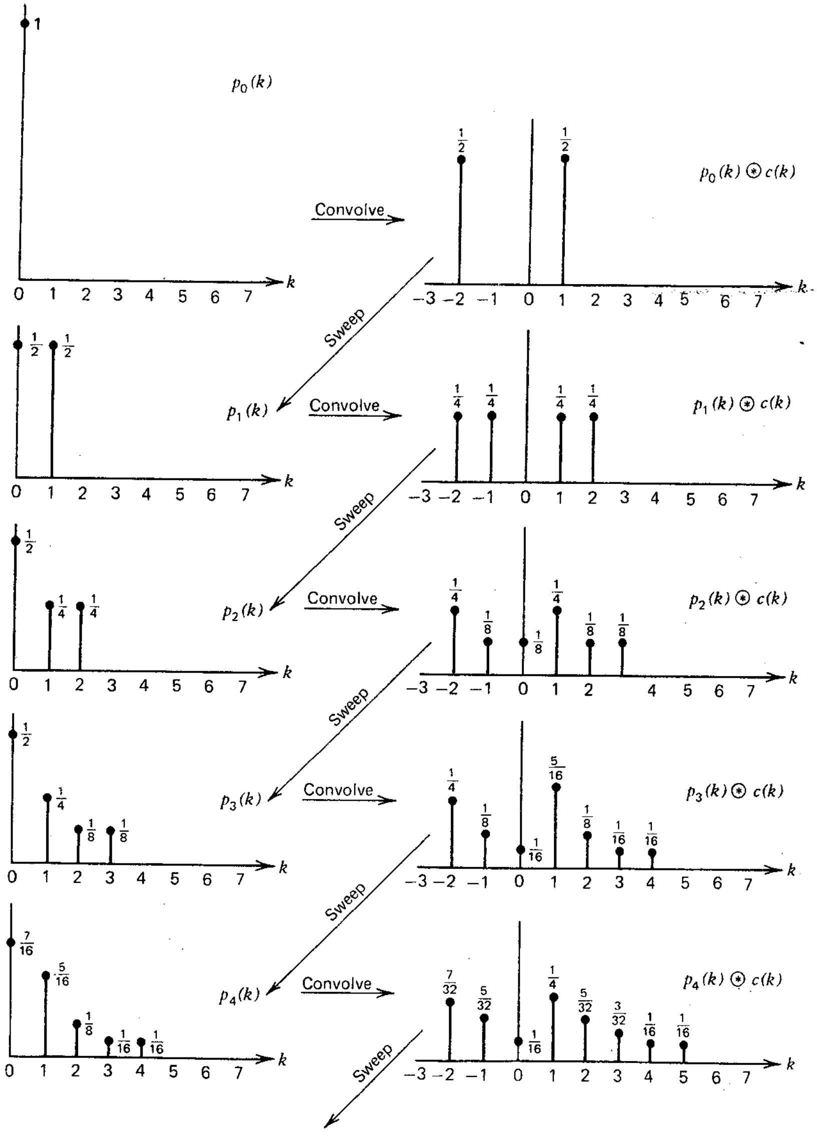

Figure 2.4 The recursion $p_{n+1}(k) = \pi(p_n(k) \circledast c(k))$.

neighbors. In particular, for our example we see that

$$p(k) = \tfrac{1}{2}p(k-1) + \tfrac{1}{2}p(k+2) \qquad k = 1, 2, 3, \ldots \tag{2.77}$$

and that the boundary equation for $p(0)$ is

$$p(0) = \tfrac{1}{2}p(0) + \tfrac{1}{2}p(1) + \tfrac{1}{2}p(2) \qquad k = 0 \tag{2.78}$$

We are now faced with the familar problem of solving a set of linear difference equations. The rest is straightforward (see Exercise 2.12). Certainly, the spectrum factorization method summarized in Section 1.10 leads to the same solution. The virtue of the method given here is that it explicitly takes advantage of the discrete nature of our random variables. In both cases the difficult part of the solution is in finding the roots of a polynomial [in the case here it is the roots of the denominator of $P(z) = \sum_k p(k)z^k$, whereas with the spectrum factorization method it is finding the roots of $A^*(-s)B^*(s) - 1$]. The point is, however, that we are suggesting an approximation scheme to convert continuous problems to discrete ones for which rather simple methods apply; the important question regarding how one generates an adequate approximation has not been discussed here. This issue of approximation has recently been studied by Wong [WONG 74]; he investigated how well matched were the distributions and moments of the input and waiting time variables. His results comparing the exact mean wait W with that of W_A as obtained from the iteration shown in Figure 2.4 and with Kingman's upper bound W_U are given in the table below; we note the excellent match between W and W_A.

System	ρ	W	W_A	W_U
M/M/1	2/3	0.13	0.13	0.22
M/E2/1	2/3	0.10	0.10	0.18
M/E3/1	2/3	0.09	0.09	0.17
M/E5/1	2/3	0.08	0.08	0.16
M/E10/1	2/3	0.07	0.07	0.16
E2/M/1	2/3	0.09	0.09	0.14
E3/M/1	2/3	0.08	0.08	0.12
E5/M/1	2/3	0.06	0.07	0.10
E10/M/1	2/3	0.06	0.06	0.08
M/D/1	2/3	0.07	0.07	0.15
D/M/1	2/3	0.05	0.05	0.07
E2/E2/1	2/3	0.06	0.06	0.11
M/H2/1	5/6	0.43	0.40	0.53
H2/M/1	4/5	0.28	0.29	0.36
H2/H2/1	5/9	0.12	0.14	0.26

We also comment that Cohen [COHE 69] gives a procedure for handling the case G/G/1 where he truncates the service time distribution and considers a new distribution $B_c(x)$ such that $B_c(x) = B(x)$ for $x \le x_c$ and $B_c(x) = 1$ for $x > x_c$. Taking advantage of the simplifications derived from this truncated distribution, he then presents a method of solution and considers the implications as $x_c \to \infty$ to remove the effect of truncation. Neuts and Klimko also consider a G/G/1 system with truncated service times and further restrict their attention to discrete time (as we have done here) [NEUT 73]; they study the ease of numerical analysis for this case.

2.7. THE FLUID APPROXIMATION FOR QUEUES

When an engineer is faced with a systems analysis problem, the first thing he attempts to do is to estimate the gross behavior of the system, however crude that estimate may be. That is, he attempts to "size" the system behavior so that he may make some first-order engineering calculations. Once this is done his task is then to refine his estimates and his approximate analysis; this refinement need be carried only so far as is necessary to insure satisfactory operation within some bounds. The point is that he must come up with answers (estimates) on *all* aspects of the system behavior including transient response, overload conditions, and so on, and not only "nice" equilibrium results. Much of queueing theory is obsessed with "nice" results; even in the early parts of this chapter our bounds and inequalities have applied only to the equilibrium conditions. In this and the three following sections we adopt a different point of view, treating queueing systems as *continuous fluid flow* rather than as discrete customer flow. This enables us to study transients and overloads.

We know for any queueing system that both the number of customers and the unfinished work as functions of time are each stochastic processes with *discontinuous* jumps (for example, at instants of customer arrivals to the system). The approximation we wish to study takes advantage of the following observation: when the system is in the heavy-traffic condition (namely when the queue sizes are large compared to unity and when the waiting times are large compared to average service times) then it appears reasonable to replace these discontinuities by smooth continuous functions of time. The usefulness of this approximation lies in the fact that the magnitude of the original discontinuities is small relative to the average value of these functions. We are dealing with a case of small relative increments. Thus we are led to a continuous stochastic fluid flow approximation for the original discrete queueing system.* Much of the work reported upon here was developed by Newell [NEWE 65, 68, 71] and Gaver [GAVE 68]; others who have studied these approximation methods include Borovkov [BORO 64, 65], Iglehart [IGLE 69], Iglehart and Whitt [IGLE 70], Kingman [KING 64], Prohorov [PROH 63], and others.

Among the fundamental stochastic processes for queues are the arrival process and the departure process defined as

$$\alpha(t) \triangleq \text{Number of arrivals in } (0, t) \tag{2.79}$$

$$\delta(t) \triangleq \text{Number of departures in } (0, t) \tag{2.80}$$

A typical realization for these step-wise increasing processes is shown in Figure 2.5. Clearly at any instant of time their difference must represent $N(t)$, the number of customers present in the system (for $N(0) = 0$), that is,

$$N(t) = \alpha(t) - \delta(t) \tag{2.81}$$

When $\alpha(t)$ gets large compared to unity then we expect only small percentage deviations from its average value $E[\alpha(t)] \triangleq \overline{\alpha(t)}$; that is by the law of large numbers we have

$$\lim_{t\to\infty} \frac{\alpha(t) - \overline{\alpha(t)}}{\overline{\alpha(t)}} = 0 \tag{2.82}$$

with probability one. This suggests that a first-order approximation to the stochastic process is to *replace it by its average value as a function of time.* This amounts to what is known as the *fluid approximation* for queues in which the discontinuous *stochastic* process $\alpha(t)$ is replaced by the

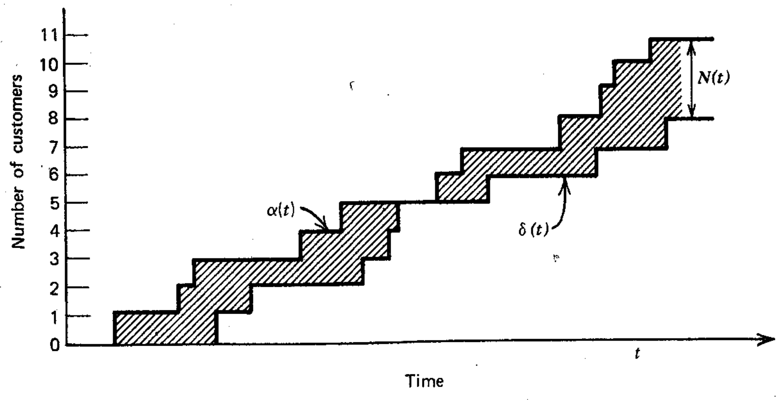

Figure 2.5 The discontinuous arrival and departure processes.

* We comment that usually it is in the case of large queues and long delays that the analysis of queueing systems is important; the case of small queues and delays is usually less interesting since typically they pose no serious problem to system performance.

continuous *deterministic* process $\overline{\alpha(t)}$. As we show below, there is considerable merit in this approach. Similarly we let the discontinuous stochastic departure process $\delta(t)$ be replaced by its mean value $\overline{\delta(t)}$. Consequently, if we assume that $N(0)=0$, the fluid approximation predicts that the number in system at time t must be given by

$$N(t)=\overline{\alpha(t)}-\overline{\delta(t)} \tag{2.83}$$

which also is a deterministic continuous function of time. The analogy with fluid flow is complete and may be thought of in terms of the following example [see Figure 2.6(*a*)]. Consider a funnel with an adjustable valve controlling the rate at which fluid may pass out of this funnel. We pour fluid in at the top at a rate $d\overline{\alpha(t)}/dt \triangleq \lambda(t)$ (the arrival rate of "customers") and we permit fluid to discharge at a rate $d\overline{\delta(t)}/dt \triangleq \mu(t)$ (the service rate for queues); of course the total amount discharged must never exceed the total amount fed in. In this case, then, we see that the total fluid accumulated in the funnel by time t will be given through Eq. (2.83). Thus we have

$$\overline{\alpha(t)}=\overline{\alpha(0)}+\int_0^t \lambda(y)\,dy$$

$$\overline{\delta(t)}=\overline{\delta(0)}+\int_0^t \mu(y)\,dy$$

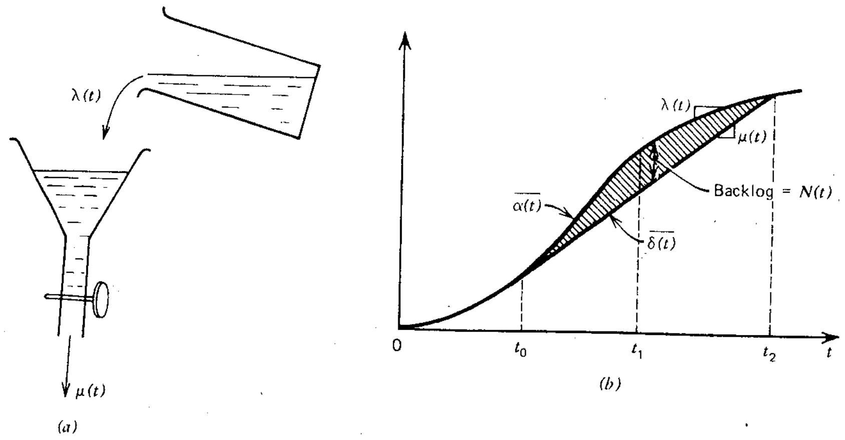

Figure 2.6 Fluid approximation to queues.

Consider, for the moment, the case where $\lambda(t)$ varies with time but where $\mu(t)$ is fixed* at μ; an example of this is shown in Figure 2.6(*b*). Here we see the case where $\lambda(t)$ increases from a small value at time 0 until it equals μ for the first time at t_0. At this instant the backlog begins to grow, reaching its maximum value at time t_1 when once again $\lambda(t)=\mu$; thereafter it decreases to 0 at time t_2. This simple approximation has a number of serious drawbacks. For example, we see that it claims that no queues form as the system approaches saturation from the left at time t_0; certainly we are aware that the size of the backlog at this time is strongly dependent on the manner in which $\lambda(t)$ approaches μ in the interval prior to t_0. As a result of this approximation, queue lengths may be badly underestimated.

In spite of the crudeness of this fluid approximation, it does lead to some worthwhile qualitative aspects of queueing behavior, which we now explore. Much of this material follows that of Newell [NEWE 71]. One of the important questions in the study of queueing theory is the way in which queues and delays grow during and after a "rush hour." The exact queueing analysis in these cases is abominably difficult even for the simplest assumptions for our stochastic processes. However, we can give a very gross picture through our fluid approximation (which will be refined in the following three sections). For example, consider Figure 2.7; here we show an idealized model of a rush-hour situation. In part (*a*) we show the arrival rate constant at one customer per second up until $t=2$; then the arrival rate rises linearly at the onset of the rush hour, levels off at a constant value for a short while, drops linearly at the end of the rush hour to its former value at which time it levels off again, and maintains that value. We show the service rate constant (see footnote below) at a value $\mu=2$. During the time interval $(2\frac{1}{2}, 7\frac{1}{2})$ we see that the system is overloaded in a serious way. In part (*b*) we show $\overline{\alpha(t)}$, the continuous arrival process, and $\overline{\delta(t)}$, the continuous departure process. The growth in the number of arrivals in the vicinity of the rush hour is evident [we assume $\alpha(0)=\delta(0)=0$]. Up until $t=2\frac{1}{2}$ we see that $\overline{\delta(t)}=\overline{\alpha(t)}$; however, for the next 5 units of time, the arrival rate exceeds the maximum departure rate and so the two curves separate forming a backlog $N(t)$. This backlog is shown in part (*c*) [on a scale twice as large as that in part (*b*)] and we observe that it grows quickly reaching its peak value at the instant when the arrival rate once again falls below the service rate at $t=7\frac{1}{2}$. We emphasize that at the "end" of the rush hour [when once again

* That is,

$$\mu(t)=\begin{cases}\mu & \text{for } N(t)>0\\ \lambda(t) & \text{for } N(t)=0\end{cases}$$

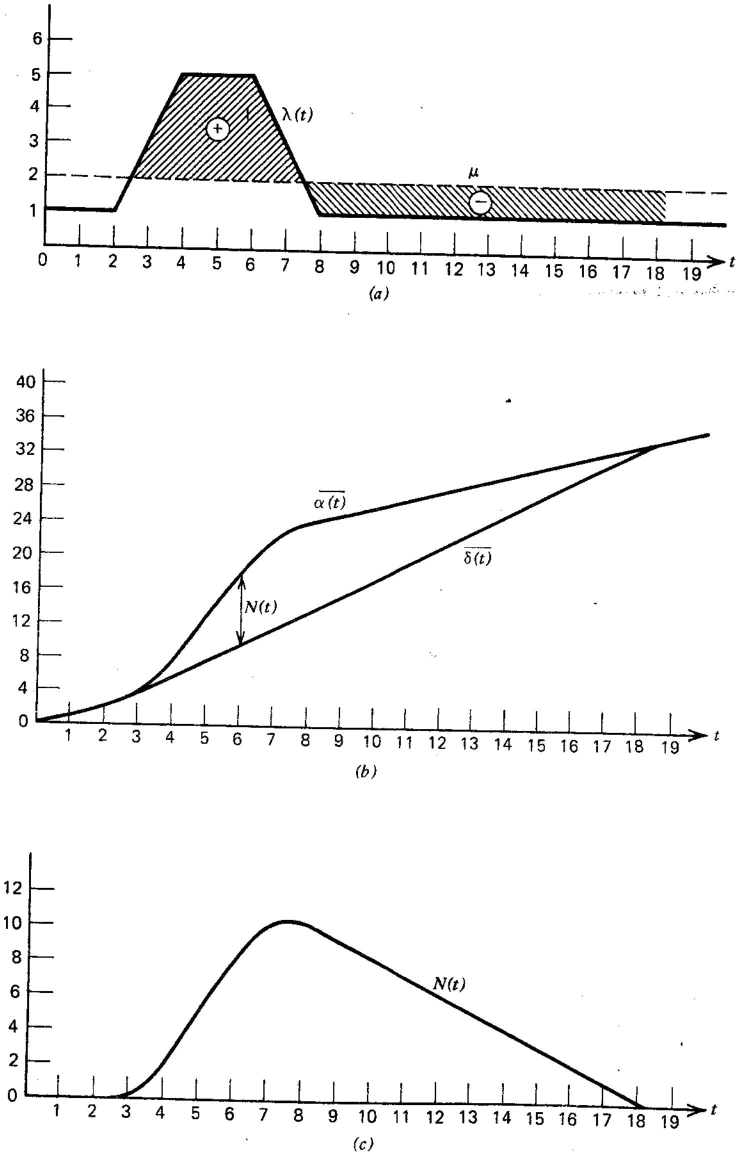

Figure 2.7 Fluid approximation to rush hour. (*a*) Rates of flow; (*b*) arrivals and departures; (*c*) number in system.

