S. B. Calo

Delay Analysis of a Two-Queue, Nonuniform Message Channel

A Message Channel is defined as a tandem connection of single server queues in which the successive service times experienced by any particular customer are scaled versions of the same random variable, and thus it serves as a model for sparsely connected store-and-forward data communications networks (or network segments) where messages typically preserve their lengths as they traverse the system. A particular instance of such a nonstandard queueing model is analyzed in this paper. The system consists of two single server queues in tandem subject to a Poisson arrival process (at the first queue) and providing service according to scaled versions of a sequence of two-level, discrete random variables. A set of recursive equations that can be used to solve the model for any given scaling factor at the second queue (normalized with respect to the first queue service) is explicitly derived. In addition, complete solutions are displayed for several cases of interest, and the equilibrium mean cumulative waiting times for these instances are compared as a method of indicating the impact of the scaling factor on the operation of the system. The extension of several results to systems with more general service time processes is discussed.

1. Introduction

One of the inherent complexities associated with queueing models for store-and-forward data communications networks arises from the fact that messages typically preserve their lengths as they traverse the system. The interarrival and service sequences at queues internal to the system are thus dependent, making standard methods of analysis realistically inappropriate. In an effort to find methods for dealing with such nonstandard queueing systems, a model for sparsely connected networks (or network segments) called a Message Channel has been studied. A Message Channel is defined as a tandem connection of single server queues in which the successive service times experienced by any particular customer are scaled versions of the same random variable. When the scaling factors are identical, the system is called a Uniform Message Channel (UMC) or Repeated-Service Tandem Connection. Some general properties of such queueing models have recently been reported (Calo [1]), and integral equations for the equilibrium distribution function of the cumulative waiting time process in Uniform Message Channels, which have, in certain instances, been explicitly solved, have been obtained (Calo [2]).

When the scaling factors are not identical, the problem becomes considerably more complex. The general Message Channel has thus proved to be much more difficult to characterize than the UMC. In this paper we consider a specific instance of such a Nonuniform Message Channel, whose structure is yet amenable to analysis.

Our model consists of two single-server queues in tandem subject to a Poisson arrival stream at the first queue and providing service according to scaled versions of a sequence of two-level discrete random variables. The interarrival sequence at the first queue, $\{\tau_n\}_{n=1}^{\infty}$, thus consists of independent, identically, negative-exponentially distributed random variables with mean value $E\{\tau\} = 1/\lambda < \infty$, and the underlying service sequence $\{S_n\}_{n=1}^{\infty}$ consists of independent, identically distributed random variables such that (for each n) S_n equals b_1 with probability p_1 or b_2 with probability $p_2 = 1 - p_1$, where for convenience we require that $0 \leq b_1 \leq b_2 < \infty$. We also denote the mean value of the elements of this service sequence by $E\{S_n\} = p_1 b_1 + p_2 b_2 = 1/\mu < \infty$. The service process at the first queue is taken as $\{S_n\}_{n=1}^{\infty}$, while the service process at the second queue is $\{aS_n\}_{n=1}^{\infty}$, where

the scaling factor is a nonnegative real number ($a \geq 0$). With $a = 1$ the model reduces to that of a two-queue UMC and is indeed a special instance of one of the examples given in [2] for which explicit solution procedures exist (two packet classes under Poisson arrivals).

While the analysis of the two-queue system described above is the principal topic of this paper, we have also undertaken to relate the effort to more expansive attempts at the development of solution procedures for problems involving tandem interconnections of message queues. Thus in Section 2, general expressions for the cumulative waiting time process in a two-queue tandem connection are formulated, and a widely applicable result concerning front-end dominated systems is exhibited. Then in Section 3 the development is continued for two-queue, nonuniform, message channels, and an integral equation for a limiting conditional distribution function from which the distribution of the equilibrium cumulative waiting time can be readily obtained is derived. The particular form of this equation that is applicable to systems that are rear-end dominated is also explicitly exhibited.

Section 4 deals with furthering our analysis when the arrivals to the system are assumed to constitute a Poisson process. In this case it is shown that a complete solution can be obtained for rear-end dominated systems without making any additional assumptions regarding the service time distribution.

Finally in Section 5 we return to the analysis of our specific model utilizing the more general results of previous sections as appropriate. For $a \in [0, b_1/b_2]$, the system is front-end dominated (the service time distributions in the separate queues are nonoverlapping with largest service always given in the first queue), and the total waiting time in the system is simply the waiting time at the first queue, which under our assumptions is M/G/1 (with our particular service process) and thus easily solved. For $a \in [b_2/b_1, \infty)$, the system is rear-end dominated (the service time distributions in the separate queues are nonoverlapping with longest service given in the last queue), and the expression for the Laplace-Stieltjes transform (LST) of the equilibrium distribution function of the cumulative waiting time process previously derived for this instance (in Section 4) then applies.

When this nonoverlapping characteristic no longer prevails, *i.e.*, for $a \in (b_1/b_2, b_2/b_1)$ in our case, the complexity of the analysis greatly increases. We show that for discrete service time random variables a complete solution can still be obtained. The simplest (two-level) such instance is used here as an example of the more general procedure. Utilizing the integral equation development of Sections 3 and 4 with the added assumption of the discrete nature of the service process, sets of recursive equations with easily identified boundary conditions are derived from which the LST of the distribution function of the equilibrium cumulative waiting time can be obtained for the scaling factor (a) in appropriate subregions of the intervals $(b_1/b_2, 1)$ and $(1, b_2/b_1)$. The number of such transform equations that must be solved in any particular instance depends upon relationships among the values of the scaling factor and the allowable service times, as one would expect. The closer the system is to being dominated (front or rear), the simpler the solution procedure.

Complete solutions for the model are displayed for the instances when the scaling factor is such that the system is (1) front-end dominated; (2) near-front-end dominated; (3) uniform ($a = 1$); (4) near-rear-end dominated; and (5) rear-end dominated. The equilibrium mean cumulative waiting times for these various instances are also compared within the context of a specific example (under which the parameters of the service time distribution are chosen so that it will have the same first three moments as a negative-exponential distribution with parameter μ) as a means of indicating the impact of the scaling factor on the operation of such a system. The extension of these results to systems with more general service time processes is also discussed.

2. The two-queue tandem connection

We consider a system of two queues in tandem where each station operates as a single channel (server) facility and services messages (customers) in their order of arrival (first-come-first-served priority discipline). Each queue in the serial connection is assumed to have potentially infinite waiting space; *i.e.*, there is no limit on queue size. We also assume for convenience that both queues are originally empty so that the first customer to arrive suffers zero waiting time in each. (See Fig. 1.)

The stochastic properties of such a tandem connection are completely determined by the interarrival time process at the first queue, designated $\{\tau_n\}_{n=1}^{\infty}$, and the two service time processes, designated $\{S_n^{(1)}\}_{n=1}^{\infty}$ and $\{S_n^{(2)}\}_{n=1}^{\infty}$, respectively. These establish the evolution of the waiting time processes at the individual queues of the system according to the nonlinear recursions

$$W_1^{(1)} = 0; \; W_{n+1}^{(1)} = (W_n^{(1)} + S_n^{(1)} - \tau_n)^+ \qquad (n \geq 1),$$

$$W_1^{(2)} = 0; \; W_{n+1}^{(2)} = (W_n^{(2)} + S_n^{(2)} - \tau_n + W_n^{(1)} + S_n^{(1)} - W_{n+1}^{(1)} - S_{n+1}^{(1)})^+ \qquad (n \geq 1), \quad (1)$$

where $W_n^{(k)}$ represents the waiting time of the nth custom-

er at the kth queue ($k = 1, 2$) and $(Y)^+ = \max\{0, Y\}$ is the positive rectification function.