$\lambda(t) \le \mu$] we have merely reached the *peak* of the backlog and the effect of the rush hour will continue for a (possibly long) while. From the figure we see that it takes until $t = 18\frac{1}{4}$ before the backlog disappears! It is easy to see what is happening by referring back to part (*a*) where the cross-hatched area labeled ⊕ is equal to the "deficit" between service rate and arrival rate and therefore represents the total number of customers backlogging in our system; on the other hand, once the arrival rate drops below the departure rate we can make up this deficit with the excess capacity shown as the cross-hatched area labeled ⊖. Only when the total negative area equals the total positive area will our backlog drop to zero. If the nonrush-hour value for $\lambda(t)$ is only slightly less than the departure rate μ, we see it will take quite a while for us to make up the deficit; this produces the "long tail" on the backlog $N(t)$. Conversely, if the rate of accumulation of negative area is large compared to that for the positive area, then the backlog will fall off rather quickly. Of course, these comments apply to any arrival and departure process such as, for example, shown in Figure 2.8; here we assume that the backlog is 0 just prior to time t_1 but that at this time it begins to grow since the arrival rate exceeds the departure rate. $N(t)$ will grow as fast as positive area is accumulated in the figure, finally reaching its peak at t_2; it then begins to decline by an amount equal to the accumulated negative area reaching zero when the two areas are equal.

Perhaps now we understand why the freeways remain saturated so long after the close of business. We express the strong caution that although the fluid approximation correctly predicts the long tail of the rush-hour effect, the backlog we have shown is, if anything, optimistically low since we have not included queues that arise due to the *variability* in the arrival and departure processes; these may be large compared to the fluid effects. For example, in Figure 2.6 we have shown $\lambda(t)$ slowly approaching the departure rate μ and finally equaling it at time t_0; in such a case we recognize from our earlier queueing results that as we approach t_0 we are also approaching $\rho \to 1$, and we expect our queues to grow large in this vicinity. These queues arise due to the random nature of our input

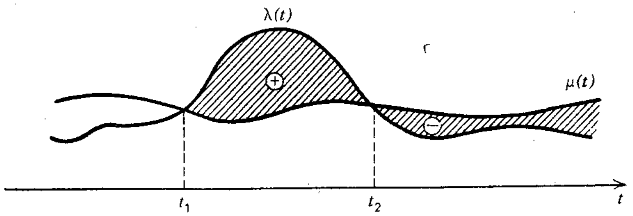

Figure 2.8 Making up the deficit.

processes. The fluid approximation (which we see is really a "continuous" D/D/1 approximation) assures us that there will be no backlog until after time t_0 and it is in this important sense that the approximation is deficient. In the next sections we improve our approximation to take account of such random effects.

2.8. DIFFUSION PROCESSES

In the previous section we used a first-order approximation for queues in which we replaced the arrival and departure processes by their mean values, thereby creating a deterministic continuous process—that is, the fluid flow approximation to queues. We realize, of course, that these processes are random in nature and in this section (and the following two sections) we improve that approximation by permitting $\alpha(t)$ and $\delta(t)$ to have *variations* about the mean. We do this by introducing the variances $\sigma^2_{\alpha(t)}$ and $\sigma^2_{\delta(t)}$ for the arrival and departure processes, respectively. A natural way for introducing these fluctuations about the mean value of our process is to represent these fluctuations by normal (Gaussian) distributions. We may justify this as follows. Observe that $\alpha(t)$ represents the total number of arrivals up to time t. If we ask for the probability that $\alpha(t) \geq n$, then that is the same as asking that customer C_n arrive at a time τ_n that occurs at or before t; that is,

$$P[\alpha(t) \geq n] = P[\tau_n \leq t] \tag{2.84}$$

This is an important equivalence and is used extensively when one considers ladder indices and combinatorial methods in queueing theory [PRAB 65]. Here, however, we take advantage of this equivalence in the following way: The arrival time of C_n is merely the sum of n interarrival times, that is $\tau_n = t_1 + t_2 + \cdots + t_n$ where we assume $\tau_0 = 0$. For G/G/1 we assume that the set $\{t_i\}$ is a set of independent identically distributed random variables [each with a distribution function $A(t)$]. When the time t, and therefore the number n, get large, then τ_n is the sum of a large number of independent (and identically distributed) random variables. Thus we expect that the central limit theorem should apply and permits us to describe the random variable τ_n and therefore also the random process $\alpha(t)$ as Gaussian functions. This assumption of normality for $\alpha(t)$ [and for $\delta(t)$] is the cornerstone of our diffusion approximation and its details are derived in this section.

For the diffusion approximation, we propose that the arrival process $\alpha(t)$ and the departure process $\delta(t)$ are both to be approximated by continuous random processes (with independent increments) which at time t are normally distributed with means $\alpha(t)$ and $\delta(t)$ and variances $\sigma^2_{\alpha(t)}$ and $\sigma^2_{\delta(t)}$, respectively. Once we determine these four parameters (the two means and two variances), then we will have completely described these two random processes since the Gaussian process with independent increments (yielding a trivial covariance function) is a two-parameter process (indeed, we have a Brownian motion process, i.e., integrated "white noise" with a nonzero mean [ITO 65]). As mentioned above, the variance terms are introduced in order to represent the random fluctuations of these processes about their means. We intend to use this approximation to make statements about the number of customers in the system $N(t)$ and the unfinished work in the system $U(t)$. As is well known [PAPO 65], if we have two independent normally distributed random processes, say $\alpha(t)$ and $\delta(t)$, then any linear combination of these two is also a normally distributed process (with some appropriate mean and variance). Of course one linear combination we are interested in is $\alpha(t) - \delta(t)$, which represents $N(t)$, the backlog expressed in number of customers. [We are also very much interested in the unfinished work, $U(t)$, which represents the backlog in units of time.] Indeed, we shall demonstrate in Eq. (2.132), below, that the backlog has a transient distribution that is the weighted difference of two Gaussian distributions. For $\rho < 1$, an equilibrium distribution exists which turns out to be the exponential distribution we obtained in Eq. (2.9) for the heavy-traffic approximation! For $\rho > 1$, of course, no equilibrium distribution exists; in this case, however, it turns out, for example, that the properly shifted and scaled waiting time w_n, namely,

$$\frac{w_n - n\bar{u}}{\sigma_{\bar{u}}\sqrt{n}}$$

satisfies the central limit theorem [KING 62b], permitting us to talk about this special kind of convergence.

When we attempt to take advantage of the linear combination $[N(t) = \alpha(t) - \delta(t)]$ of two independent Gaussian processes, a difficulty arises immediately: $\delta(t)$, the departure process, clearly is dependent upon the arrival process [that is, $\delta(t) \leq \alpha(t)$]. Fortunately, however, when $N(t) > 0$, then the departure process increases by one each time a service is completed and so the interdeparture times are distributed as service times, namely $B(x)$, independent of the arrival process [so long as $N(t) > 0$]. Thus when $N(t)$ is large, we have a departure process that is approximately independent of the arrival process, and it is this case which interests us. As a result we might expect that the approximation we are making is poor when the system is lightly loaded.

Thus we have the framework for our second-order approximation (the diffusion approximation) to our queueing system. Replacing $\alpha(t)$ by its mean $\overline{\alpha(t)}$ and its variance $\sigma^2_{\alpha(t)}$ is equivalent to making a Taylor expansion of this process about its mean value and throwing away all but the first two terms in this expansion [see Eq. (2.102)]. Let us now establish the relationship between the means and variances of our arrival and departure processes and the parameters of $A(t)$ and $B(x)$. We have already shown that $P[\alpha(t) \geq n] = P[\tau_n \leq t]$ for any t and n. Similarly if we let $X_n = x_1 + x_2 + \cdots + x_n$ represent the total time to service the first n customers then it is clear that $P[\delta(t) \geq n] = P[X_n \leq t]$, where we are assuming that the system never empties. Both the arrival and departure processes are similar in this sense and so the study of one gives us results for the other. Let us use the departure process in the following calculations. Given n, we have that

$$\bar{X}_n = n\bar{x}$$

$$\sigma_{X_n}{}^2 = n\sigma_b{}^2$$

using our former notation. Applying the central limit theorem we have that the normalized sum $(X_n - n\bar{x})/\sigma_b\sqrt{n}$ must be normally distributed with zero mean and unit variance as $n \to \infty$; that is, for $n \gg 1$,

$$P\left[\frac{X_n - n\bar{x}}{\sigma_b\sqrt{n}} \leq x\right] \cong \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy \tag{2.85}$$

where the right-hand side is defined as $\Phi(x)$. If we say X is $N(m, \sigma^2)$, that is shorthand for saying that X is a random variable with a normal distribution of mean m and variance σ^2. Thus we say that X_n is $N(n\bar{x}, n\sigma_b{}^2)$ and that $(X_n - n\bar{x})/\sigma_b\sqrt{n}$ is $N(0, 1)$; we see that this second random variable is a shifted and scaled version of X_n so as to give us a new normalized variable with zero mean and unit variance. We observe that X_n itself has a mean that grows linearly with n and a standard deviation that grows like $\sqrt{n}$; therefore the ratio of the standard deviation to the mean is a decreasing function of n. This implies that the fluctuations about the mean become insignificant with respect to the mean as n approaches infinity.

Now the event $(X_n - n\bar{x})/\sigma_b\sqrt{n} \leq x$ is the same as the event $X_n \leq x\sigma_b\sqrt{n} + n\bar{x}$. Let us define

$$t \triangleq x\sigma_b\sqrt{n} + n\bar{x} \tag{2.86}$$

By our former arguments we have

$$P[X_n \leq t] = P[\delta(t) \geq n] \to \text{Gaussian} \tag{2.87}$$

for large n. We are interested in showing that $\delta(t)$ is indeed a normally distributed random process; this last equation almost does that but expresses it in terms of a given quantity n, and we must now use the relationship between t and n in Eq. (2.86) to express this properly. We note for large n that the dominant term is $t \cong n\bar{x}$ and so as an approximation [with the correction term from Eq. (2.86)], we write

$$n \cong \frac{t}{\bar{x}} - x\left(\frac{\sigma_b}{\bar{x}}\right)\sqrt{\frac{t}{\bar{x}}} \tag{2.88}$$

Thus the event $\delta(t) \geq n$ may be written using this last approximation as

$$\frac{\delta(t) - (t/\bar{x})}{(\sigma_b/\bar{x})\sqrt{t/\bar{x}}} \gtrsim -x$$

If we apply this to Eq. (2.87) and use the symmetry given by $\Phi(x) = 1 - \Phi(-x)$ we find that

$$P\left[\frac{\delta(t) - (t/\bar{x})}{(\sigma_b/\bar{x})\sqrt{t/\bar{x}}} \leq x\right] \cong \Phi(x) \tag{2.89}$$

for large t. Thus we conclude that the departure process is a normal random variable with mean $t/\bar{x}$ and standard deviation $(\sigma_b\sqrt{t/\bar{x}})/\bar{x}$; that is, $\delta(t)$ is $N(t/\bar{x}, \sigma_b{}^2 t/(\bar{x})^3)$ for large t. This same result applies for our arrival process $\alpha(t)$ if the mean and variance of service time are replaced by the mean and variance of interarrival time; that is, $\alpha(t)$ is $N(t/\bar{t}, \sigma_a{}^2 t/(\bar{t})^3)$. Note that these are both approximations that are good only for large t and for moderate-to-heavily loaded queueing systems.

Thus we conclude that the number of customers in the system at time t, given by $N(t) = \alpha(t) - \delta(t)$, is also given as a normal random process whose mean is $\overline{N(t)} = \overline{\alpha(t)} - \overline{\delta(t)}$ and with variance $\sigma^2_{N(t)} = \sigma^2_{\alpha(t)} + \sigma^2_{\delta(t)}$ since variances for independent processes must add. We note that the mean number in system for this second-order approximation is exactly the result for the first-order approximation (the fluid approximation) in the previous section. Thus, with this approximation, we have shown for the case G/G/1 that the mean and variance of $N(t)$ are

$$\overline{N(t)} = \frac{t}{\bar{t}} - \frac{t}{\bar{x}} = \left(\frac{\rho - 1}{\bar{x}}\right) t \tag{2.90}$$

$$\sigma^2_{N(t)} = \left[\frac{\sigma_a{}^2}{(\bar{t})^3} + \frac{\sigma_b{}^2}{(\bar{x})^3}\right] t \tag{2.91}$$

We note that both the mean and variance grow linearly with time. However, for $\rho < 1$ our approximation shows that the mean number in system becomes negative; clearly we cannot tolerate this, and below we repair that defect by placing a "reflecting boundary" at the origin for $N(t)$. However for $\rho > 1$ we see that a reasonable approximation has developed where the mean number in system grows linearly with t and where the standard deviation of this number grows with $\sqrt{t}$ to represent the fluctuations about that growing mean.

Of course this approximation as a normal random process conceals the discrete nature of the arrival and departure processes. Nevertheless, in the case when $N(t)$ is large compared to unity, this approximation is useful. In order to gain more information about this diffusion approximation for the process $N(t)$, we will now study the partial differential equations for its probability distribution function. This will permit us to properly include the reflecting boundary at the origin as well as to make more explicit statements regarding transient and equilibrium behavior under this approximation. In addition to studying $N(t)$ we will also study $U(t)$, the unfinished work at time t, which will also be approximated as a normal random process. In particular we are interested in the way in which these random processes change during a small time interval; this time interval must be small enough so that the random process changes by a small fraction of its value but it must be large enough to permit enough discrete jumps to take place so that these two processes may be approximated by a continuum. We are thus led to the study of *continuous-time continuous-state Markov processes*. That is, we assume the processes to be Markovian in this time frame. We now launch into a derivation of the underlying partial differential equations for these continuous-time continuous-state Markov processes, and if the reader chooses to pass over this derivation he may do so and immediately proceed to Eq. (2.113), which is the key result of the following development.

In Section 1.3 we considered Markov processes that were continuous in time but with discrete state spaces and observed that the state-transition probabilities obeyed the Chapman–Kolmogorov equation given in Eq. (1.44). Since we intend to replace discrete and mixed random processes [that is, $N(t)$ and $U(t)$] with continuous ones we are naturally led to the consideration of a continuous-time continuous-state Markov process, which we denote by $X(t)$. In analogy with Eq. (1.43) for the conditional discrete-state transition probability, we consider the following conditional continuous-state transition probability:

$$F(x, t; y, \tau) \triangleq P[X(\tau) \leq y \mid X(t) = x] \qquad \text{for} \quad t < \tau \tag{2.92}$$

Thus $F = F(x, t; y, \tau)$ merely gives the probability that the process takes on a value $\leq y$ at time τ given that it took on the value x at time t. This (possibly time-dependent) transition probability obviously satisfies the following Chapman–Kolmogorov equation:

$$F(x, t; y, \tau) = \int_{-\infty}^{\infty} F(w, u; y, \tau)\, d_w F(x, t; w, u)$$

where $t < u < \tau$.

We now introduce certain assumptions on the stochastic process $X(t)$ that are related to its continuity (as suggested above); in particular, we assume

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x| \geq \varepsilon} d_y F(x, t - \Delta t; y, t) = 0 \tag{2.93}$$

for any $\varepsilon > 0$. Furthermore, we assume that $\partial F/\partial x$ and $\partial^2 F/\partial x^2$ exist and are continuous. We now introduce the conditional mean $M(x, t; \tau)$ and the conditional variance $V(x, t; \tau)$, where the condition is on the position (x) of the process at some previous time t $(t < \tau)$; these we define as follows:

$$M(x, t; \tau) \triangleq E[X(\tau) \mid X(t) = x] \tag{2.94}$$

$$V(x, t; \tau) \triangleq E[\{X(\tau) - M(x, t; \tau)\}^2 \mid X(t) = x] \tag{2.95}$$

We note that $M(x, t; t) = x$ and $V(x, t; t) = 0$. More interesting than this mean and variance are the *infinitesimal mean* $m(x, t)$ and *infinitesimal variance* $\sigma^2(x, t)$, which give the *rate of change* of these quantities with respect to τ at the point $\tau = t$, namely

$$m(x, t) \triangleq \left.\frac{\partial M(x, t; \tau)}{\partial \tau}\right|_{\tau = t} \tag{2.96}$$

$$\sigma^2(x, t) \triangleq \left.\frac{\partial V(x, t; \tau)}{\partial \tau}\right|_{\tau = t} \tag{2.97}$$

In these last two equations derivatives are taken for $\tau \geq t$; see Fig. 2.9. These infinitesimal quantities may be expressed in terms of the transition

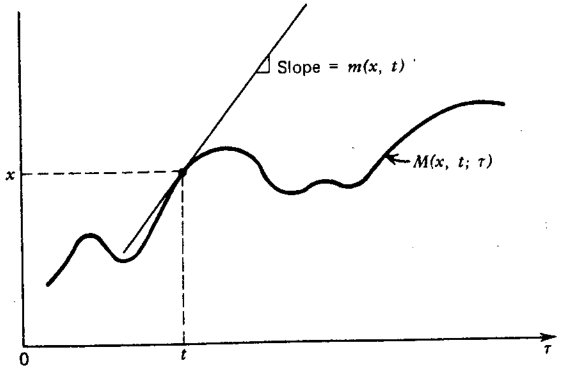

Figure 2.9 Relationship between the conditional mean and the infinitesimal mean.

probabilities as follows:

$$m(x, t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (y-x)\, d_y F(x, t-\Delta t; y, t) \tag{2.98}$$

$$\sigma^2(x, t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\infty}^{\infty} (y-x)^2\, d_y F(x, t-\Delta t; y, t) \geq 0 \tag{2.99}$$

Thus for small positive Δt we have the following approximation:

$$M(x, t; t+\Delta t) \cong x + m(x, t)\, \Delta t$$

$$V(x, t; t+\Delta t) \cong \sigma^2(x, t)\, \Delta t$$

With this description of a continuous Markov process we now wish to derive the backward and forward equations for F. We begin with the backward equation. From the Chapman–Kolmogorov equation, we have

$$F(x, t-\Delta t; y, \tau) = \int_{-\infty}^{\infty} F(w, t; y, \tau)\, d_w F(x, t-\Delta t; w, t)$$

In a trivial way we may write F in the form

$$F(x, t; y, \tau) = \int_{-\infty}^{\infty} F(x, t; y, \tau)\, d_w F(x, t-\Delta t; w, t)$$

since upon factoring F out of the integral we are left with the integral of a pdf that goes to unity. Subtracting these last two equations and dividing by Δt we have

$$\frac{F(x, t-\Delta t; y, \tau) - F(x, t; y, \tau)}{\Delta t}$$

$$= \frac{1}{\Delta t} \int_{-\infty}^{\infty} [F(w, t; y, \tau) - F(x, t; y, \tau)]\, d_w F(x, t-\Delta t; w, t) \tag{2.100}$$

Now the integral on the right-hand side may be broken into two integrals, the first over the region $|w-x| \geq \varepsilon$ for $\varepsilon > 0$, and the second over the region $|w-x| < \varepsilon$. By Eq. (2.93) the first of these two integrals vanishes as $\Delta t \to 0$ and we may replace the integrand in the second of these two integrals by the following Taylor expansion:

$$F(w, t; y, \tau) - F(x, t; y, \tau)$$

$$= (w-x) \frac{\partial F}{\partial x} + \frac{1}{2}(w-x)^2 \frac{\partial^2 F}{\partial x^2} + o((w-x)^2) \tag{2.101}$$

If we now substitute Eq. (2.101) into Eq. (2.100) and take the limit as $\Delta t \to 0$ then from Eqs. (2.98) and (2.99) we arrive at the following partial differential equation for F:

$$-\frac{\partial F}{\partial t} = m(x, t) \frac{\partial F}{\partial x} + \frac{1}{2} \sigma^2(x, t) \frac{\partial^2 F}{\partial x^2} \tag{2.102}$$

This is the backward Kolmogorov equation for our continuous-time continuous-state Markov process; F will satisfy this equation except at points of accumulation (such as at the origin, $y = 0$).

We now derive the diffusion equation (also known as the *Fokker–Planck* equation) that is the forward equation for our process. The function satisfying this equation is the pdf associated with the final state y, namely,

$$f(x, t; y, \tau) \triangleq \frac{\partial F(x, t; y, \tau)}{\partial y} \tag{2.103}$$

This density f also satisfies the Chapman–Kolmogorov equation

$$f(x, t; y, \tau) = \int_{-\infty}^{\infty} f(w, u; y, \tau) f(x, t; w, u)\, dw \tag{2.104}$$

where $t < u < \tau$. We now consider an arbitrary function $Q(y)$ that (along with its derivatives) vanishes rapidly enough at $\pm\infty$ for the integral I to converge, where

$$I \triangleq \int_{-\infty}^{\infty} Q(y) \frac{\partial f(x, t; y, \tau)}{\partial \tau}\, dy \tag{2.105}$$

Now from the definition of a derivative and from Eq. (2.104) we may rewrite I as

$$I = \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \int_{-\infty}^{\infty} Q(y)[f(x, t; y, \tau+\Delta\tau) - f(x, t; y, \tau)]\, dy$$

$$= \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \left[\int_{-\infty}^{\infty} Q(y) \int_{-\infty}^{\infty} f(x, t; w, \tau) f(w, \tau; y, \tau+\Delta\tau)\, dw\, dy \right.$$

$$\left. - \int_{-\infty}^{\infty} Q(w) f(x, t; w, \tau)\, dw \right] \tag{2.106}$$

We examine the double integral in this last equation; interchanging orders of integration and using a Taylor expansion for Q about w we may write this double integral (denoted by I_2) as

$$I_2 = \int_{-\infty}^{\infty} f(x, t; w, \tau) \left[\sum_{n=0}^{\infty} \frac{d^n Q(w)}{dw^n} \frac{1}{n!} \right.$$

$$\left. \times \int_{-\infty}^{\infty} f(w, \tau; y, \tau+\Delta\tau)(y-w)^n\, dy \right] dw \tag{2.107}$$

Now just as we defined the infinitesimal mean and variance in Eqs. (2.98) and (2.99) we now define the infinitesimal nth moments as

$$A_n(w, \tau) \triangleq \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \int_{-\infty}^{\infty} (y-w)^n f(w, \tau; y, \tau+\Delta\tau)\, dy \tag{2.108}$$

which are assumed to exist as finite quantities. Clearly $A_1(w,\tau)=m(w,\tau)$ and $A_2(w,\tau)=\sigma^2(w,\tau)$. We note that the term for $n=0$ in Eq. (2.107) involving $A_0(w,\tau)$ cancels the second (single) integral in Eq. (2.106). Now using the definition $A_n(w,\tau)$ and the expansion of I_2 we arrive at the following expression:

$$I=\sum_{n=1}^{\infty}\frac{1}{n!}\int_{-\infty}^{\infty} f(x,t;w,\tau)A_n(w,\tau)\frac{d^nQ(w)}{dw^n}\,dw$$

Let us integrate this last equation by parts (n times for the nth term) where $u=f(x,t;w,\tau)A_n(w,\tau)$ and $dv=(d^nQ(w)/dw^n)\,dw$. Since we have assumed that Q and all of its derivatives vanish rapidly enough at ±∞, the terms $uv|_{w=-\infty}^{w=+\infty}$ drop out in our integration by parts. Thus only the terms $-\int v\,du$ (and so on) remain, giving us

$$I=\int_{-\infty}^{\infty} Q(w)\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial w^n}[A_n(w,\tau)f(x,t;w,\tau)]\,dw \qquad (2.109)$$

Subtracting Eq. (2.109) from (2.105) we have

$$I-I=0=\int_{-\infty}^{\infty} Q(w)\left\{\frac{\partial f(x,t;w,\tau)}{\partial\tau} -\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial w^n}[A_n(w,\tau)f(x,t;w,\tau)]\right\}dw \qquad (2.110)$$

Now recall that $Q(w)$ was an arbitrary function, and so if Eq. (2.110) is to be satisfied, it must be that $f=f(x,t;w,\tau)$ satisfies

$$\frac{\partial f}{\partial\tau}=\sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial w^n}[A_n(w,\tau)f] \qquad (2.111)$$

Equation (2.111) holds for our continuous-time continuous-state Markov process, and is the basic equation for our process. It is from this equation that we may derive various of our approximations for queues. For example as we show below, we easily develop the fluid approximation by taking only the first term in this equation; that is, we assume $A_n(w,t)=0$ for $n=2,3,4,\ldots$, and this gives us

$$\frac{\partial f}{\partial t}=-\frac{\partial}{\partial w}[m(w,t)f] \qquad (2.112)$$

[recall that $A_1(w,t)=m(w,t)$], where now we have replaced τ by t for convenience. (Following the next paragraph, we study the solution of this equation for the case of constant arrival and departure rates.)

The approximation of more interest to this section is the second-order approximation in which we take the first two terms in our series to be nonzero and assume[*] $A_n(w,t)=0$ for $n=3,4,5,\ldots$, giving

$$\frac{\partial f}{\partial t}=-\frac{\partial}{\partial w}[m(w,t)f]+\frac{1}{2}\frac{\partial^2}{\partial w^2}[\sigma^2(w,t)f] \qquad (2.113)$$

[recall that $A_2(w,t)=\sigma^2(w,t)$]. This is known as a one-dimensional Fokker–Planck equation. Both equations (2.113) and (2.102) are referred to as *diffusion equations*, and the reader may observe that they are in fact the forward and backward Kolmogorov equations, respectively. Equation (2.113) is the one we use[†] in this section; it is satisfied except where f contains impulse functions. We comment that even were we to consider all terms in Eq. (2.111), this would still be only an approximation for queues since we have assumed that the underlying processes are continuous (which we know they are not)!

Let us begin simply by considering our deterministic fluid flow approximation to queues, namely Eq. (2.112). Here we see that only the infinitesimal mean of our process enters the picture, and this of course is equivalent to that in the previous section where we replaced our stochastic processes by their mean values. Since we studied the number in system $N(t)$ in that section, let us now take the unfinished work $U(t)$ as the related stochastic process of interest. Thus we consider

$$F(w_0,0;w,t)=P[U(t)\le w\mid U(0)=w_0]$$

which describes the time-dependent distribution of the unfinished work and where we have assumed an initial unfinished work of size w_0 at time $t=0$. For simplicity we will assume that the mean arrival rate $\lambda(t)=\lambda$ and that the average departure rate $\mu(t)=\mu$ (both constant in time); as a result, $m(w,t)$ is also constant and independent of both w and t. Now, $m(w,t)$ may be calculated (in terms of the system parameters) as the average net rate of work accumulating in the system (for $w>0$) as follows. Since we have on the average λ arrivals per second, and since each arrival carries with it an average unfinished work of magnitude $\bar{x}$ (the average service time), and since we assume that the service facility operates continuously and therefore clears work at a rate of 1 sec/second (this, too, is part of our approximation, namely, that the service facility

[*] The justification for neglecting these higher-order terms is that we expect the conditional pdf to be tightly concentrated around the value w. See Exercise 2.23 for a third-order approximation.

[†] Note that if $m(w,t)=m(t)$ and $\sigma(w,t)=\sigma(t)$, then these parameters may be moved outside the differential operators. This is the case where the arrival and service processes are independent of the backlog in the system, although they are permitted to vary with time.

never goes idle), then we find that $m(w, t)$ is a constant m given by

$$\begin{aligned} m(w,t)\,\Delta t \triangleq m\,\Delta t &= E[U(t+\Delta t)-U(t)\mid U(t)] \\ &= (\lambda\bar{x}-1)\,\Delta t \end{aligned}$$

and so

$$m = \rho - 1 \tag{2.114}$$

Now, not only f, but also F (with different boundary conditions) must satisfy Eq. (2.112), and so we are asked to solve

$$\frac{\partial F(w,t)}{\partial t} = (1-\rho)\frac{\partial F(w,t)}{\partial w} \qquad \blacksquare \tag{2.115}$$

where we have simplified our notation by suppressing the initial condition; that is, we write $F(w_0, 0; w, t) = F(w, t)$. In addition, we have the two natural boundary conditions, which are good for all t:

$$F(w,t) = 0 \qquad \text{for} \quad w<0 \tag{2.116}$$

$$F(\infty, t) = 1 \tag{2.117}$$

We have already assumed the initial condition, which states that the waiting time at $t=0$ is w_0 with probability 1, namely,

$$F(w,0) = F_0(w) \triangleq \begin{cases} 0 & w < w_0 \\ 1 & w \geq w_0 \end{cases} \tag{2.118}$$

The solution to Eq. (2.115) is an arbitrary function of $(w-mt)$, and once we apply the additional conditions stated above, we find the unique solution

$$F(w,t) = \begin{cases} F_0(w+(1-\rho)t) & w \geq 0 \\ 0 & w < 0 \end{cases} \tag{2.119}$$

If we examine this solution, we see for $\rho < 1$ that the unfinished work (namely, the virtual waiting time) begins with probability 1 at a value w_0 and decreases toward 0 at a rate $1-\rho$, and finally at time $t = w_0/(1-\rho)$ yields a zero backlog that persists forever (i.e., we force this boundary condition). On the other hand, for $\rho > 1$ we find that the backlog begins at a value w_0 and increases without bound at a rate of $\rho - 1$ sec/sec. Clearly, we have correctly described the transient behavior of deterministic queues as described in Section 2.7 above.

Let us temporarily leave behind the fluid approximation and proceed with an investigation of the diffusion approximation, which as we have seen is a second-order approximation including the mean and variance of the original process. (This corresponds to replacing the stochastic process with Brownian motion [ITO 65].) Once again the process we choose to look at is the time-dependent distribution of the unfinished work, that is, $F = F(w, t) = P[U(t) \leq w]$, where we have suppressed the initial condition and will introduce it only as needed. Equation (2.113) is the basic partial differential equation of motion that we must solve; we note that it is also satisfied by F with the appropriate boundary conditions. Again we are able to find the solution to this equation in the case when both the infinitesimal mean and the infinitesimal variance are independent of both w and t; therefore we assume

$$m(w,t) = m$$

$$\sigma^2(w,t) = \sigma^2$$

and we will assume that these are constants both in the original stochastic process $U(t)$ and in our diffusion approximation to it. We now have that F must satisfy the Fokker–Planck equation as follows:

$$\frac{\partial F}{\partial t} = -m\frac{\partial F}{\partial w} + \frac{1}{2}\sigma^2\frac{\partial^2 F}{\partial w^2} \qquad \blacksquare \tag{2.120}$$

We have already calculated the value for m as given in Eq. (2.114). For σ^2 we carry out the following computation:

$$\begin{aligned} \sigma^2\,\Delta t &\triangleq \operatorname{Var}[U(t+\Delta t)-U(t)\mid U(t)] \\ &= E\{[(U(t+\Delta t)-U(t))-m\,\Delta t]^2 \mid U(t)\} \\ &= E\{[U(t+\Delta t)-U(t)]^2 \mid U(t)\} - m^2(\Delta t)^2 \\ &= \lambda\,\Delta t\overline{x^2} + o(\Delta t) \end{aligned} \tag{2.121}$$

This last line comes from the fact that with probability $\lambda\Delta t$ (that is, the probability of an arrival) the second moment of the change in the unfinished work during $(t, t+\Delta t)$ will be $\overline{x^2}$ (namely, the second moment of service time). Thus we have

$$\begin{aligned} m &= \rho - 1 \\ \sigma^2 &= \lambda\overline{x^2} \end{aligned} \tag{2.122}$$

A general solution to Eq. (2.120) is

$$F(w,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(w-w_0-mt)/\sigma\sqrt{t}} e^{-x^2/2}\,dx = \Phi\left(\frac{w-w_0-mt}{\sigma\sqrt{t}}\right) \tag{2.123}$$

where $\Phi(x)$ is again the PDF for a standardized normal random variable as given in Eq. (2.85). However, this solution is unsatisfactory since it violates our boundary condition given in Eq. (2.116).