We note that the first station of our tandem connection can be treated as an isolated single server queue, since all the information concerning its arrival process is contained in the sequence $\{\tau_n\}_{n=1}^{\infty}$, and the subsequent history of any customer leaving the queue has no effect upon its operation. The waiting time process at the first queue as indicated in (1) thus follows the usual single server queueing recursion, which can be formulated as (see, for example, Loynes [3])

$$W_1^{(1)} = 0; \quad W_{n+1}^{(1)} = \left(\max_{1\le r\le n}\left\{\sum_{j=r}^{n}(S_j - \tau_j)\right\}\right)^+ \qquad (n \ge 1). \quad (2)$$

The operation of the second queue of our tandem connection, however, is very much affected by the prior history of its customers. The arrival process to this queue is imparted a particular correlation structure by the queueing process at the previous station, as indicated by the defining recursions of (1). This correlation structure is in general quite complicated and difficult to characterize, thus making any attempt at a direct analysis of this nonstandard system as an isolated queue prohibitively complex.

If, however, the performance characteristic of primary interest is the total delay suffered by a message in passing through the tandem connection, one need not be concerned with the details of internal operations but can deal with global processes directly. We therefore concentrate on formulations for the cumulative waiting time of customers in the system, from which their delays can be readily determined.

Letting t_n denote the arrival epoch of the nth customer to the system ($0 \le t_1 \le t_2 \le \cdots$), it can be easily shown (see, for example, Tembe and Wolff [4]) that the departure epoch of the nth customer from the kth queue of a tandem connection of length M ($k \le M$, $M \ge 1$) can be represented as

$$T_n^{(k)} = \max_{1\le i_1\le i_2\le\cdots\le i_k\le n}\left\{t_{i_1} + \sum_{j=i_1}^{i_2} S_j^{(1)} + \sum_{j=i_2}^{i_3} S_j^{(2)} + \cdots + \sum_{j=i_k}^{n} S_j^{(k)}\right\}. \quad (3)$$

The total amount of time that the nth customer spends waiting in a two-queue tandem connection (the cumulative waiting time in our system) is then clearly

$$A_n = W_n^{(1)} + W_n^{(2)} = T_n^{(2)} - t_n - S_n^{(1)} - S_n^{(2)}. \quad (4)$$

The last equality of (4) can be expanded upon using (3) to yield ($n \ge 1$)

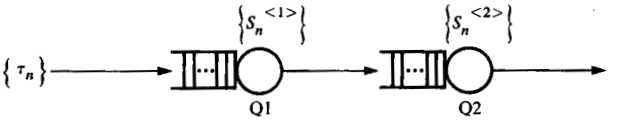

Figure 1 The two-queue tandem connection.

$$A_{n+1} = \left(\max_{1\le r\le n}\left\{\max_{r\le\nu\le n+1}\left[\sum_{r}^{\nu} S_j^{(1)} + \sum_{\nu}^{n+1} S_j^{(2)}\right] - S_{n+1}^{(1)} - S_{n+1}^{(2)} - \sum_{r}^{n}\tau_j\right\}\right)^+, \quad (5)$$

where $\tau_n = t_{n+1} - t_n$ denotes the time between the arrivals of the nth and $(n+1)$th customers (interarrival time) as before.

While the given expression (5) provides an algebraic characterization, the stochastic behavior of the cumulative waiting time process remains difficult to ascertain for general service and interarrival time sequences. Indeed, solutions to problems of this type have been obtained only under very specific assumptions concerning these underlying processes and their interrelationships (Burke [5], Friedman [6], Rubin [7, 8], Boxma [9], Calo [1, 2]). Some additional cases of interest are pursued here.

One particular and quite general result that is related to several previous studies follows quite readily from (5) when the first server is known to "dominate" the system, *i.e.*, $S_n^{(1)} \ge S_m^{(2)}$ (a.s.) for all $n \ge 1$ and $m \ge 1$. In this instance we have that

$$\max_{r\le\nu\le n+1}\left\{\sum_{r}^{\nu} S_j^{(1)} + \sum_{\nu}^{n+1} S_j^{(2)}\right\} = \sum_{r}^{n+1} S_j^{(1)} + S_{n+1}^{(2)},$$

which implies that

$$A_{n+1} = \left(\max_{1\le r\le n}\left\{\sum_{r}^{n}(S_j^{(1)} - \tau_j)\right\}\right)^+ = W_{n+1}^{(1)}. \quad (6)$$

Hence for a front-end dominated tandem connection the cumulative waiting time in the system is simply the waiting time at the first queue (*cf.* Tembe and Wolfe [4]). The waiting time at the second queue will be precisely zero for all customers. We note that this result can be generalized to the case of a tandem connection of any given length, wherein, if the system is front-end dominated, the waiting times at all queues but the first will be precisely zero for all customers (Calo [10]). We note also that no specific assumptions need be made regarding the stochastic nature of the interarrival time or service processes (except, of course, for the dominance relation) for the indicated result (6) to hold. An instance of this very general relationship arises naturally in a subsequent section of this paper.

3. The two-queue, nonuniform message channel

Our principal interest is in tandem connections of a particular type, *viz.*, message channels, for which the successive service times experienced by any particular customer are scaled versions of the same random variable. In the present context this means that if we denote the service sequence at our first queue by $\{S_n\}_{n=1}^{\infty}$, *i.e.*, $S_n^{(1)} = S_n$ for each n, then $S_n^{(2)} = aS_n$ for each n, where a is a positive real constant. When $a = 1$, we have a Uniform Message Channel (UMC) or Repeated-Service Tandem Connection. Such systems have recently been studied, and numerous results have been reported (Boxma [9] deals with a two-queue UMC under Poisson arrivals, and Calo [1, 2, 10, 11] deals with UMCs of any given length $M \geq 1$ under several sets of assumptions). We concentrate here on specific instances of the two-queue, nonuniform message channel described above.

In addition, we assume that the service sequence $\{S_n\}_{n=1}^{\infty}$ consists of independent, identically distributed (i.i.d.) random variables with common distribution function $B(y) = Pr\{S_n \leq y\}$ and mean value $E\{S\} = (1/\mu) < \infty$, the interarrival sequence $\{\tau_n\}_{n=1}^{\infty}$ consists of i.i.d. random variables with common distribution function $\mathcal{A}(y) = Pr\{\tau_n \leq y\}$ and mean value $E\{\tau\} = (1/\lambda) < \infty$, and that these two sequences are mutually independent. Our point of interest is in characterizing the cumulative waiting time process of (5) under this set of assumptions.

The basic recursion that we shall be considering is thus

$$A_1 = 0;$$

$$A_{n+1} = \left(\max_{1 \leq r \leq n} \left\{ \max_{r \leq \nu \leq n+1} \left[\sum_{r}^{\nu} S_j + a \sum_{\nu}^{n+1} S_j \right] - (a+1)S_{n+1} - \sum_{r}^{n} \tau_j \right\} \right)^+ \qquad (n \geq 1). \tag{7}$$

We denote the elements of the sequence of cumulative waiting time distribution functions by $F_{n+1}(x) = Pr\{A_{n+1} \leq x\}$, $n \geq 0$, where, clearly, $F_{n+1}(x) = 0$ for $x < 0$, and, for $x \geq 0$, we have

$$F_{n+1}(x) = \int_{[0,\infty)} dB(y) G_n(x, y), \tag{8}$$

$$G_n(x, y) = Pr\{A_{n+1} \leq x \mid S_{n+1} = y\} \qquad (y \geq 0). \tag{9}$$

Defining the random functions

$$\sigma_{r,n}(y) = \max \left\{ \sum_{r}^{n} S_j, \max_{r \leq \nu \leq n} \left[\sum_{r}^{\nu} S_j + a \sum_{\nu}^{n} S_j \right] - y \right\} \tag{10}$$

and

$$h_{r,n}(y) = \sigma_{r,n}(y) - \sum_{r}^{n} \tau_j, \tag{11}$$

it can be easily established from (7) and (9), by appropriately exploiting the independence properties of the underlying sequences, that

$$G_n(x, y) = Pr \left\{ \max_{1 \leq r \leq n} h_{r,n}(y) \leq x \right\}. \tag{12}$$

These conditional distribution functions play a key role in determining the characteristics of the equilibrium cumulative waiting time, as is presently indicated.