For the moment, let us simplify our task somewhat and ask merely for the *equilibrium* solution to Eq. (2.120) in the case when $\rho < 1$. That is, we seek

$$F(w) = \lim_{t\to\infty} F(w, t)$$

where F must satisfy Eqs. (2.116), (2.117), and (2.120). Clearly, the left-hand side of Eq. (2.120) goes to zero in this limiting case, and we see that the solution to the resulting first-order linear differential equation that also satisfies the boundary conditions is given by

$$F(w) = 1 - e^{2mw/\sigma^2} \qquad w \geq 0 \qquad \blacksquare \qquad (2.124)$$

How good is this solution? We expect it should be a fairly good approximation in the heavy-traffic case, for then the backlog is typically large, and so the assumption that the service facility never goes idle is a fair approximation; the continuity assumption on the backlog is then reasonable since the truly discontinuous jumps are small in magnitude compared to the backlog itself. We have, in fact, already made a calculation for the system G/G/1 in the heavy-traffic case, and the solution is given in Eq. (2.9); it is identical in form to the result given above in Eq. (2.124). In that earlier solution, the exponent had a value s_0 as given in Eq. (2.7), namely,

$$s_0 \cong \frac{-2\bar{t}(1-\rho)}{\sigma_a{}^2 + \sigma_b{}^2}$$

However, this result due to Kingman was for the waiting time, whereas Eq. (2.124) is for the unfinished work; these two quantities are the same in the case of a first-come–first-serve M/G/1 queue. In that case we have $\sigma_a{}^2 = 1/\lambda^2$ and $\bar{t} = 1/\lambda$; so

$$s_0 \cong \frac{-2(1-\rho)}{(1/\lambda) + \lambda\sigma_b{}^2} = \frac{-2(1-\rho)}{(1/\lambda) + \lambda\overline{x^2} - \lambda\bar{x}^2}$$

However, in the heavy-traffic case $\rho = \lambda\bar{x} \cong 1$, and so $1/\lambda - \lambda\bar{x}^2 \cong 1/\lambda - \bar{x} \cong 0$, which yields

$$s_0 \cong \frac{-2(1-\rho)}{\lambda\overline{x^2}} \qquad (2.125)$$

But, from Eqs. (2.122) and (2.124), we see that s_0 is approximately equal to the exponent $2m/\sigma^2 = 2(\rho - 1)/\lambda\overline{x^2}$. *Thus the diffusion approximation agrees with Kingman's heavy-traffic approximation to queues!*

Equation (2.124) gives us the diffusion approximation to the equilibrium distribution for the waiting time. An analogous result of course holds for the limiting distribution of $N(t)$, where now m and σ^2 must be calculated for the number in system rather than for the unfinished work in system; in particular, we see from Eq. (2.90) that $m = \bar{N}(t)/t = (\rho - 1)/\bar{x}$, and from Eq. (2.91) that $\sigma^2 = \sigma^2_{N(t)}/t = (C_a{}^2/\bar{t}) + (C_b{}^2/\bar{x})$. Then, we find that the continuous diffusion approximation $F(w)$ for the equilibrium distribution for number in system is given again in Eq. (2.124). As suggested by Kobayashi [KOBA 74a], we may discretize this to obtain an approximation $\hat{p}_k$ to the distribution for the number in system:

$$\begin{aligned}\hat{p}_k &= F(k+1) - F(k)\\ &= (1-\hat{\rho})(\hat{\rho})^k\end{aligned}$$

where

$$\hat{\rho} = e^{-2(1-\rho)/(\rho C_a{}^2 + C_b{}^2)}$$

Note that the solution for $\hat{p}_k$ reminds us of the M/M/1 solution given in Eq. (1.56) and, in fact, when $C_a = C_b = 1$ (M/M/1) then $\hat{\rho}$ is very close to ρ [KOBA 74a]. The predicted server utilization factor is $1 - \hat{p}_0 = \hat{\rho}$; however, we know from Eq. (1.26) that the exact value for the server utilization is ρ. From this observation, Kobayashi recommends an adjustment to $\hat{p}_k$ for $k = 0$, namely,

$$\hat{p}_k = \begin{cases} 1-\rho & k = 0\\ \rho(1-\hat{\rho})(\hat{\rho})^{k-1} & k \geq 1\end{cases}$$

In [REIS 74], it is shown that the error in the equilibrium mean number in system, $\bar{N}$, due to this approximation is small for M/G/1 systems with $C_b \cong 1$ and that it grows as C_b deviates from one; however, the relative error in $\bar{N}$ goes to zero as $\rho \to 1$. The modification at $k = 0$ represents one method for reducing the errors that are caused by the compromise we have made in placing a reflecting barrier at the origin. Another approach to this problem is given by Gelenbe [GELE 74]; he places an absorbing boundary at the origin which collects mass (i.e., probability), allows the mass to remain absorbed for an exponentially distributed amount of time (with parameter λ, the arrival rate), and after this time, the mass "jumps" to unity on the real line. The mass collected at the origin corresponds to the probability of an empty system, and the jump to unity (at rate λ) corresponds to an arrival (of one customer) to the system. The solution to this diffusion approximation for the distribution of number in an M/G/1 system is

$$p_k{}^* = \begin{cases} 1-\rho & k = 0\\ K_1\hat{\rho} & k = 1\\ K_2(\hat{\rho})^k & k \geq 2\end{cases}$$

where K_1 and K_2 are appropriate constants and $\hat{\rho}$ is as given earlier. It is interesting to note that $1 - p_0{}^* = \rho$ is the correct exact value for the server utilization. Also the mean number in the system, $\bar{N}$, predicted by this approximation differs from the known value [as given by the P-K mean value formula in Eq. (1.83)] by $\rho/2C_b^2$.

Of more interest to us is the time-dependent behavior of the mean wait. We have seen that Eq. (2.123) begins to yield the *transient* solution for the waiting time distribution, but, as we observed, it violates the boundary condition $F(w, t) = 0$ for $w < 0$; this boundary condition will be accounted for by the use of our reflecting boundary at the origin. Before proceeding, however, let us consider a convenient *scaling* transformation.

We consider once again the basic diffusion equation for the case of constant arrival and departure rates (that is, $\lambda(t) = \lambda$, $\mu(t) = \mu$), namely,

$$\frac{\partial F}{\partial t} = -m\frac{\partial F}{\partial w} + \tfrac{1}{2}\sigma^2 \frac{\partial^2 F}{\partial w^2} \qquad (2.126)$$

subject to the boundary equations (2.116) and (2.117) and the initial condition (2.118). Our objective is to render this equation dimensionless, solve for the pertinent system behavior, and then recover the appropriate performance measures in their original unscaled form. The basic transformation is to consider the following change of time variable and space variable:

$$t' = \frac{m^2}{\sigma^2} t \qquad (2.127)$$

$$w' = \frac{-m}{\sigma^2} w \qquad (2.128)$$

Thus we measure time in units of σ^2/m^2 and work in units of $-\sigma^2/m$ (recall that $m < 0$ for $\rho < 1$). The scaling operation produces the following dimensionless equation:

$$\frac{\partial F}{\partial t'} = \frac{\partial F}{\partial w'} + \frac{1}{2} \frac{\partial^2 F}{(\partial w')^2} \qquad \blacksquare \quad (2.129)$$

where now $F = F(w', t')$ is the distribution of $U'(t') \triangleq -(m/\sigma^2)U(m^2 t/\sigma^2)$. Once we solve this last equation we will have solved all equations of the form (2.126).

We will soon give the solution to this dimensionless diffusion equation. At this point, however, we may draw an important conclusion from the fact that our transformation given in Eqs. (2.127) and (2.128) did indeed yield a dimensionless equation independent of m and σ^2. Observe that the natural unit in which we should measure "significant" values for the unfinished work is $-\sigma^2/m$ and that the natural unit for measuring time is σ^2/m^2. That is, significant changes ($\approx -\sigma^2/m$) in $U(t)$ occur during natural time units ($\approx \sigma^2/m^2$). This leads us to the conclusion that the basic

"relaxation time" of the system is approximately

$$\text{Relaxation time} \cong \frac{\sigma^2}{m^2} = \frac{\lambda \overline{x^2}}{(1-\rho)^2} \qquad \blacksquare \quad (2.130)$$

A result of this form was noted early by Morse [MORS 55]. It is clear that when ρ is near unity, then the relaxation time may easily exceed the duration of a "rush hour" in practical problems. Also, note that this relaxation time is related to the average wait, W, in an M/G/1 queue as follows: $\sigma^2/m^2 = 2W/(1-\rho)$.

If we consider the equilibrium solution, $F(w') = \lim_{t' \to \infty} F(w', t')$ (assuming $m < 0$, that is, $\rho < 1$), then the dimensionless diffusion equation (2.129) yields

$$F(w') = 1 - e^{-2w'} \qquad w' \geq 0 \qquad (2.131)$$

with a mean value equal to one-half; this, of course, is the same as our previous result in Eq. (2.124).

Let us now return to the transient solution of Eq. (2.120). We saw earlier that $\Phi([w - w_0 - mt]/\sigma\sqrt{t})$ given in Eq. (2.123) satisfied this equation but unfortunately violated our boundary conditions. Let us denote this solution by $\alpha(w, t)$. Furthermore, we saw that the equilibrium solution to Eq. (2.120) was as given in Eq. (2.124). This leads us to consider the function $e^{2mw/\sigma^2}\alpha(w, t)$. If we let F take on this value, then Eq. (2.120) gives

$$\frac{\partial \alpha(w, t)}{\partial t} = m\frac{\partial \alpha(w, t)}{\partial w} + \frac{1}{2}\sigma^2 \frac{\partial^2 \alpha(w, t)}{\partial w^2}$$

But this is the same as Eq. (2.120) except for the sign of the first term on the right-hand side! This sign variation can be corrected by making the change of variable from w to $-w$. Thus we see that for any function $F(w, t)$ satisfying Eq. (2.120) there must correspond another solution of the form $e^{2mw/\sigma^2}F(-w, t)$. Now since the diffusion equation (2.120) is linear, it must be that any linear combination of these two solutions must also be a solution. In particular we wish to consider the combination $F(w, t) - e^{2mw/\sigma^2}F(-w, t)$. Now we are in a position to take advantage of our earlier solution $\alpha(w, t)$ in Eq. (2.123), which, when in this combination, quite fortunately also satisfies the previously violated boundary condition (2.116). Thus our time-dependent solution to the diffusion Eq. (2.120) is [NEWE 71]

$$F(w, t) = \Phi\left(\frac{w - w_0 - mt}{\sigma\sqrt{t}}\right) - e^{2mw/\sigma^2}\Phi\left(\frac{-w - w_0 - mt}{\sigma\sqrt{t}}\right) \qquad \blacksquare \quad (2.132)$$

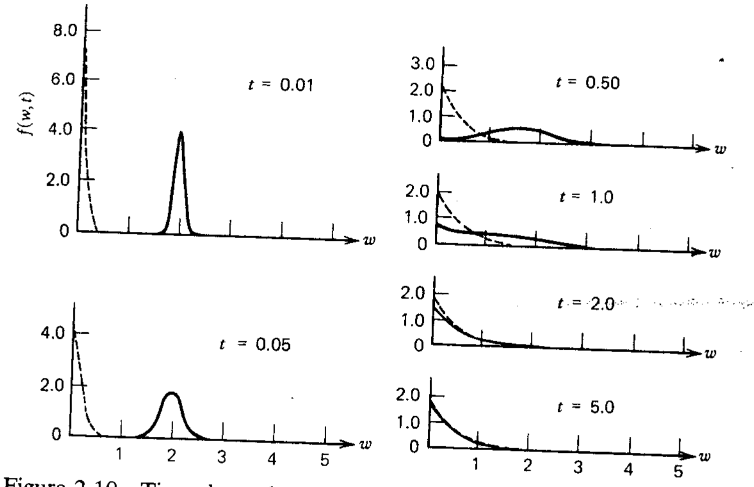

Figure 2.10 Time-dependent behavior of $f(w, t)$ for $\rho < 1$.

This solution corresponds to the case of constant-arrival-rate and constant-departure-rate processes and is good both for $\rho \leq 1$ and $\rho \geq 1$. (If w_0 is given in terms of a distribution function then our solution is obtained by integrating over this distribution.) Kobayashi [KOBA 74b] has evaluated the pdf $f(w, t) = \partial F(w, t)/\partial w$ for some examples of $\rho < 1$ (Figure 2.10) and the PDF $F(w, t)$ for $\rho > 1$ (Figure 2.11). In Figure 2.10 we see the approach of $f(w, t)$ to its exponential limit for $0 < t \leq 5$ with the two cases $w_0 = 0$ (dashed curves) and $w_0 = 2$ (solid curves). For $\rho > 1$, we see the unstable transient behavior of $F(w, t)$ in Figure 2.11; the same two cases are shown ($w_0 = 0$ as dashed curves and $w_0 = 2$ as solid curves) in the range $0 < t \leq 10$.

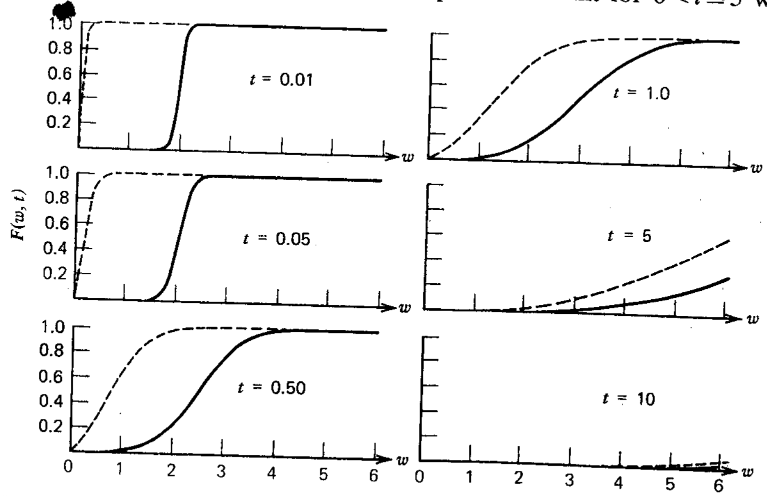

Figure 2.11 Time-dependent behavior of $F(w, t)$ for $\rho > 1$.

It is truly amazing that such a simple solution to our diffusion equation exists in this case since it must contain the elements of our fluid approximation, as well as our limiting equilibrium distribution, and must give the time-dependent solution for all values of ρ. We note that the first term in the solution involves a normal distribution with variance $\sigma^2 t$ and with mean $w_0 + mt$. For $\rho < 1$ $(m < 0)$ we see that this mean drifts to the left and corresponds to our earlier solution using the fluid approximation as given in Eq. (2.119); for $\rho > 1$ $(m > 0)$ we see that the mean drifts to the right again as in Eq. (2.119). The second term in our solution corresponds to a normal distribution drifting in a direction opposite to the first one but with an exponentially decreasing (increasing) weight for the case $\rho < (>) 1$. It is the second term that provides the reflection off the boundary at the origin. Moreover we note that as $t \to \infty$, and for $\rho < 1$ $(m < 0)$, both of the Φ functions go to unity leaving us with the equilibrium distribution as calculated in Eq. (2.124)! On the other hand, when $\rho > 1$ $(m > 0)$ we get behavior that does not settle down; in the following section we discuss this behavior in the context of M/G/1. The reader is referred to the excellent monograph by Newell [NEWE 71] for considerably more discussion of these matters. For now we wish to apply our diffusion approximation to M/G/1.

2.9. DIFFUSION APPROXIMATION FOR M/G/1 [GAVE 68]

In this section we study the time-dependent behavior of the unfinished work $U(t)$ for the first-come-first-serve M/G/1 system. In this system, $U(t)$ has a distribution that is the same as the waiting time distribution $W(y)$. We will use the continuous diffusion process as an approximation to our random sawtooth process $U(t)$, and in order to distinguish the approximation from the true process, we denote the former by $U_d(t)$. In the case when ρ is close to unity this is a good approximation. Much of this material is based upon the work of Gaver.

The general solution we obtained in Eq. (2.132) in the preceding section certainly provides the solution for the queue M/G/1. Nevertheless in this section we choose to study the behavior of M/G/1 using *transform* techniques for two reasons: first, because it gives an alternative approach to the solution and moreover provides additional insight into that solution; secondly, so that we may compare it with the exact time-dependent solution for M/G/1, which is given only in terms of transforms. Much of the algebra and "routine" development of these results is relegated to Exercise 2.21.

Using the notation and results developed in the previous section we therefore proceed to examine the behavior of the time-dependent distribution of waiting time $F(w, t)$ where F and its associated initial and boundary conditions are repeated here:

$$F(w, t) = P[U_d(t) \le w \mid U_d(0) = w_0]$$

$$F(w, 0) = \begin{cases} 0 & w < w_0 \\ 1 & w \ge w_0 \end{cases}$$

$$F(w, t) = 0 \quad \text{for} \quad w < 0$$

$$F(\infty, t) = 1$$

Also, as above, the infinitesimal mean m and variance σ^2 are independent of both w and t, and have values

$$m = \rho - 1$$

$$\sigma^2 = \lambda \overline{x^2}$$

where

$$\rho = \lambda \bar{x}$$

Furthermore, except where F takes jumps, this diffusion process must satisfy the Fokker-Planck equation, namely,

$$\frac{\partial F}{\partial t} = -m \frac{\partial F}{\partial w} + \tfrac{1}{2}\sigma^2 \frac{\partial^2 F}{\partial w^2} \tag{2.133}$$

In order to extract the behavior of F we find it convenient to use transform methods much as we did in Section 1.7. Therefore, as earlier, we define the double Laplace transform (on both w and t) as

$$F^{**}(r, s) \triangleq \int_0^\infty e^{-st} \int_{0^-}^\infty e^{-rw}\, d_w F(w, t)\, dt \tag{2.134}$$

which we assume exists certainly for Re $(s) > 0$ and Re $(r) > 0$. We must now apply this double transform to our partial differential equation (2.133). As developed in Exercise 2.21 we arrive at the following expression for this transform:

$$F^{**}(r, s) = \frac{2}{\sigma^2}\left[\frac{(r/\eta)e^{-\eta w_0} - e^{-rw_0}}{(r - r_1)(r - r_2)}\right] \tag{2.135}$$

where

$$r_1, r_2 = \frac{m}{\sigma^2}\left[1 \pm \left(1 + \frac{2s\sigma^2}{m^2}\right)^{1/2}\right] \tag{2.136}$$

and where r_1 takes the positive square root and r_2, the negative; also

$$\eta = \begin{cases} r_2 & \rho < 1 \\ r_1 & \rho > 1 \end{cases} \tag{2.137}$$

It is worthwhile to compare $F^{**}(r, s)$, which is the solution for our diffusion process, with the analogous result for the exact M/G/1 queueing system. We denote the latter by $F^{**}_{M/G/1}(r, s)$, which is defined as

$$F^{**}_{M/G/1}(r, s) \triangleq \int_0^\infty e^{-st} E[e^{-rU(t)} \mid U(0) = w_0]\, dt \tag{2.138}$$

This result was given in Section 1.7 as Eq. (1.97) and takes the form

$$F^{**}_{M/G/1}(r, s) = \frac{(r/\eta)e^{-\eta w_0} - e^{-rw_0}}{\lambda B^*(r) - \lambda + r - s} \tag{2.139}$$

where now η is the positive real root of the denominator. We notice the remarkable similarity between the solution for the original discrete stochastic process in Eq. (2.139) and the solution to the diffusion approximation given in Eq. (2.135).

We are now interested in inverting $F^{**}(r, s)$ on the transform variable r in order to obtain the time-transformed density for $U(t)$; this we define as

$$F^{\bullet *}(w, s) \triangleq \int_0^\infty e^{-st} \frac{\partial F(w, t)}{\partial w}\, dt \tag{2.140}$$

In Exercise 2.21 we show that this leads to

$$sF^{\bullet *}(w, s) = \begin{cases} -r_1 e^{r_1 w} & r_1 < 0 \quad \rho < 1 \\ -r_2 e^{r_2 w} & r_2 < 0 \quad \rho > 1 \end{cases} \tag{2.141}$$

where we have assumed temporarily that $w_0 = 0$. The reader should note that the dependence of this last equation upon s is through the value of the root r_i ($i = 1$ or 2) as given in Eq. (2.136).

Let us now comment on these last results. We begin by recalling from the method of collective marks (see Section 1.7) that the Laplace transform of a pdf evaluated at some (real) value s is equal to the probability that no Poisson-generated catastrophe will occur (where catastrophes occur at a rate of s per second) during a time interval whose duration is a random variable chosen from the pdf. In a similar fashion it is easily seen that the quantity $sF^{\bullet *}(w, s)$ in Eq. (2.141) may be interpreted as the pdf for the state of our diffusion process if it is observed *at the instant* of a catastrophe where catastrophes occur at the rate of s per second; in particular, if $w_0 = 0$ then the state (value) of our diffusion process $U_d(t)$ at the instant of a catastrophe has an exponential density as given by the right-hand side of Eq. (2.141). In the case when $w_0 > 0$ then a similar statement may be made which provides an explicit expression for the density of our diffusion process where it can be shown that this pdf will be a linear combination of exponentials. [We see from Eq. (2.139) that the solution to our exact process, namely the distribution of $U(t)$, enjoys no such simple interpretation.] Now we permit the rate of catastrophes to approach zero; therefore our random observation time (i.e., the time of occurence of a catastrophe) approaches infinity. This implies that $s \to 0$ and so we are interested in

$$\lim_{s\to 0} sF^{**}(w, s) = \lim_{t\to\infty} \frac{\partial}{\partial w} P[U_d(t) \le w \mid U_d(0) = w_0]$$

which is nothing more than the final value theorem for Laplace transforms [KLEI 75]. In Exercise 2.21 we find for $\rho < 1$ that this leads to

$$\lim_{s\to 0} sF^{**}(w, s) = -\frac{2m}{\sigma^2} e^{2mw/\sigma^2} \qquad (\rho < 1) \tag{2.142}$$

which is the pdf corresponding to the equilibrium solution we found earlier in Eq. (2.124) as of course it must. (For $\rho > 1$ we find that the limit is zero indicating that no equilibrium solution exists.) Substituting in the values for m and σ^2 we see that the equilibrium distribution for the diffusion approximation to the unfinished work is given by

$$\lim_{t\to\infty} P[U_d(t) \le w \mid U_d(0) = w_0] = 1 - e^{-2(1-\rho)w/\lambda\overline{x^2}} \tag{2.143}$$

so long as $\rho < 1$ [Eq. (2.124) again]. This distribution of course is independent of w_0 and corresponds to the diffusion approximation to the equilibrium distribution of waiting time for the stable M/G/1 system under a first-come–first-serve queueing discipline.

From Eq. (2.143) we immediately recognize that the mean unfinished work $E[U_d(t)]$ for our diffusion process, which also represents our approximation to the mean waiting time, has a limit for $\rho < 1$,

$$\lim_{t\to\infty} E[U_d(t)] = \frac{\lambda\overline{x^2}}{2(1-\rho)} \tag{2.144}$$

and this is exactly the P-K formula for the mean wait W in M/G/1 as given in Eq. (1.82). Thus the limiting mean wait for our diffusion process is identical to the limiting mean wait for the exact process (for all $\rho < 1$)! The limiting value of the variance for the wait in our diffusion process, which we denote by $\sigma_{U_d}{}^2$ is easily calculated from the equilibrium distribution in Eq. (2.143) as

$$\sigma_{U_d}{}^2 = \left[\frac{\lambda\overline{x^2}}{2(1-\rho)}\right]^2 \tag{2.145}$$

However, from Eq. (1.87) we have that the variance of the equilibrium waiting time, which we denote by $\sigma_U{}^2$, for our exact process is given by

$$\sigma_U{}^2 = \left[\frac{\lambda\overline{x^2}}{2(1-\rho)}\right]^2 + \frac{\lambda\overline{x^3}}{3(1-\rho)} \tag{2.146}$$

Now as $\rho \to 1$ we see that $\sigma_U{}^2 \to \sigma_{U_d}{}^2$. Thus our diffusion approximation gives an exact answer for the limiting *mean* wait and an answer for the *variance* of this wait that improves (to perfection) as ρ approaches 1 from below.

Let us now make use of our results to examine the *time-dependent* behavior of the mean waiting time, namely $E[U_d(t) \mid U_d(0) = w_0]$. To obtain this expression we first look at its Laplace transform,

$$\int_0^\infty e^{-st} E[U_d(t) \mid U_d(0) = w_0]\, dt$$

From the definition given in Eq. (2.134) and from the moment-generating properties of Laplace transforms we see that the expression we are looking for is obtainable from $-\partial F^{**}(r, s)/\partial r\,|_{r=0}$. Thus taking the partial differential with respect to r in Eq. (2.135) and then setting $r = 0$ we obtain (see Exercise 2.21)

$$\int_0^\infty e^{-st} E[U_d(t) \mid U_d(0) = w_0]\, dt = \frac{m}{s^2} + \frac{w_0}{s} + \frac{e^{-\eta w_0}}{s\eta} \tag{2.147}$$

Let us study Eq. (2.147) by first considering the case $\rho < 1$. We already know that as $t \to \infty$ the *equilibrium* mean waiting time is given through Eq. (2.144). The time-dependent behavior in this case is obtained through Eq. (2.147). For the case $\rho < 1$ and $w_0 = 0$, we have from Exercise 2.21 that

$$\begin{aligned}\int_0^\infty e^{-st} E[U_d(t) \mid U_d(0) = 0]\, dt &= -\frac{1}{r_1 s} \\ &= -\frac{\sigma^2}{sm\{1 + [1 + (2\sigma^2/m^2)s]^{1/2}\}}\end{aligned} \tag{2.148}$$

If we invert Eq. (2.148) we will in fact obtain the time-dependent behavior of the mean waiting time in our diffusion approximation to the system M/G/1. It is appropriate to compare this approximation to exact results calculable for M/G/1 by inverting Eq. (2.139). Gaver [GAVE 68] has made this comparison and we reproduce his numerical results in the examples below. In these examples the Poisson arrival rate is taken at $\lambda = 0.95$ customers per minute and the expected service time is $\bar{x} = 1.0$

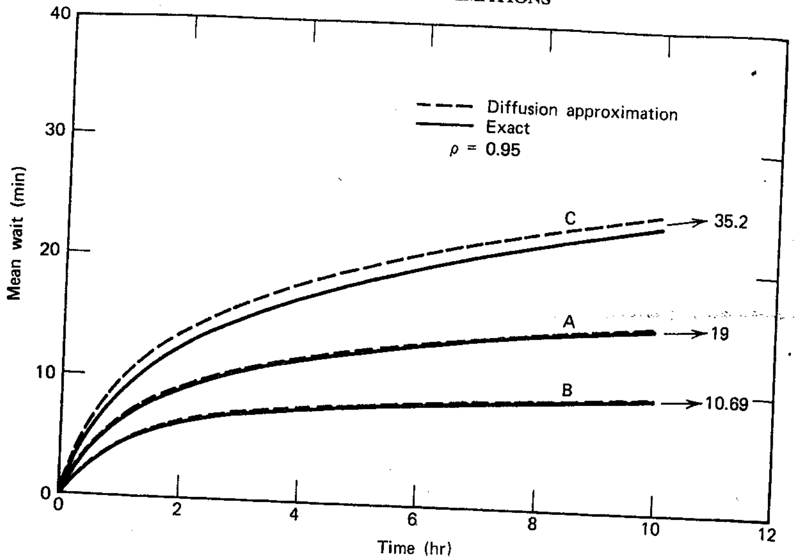

Figure 2.12 Comparison of mean wait versus elapsed time for exact and diffusion analyses: $\rho < 1$.

minute; therefore $\rho = 0.95$. Specifically, the three cases are

$$\text{A.} \qquad b(x) = e^{-x} \qquad x \geq 0 \tag{2.149}$$

$$\text{B.} \qquad b(x) = \frac{8(8x)^7}{7!} e^{-8x} \qquad x \geq 0 \tag{2.150}$$

$$\text{C.} \qquad b(x) = 0.1\left[\frac{1}{4} e^{-x/4}\right] + 0.9[(6x)^3 e^{-6x}] \qquad x \geq 0 \tag{2.151}$$

That is, A is exponential, B is Erlang distributed (with eight stages), and C is a mixture of an exponential (mean $= 4$) and a four-stage Erlang (mean $= \frac{2}{3}$). In Figure 2.12 we plot the mean wait in minutes as a function of time in hours for the diffusion approximation and for the exact result for the three cases. We also show the asymptotic value of the mean wait, which, as we have demonstrated above, coincides for the diffusion approximation and the exact analysis. In this figure we have taken $w_0 = 0$. We note how excellent the diffusion approximation is for the case $\rho < 1$. We observe once again for these M/G/1 systems that the approach to the equilibrium waiting time is rather slow [that is, from Eq. (2.130) we see that the relaxation time is 12.67, 7.13 and 23.43 hours for cases A, B, C, respectively].

Let us now consider the time-dependent behavior of the mean waiting time for the case $\rho > 1$. Temporarily we will consider once again the more general case of arbitrary w_0. For $\rho > 1$ we have from Eq. (2.137) that $\eta = r_1$. Thus from Eq. (2.147) we seek to invert

$$\frac{m}{s^2} + \frac{w_0}{s} + \frac{e^{-r_1 w_0}}{s r_1}$$

In order to simplify our task we will study the asymptotic behavior as $t \to \infty$ by permitting $s \to 0$; this allows us to make the replacement $r_1 = 2m/\sigma^2$ from Exercise 2.21. Thus inverting our expression we get the asymptotic ($t \to \infty$) time-dependent behavior for the mean wait conditioned on an initial wait of w_0, namely,

$$E[U_d(t) \mid U_d(0) = w_0] \to (\rho - 1)t + w_0 + \frac{\lambda\overline{x^2} e^{-2(\rho-1)w_0/\lambda\overline{x^2}}}{2(\rho - 1)} \qquad \blacksquare \tag{2.152}$$

This result demonstrates the linear growth of the mean wait predicted by the fluid approximation for the case $\rho > 1$ and large t, and provides an interpretation for the effect of the second term in Eq. (2.129) on the mean wait when $\rho > 1$. As we did for the case $\rho < 1$ we wish to provide some examples for this case to compare exact time-dependent behavior with our diffusion approximation to that behavior. Again we choose $w_0 = 0$, and therefore Eq. (2.152) becomes

$$E[U_d(t) \mid U_d(0) = 0] \to (\rho - 1)t + \frac{\lambda\overline{x^2}}{2(\rho - 1)} \tag{2.153}$$

Observe that this approximation includes the effect of the variance of the service time, an effect often omitted in such approximations (as for example in [COX 61, p. 66]). The examples here once again come from Gaver [GAVE 68], and we consider the case $\lambda = 1.1$, $\bar{x} = 1.0$, and therefore $\rho = 1.1$. In Figure 2.13 we show the time-dependent mean waiting time (in minutes) versus time (in hours) for cases A and B from our previous examples [see Eqs. (2.149) and (2.150)] and compare the exact results from Eq. (2.139) with the diffusion approximation given in Eq. (2.153). Once again we note the excellent approximation provided by our diffusion process. Note that cases A and B give slightly different results due to the variance term in Eq. (2.153).