It can be shown that there exists a function $G(x, y)$ such that

$$\lim_{n \to \infty} G_n(x, y) = G(x, y)$$

and

$$F(x) = \lim_{n \to \infty} F_n(x) = \int_{[0,\infty)} dB(y) G(x, y) \tag{13}$$

under appropriate stability conditions (see Appendix 1). Further, from the defining equation for $G_n(x, y)$ we can develop an integral recursion in the following manner:

$$\begin{aligned}
G_{n+1}(x, y) &= Pr \left\{ \max_{1 \leq r \leq n+1} h_{r,n+1}(y) \leq x \right\} \\
&= Pr \left\{ \max \left\{ h_{n+1,n+1}(y), \max_{1 \leq r \leq n} h_{r,n+1}(y) \right\} \leq x \right\} \\
&= Pr \left\{ h_{n+1,n+1}(y) + \left(\max_{1 \leq r \leq n} h_{r,n+1}(y) - h_{n+1,n+1}(y) \right)^+ \leq x \right\} \\
&= Pr \left\{ h_{n+1,n+1}(y) + \left(\max_{1 \leq r \leq n} h_{r,n}(S_{n+1} + (y - aS_{n+1})^+) \right)^+ \leq x \right\} \\
&= Pr \left\{ S_{n+1} + (aS_{n+1} - y)^+ - \tau_{n+1} + \left(\max_{1 \leq r \leq n} h_{r,n}(S_{n+1} + (y - aS_{n+1})^+) \right)^+ \leq x \right\} \\
&= \int_{[0,\infty)} dB(\nu) \cdot \int_{[(\nu + (a\nu - y)^+ - x)^+, \infty)} d\mathcal{A}(\tau) G_n(x + \tau - \nu - (a\nu - y)^+, \nu + (y - a\nu)^+),
\end{aligned}$$

and by applying the Dominated Convergence Theorem to the iterated integrals (on the product space) it can be shown that the limiting conditional distribution function must obey

$$G(x, y) = \int_{[0,\infty)} dB(\nu)$$

$$\cdot \int_{[(v+(av-y)^+-x)^+,\infty)} d\mathcal{A}(\tau)G(x+\tau-v-(av-y)^+, \quad v+(y-av)^+). \tag{14}$$

We have then an integral equation which along with (13) determines the equilibrium cumulative waiting time distribution. If we consider the case $a = 1$, *i.e.*, a two-queue uniform message channel, we obtain the appropriate form of the integral equation developed in [2]. The solution procedures for the uniform case, however, are not directly applicable to the more general, nonuniform, message channel characterized by (14). Indeed, we have as yet not been successful in developing methods of solving such equations except under very specific assumptions regarding the underlying interarrival and service processes.

An alternate form of Eq. (14) which lends itself more readily to analysis arises when we consider systems that are rear-end dominated. For this instance $S_n^{(2)} \geq S_m^{(1)}$ for all $n \geq 1$ and $m \geq 1$ so that

$$\max_{r\leq v\leq n+1}\left\{\sum_r^v S_j^{(1)} + \sum_v^{n+1} S_j^{(2)}\right\} = \sum_r^{n+1} S_j^{(2)} + S_r^{(1)},$$

and thus from (5) the elements of the cumulative waiting time process can be written ($n \geq 1$)

$$A_{n+1} = \left(\max_{1\leq r\leq n}\left\{\sum_r^n (S_j^{(2)} - \tau_j) + S_r^{(1)} - S_{n+1}^{(1)}\right\}\right)^+. \tag{15}$$

Incorporating our assumptions concerning the interarrival and service processes, and following the same procedure as in the development of (14) but now instead of (11) using

$$h_{r,n}^*(y) = \sum_r^n (aS_j - \tau_j) + S_r - y, \tag{16}$$

we can establish that the limiting conditional distribution function must obey

$$G^*(x, y) = \int_{[0,\infty)} dB(v) \cdot \int_{[(v(a+1)-y-x)^+,\infty)} d\mathcal{A}(\tau)G^*(x+\tau-(a+1)v+y, v) \tag{17}$$

for rear-end dominated tandem connections. We note that informally (17) can be obtained from (14) by simply asserting that $av \geq y$ over all appropriate regions of the probability space. We pursue later the analysis of these equations by the use of Laplace-Stieltjes transforms.

4. Poisson arrivals

When the additional assumption is imposed that the arrivals to the system constitute a Poisson process, further progress can be made in solving the integral equations of the previous section. In particular, expressions for the Laplace-Stieltjes Transform (LST) of the distribution function of the equilibrium cumulative waiting time can be obtained that in certain instances yield explicit results. Thus, as in the analysis of uniform message channels or even standard GI/G/1 queues, the memoryless nature of the negative-exponential interarrival time distribution greatly simplifies the appropriate analytical methods.

If we let $\Omega(z)$ denote the Laplace-Stieltjes transform of the equilibrium cumulative waiting time distribution and define

$$H(z, y) = \int_{[0,\infty)} e^{-zx}G(x, y)dx, \tag{18}$$

then it follows from (13) that

$$\Omega(z) = z\int_{[0,\infty)} dB(y)H(z, y). \tag{19}$$

By transforming both sides of the integral equation of (14), identifying the interarrival time distribution explicitly as $\mathcal{A}(\tau) = 1 - e^{-\lambda\tau}$ ($\tau \geq 0$), and using (18) we obtain

$$(z-\lambda)H(z, y) = \lambda\int_{[0,\infty)} dB(v)e^{-\lambda[v+(av-y)^+]}H(\lambda, v+(y-av)^+) - \lambda\int_{[0,\infty)} dB(v)e^{-z[v+(av-y)^+]}H(z, v+(y-av)^+). \tag{20}$$

This not particularly transparent relationship forms the basis of our subsequent analysis.

If we consider for the moment the special case of a rear-end dominated system, then from an appropriate transformation of (17) we obtain the somewhat simpler version below:

$$(z-\lambda)H(z, y) = \lambda e^{\lambda y}\int_{[0,\infty)} dB(v)e^{-\lambda(a+1)v}H(\lambda, v) - \lambda e^{zy}\int_{[0,\infty)} dB(v)e^{-z(a+1)v}H(z, v). \tag{21}$$