Let us return once again to the time-dependent mean waiting time conditioned on an initial wait of zero for the case $\rho < 1$; that is, we are interested in the expression $E[U_d(t) \mid U_d(0) = 0]$ whose transform is given

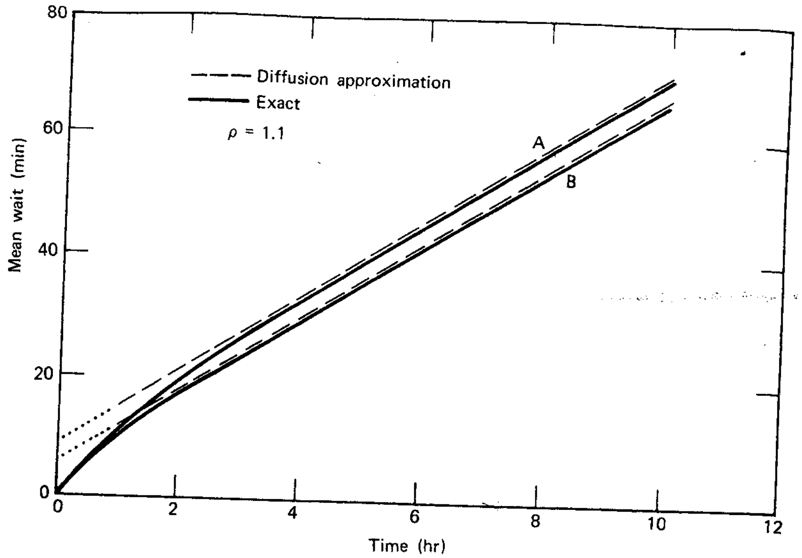

Figure 2.13 Comparison of mean wait versus elapsed time for exact and diffusion analyses: $\rho > 1$.

in Eq. (2.148). In particular we now wish to give a dimensionless form for this equation. Thus we will scale our time and unfinished work functions according to Eqs. (2.127) and (2.128), respectively, and in addition will scale the transform variable s (whose dimensions are 1/sec) as follows

$$s' \triangleq \frac{\sigma^2}{m^2} s \tag{2.154}$$

Thus we may rewrite Eq. (2.148) involving only properly scaled quantities as

$$\int_0^\infty e^{-s't'} E[U_d'(t') \mid U_d'(0) = 0] \, dt' = \frac{1}{s'[1 + \sqrt{1 + 2s'}]} \tag{2.155}$$

Using the final value theorem as earlier, we obtain the limiting expression for the mean wait in system,

$$\lim_{t \to \infty} E[U_d'(t') \mid U_d'(0) = 0] = \lim_{s' \to 0} s' \frac{1}{s'[1 + \sqrt{1 + 2s'}]} = \frac{1}{2} \tag{2.156}$$

Thus, as in Eq. (2.131), we see again that the equilibrium mean wait is equal to one-half (in scaled time units). If we invert expression (2.155) we obtain the time-dependent mean wait for our scaled diffusion approximation to the M/G/1 system as given by (see Exercise 2.22)

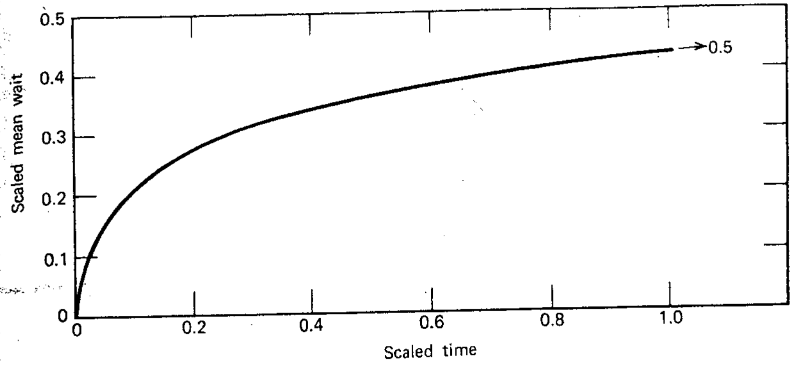

Figure 2.14 Mean wait in the scaled diffusion approximation to M/G/1 ($\rho < 1$).

$$E[U_d'(t') \mid U_d'(0) = 0] = \left(1 + \frac{t'}{2}\right)[2\Phi(\sqrt{t'}) - 1] - \frac{t'}{2} - \frac{1}{2} P\left(\frac{3}{2}, \frac{t'}{2}\right) \quad \blacksquare \tag{2.157}$$

where $\Phi(x)$ is given in Eq. (2.85) and $P(a, x)$ is the incomplete gamma function defined as

$$P(a, x) \triangleq \frac{1}{\Gamma(a)} \int_0^x e^{-y} y^{a-1} \, dy \tag{2.158}$$

and $\Gamma(x)$ is the usual gamma function (see, for example, [ABRA 64]). This scaled mean wait is plotted in Figure 2.14. This figure gives a *universal* curve for the diffusion approximation to the scaled mean wait in the queue M/G/1; from it, we could have obtained the diffusion approximations in Figure 2.12.

A compact and interesting discussion of these and other asymptotic relations is given by Cohen [COHE 73]; for example, he studies further details regarding the approach to equilibrium.

This ends our specific investigation of the diffusion approximation for the stationary M/G/1 queue. In the following section, we return to the case where the arrival process and departure process are permitted to vary with time.

2.10. THE RUSH-HOUR APPROXIMATION [NEWE 68, 71]

In applying the diffusion equation (2.113) we have so far emphasized the case where the infinitesimal mean and infinitesimal variance are both constant, that is, $m(w, t) = m$ and $\sigma^2(w, t) = \sigma^2$. Of course, this implies that the arrival and departure processes have average rates that are not time dependent, that is, $\lambda(t)=\lambda$, $\mu(t)=\mu$, and similarly for their variances. Moreover, from our previous studies, we know that in the case $\rho(=\lambda/\mu)<1$, an equilibrium distribution will exist; much of Volume I [KLEI 75] was devoted to a discussion of this equilibrium behavior. On the other hand, most interesting queueing systems are not stationary in time, and it is these we wish to discuss more fully in this section.

In particular, if we assume that the infinitesimal mean and variance are functions only of t, but not of w, then we may rewrite the Fokker-Planck equation as

$$\frac{\partial f}{\partial t}=-m(t)\frac{\partial f}{\partial w}+\frac{\sigma^2(t)}{2}\frac{\partial^2 f}{\partial w^2} \tag{2.159}$$

where we have set $m(w,t)=m(t)$ and $\sigma^2(w,t)=\sigma^2(t)$. We have already seen an "analysis" for time-varying arrival and departure rates if we are willing to accept the fluid approximation of Section 2.7; however, the diffusion equation (2.159) permits us to include the effect of fluctuations about the mean with time-varying rates. Let us begin by estimating the queueing behavior for some extreme cases that fall into the categories we have so far analyzed. First, if $m(t)<0$ (that is, the departure rate can keep up with the arrival rate) and if the variation in this infinitesimal mean is slow compared to $[\sigma(t)/m(t)]^2$, the relaxation time of our system [see Eq. (2.130)], then we expect that the queues and waiting times will "follow" this slow variation such that over a small number of relaxation times the system appears "stationary"; we shall refer to such a situation as being "quasistationary." The quasistationary distribution for the unfinished work is given as

$$F(w,t)=1-e^{2m(t)w/\sigma^2(t)} \qquad w\geq 0 \tag{2.160}$$

which is the natural extension of our former equilibrium solution in Eq. (2.124). Recall the definitions $m(t)=[\lambda(t)/\mu(t)]-1$ and $\sigma^2(t)=\lambda(t)\overline{x^2(t)}$ [where $\overline{x^2(t)}$ is the second moment of the service time duration at the instant t]. Note that the relaxation time is a sensitive function of

$$\rho(t)=\frac{\lambda(t)}{\mu(t)}$$

To be a bit more precise, the quasistationary result in Eq. (2.160) will hold if no significant changes in the waiting time [as measured on the scale in Eq. (2.128), that is, changes in unfinished work of approximately $-\sigma^2(t)/m(t)$ sec] occur in intervals on the order of a relaxation time; thus our condition for quasistationarity becomes

$$\frac{|d\bar{U}(t)/dt|\,(\text{relaxation time})}{\bar{U}(t)}\ll 1 \tag{2.161}$$

where $\bar{U}(t)$ is the average unfinished work at time t. From Eq. (2.160) we see that $\bar{U}(t)=-\sigma^2(t)/2m(t)$, which confirms our earlier choice of scale in Eq. (2.128). Thus our condition becomes

$$\left|2\sigma(t)\frac{d\sigma(t)}{dt}-\frac{\sigma^2(t)}{m(t)}\frac{dm(t)}{dt}\right|[m^2(t)]^{-1}\ll 1 \tag{2.162}$$

The interesting behavior of course is when $\rho(t)$ is close to 1 [that is, $m(t)$ close to zero], in which case the first term is usually insignificant compared to the other; using our expression for $m(t)$ the condition therefore becomes

$$\frac{\sigma^2(t)}{[1-\rho(t)]^3}\left|\frac{d\rho(t)}{dt}\right|\ll 1 \tag{2.163}$$

We note for $\rho(t)$ close to unity that the left-hand side grows arbitrarily large, and so we can readily imagine situations in which this condition will not hold. Then of course our quasistationary solution will not describe the true picture; in fact the solution for $\bar{U}(t)$ given above for this case predicts that the average waiting time will grow to infinity at an enormous rate as $\rho(t)\to 1$. We know this cannot be the case since *the waiting time can grow no faster than the rate at which work enters the system* and this rate is finite (in spite of the fact that the system is overloaded).

Thus we see that as long as the time variations are slow and small, our quasistationary solution Eq. (2.160) will approximately describe the waiting time distribution in the case $\rho(t)<1$. However, as ρ approaches and then perhaps exceeds 1, we find that the actual waiting time cannot grow as fast as the quasistationary solution would predict. Second, we observe that for stable queueing systems most of the delays arise because of the stochastic effects of the variability in the arrival and service processes. Recall from Chapter 1 that as long as the input rate is less than the capacity, no backlog should form when the flow is steady; thus, the backlog we see with unsteady flow is caused by the "random effects." On the other hand, from Chapter 1 we see that if the input rate exceeds the capacity ($\rho>1$) then a huge backlog will develop in time and in such a situation *the stochastic effects become unimportant*! We might therefore expect that the fluid approximation given in Section 2.7 should describe the major part of the growth of delays and of queues for $\rho>1$. In some sense we have already anticipated this result for the constant-parameter solution in Eq. (2.132) when $\rho>1$. There we saw that the waiting time behaves as a normal distribution with a linearly growing mean and with a standard deviation that grows only as $\sqrt{t}$; thus the dominant effect is given by the mean with the fluctuations about that mean reducing in relative size. In the case of nonconstant parameters such as we are considering in this section, a similar statement can be made for queues that have been saturated for some time. Thus we can estimate the system behaviór at both extremes [$\rho(t) \cong \rho < 1$ and $\rho(t) > 1$], and it is our intention in this section to describe what happens between these extremes.

In Figure 2.7 we illustrated a rush-hour condition and in parts *b* and *c* of that figure we gave the fluid approximation to its queueing behavior. We wish now to study the diffusion approximation to the onset of a rush hour; that is, we wish to consider the case where $\rho(t)$ grows with t from the stable case [$\rho(t) < 1$] through the critical value [$\rho(t) = 1$] and on into the overloaded case [$\rho(t) > 1$]. This "transition through saturation" as described by Newell [NEWE 68] (whose development we follow here) corresponds to the onset of a rush hour. Let us define our time axis such that $\rho(t)$ passes through the critical value (unity) at $t = 0$; therefore for $t < 0$ we have a stable case, whereas for $t > 0$ we clearly have an unstable situation. As described in the previous paragraph as long as $\rho(t)$ grows slowly toward its critical value then the quasistationary distribution given in Eq. (2.160) will approximately describe this system behavior. To grow slowly enough we mean that condition (2.163) is satisfied. However, as mentioned above, as $\rho(t)$ approaches unity this condition must certainly be violated and then the system behavior will depart from that of the quasistationary solution and the waiting time will not be able to grow as quickly as that result would imply. Well beyond $t = 0$ we expect the waiting time to grow much as the fluid approximation describes.

As Newell suggests [NEWE 68], let us make a Taylor expansion for $\rho(t)$ about the time origin, that is

$$\rho(t) = 1 + \alpha t + \frac{\alpha^2 t^2}{2!} + \cdots$$

In the vicinity of the critical value ($t = 0$) ρ behaves approximately as $1 + \alpha t$, and we will assume that this is a good approximation in the transition region between that point where the waiting time behavior begins to depart from the quasistationary solution and before it begins to behave as the fluid approximation would indicate. If we use this linear expression for $\rho(t)$ then condition (2.163) becomes

$$\left|\frac{\sigma^2(t)}{\alpha^2 t^3}\right| \ll 1$$

that is,

$$|t| \gg \left(\frac{\sigma^2}{\alpha^2}\right)^{1/3} \triangleq t_0 \tag{2.164}$$

where we have assumed that $\sigma^2(t) \cong \sigma^2$ in this vicinity. When both sides of this inequality are approximately of the same order of magnitude, then we find that the quasistationary prediction for the waiting time will begin to diverge from its true value. At the time $t = -t_0$ we expect the average unfinished work to be approximately the mean of Eq. (2.160), namely, $-\sigma^2/2m(-t_0)$; however, since we found it convenient earlier to use $-\sigma^2/m$ as the unit of unfinished work in our scaled equations [see Eq. (2.128)], and since Newell also makes this choice, we too choose the following approximation for the mean work at $t = -t_0$,

$$\bar{U}(-t_0) \cong \frac{-\sigma^2}{m(-t_0)} = \left(\frac{\sigma^4}{\alpha}\right)^{1/3} \triangleq \bar{U}_0 \tag{2.165}$$

During the time from $-t_0$ to 0 the average backlog should grow by an amount approximately equal to the work that arrives during $(-t_0, 0)$ less the work discharged (1 sec/sec) during this interval. Thus the expected increase in $\bar{U}$ during $(-t_0, 0)$ is

$$[\lambda(t)/\mu(t)]t_0 - t_0 = \alpha t_0^2 = \bar{U}_0$$

Therefore, we have defined a natural time unit t_0 and a natural backlog (virtual waiting time) unit $\bar{U}_0$. It is significant to note that the average backlog changes in proportion to $\alpha^{-1/3}$.

Now let us get a feeling for what the diffusion predicts in this transition through saturation. The equation of motion is given in Eq. (2.159); since $m(t) = \rho(t) - 1$ our approximation in the vicinity $t = 0$ is $m(t) \cong \alpha t$ (and moreover we have already assumed that the infinitesimal variance is essentially independent of t in this region, that is, $\sigma^2(t) \cong \sigma^2$), which then gives

$$\frac{\partial F}{\partial t} = -\alpha t \frac{\partial F}{\partial w} + \frac{\sigma^2}{2} \frac{\partial^2 F}{\partial w^2}$$

We have written this in terms of the distribution function rather than the pdf. We have prepared the way for transforming this last equation into a dimensionless equation and so we define the new scaled variables $t' = t/t_0$ and $w' = w/\bar{U}_0$; thus we now have $F(w', t')$ as giving the distribution of $U'(t') = U(t/t_0)/\bar{U}_0$. The scaled equation then becomes

$$\frac{\partial F}{\partial t'} = -t' \frac{\partial F}{\partial w'} + \frac{1}{2} \frac{\partial^2 F}{(\partial w')^2} \tag{2.166}$$

We see that we have successfully scaled this equation to eliminate its dependence on the specific parameters of the problem; these parameters are now contained in the scaling factors $t_0 = \sigma^{2/3}/\alpha^{2/3}$ and $\bar{U}_0 = \sigma^{4/3}/\alpha^{1/3}$. This equation of course is subject to the boundary conditions that $F(\infty, t') = 1$, $F(0^-, t') = 0$; also for $t' \ll -1$ the solution must be the quasi-stationary solution $F(w', t') = 1 - e^{2w't'}$. As Newell points out, there is no

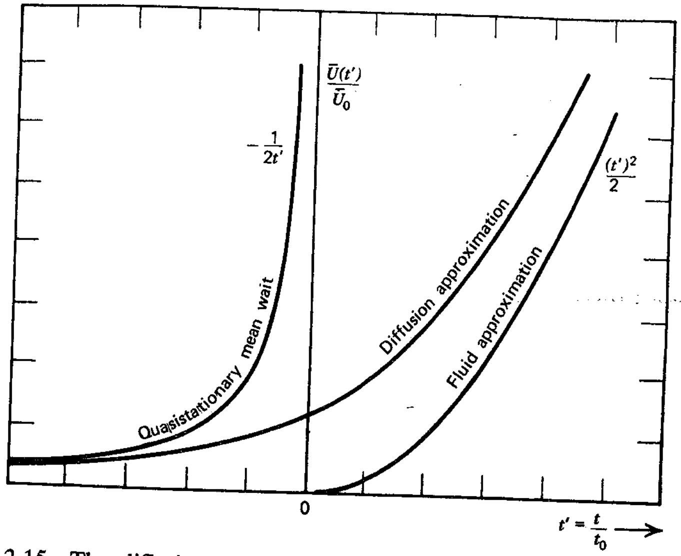

Figure 2.15 The diffusion approximation to the mean wait for the transition through saturation.

known simple analytic solution for Eq. (2.166) subject to these conditions. Fortunately this is a "universal" equation (that is, it is scaled to represent any set of parameter values), and thus we need solve it only once. Furthermore, if we carry out this computation numerically (as in Newell [NEWE 71]), we obtain the required system behavior in the transition through saturation. One way to demonstrate this behavior is to plot the average (virtual) waiting time (in units of $\bar{U}_0$ seconds) versus the normalized time scale t_0. A diagram showing the results of Newell's numerical computation is given in Figure 2.15. Here we see three curves, of which one, the quasistationary mean wait in the region $t' < 0$, forms the left asymptote for our diffusion approximation (which is the second curve); we also see the fluid approximation to the mean wait in the region $t' > 0$, which forms an approximate right asymptote to the diffusion approximation. From Eq. (2.160) we see that the quasistationary mean wait is merely

$$\bar{U}(t) = \frac{-\sigma^2(t)}{2m(t)} \tag{2.167}$$

which when normalized with respect to $\bar{U}_0$ and expressed in terms of the scaled time t' becomes merely [for $\rho(t) = 1 + \alpha t$]

$$\frac{\bar{U}(t')}{\bar{U}_0} = -\frac{1}{2t'} \tag{2.168}$$

Equation (2.168) could also have easily been obtained by inspecting the form for the quasistationary solution $F(w', t')$ given earlier as an exponential. For the fluid approximation we see for $t' > 0$ that we are aggravating the work deficit at a rate αt sec/sec, which gives an accumulated work backlog of size $\alpha t^2/2$ by time t; normalizing this with respect to $\bar{U}_0$ and scaling the time axis to t' we easily see that the scaled work backlog has a value $(t')^2/2$ at time t'. This then gives us the shape for the fluid approximation to the mean wait. Once we are deep into saturation (when the probability of the system emptying is insignificant) then our remarks from previous sections assure us that the change in the work backlog will be normally distributed over any interval of time. In fact we see that the average change in this backlog (under the diffusion approximation) will be the same as the average change predicted by the fluid approximation, and this calculation is just the integral of the overload during this time interval; this last statement does not depend upon the fact that the overload grows linearly in this region. As Newell has shown, the distribution F for $t' < -1$ is essentially of the exponential form given earlier, whereas for $t' > 1$, F begins to approach a normal distribution whose mean grows as predicted by the fluid approximation! The reader is urged to consult the fine monograph by Newell [NEWE 71] as well as his earlier article [NEWE 68] for more details.

And so we have a rather good understanding of the behavior of the waiting time as the system enters and continues in the rush hour. We see the important role played by the fluid approximation in this case. However, we again caution the reader that when the approach to saturation is slow, then the zero waiting time predicted by the fluid approximation prior to saturation is badly in error since the system has time to follow the quasistationary mean wait, which grows to large values in such a case. Nevertheless, the average change in queue size will follow the fluid approximation once we have been in saturation for a time on the order of one normalized time unit; the effect of a slow approach to saturation will be an offset between the diffusion and fluid asymptotes for $t' > 1$.

We now inquire into the waiting time behavior for the entire rush-hour cycle. This in some sense requires that we investigate the inverse to the problem we have just studied. We expect that the fluid approximation will be an accurate prediction while $\rho(t) > 1$ and as the system makes the transition back down from saturation into the range $\rho(t) < 1$ then it will settle down into a quasistationary mode. However, at the time when $\rho(t)$ first falls below 1 we recognize from the fluid approximation as shown in Figure 2.7 that the backlog at that time has a maximum value and therefore there will be a "long-tail" effect until the system has a possibility of going idle. During this long tail the behavior will be dominated by the fluid approximation, that is, there will be a normally distributed waiting time with a decreasing mean given by that approximation. After the tail expires then the quasistationary solution takes over and once again the stochastic effects are responsible for the occurrence of queues and delays. In his monograph Newell postulates a parabolic form for $\rho(t)$ during the rush hour merely as an example and shows in fact that the behavior just described does obtain; he derives another dimensionless diffusion equation to examine the transition region for this case as well and presents curves to display this behavior, which is very much what one would expect.

McNeil [McNE 73] nicely summarizes some diffusion models, including the rush-hour congestion process as well as some "almost stationary" situations.

Newell [NEWE 73] has also studied the approximate behavior of the G/G/m queue for $m \gg 1$. He classifies types of behavior, describes the qualitative properties of these types and discusses graphical and analytic (e.g., diffusion approximation) methods that might be used to obtain more quantitative behavior. He observes when the typical queue size is large and all servers are kept busy most of the time, that the system behaves like an effective G/G/1 system with service times $\{x_n/m\}$. Further, when $\rho(t)$ remains less than (approximately) $1-1/m$, then the system behaves like a G/G/∞ system. Between these two extremes, many different types of behavior may be observed, depending upon how $\rho(t)$ passes through the transition region, and these are discussed in [NEWE 73].

From a philosophical point of view the fluid approximation we have considered is extremely appealing in its simplicity both for calculation as well as for physical intuition. We find that it is not a bad approximation in the overloaded case, but we wonder if it has anything of interest to suggest about the stable case. We have seen that for the stationary case with $\rho < 1$ then the fluid approximation predicts a zero waiting time. This comes about because we have averaged the rate at which work arrives over the infinite time axis to find that on the average $\lambda\bar{x} < 1$, which means that work arrives at a rate less than the capacity of the system (which is 1 sec of work per second of elapsed time) and therefore no "fluid" will accumulate in our funnel. Of course one need not have averaged over the entire time axis, and it is this point of view we wish to take now. For example one may argue that as soon as a customer arrives to an empty system he "overloads" it in the sense that more work has arrived in the differential time interval surrounding his arrival than can be discharged in that differential time interval; thus the time derivative of work arriving is unbounded in this differential time interval, giving rise to an impulse whose area is equal to the service time of this customer and this represents the work backlog at that instant. If no other work arrives then this work is discharged at a unit rate until the customer departs; his time in system corresponds to the "long tail" of our fluid approximation. This is much like Figure 2.7 where the total positive area under the curve is considered to have arrived in zero time, giving rise to a step in Figure 2.7(*c*) rather than the smooth rise to its peak. Of course if more customers arrive before the backlog is discharged then the overload continues and takes vertical jumps equal to the service time of each arriving customer. What we are in fact describing in terms of this "instantaneous" fluid approximation is the unfinished work $U(t)$ itself! In Figure 2.16 we give an example of $U(t)$ and directly below it we show the impulses describing the arrival of work at the customer arrival instants; the number next to the impulse gives its area and is equal to the number of seconds of service brought in by each arrival.* Now, let $\omega(t)$ be the

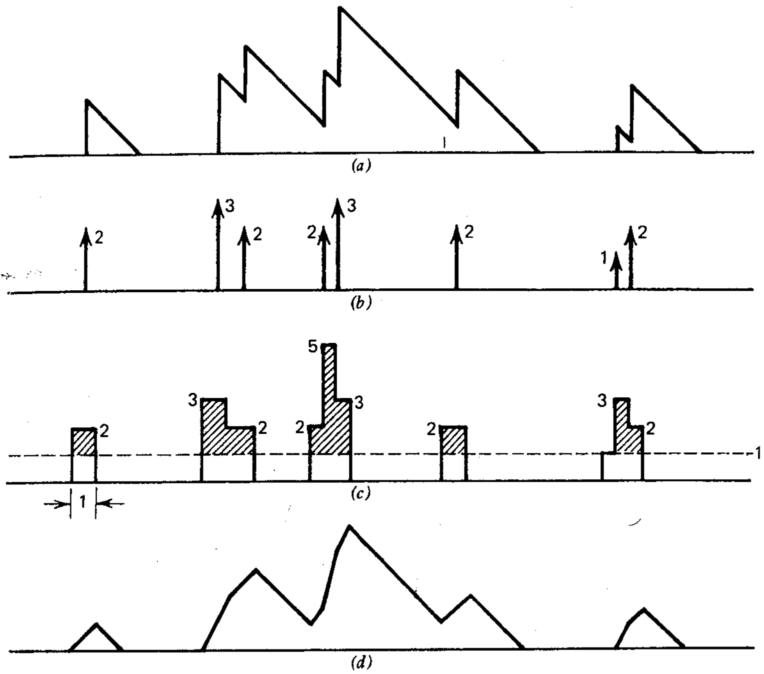

Figure 2.16 An intermediate fluid approximation. (*a*) $U(t)$; (*b*) true burst arrivals; (*c*) "smoothed" input; (*d*) intermediate fluid approximation.

* Thus $U(t)$ may be thought of as the output of a linear system whose impulse response is a linearly decaying ramp (slope, -1) with unit height (that is, a small triangle) and whose input is the sequence of work arrivals as shown in part (*b*) of Figure 2.16.

stochastic process representing the arrival of work to the system, that is, a sequence of impulses such as in Figure 2.16(*b*); to be precise we have

$$\omega(t) = \sum_{n=0}^{\infty} x_n u_0(t - \tau_n)$$

where, as usual, x_n and τ_n represent the service time and arrival time for C_n, and $u_0(y)$ is the unit impulse function occurring at the instant $y = 0$. Let us now consider a continuum of "intermediate" fluid approximations that lie between the original extreme fluid approximation in which the burst arrivals are averaged over an infinite interval and the *exact* situation in which the burst arrivals are averaged over an infinitesimal interval (that is, no average at all). Thus let us consider an averaging interval of length A. The continuum of smoothed input functions is defined as follows:

$$\omega_A(t) \triangleq \frac{1}{A}\int_{-\infty}^{\infty} \left[\omega\left(y+\frac{A}{2}\right) - \omega\left(y-\frac{A}{2}\right)\right] dy \qquad (2.169)$$

Thus we are taking the impulses and uniformly spreading their area over an interval of length A centered about the instant of their occurrence. Figure 2.16(*c*) shows an example with $A = 1$ for the arrival patterns shown in part (*b*). If this smoothed input is considered as the instantaneous rate of fluid flow into a queueing system whose departure capacity is 1 sec of work per elapsed second then we see that the cross-hatched region represents the short-term overloads to this system much as the positive area in Figure 2.7 did. Integrating this overload (which goes negative, with value -1 when the input drops to zero) gives us Figure 2.16(*d*), which corresponds to this intermediate fluid approximation for the original unfinished work $U(t)$ in Figure 2.16(*a*). Not a bad approximation at all! This relatively good fit results because A is on the order of an average service time; were A considerably larger than an average service time then we would begin to approach the original fluid approximation, which would have predicted zero backlog over most of the time axis. Of course one need not choose a uniform averaging as in Eq. (2.169). The usefulness of these intermediate fluid approximations is that they provide another point of view for understanding queueing systems and the formation of queues and delays.

In the remaining chapters of this book, we find use for many of these approximation techniques. It is perhaps fair to say that this field of generating clever approximations to the complex stochastic processes involved in queueing systems will provide the greatest impetus to the advancement of queueing theory and its applications in the next few years. There is great challenge and reward lying in that direction and the reader is urged to meet that challenge (and thus reap the reward).

REFERENCES

ABRA 64 Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards (Wash., D.C.), 1964.

BARL 65 Barlow, R. E., and F. Proschan, *Mathematical Theory of Reliability*, Wiley (New York), 1965.

BORO 64 Borovkov, A., "Some Limit Theorems in the Theory of Mass Service, I," *Theory of Probability and Its Applications*, **9,** 550–565 (1964).

BORO 65 Borovkov, A., "Some Limit Theorems in the Theory of Mass Service, II," *Theory of Probability and Its Applications*, **10,** 375–400 (1965).

BROC 48 Brockmeyer, E., H. L. Halstrøm, and A. Jensen, "The Life and Works of A. K. Erlang," *Transactions of the Danish Academy of Technology and Science*, **2** (1948).

BRUM 71 Brumelle, S. L., "Some Inequalities for Parallel Server Queues," *Operations Research*, **19,** 402–413 (1971).

BRUM 73 Brumelle, S. L., "Bounds on the Wait in a GI/M/k Queue," *Management Science*, **19,** No. 7, 773–777 (1973).

BUZE 74 Buzen, J. P., "Structural Considerations for Computer System Models," *Proceedings of the Eighth Annual Princeton Conference on Information Sciences and Systems*, March 1974.

COHE 69 Cohen, J. W., *The Single Server Queue*, Wiley-Interscience (New York), 1969.

COHE 73 Cohen, J. W., "Asymptotic Relations in Queueing Theory," *Stochastic Processes and Their Applications*, **1,** No. 2, 107–124 (1973).

COX 61 Cox, D. R., and W. L. Smith, *Queues*, Methuen (London) and John Wiley and Sons (New York), 1961.

COX 65 Cox, D. R., and H. D. Miller, *The Theory of Stochastic Processes*, John Wiley (New York) 1965.

GAVE 68 Gaver, D. P., Jr., "Diffusion Approximations and Models for Certain Congestion Problems," *Journal of Applied Probability*, **5,** 607–623 (1968).

GELE 74 Gelenbe, E., "On Approximate Computer Systems Models," in E. Gelenbe and R. Mahl, eds., *Computer Architectures and Networks*, North-Holland Publishing Company (Amsterdam), 187–206 (1974).

GROS 73 Gross, D., "Sensitivity of Queueing Models to the Assumption of Exponentiality: I—Single-channel Queues," Technical Memorandum Serial 64121, Institute for Management Science and Engineering, The George Washington University (1973).

HARR 73 Harrison, J. M., "The Heavy Traffic Approximation for Single Server Queues in Series," *Journal of Applied Probability*, **10**, No. 3, 613–629 (1973).

HEYM 74 Heyman, D. P., "An Approximation for the Busy Period of the M/G/1 Queue Using a Diffusion Model," *Journal of Applied Probability*, **11**, 159–169 (1974).

IGLE 69 Iglehart, D., "Multiple Channel Queues in Heavy Traffic, IV: Law of the Iterated Logarithm," Technical Report No. 8, Dept. Operations Research, Stanford University, 1969.

IGLE 70 Iglehart, D., and W. Whitt, "Multiple Channel Queues in Heavy Traffic, I," *Advances in Applied Probability*, **2**, 150–177 (1970); "Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks, and Batches," *Advances in Applied Probability*, **2**, 355–369 (1970).

ITO 65 Itô, K., and H. P. McKean, Jr., *Diffusion Processes and Their Sample Paths*, Academic Press (New York) 1965.

KIEF 55 Kiefer, J., and J. Wolfowitz, "On the Theory of Queues with Many Servers," *Transactions of the American Mathematics Society*, **78**, 1–18 (1955).

KING 61 Kingman, J. F. C., "The Single Server Queue in Heavy Traffic," *Proceedings of the Cambridge Philosophical Society*, **57**, 902–904 (1961).

KING 62a Kingman, J. F. C., "On Queues in Heavy Traffic," *Journal of the Royal Statistical Society, Series B*, **24**, 383–392 (1962).

KING 62b Kingman, J. F. C., "Some Inequalities for the Queue GI/G/1," *Biometrika*, **49**, 315–324 (1962).

KING 64 Kingman, J. F. C., "The Heavy Traffic Approximation in the Theory of Queues," in W. L. Smith and R. I. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*, Univ. of North Carolina (Chapel Hill), 137–169 (1964).

KING 70 Kingman, J. F. C., "Inequalities in the Theory of Queues," *Journal of the Royal Statistical Society, Series B*, **32**, 102–110 (1970).

KLEI 75 Kleinrock, L., *Queueing Systems, Volume I: Theory*, Wiley-Interscience (New York) 1975.

KOBA 74a Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks, I. Equilibrium Queue Distributions," *Journal of the Association for Computing Machinery*, **21**, No. 2, 316–328 (1974).

KOBA 74b Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks, II. Nonequilibrium Distributions and Computer Modeling," *Journal of the Association for Computing Machinery*, **21**, No. 3, 459–469 (1974).