An explicit solution for $\Omega(z)$ is now most readily obtained by introducing an additional transformation of the form

$$\Gamma(z, s) = \int_{[0,\infty)} dB(y)e^{-sy}H(z, y), \tag{22}$$

so that from (19)

$$\Omega(z) = z\Gamma(z, 0). \tag{23}$$

Transforming both sides of (21) then according to (22) we have

$$(z - \lambda)\Gamma(z, s) = \lambda\eta(s - \lambda)\Gamma(\lambda, \lambda(a + 1)) - \lambda\eta(s - z)\Gamma(z, z(a + 1)), \quad (24)$$

where

$$\eta(z) = \int_{[0,\infty)} dB(y)e^{-zy} \quad (25)$$

represents the LST of the service time distribution. Since (24) holds for general values of s (provided the given transforms remain well defined), it holds for $s = z(a + 1)$, in which case we must have

$$\Gamma(z, z(a + 1)) = \frac{\lambda\eta(z(a + 1) - \lambda)}{D_a(z)} \Gamma(\lambda, \lambda(a + 1)),$$

where

$$D_a(z) = z - \lambda + \lambda\eta(az). \quad (26)$$

This then determines $\Gamma(z, s)$ as

$$\Gamma(z, s) = \frac{\lambda}{z - \lambda} \cdot \left[\eta(s - \lambda) - \eta(s - z)\frac{\lambda\eta(z(a + 1) - \lambda)}{D_a(z)}\right]\Gamma(\lambda, \lambda(a + 1)) \quad (27)$$

and allows us to write $\Omega(z)$ from (23) as

$$\Omega(z) = \frac{z\Gamma(\lambda, \lambda(a + 1))}{D_a(z)}\frac{\lambda}{z - \lambda} \cdot \{D_a(z)\eta(-\lambda) - \eta(-z)\lambda\eta(z(a + 1) - \lambda)\} \quad (28)$$

for all values of z for which the relevant transforms exist. Noting that $\Omega(z)$ must obey the boundary condition $\Omega(0) = 1$, and applying this fact to (28) by taking the limit as z goes to zero, we find that

$$\Gamma(\lambda, \lambda(a + 1)) = \frac{(1 - a\rho)}{\lambda\eta(-\lambda)},$$

where $\rho = (\lambda/\mu)$ denotes the traffic intensity at the first queue of our tandem connection. This finally determines the LST of the equilibrium cumulative waiting time distribution as

$$\Omega(z) = \Omega_a(z)\frac{D_a(z)\eta(-\lambda) - \eta(-z)\lambda\eta(z(a + 1) - \lambda)}{(z - \lambda)\eta(-\lambda)}, \quad (29)$$

where

$$\Omega_a(z) = \frac{z(1 - a\rho)}{D_a(z)} \quad (30)$$

represents the LST of the equilibrium waiting time distribution in a standard M/G/1 queue with the same service time distribution as our dominant server (second queue).

Equation (29) provides the solution that we had sought in a quite usable form. We can, for example, differentiate (29) and take the limit as z goes to zero to obtain the equilibrium mean cumulative waiting time as

$$E\{A_\infty\} = \frac{\lambda a^2 E\{S^2\}}{2(1 - a\rho)} + (a + 1)\left[\frac{-\eta'(-\lambda) - \eta(-\lambda)E\{S\}}{\eta(-\lambda)}\right], \quad (31)$$

where

$$\eta'(-\lambda) = -\int_{[0,\infty)} ye^{\lambda y}dB(y).$$

The equilibrium mean sojourn time in the system (end-to-end delay) is then simply

$$E\{J_\infty\} = \frac{\lambda a^2 E\{S^2\}}{2(1 - a\rho)} + (a + 1)\frac{-\eta'(-\lambda)}{\eta(-\lambda)}. \quad (32)$$

The equilibrium mean waiting times or sojourn times at each of the individual queues of the connection are also easily derivable from the above as are other parameters of interest.

We have thus obtained a fairly complete characterization of the operation of a rear-end dominated system faced with Poisson arrivals without having to make any assumptions concerning the service time distribution. Unfortunately, the same techniques of solution are not as effective in dealing with more general cases.

5. Analysis of the simplified model

In the previous sections of this paper we have presented various results at various levels of generality concerning the behavior of two-queue tandem connections. Now we consider the analysis of a specific model in which we incorporate the assumptions of previous sections along with a further assumption concerning the nature of the service process. Much of our prior effort is directly applicable, but quite a bit of additional work is also required. It is a characteristic of these types of queueing problems that they do not yield easily to analysis.

As before, our model consists of two single-server queues in tandem subject to a Poisson arrival stream at the first queue and providing service according to scaled versions of the same sequence of random variables. We now specifically assume as well that the service time distribution function is that of a two-level discrete random variable, *i.e.*, S_n equals b_1 with probability p_1 or b_2 with probability $p_2 = 1 - p_1$, where for convenience we require that $0 \le b_1 \le b_2 < \infty$. The mean value of the elements of the service sequence will thus obey

$$E\{S_n\} = p_1b_1 + p_2b_2 = 1/\mu \quad (33)$$

according to our definitions.

The Laplace-Stieltjes transform of the distribution function of the equilibrium cumulative waiting time for messages in the system in question can be written as

$$\Omega(z) = z\{p_1 H_1(z) + p_2 H_2(z)\} \tag{34}$$

following Eq. (19), where the notation $H_j(z) = H(z, b_j)$, $j = 1, 2$, has been used for convenience. Relationships involving these transforms can be developed from (20), which under our particular service time assumptions becomes

$$\begin{aligned}(z - \lambda)H(z, y) = {} & \lambda p_1 e^{-\lambda[b_1 + (ab_1 - y)^+]} H(\lambda, b_1 + (y - ab_1)^+) \\ & + \lambda p_2 e^{-\lambda[b_2 + (ab_2 - y)^+]} H(\lambda, b_2 + (y - ab_2)^+) \\ & - \lambda p_1 e^{-z[b_1 + (ab_1 - y)^+]} H(z, b_1 + (y - ab_1)^+) \\ & - \lambda p_2 e^{-z[b_2 + (ab_2 - y)^+]} H(z, b_2 + (y - ab_2)^+).\end{aligned} \tag{35}$$

The precise form that this equation takes in any given situation obviously depends upon relationships among the parameter y, the scaling factor a, and the service levels b_1 and b_2. Several different cases must be examined individually.

If $a \leq 1$, *i.e.*, the first queue gives longer service times to each customer than the second, then (10) can be conveniently rewritten as

$$\sigma_{r,n}(y) = \sum_{r}^{n} S_j + \left(\max_{r \leq \nu \leq n} \left\{ S_\nu - y - (1 - a) \sum_{\nu}^{n} S_j \right\} \right)^+.$$

The second term in the above formulation can be easily shown to obey the following inequalities:

$$\begin{aligned}&\left(\max_{r \leq \nu \leq n} \left\{ S_\nu - y - (1 - a) \sum_{\nu}^{n} S_j \right\} \right)^+ \\ &\quad \leq \left(\max_{r \leq \nu \leq n} \{ aS_\nu - y \} \right)^+ \leq (ab_2 - y)^+,\end{aligned}$$

the latter resulting from the fact that for every n, $S_n \leq b_2$. Hence, for $y \geq ab_2$, we must have that

$$\sigma_{r,n}(y) = \sum_{r}^{n} S_j,$$

which in turn implies from (11) and (12) that

$$G_n(x, y) = Pr\{W_{n+1}^{(1)} \leq x\}.$$

The above indicates that, conditioned on sufficiently large services being given in the first queue, the waiting time at the second queue will be zero. Therefore, in this specific region of the underlying probability space, the conditional distribution function sequence will converge to the distribution function of the waiting time in the first queue (which under our assumptions is an instance of a standard M/G/1 system). From (18) then, we can now write

$$H(z, y) = \frac{(1 - \rho)}{D^*(z)} = H^*(z) \qquad (y \geq ab_2), \tag{36}$$

where

$$D^*(z) = z - \lambda + \lambda p_1 e^{-zb_1} + \lambda p_2 e^{-zb_2}, \tag{37}$$

and

$$\rho = \lambda E\{S\} = \lambda p_1 b_1 + \lambda p_2 b_2 \tag{38}$$

is simply the traffic intensity at the first queue. Equation (36) follows directly from classical results for the LST of the equilibrium waiting time in a standard M/G/1 queue (see, for example, Kleinrock [12]).

Using the above development and (35), we can determine that for $a \leq 1$,

$$H_2(z) = H^*(z) \tag{39}$$

and

$$\begin{aligned}(z - \lambda)H_1(z) = {} & \lambda p_1 e^{-\lambda b_1} H(\lambda, b_1[2 - a]) \\ & - \lambda p_1 e^{-zb_1} H(z, b_1[2 - a]) \\ & + \lambda p_2 e^{-\lambda[b_2 + (ab_2 - b_1)^+]} H^*(\lambda) \\ & - \lambda p_2 e^{-z[b_2 + (ab_2 - b_1)^+]} H^*(z).\end{aligned} \tag{40}$$

This latter relation (40) cannot yet be used to establish an explicit expression for $H_1(z)$. The effects on our analysis of the size of the scaling factor, a, with respect to the sizes of the service levels, b_1 and b_2, must be considered in some detail before we can proceed further.

Clearly, the support of the random variable S_n is the closed interval $[b_1, b_2]$, while the support of the random variable aS_n is the closed interval $[ab_1, ab_2]$. Hence, if our scaling factor, a, has a value in the interval $[0, b_1/b_2]$, the system is front-end dominated and readily solved. For this situation, then, the LST we have been seeking is simply [refer to (6)]

$$\Omega(z) = \Omega_1(z) = \frac{z(1 - \rho)}{D^*(z)}, \qquad a \in [0, b_1/b_2] \tag{41}$$

with concomitant mean value

$$E\{A_\infty\} = E\{W_\infty^{(1)}\} = \frac{\lambda E\{S^2\}}{2(1 - \rho)}, \qquad a \in [0, b_1/b_2]. \tag{42}$$

The second moment of the service time distribution is, of course, $E\{S^2\} = p_1 b_1^2 + p_2 b_2^2$ under our assumptions.

If we now consider the system scaling factor to have a value such that

$$a \in \left(\frac{b_1 + Nb_1}{b_2 + Nb_1} , \frac{b_1 + (N+1)b_1}{b_2 + (N+1)b_1} \right] \tag{43}$$

for some nonnegative integer N, and we define the indexed functions

$$h_n(z) = H(z, y_n), \tag{44}$$

where $y_n = b_1 + nb_1(1 - a)$, then we have from (35) that

$$\begin{aligned}(z - \lambda)h_n(z) = {} & \lambda p_1 e^{-\lambda b_1} h_{n+1}(\lambda) + \lambda p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - y_n)} H^*(\lambda) \\ & - \lambda p_1 e^{-zb_1} h_{n+1}(z) \\ & - \lambda p_2 e^{-zab_2} e^{-z(b_2 - y_n)} H^*(z)\end{aligned} \tag{45}$$

for $0 \le n \le N$ and

$$h_n(z) = H^*(z) \tag{46}$$

for $n \ge (N + 1)$. We have thus constructed a recursion with a known boundary condition that can theoretically be solved quite straightforwardly for $h_0(z)$, which by definition equals the desired $H_1(z)$, for a in any of a collection of disjoint intervals whose union is the interval $(b_1/b_2, 1)$. The difficulty with this procedure is that the closer the scaling factor gets to the value "one," the larger the number of elements in the finite recursion. The procedure can thus become quite tedious.

For small values of N, however, the calculation is quite readily performed. For example, with $N = 0$, *i.e.*,

$$a \in \left(\frac{b_1}{b_2} , \frac{2b_1}{b_1 + b_2} \right],$$

it follows from (46) that $h_1(z) = H^*(z)$, so that (45) immediately yields

$$\begin{aligned}(z - \lambda)h_0(z) = {} & [\lambda p_1 e^{-\lambda b_1} + \lambda p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - b_1)}] H^*(\lambda) \\ & - [\lambda p_1 e^{-zb_1} + \lambda p_2 e^{-zab_2} e^{-z(b_2 - b_1)}] H^*(z),\end{aligned}$$

which in turn establishes from (34) that

$$\begin{aligned}\Omega(z) = \Omega_1(z) \Bigg\{ & p_2 + \frac{p_1}{(z - \lambda)} \\ & \left\{ \frac{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - b_1)}}{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}} D^*(z) \right. \\ & \left. - \lambda [p_1 e^{-zb_1} + p_2 e^{-zab_2} e^{-z(b_2 - b_1)}] \right\} \Bigg\},\end{aligned} \tag{47}$$

where $\Omega_1(z)$ is the LST of the distribution of the equilibrium waiting time at the first queue, as in (41). The equilibrium mean cumulative waiting time can then be determined from (47) as

$$\begin{aligned}E\{A_\infty\} = {} & E\{W_\infty^{(1)}\} \\ & + p_1 \left\{ (ab_2 - b_1)p_2 \right. \\ & \left. - \frac{(1 - \rho)}{\lambda} \frac{p_2 e^{-\lambda b_2} [1 - e^{-\lambda(ab_2 - b_1)}]}{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}} \right\},\end{aligned} \tag{48}$$

where $E\{W_\infty^{(1)}\}$ is the mean waiting time at the first queue, as in (42). A more general development is given in Appendix 2, where an explicit solution is also exhibited for the case $N = 1$, and this is already seen to be somewhat complicated algebraically.

If $a \ge 1$, a similar series of considerations must be made concerning the relative values of the scaling factor and the service levels. With $a \in [b_2/b_1, \infty)$ the system is rear-end dominated and the more general results of Section 4 apply, but with $B(y) = p_1 U(y - b_1) + p_2 U(y - b_2)$, where $U(x)$ denotes the unit step function. Hence, from (29) we have

$$\begin{aligned}\Omega(z) = {} & \Omega_a(z) \frac{1}{z - \lambda} \\ & \cdot \left\{ D_a(z) - \lambda \frac{p_1 e^{zb_1} + p_2 e^{zb_2}}{p_1 e^{\lambda b_1} + p_2 e^{\lambda b_2}} \right. \\ & \left. \cdot [p_1 e^{\lambda b_1} e^{-z(a+1)b_2} + p_2 e^{\lambda b_2} e^{-z(a+1)b_2}] \right\},\end{aligned} \tag{49}$$

where $\Omega_a(z)$ has been previously defined in (30); and, from (31), the equilibrium mean cumulative waiting time becomes

$$\begin{aligned}E\{A_\infty\} = {} & \frac{\lambda a^2 E\{S^2\}}{2(1 - a\rho)} \\ & + (a + 1)(b_2 - b_1) \frac{p_1 p_2 [1 - e^{-\lambda(b_2 - b_1)}]}{p_2 + p_1 e^{-\lambda(b_2 - b_1)}} .\end{aligned} \tag{50}$$

We note that the above, particularly Eq. (50), can be easily renormalized so that in effect the service process at the second queue is taken as $\{S_n\}_{n=1}^{\infty}$, while the service process at the first queue is $\{(1/a)S_n\}_{n=1}^{\infty}$, an operation that is often convenient when comparing different systems. The equilibrium mean value, for instance, then becomes

$$\begin{aligned}E\{A_\infty\} = {} & \frac{\lambda E\{S^2\}}{2(1 - \rho)} \\ & + \left(1 + \frac{1}{a}\right)(b_2 - b_1) \frac{p_1 p_2 [1 - e^{-\frac{\lambda}{a}(b_2 - b_1)}]}{p_2 + p_1 e^{-\frac{\lambda}{a}(b_2 - b_1)}} ,\end{aligned} \tag{51}$$

and this expression can now be compared with (42), which can be interpreted as representing the same system but with the positions of the two servers reversed.