KOBA 74c Kobayashi, H., "Bounds for the Waiting Time in Queueing Systems," in E. Gelenbe and R. Mahl, eds., *Computer Architectures and Networks*, North-Holland Publishing Company (Amsterdam), 263–274 (1974).

KOLL 74 Köllerström, J., "Heavy Traffic Theory for Queues with Several Servers. I," *Journal of Applied Probability*, **11**, 544–552 (1974).

MARC 74 Marchal, W. G., "Some Simple Bounds and Approximations in Queueing," Technical Memorandum Serial T-294, Institute for Management Science and Engineering, The George Washington University, January 1974.

MARS 68a Marshall, K. T., "Some Inequalities in Queueing," *Operations Research*, **16**, 651–665 (1968).

MARS 68b Marshall, K. T., "Bounds for Some Generalizations for the GI/G/1 Queue," *Operations Research*, **16**, 841–848 (1968).

MARS 68c Marshall, K. T., "Some Relationships between the Distributions of Waiting Time, Idle Time and Interoutput Time in the GI/G/1 Queue," *SIAM Journal of Applied Mathematics*, **16**, 324–327 (1968).

McNE 73 McNeil, D. R., "Diffusion Limits for Congestion Models," *Journal of Applied Probability*, **10**, 368–376 (1973).

MORS 55 Morse, P. M., "Stochastic Properties of Waiting Lines," *Operations Research*, **3**, 255–261 (1955).

NEUT 73 Neuts, M. F., "The Single Server Queue in Discrete Time-Numerical Analysis," *Naval Research Logistics Quarterly*, Part I: **20**, No. 2, 297–304 (1973); Part II (with E. M. Klimko): **20**, No. 2, 305–320 (1973); Part III (with E. M. Klimko): **20**, No. 3, 557–568 (1973).

NEWE 65 Newell, G. F., "Approximate Methods for Queues with Application to the Fixed-Cycle Traffic light," *SIAM Review*, **7**, 223–240 (1965).

NEWE 68 Newell, G. F., "Queues with Time-Dependent Arrival Rates I–III," *Journal of Applied Probability*, **5**, 436–451, 579–606 (1968).

NEWE 71 Newell, G. F., *Applications of Queueing Theory*, Chapman and Hall, Ltd. (London), 1971.

NEWE 73 Newell, G. F., *Approximate Stochastic Behavior of n-Server Service Systems with Large n*, Springer-Verlag (Berlin), 1973.

PAPO 65 Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill (New York), 1965.

PROH 63 Prohorov, Y., "Transient Phenomena in Processes of Mass Service," (in Russian), *Litovsk. Mat. Sb.*, **3**, 199–205 (1963).

REIS 74 Reiser, M., and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Queuing Systems," *IBM Journal of Research and Development*, **18**, 110–124 (1974).

ROSS 74 Ross, S. M., "Bounds on the Delay Distribution in GI/G/1 Queues," *Journal of Applied Probability*, **11**, 417–421 (1974).

SUZU 70 Suzuki, T., and Y. Yoshida, "Inequalities for Many-Server Queue and Other Queues," *Journal of the Operations Research Society of Japan*, **13**, 59–77 (1970).

SYSK 62 Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd (London), 1962.

WONG 74 Wong, D. K., "A Discrete Approximation for G/G/1 Queue," M.S. Thesis, Computer Science Department, School of Engineering and Applied Science, University of California at Los Angeles, December 31, 1974.

EXERCISES

2.1. Show that the mean waiting time obtained from the heavy traffic approximation is too large by no more than a mean interarrival time for M/G/1.

2.2. Consider a G/D/1 system in which the constant service time has value $a+c$ and the interarrival time is uniformly distributed between c and $a+b+c$, where a, b, and c are all non-negative constants.

(a) What relationship must exist among the constants if the system is to be stable?

(b) Find W_U.

(c) Find W_M.

(d) In solving part (c), prove that the solution to Eq. (2.35) is unique in this case.

2.3. Joe is hired to measure the average length of the queue in front of the factory emergency room. Emergencies occur at an average of 6 emergencies per hour (with second moment 100 min^2) and they take 5 min on the average to treat (with variance 50 min^2). Assume this queue behaves as a stationary G/G/1 system.

After many weeks, Joe puts out a report in which he claims that the average queue length he measured is 1.05. Do you believe him? Why?

2.4. Consider a G/G/1 system such that, for $0 \le \alpha \le 1$,

$$A(t)=\begin{cases}0 & t<0\\ \alpha & o\le t<T_0\\ 1 & T_0\le t\end{cases}\qquad B(x)=\begin{cases}0 & x<-(1-\alpha)\log\alpha\\ 1 & -(1-\alpha)\log\alpha\le x\end{cases}$$

(a) Find $\bar{t}$, σ_a^2, $\bar{x}$, σ_b^2, and ρ in terms of α and T_0. What relation between α and T_0 must be true for stability?

(b) Express the upper bound W_U in terms of α and T_0.

(c) For a given value of T_0, find that value of α which minimizes W_U.

(d) For this value of α, find W_U in terms of T_0.

(e) For this value of α, find that value of T_0 which maximizes W_U. What value do we now get for W_U?

2.5. Consider a G/G/1 system with bulk arrivals, where the average bulk size is $\bar{g}$ and the variance is σ_g^2. Assume that $A(t)$ is $\bar{t}$-MRLA, where $\lambda = 1/\bar{t}$ is the mean arrival rate of groups. Find upper and lower bounds on W_g, the mean time a group spends in the queue until the first of the group's members begins service.

2.6. Let us derive W_U in an alternate fashion.

(a) It is clear for G/G/1 that a_0 ($=P$[arrival finds system empty]$=P[\tilde{y}>0]$) is such that $0<a_0<1$. Show that $E[\tilde{y}^k]$ is such that $\overline{y^k}=a_0\overline{I^k}$. From these, establish a simple lower bound on $\bar{I}$. Also give a simple lower bound on $\overline{I^2}$ in terms of $\bar{I}$.

(b) From (a), establish a lower bound on the mean residual life of the idle time in terms of λ and ρ only.

(c) Using (b) in Eq. (1.132) prove the basic upper bound in Eq. (2.22).

2.7. Consider the waiting time variance, $\sigma_{\tilde{w}}^2$.

(a) By first cubing Eq. (1.125) and then forming expectations in the limit as $n\to\infty$, express $\sigma_{\tilde{w}}^2$ in terms of the first three moments of $\bar{t}$, $\tilde{x}$, and I.

(b) From (a), proceed as in Sections 2.2 and 2.3 to show Eq. (2.42).

2.8. We wish to prove Eq. (2.47) for the queue γ-MRLA/G/1.

(a) Let $\tilde{w}+\tilde{x}$ have the PDF $S(x)=P[\tilde{w}+\tilde{x}\le x]$. Prove that

$$P[\tilde{y}>y]=\int_0^\infty [1-A(y+x)]\,dS(x)$$

(b) Show that the idle time I must obey

$$P[I>y]=\frac{1}{a_0}P[\tilde{y}>y]$$

(c) Using the γ-MRLA properties of $A(t)$ show the following, using (a) and (b) above:

$$\int_t^\infty P[I>x]\,dx \le \gamma P[I>t]$$

(d) Form the mean residual life for I from (c) and show

$$\frac{\overline{I^2}}{2\bar{I}}\le\gamma$$

2.9. Using an approach similar to that in Exercise 2.8, prove for G/G/1 where $A(t)$ has DMRL, that when $t \geq 0$,

$$\int_t^\infty \frac{P[I>x]}{P[I>t]}\,dx \leq \int_t^\infty \frac{1-A(x)}{1-A(t)}\,dx$$

2.10. Now we consider a G/G/1 system where $A(t)$ has IFR.

(a) Beginning with the expression for $P[I>t]$ from part (b) of Exercise 2.8, find the failure rate for I.

(b) Using the IFR property for $A(t)$, show for $\varepsilon>0$ that

$$\frac{P[I>t+\varepsilon]}{1-A(t+\varepsilon)} \leq \frac{P[I>t]}{1-A(t)}$$

(c) From (b), show that the following determinant must be nonpositive:

$$\det\begin{bmatrix} \int_t^\infty P[I>x]\,dx & \int_0^t P[I>x]\,dx \\ \int_t^\infty [1-A(x)]\,dx & \int_0^t [1-A(x)]\,dx \end{bmatrix} \leq 0$$

(d) From (c) show the final result for IFR/G/1,

$$\int_t^\infty \frac{P[I>x]}{\bar{I}}\,dx \leq \int_t^\infty \frac{1-A(x)}{\bar{t}}\,dx$$

2.11. Once again, we consider an IFR/G/1 system.

(a) Prove for any random variable X with second moment $\overline{X^2}$ and PDF $F(x)$, that

$$\int_0^\infty \int_t^\infty [1-F(x)]\,dx\,dt = \frac{\overline{X^2}}{2}$$

(b) Using the final result from Exercise 2.10, and (a) above, prove Eq. (2.50).

2.12. We wish to solve the system given in Eqs. (2.77)–(2.78).

(a) Find $P(z)=\sum_{k=0}^{\infty} p(k)z^k$ in terms of $p(0)$.

(b) Evaluate $p(0)$ and find $\{p(k)\}$ explicitly.

2.13. Repeat the solution of Exercise 2.12 (for the example of Section 2.6) using the method of spectrum factorization.

2.14. As in Section 2.6, consider a discrete queue for which

$$A(t)=\begin{cases} 0 & t<0 \\ \alpha & 0\leq t<2 \\ 1 & 2\leq t \end{cases} \qquad B(x)=\begin{cases} 0 & x<1 \\ 1 & 1\leq x \end{cases}$$

(a) Find ρ.

(b) Find $c(k)$.

(c) Assume $w_0=0$ with probability one. Find and draw $p_n(k)$ for $n=0,1,2,3$.

(d) Write the equilibrium equations for $p(k)$.

(e) Using z-transforms, solve for $p(k)$ explicitly.

2.15. Repeat the previous exercise for

$$A(t)=\begin{cases} 0 & t<1 \\ \frac{1}{2} & 1\leq t<2 \\ 1 & 2\leq t \end{cases} \qquad B(x)=\begin{cases} 0 & x<0 \\ \frac{1}{3} & 0\leq x<1 \\ \frac{2}{3} & 1\leq x<2 \\ 1 & 2\leq x \end{cases}$$

2.16. Repeat Exercise 2.14 for

$$a(k)=\begin{cases} \frac{1}{3} & k=0 \\ \frac{2}{3} & k=1 \end{cases} \qquad b(k)=\begin{cases} \frac{1}{2} & k=0 \\ \frac{1}{2} & k=1 \end{cases}$$

2.17. For the example of Section 2.6,

(a) Give an upper bound on W.

(b) Give the strongest lower bound you can for W, using the techniques from Section 2.3.

(c) Find W exactly (use the results from Exercise 2.12).

2.18. Consider the G/G/1 system of Exercise 2.4 with the values for α and T_0 as found in parts (c) and (e). If we try to solve this system by the method of Section 2.6, what problems do we encounter?

2.19. Express Eq. (2.132) in dimensionless form.

2.20. **(a)** Find an upper bound in terms of W_U for the root σ associated with G/M/1.

(b) Repeat for G/M/m.

2.21. In this exercise, we develop some of the material for the diffusion approximation to M/G/1.

(a) From the Fokker–Planck equation (2.133) form $F^{**}(r,s)$ given in Eq. (2.134) and show that

$$F^{**}(r,s)=\frac{2}{\sigma^2}\left[\frac{C_1+rC_2-e^{-rw_0}}{r^2-(2m/\sigma^2)r-(2/\sigma^2)s}\right]$$

where w_0 is the initial backlog, and C_1 and C_2 are constants with respect to r.

(b) Clearly, r_1 and r_2 as given in Eq. (2.136) are the denominator roots. Also define η as in Eq. (2.137). Establish that $\eta > 0$ for $0 \leq \rho$.

(c) By setting $r = \eta$, find a relation between C_1 and C_2.

(d) What value must $sF^{**}(0, s)$ take on?

(e) Using (d), solve for C_1. From C_1 and (c), solve for C_2. We have now proven Eq. (2.135).

(f) Let $w_0 = 0$. Expand $F^{**}(r, s)$ in partial fractions. Observing that $r_1 r_2 = -2s/\sigma^2$, prove Eq. (2.141).

(g) Show that $\lim_{s \to 0} r_2 = -s/m$ and $\lim_{s \to 0} r_1 = 2m/\sigma^2$.

(h) From (g) show that Eq. (2.142) must hold for $\rho < 1$.

(i) By direct calculation, prove Eq. (2.147).

(j) For $w_0 = 0$ and $\rho < 1$, prove Eq. (2.148), noting that $\eta = r_2$ and $r_1 r_2 = -2s/\sigma^2$ again.

(k) Show that the scaled version of Eq. (2.148) is as given in Eq. (2.155).

(l) For $\rho > 1$, prove Eq. (2.152).

2.22. Show that Eq. (2.157) is the inverse of the transform given in Eq. (2.155). For this, use the common properties of transforms (see [KLEI 75] Table I.3) and the helpful transform pair

$$\frac{1}{\sqrt{\pi t}} - 2ae^{a^2 t}[1 - \Phi(a\sqrt{2t})] \Leftrightarrow \frac{1}{a + \sqrt{s}}$$

2.23. Consider the *third*-order approximation to Eq. (2.111) in which we permit the first three terms $A_n(w, t)$, for $n = 1, 2, 3$ to be nonzero, and assume all the rest to be zero ($n > 3$). Let us study this solution for the unfinished work in an M/G/1 system in equilibrium ($\rho = 1 - \varepsilon$ where $1 \gg \varepsilon > 0$) [COHE 73].

(a) Show that the general dimensionless [see Eqs. (2.127)–(2.128)] equilibrium solution must be

$$F(w') = C_1 + C_2 e^{s_1 w'} + C_3 e^{s_2 w'}$$

where s_1 and s_2 are the roots of

$$\frac{\gamma}{4} s^2 - \frac{1}{2} s - 1 = 0$$

and

$$\gamma = \frac{2}{3} \frac{\overline{x^3}}{\lambda (\overline{x^2})^2} (1 - \rho)$$

(b) Clearly, the two roots have opposite sign. Let $s_1 < 0$, $s_2 > 0$. Show that

$$F(w') = 1 - e^{s_1 w'} \qquad w' \geq 0$$

(c) Consider the approximation

$$s_1 \cong a + b\gamma + c\gamma^2$$

and find the best values a, b, c (note that $|\gamma| \ll 1$).

(d) From parts (b) and (c), find an explicit form for $F(w')$ and compare to Eq. (2.131).

2.24. Consider a diffusion approximation to $U(t)$ for M/G/1 [HEYM 74]. Consider a busy period initiated by a customer whose service time is x sec. Let $g(y; x)$, $G^*(s; x)$, and $g_k(x)$ be the pdf, its Laplace transform, and the kth moment of the duration of such a busy period when it is approximated by a diffusion process with mean $m = \rho - 1$ and variance $\sigma^2 = \lambda \overline{x^2}$. It has been shown [COX 65] that

$$G^*(s; x) = \exp\left\{-\frac{mx}{\sigma^2}\left[1 - \sqrt{1 + \frac{2\sigma^2 s}{m^2}}\right]\right\}$$

(a) Find $g_k(x)$ for $k = 1, 2, 3$.

(b) Let $g(y)$, $G^*(s)$, and g_k be the pdf, its Laplace transform, and the kth moment of the unconditional busy period duration under the diffusion approximation. Express $g(y)$ in terms of $g(y; x)$.

(c) From (a) find g_k for $k = 1, 2, 3$ and compare to the known values of the moments of the exact M/G/1 busy period.

(d) Express $G^*(s)$ in terms of $B^*(s)$.

(e) In what way is the expression in (d) superior to the corresponding expression for the exact M/G/1 system?

2.25. For the input work stream shown in Figure 2.16(b), redraw parts (c) and (d) of that figure in the case when the smoothing "filter" is such that it spreads a unit impulse uniformly over the two unit time slots surrounding that impulse (the height of this rectangular pulse will therefore be $\frac{1}{2}$. Repeat in the case where the unit impulse gets spread as a small triangle rising linearly from zero at $\frac{1}{2}$ sec prior to the impulse, to a value of 2 at the time of occurrence of the impulse, and then dropping linearly to zero at $\frac{1}{2}$ sec following the impulse.