If we now consider the system scaling factor to have a value such that

$$a \in \left[\frac{b_2 + (N+1)b_1}{b_1 + (N+1)b_1}, \frac{b_2 + Nb_1}{b_1 + Nb_1}\right) \tag{52}$$

for some nonnegative integer N, and we define the indexed functions

$$h_n(z) = H(z, y_n), \tag{53}$$

where now, however, $y_{n+1} = b_1 + (b_2 - ab_1 - n(a-1)b_1)^+$ $(n \geq 0)$ and $y_0 = b_2$, then we have from (35) that

$$\begin{aligned}(z-\lambda)h_n(z) = {} & \lambda p_1 e^{-\lambda a b_1} e^{\lambda(y_n - y_{n+1})} h_{n+1}(\lambda) \\ & + \lambda p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - y_n)} H_2(\lambda) \\ & - \lambda p_1 e^{-z a b_1} e^{z(y_n - y_{n+1})} h_{n+1}(z) \\ & - \lambda p_2 e^{-z a b_2} e^{-z(b_2 - y_n)} H_2(z)\end{aligned} \tag{54}$$

for $0 \leq n \leq (N+1)$; also $h_0(z) = H_2(z)$, and

$$h_n(z) = H_1(z) \tag{55}$$

for $n \geq N+2$. This recursion thus involves both $H_1(z)$ and $H_2(z)$ as boundary conditions, and indeed must be solved for both transforms. Hence its analysis is somewhat more complex than that of the previously considered recursion of (45). Nevertheless, theoretically we can obtain explicit expressions for the relevant transforms for the scaling factor, a, in any of a collection of disjoint intervals whose union is the interval $(1, b_2/b_1)$. The difficulty as before is that the closer the scaling factor approaches unity the more complex the analysis.

A general development of a solution procedure is given in Appendix 3, where explicit expressions are also derived for the case $N = 0$ for both the Laplace-Stieltjes transform, $\Omega(z)$, and the equilibrium mean value, $E\{A_\infty\}$. These are seen to be considerably more complicated than the complementary results for the regions where $a < 1$. We note that, as before, equations like (A47) can be easily renormalized so that they are referenced to the service time of the more dominant server (in this case the second). Solutions for larger values of N can be developed in much the same way, but no illuminating results have as yet been obtained by doing so. The complexities of the resulting equations tend to mask any evolutionary factors that might be exploited in projecting the solution closer to its limiting form at the value $a = 1$.

In Table 1 we present a compendium of results concerning the manner in which the equilibrium mean cumulative waiting time depends upon the scaling factor (a). Each of the indicated equations has been developed above (or in the accompanying Appendices) except for the one describing the uniform case $(a = 1)$, which is easily obtainable from results in [2]. All the equations have also been normalized so that service times are taken with respect to the more dominant server; hence, the first two describe a two-queue tandem connection with $S_n^{\langle 1 \rangle} = S_n$ and $S_n^{\langle 2 \rangle} = aS_n$ $(a < 1)$, while the last two describe such a system with $S_n^{\langle 1 \rangle} = (1/a)S_n$ and $S_n^{\langle 2 \rangle} = S_n$ $(a > 1)$, using the notation of Section 2. These pairs of equations as exhibited are thus complementary and can be used to compare instances of the same system but with the positions of the two servers reversed.

In order to provide more concrete performance comparisons, the equilibrium mean cumulative waiting time was evaluated numerically by computer program for a particular example of interest. The results of such a calculation for differing values of the service time scaling factor are shown in Table 2. For the purposes of this example, the parameters of the service time distribution have been chosen so that it will have the same first three moments as a negative-exponential distribution with parameter μ. In particular this determines p_1, p_2, b_1, and b_2 as indicated below:

$$p_1 = \frac{2 + \sqrt{2}}{4}\ ;\ p_2 = \frac{2 - \sqrt{2}}{4}\ ;$$

$$b_1 = \frac{1}{\mu}\left[2 - \sqrt{2}\right];\ b_2 = \frac{1}{\mu}\left[2 + \sqrt{2}\right].$$

The table compares the normalized mean wait $\mu E\{A_\infty\}$ as a function of the traffic intensity ρ for five values of the scaling factor a, each in a different performance subinterval of its range.

As can be seen from the table, the waiting time is smallest for the front-end dominated system. Also, for any given scaling factor, the waiting time is less when the first server tends to give longer service times than the second, than for the complementary case (*e.g.*, compare the columns for $a = 1/4$ and $a^{-1} = 1/4$ in Table 2). The maximum waiting time occurs for the uniform case, with the waiting time decreasing monotonically as the system becomes either more front-end or more rear-end dominated. The relative difference between the minimum (front-end dominated) and maximum (uniform) values, while quite significant for low traffic intensities, decreases substantially as the traffic intensity increases. Hence the value of the scaling factor tends to become less and less important as the system tends toward instability, as one would expect.

The same general types of behavior exhibited by the equilibrium mean cumulative waiting time in the above example would probably apply in systems with more complex service time distributions. These remain, however, quite difficult to analyze directly. We note that in principle, once we have the integral equation of (14) and

Table 1 Equilibrium mean cumulative waiting time equations.

Front-end dominated $\quad a \in \left[0, \dfrac{b_1}{b_2}\right]$

$$E\{A_\infty\} = \frac{\lambda[p_1 b_1^2 + p_2 b_2^2]}{2(1-\rho)}$$

Near front-end dominated $\quad a \in \left(\dfrac{b_1}{b_2}, \dfrac{2b_1}{b_1+b_2}\right]$

$$E\{A_\infty\} = \frac{\lambda[p_1 b_1^2 + p_2 b_2^2]}{2(1-\rho)} + p_1 p_2 \left\{(ab_2 - b_1) - \frac{(1-\rho)}{\lambda} \frac{e^{-\lambda(b_2-b_1)}[1 - e^{-\lambda(ab_2-b_1)}]}{p_1 + p_2 e^{-\lambda(b_2-b_1)}}\right\}$$

Uniform $\quad a = 1$

$$E\{A_\infty\} = \frac{\lambda[p_1 b_1^2 + p_2 b_2^2]}{2(1-\rho)} + p_1 \left\{(b_2 - b_1) - \frac{(1-\rho)}{\lambda p_2}[1 - e^{-w_1(b_2-b_1)}]\right\}$$

$$w_1 = \lambda[1 - p_1 e^{-w_1 b_1}]$$

Near rear-end dominated $\quad a^{-1} \in \left(\dfrac{b_1}{b_2}, \dfrac{2b_1}{b_1+b_2}\right]$

$$E\{A_\infty\} = \frac{\lambda[p_1 b_1^2 + p_2 b_2^2]}{2(1-\rho)} + p_1 p_2 \left\{(1 + a^{-1})(b_2 - b_1)\right.$$

$$\cdot \frac{e^{\lambda a^{-1}(b_2-b_1)}[1 - \lambda a^{-1} p_1 e^{-\lambda b_1}(b_2 - ab_1)] + p_1[e^{\lambda a^{-1}(b_2-ab_1)} - 1] - 1}{p_1[p_1 + p_2 e^{\lambda a^{-1}(b_2-ab_1)}] + p_2 e^{\lambda a^{-1}(b_2-b_1)}[1 - \lambda a^{-1} p_1 e^{-\lambda b_1}(b_2 - ab)]}$$

$$\left. + \frac{p_1[2\lambda^{-1} - (b_2 + a^{-1}b_1)][e^{\lambda a^{-1}(b_2-ab_1)} - 1] - a^{-1} p_1 (b_2 - ab_1)}{p_1[p_1 + p_2 e^{\lambda a^{-1}(b_2-ab_1)}] + p_2 e^{\lambda a^{-1}(b_2-b_1)}[1 - \lambda a^{-1} p_1 e^{-\lambda b_1}(b_2 - ab_1)]}\right\}$$

Rear-end dominated $\quad a^{-1} \in \left(0, \dfrac{b_1}{b_2}\right]$

$$E\{A_\infty\} = \frac{\lambda[p_1 b_1^2 + p_2 b_2^2]}{2(1-\rho)} + p_1 p_2 \left\{(1 + a^{-1})(b_2 - b_1) \frac{[1 - e^{-\lambda a^{-1}(b_2-b_1)}]}{p_2 + p_1 e^{-\lambda a^{-1}(b_2-b_1)}}\right\}$$

Table 2 Mean wait for different scaling factors.

ρ	$\left(a \le 3 - 2\sqrt{2}\right)$	$\left(a = \frac{1}{4}\right)$	$(a = 1)$	$\left(a^{-1} = \frac{1}{4}\right)$	$\left(a^{-1} = \frac{1}{8}\right)$
0.1	0.11111	0.12134	0.28784	0.16199	0.12535
0.2	0.25000	0.26766	0.61611	0.32475	0.27883
0.3	0.42857	0.45152	0.99763	0.52881	0.47236
0.4	0.66667	0.69332	1.45312	0.79405	0.72577
0.5	1.00000	1.02919	2.01907	1.15624	1.07479
0.6	1.50000	1.53090	2.76745	1.68682	1.59085
0.7	2.33333	2.36536	3.86516	2.55251	2.44061
0.8	4.00000	4.03275	5.81190	4.25332	4.12408
0.9	9.00000	9.03320	11.10666	9.28927	9.14126

the concomitant transform equation of (20) for systems with Poisson arrivals, the assumption of any discrete service time distribution with a finite support will lead to equations like (35) except with correspondingly more than two levels represented. Such equations could then be solved in subintervals determined by relationships among the values of the scaling factor (a) and the allowable service times ($b_1, b_2, \cdots, b_L$ for a discrete distribution with L levels) in much that same way as has been demonstrated for the simplest, nontrivial such instance ($L = 2$). The degree of complexity of the problem, however, increases markedly as L increases.

Concluding remarks

In this paper we have presented a number of results concerning the analysis of two-queue, tandem connections. Some have been quite general, but most have dealt with message channels that incorporate discrete service time distributions in their structure and are subject to Poisson arrivals. Solution procedures have been developed in some detail for the particular case of such a system with but two service levels. This type of service time distribution arises in models for packet-switching systems that support two packet classes—one for interactive traffic (short) and the other for file transfers (long), for example.

While much of the work done here seems difficult to generalize further, it does provide a basis for a firmer understanding of the properties of nonuniform message channels. Indeed, to this author's knowledge, these efforts represent the first analytical characterization of the performance of any member of this class of nonstandard queueing models. There remain, therefore, many interesting problems to consider in subsequent studies.

Appendix 1: Convergence considerations

From their definitions as probabilities the functions $G_n(x, y)$ are obviously bounded

$$0 \le G_n(x, y) \le 1 \qquad \begin{matrix} n \ge 0 \\ x \in [0, \infty) \\ y \in [0, \infty). \end{matrix} \tag{A1}$$

They are also monotone nonincreasing in n for each (x, y) as we proceed to show. With

$$\sigma_{r,n}(y) = \max\left\{\sum_{r}^{n} S_j, \max_{r \le v \le n}\left[\sum_{r}^{v} S_j + a\sum_{v}^{n} S_j\right] - y\right\}$$

and

$$h_{r,n}(y) = \sigma_{r,n}(y) - \sum_{r}^{n} \tau_j,$$

we have that

$$G_{n+1}(x, y) = P_r\left\{\max_{1 \le r \le n+1} h_{r,n+1}(y) \le x\right\}. \tag{A2}$$

Consider for $n > 1$

$$\max_{1 \le r \le n+1} h_{r,n+1}(y) = \max\left\{h_{1,n+1}(y), \max_{2 \le r \le n+1} h_{r,n+1}(y)\right\}$$

$$\ge \max_{2 \le r \le n+1} h_{r,n+1}(y)$$

but, because of the independence of the underlying random variables,

$$\max_{2 \le r \le n+1} h_{r,n+1}(y) \sim \max_{1 \le r \le n} h_{r,n}(y),$$

i.e., they are identically distributed. This then directly implies that for each (x, y) and $n \ge 0$

$$G_{n+1}(x, y) \le G_n(x, y). \tag{A3}$$

We also have from the defining equations that

$$G_n(\infty, y) = 1,$$

$$G_n(x, \infty) = F^{\langle 1 \rangle}_{n+1}(x) = Pr\{W^{\langle 1 \rangle}_{n+1} \le x\}, \tag{A4}$$

where we note that $F^{\langle 1 \rangle}_{n+1}(x)$ is an element from the distribution function sequence of a standard GI/G/1 queue.

Since for every pair of nonnegative real numbers (x, y) the sequence $\{G_n(x, y)\}_{n=0}^{\infty}$ is bounded and monotone, it must converge. Hence there exists a $G(x, y)$ such that

$$\lim_{n \to \infty} G_n(x, y) = G(x, y)$$

for every pair (x, y). Further, since the $G_n(x, y)$ are bounded, we can apply the Dominated Convergence Theorem to obtain

$$F(x) = \lim_{n \to \infty} F_{n+1}(x) = \lim_{n \to \infty} \int_{[0,\infty)} dB(y) G_n(x, y)$$

$$= \int_{[0,\infty)} dB(y) G(x, y). \tag{A5}$$

While we now have an expression for the equilibrium cumulative waiting time distribution function in terms of the limiting conditional distribution function, we note that the condition

$$\max(1, a) E\{S\} < E\{\tau\} \tag{A6}$$

must be imposed for Eq. (A5) to be meaningful. For this case the system is stable (Loynes [3]), and the cumulative waiting time process will converge honestly in distribution (independently of initial conditions).

Appendix 2: Solution in subintervals of [0, 1)

If we consider the recursion given as (45) and introduce some subsidiary notation by defining

$$\varepsilon_n(\lambda) = \lambda p_1 e^{-\lambda b_1} h_{n+1}(\lambda) + \lambda p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - y_n)} H^*(\lambda), \tag{A7}$$

along with

$$\beta(z) = \frac{\lambda p_2 e^{-zb_1}}{z - \lambda} \tag{A8}$$

and

$$\alpha_n(z) = \frac{\lambda p_2 e^{-zab_2}}{z - \lambda} e^{-z(b_2 - b_1)} e^{zb_1(1-a)n}, \tag{A9}$$

we have that for $0 \le n \le N$

$$h_n(z) = (z - \lambda)^{-1} \varepsilon_n(\lambda) - \beta(z) h_{n+1}(z) - \alpha_n(z) H^*(z). \tag{A10}$$

In this form, the recursion admits the obvious solution

$$\begin{aligned} h_n(z) = (z - \lambda)^{-1} &\sum_{k=n}^{N} (-1)^{k-n} \beta^{k-n}(z) \varepsilon_k(\lambda) \\ &- \sum_{k=n}^{N} (-1)^{k-n} \beta^{k-n}(z) \alpha_k(z) H^*(z) \\ &+ (-1)^{N+1-n} \beta^{N+1-n}(z) h_{N+1}(z), \end{aligned} \tag{A11}$$

so that, in particular,

$$\begin{aligned} H_1(z) = (z - \lambda)^{-1} &\sum_{k=0}^{N} (-1)^k \beta^k(z) \varepsilon_k(\lambda) + H^*(z) \\ &\cdot \left\{ (-1)^{N+1} \beta^{N+1}(z) - \sum_{k=0}^{N} (-1)^k \beta^k(z) \alpha_k(z) \right\} \end{aligned} \tag{A12}$$

follows from (A11) by simply noting that $H_1(z) = h_0(z)$ and $h_{N+1}(z) = H^*(z)$ as indicated in (44) and (46). Now, using (34), (36), (39), (A12) and a considerable amount of algebra it can be straightforwardly established that

$$\Omega(z) = \Omega_1(z) \frac{Q_N(z)}{(z - \lambda)^{N+1} D_1(z)}, \tag{A13}$$

where

$$\Omega_1(z) = \frac{z(1 - \rho)}{D^*(z)} \tag{A14}$$

as in (41);

$$D^*(z) = z - \lambda + \lambda p_1 e^{-zb_1} + \lambda p_2 e^{-zb_2} \tag{A15}$$

as in (37);

$$D_1(z) = z - \lambda + \lambda p_1 e^{-zab_1} \tag{A16}$$

represents a "partial" transform as in Calo [2]; and

$$\begin{aligned} Q_N(z) = p_1 D_1(z) D^*(z) &\sum_{k=0}^{N} (-\lambda p_1 e^{-zb_1})^k (z - \lambda)^{N-k} \frac{\varepsilon_k(\lambda)}{(1 - \rho)} \\ &+ p_2 (z - \lambda)^{N+1} D_1(z) \\ &+ p_1 (-\lambda p_1 e^{-zb_1})^{N+1} D_1(z) \\ &- \lambda p_1 p_2 e^{-zab_2} e^{-z(b_2 - b_1)} \\ &\cdot [(z - \lambda)^{N+1} - (-\lambda p_1 e^{-zab_1})^{N+1}]. \end{aligned} \tag{A17}$$

We note that (A14)–(A17) establish the desired transform $\Omega(z)$, except for the set of positive real constants $\{\varepsilon_n(\lambda)\}_{n=0}^{N}$, which have yet to be determined.

These $(N + 1)$ unknowns can be obtained by appealing to analyticity arguments for the Laplace-Stieltjes transform $\Omega(z)$. Thus, since the denominator of (A13) has a zero of order $(N + 1)$ at the point $z = \lambda$, it must follow that

$$\lim_{z \to \lambda} \frac{d^\nu}{dz^\nu} Q_N(z) = 0 \qquad (\nu = 0, 1, 2, \cdots, N) \tag{A18}$$

in order for $\Omega(z)$ to remain finite at that point. The above, (A18), represents a system of $(N + 1)$ independent, linear equations in the unknowns $\{\varepsilon_n(\lambda)\}_{n=0}^{N}$ and can therefore theoretically be solved straightforwardly by classical methods. Pragmatically, the calculations involved become increasingly tedious as N increases. We note that (A13) has an additional zero at some value of z that we shall designate as w_1, where $D_1(w_1) = 0$ ($D_1(z)$ can be shown to have a unique real root for $Re(z) \ge 0$ as in Calo [2]). A direct calculation readily yields $Q_N(w_1) = 0$ as required.

For the simplest case, $N = 0$, we have only one unknown constant $\varepsilon_0(\lambda)$ to determine. This follows from the single equation $Q_0(\lambda) = 0$ as

$$\varepsilon_0(\lambda) = \frac{(1 - \rho)}{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}} [p_1 e^{-\lambda b_1} + p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - b_1)}], \tag{A19}$$

and $\Omega(z)$ is then given by

$$\begin{aligned} \Omega(z) = \Omega_1(z) \Bigg\{ p_2 + \frac{p_1}{(z - \lambda)} \Bigg\{ D^*(z) \frac{\varepsilon_0(\lambda)}{(1 - \rho)} \\ - \lambda [p_1 e^{-zb_1} + p_2 e^{-zab_2} e^{-z(b_2 - b_1)}] \Bigg\} \Bigg\} \end{aligned} \tag{A20}$$

as indicated by (A13) and (A17). We note that this of course duplicates the result of (47) obtained earlier.

For the case $N = 1$ the transform of interest, as given by the appropriate expansions of (A13) and (A17), becomes

$$\begin{aligned} \Omega(z) = \Omega_1(z) \Bigg\{ p_2 + \frac{p_1}{(z - \lambda)^2} \\ \cdot \Bigg\{ D^*(z) \left[(z - \lambda) \frac{\varepsilon_0(\lambda)}{(1 - \rho)} - \lambda p_1 e^{-zb_1} \frac{\varepsilon_1(\lambda)}{(1 - \rho)} \right] \end{aligned}$$

$$+ (-\lambda p_1 e^{-zb_1})^2 - \lambda p_2 e^{-zab_2} e^{-z(b_2-b_1)} \cdot \frac{[(z-\lambda)^2 - (-\lambda p_1 e^{-zab_1})^2]}{D_1(z)} \Big\}\Big\}. \quad \text{(A21)}$$

As indicated by (A18) we now have two equations to solve for the two constants $\varepsilon_0(\lambda)$ and $\varepsilon_1(\lambda)$; namely, $Q_1(\lambda) = 0$, and $Q_1'(\lambda) = 0$. The first gives

$$\varepsilon_1(\lambda) = \frac{(1-\rho)}{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}} \cdot [p_1 e^{-\lambda b_1} + p_2 e^{-\lambda a b_2} e^{-\lambda(b_2-b_1)} e^{\lambda(1-a)b_1}], \quad \text{(A22)}$$

and the second gives

$$\varepsilon_0(\lambda) = \frac{(1-\rho)}{p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}} \left\{ e^{-\lambda b_1} \frac{\varepsilon_1(\lambda)}{(1-\rho)} c(\lambda) + d(\lambda) \right\}, \quad \text{(A23)}$$

where

$$c(\lambda) = p_1[1 - \lambda p_1 b_1 e^{-\lambda b_1} - \lambda p_2 b_2 e^{-\lambda b_2}] - [p_1 e^{-\lambda b_1} + p_2 e^{-\lambda b_2}][\lambda p_1 b_1 (a+1) - e^{\lambda a b_1}]$$

and

$$d(\lambda) = p_1\{e^{\lambda(a-2)b_1} - \lambda p_1 b_1 (a+2) e^{-2\lambda b_1} - \lambda p_2 [ab_2 + b_2 - b_1 + 2ab_1] e^{-\lambda[ab_2 + b_2 - b_1 + ab_1]}\}.$$

The constant $\varepsilon_0(\lambda)$ can then be explicitly obtained by incorporating (A22) into (A23). A rather formidable expression for $\Omega(z)$ finally follows from (A21) by including these values for the constants in that equation.

Appendix 3: Solution in subintervals of (1, ∞)

If we consider the recursion given as (54) and introduce some subsidiary notation by defining

$$\varepsilon_n(\lambda) = \lambda p_1 e^{-\lambda a b_1} e^{\lambda(y_n - y_{n+1})} h_{n+1}(\lambda) + \lambda p_2 e^{-\lambda a b_2} e^{-\lambda(b_2 - y_n)} H_2(\lambda), \quad \text{(A24)}$$

along with

$$\beta(z) = \frac{\lambda p_1 e^{-zb_1}}{z - \lambda} \quad \text{(A25)}$$

and

$$\alpha_n(z) = \frac{\lambda p_2 e^{-zab_2}}{z - \lambda} e^{-z(a-1)b_1 n}, \quad \text{(A26)}$$

we have that for $0 \le n \le N$

$$h_n(z) = (z-\lambda)^{-1}\varepsilon_n(\lambda) - \beta(z) h_{n+1}(z) - \alpha_n(z) H_2(z). \quad \text{(A27)}$$

In this form, the recursion admits the obvious solution

$$h_n(z) = (z-\lambda)^{-1} \sum_{k=n}^{N} (-1)^{k-n} \beta^{k-n}(z) \varepsilon_k(\lambda) - \sum_{k=n}^{N} (-1)^{k-n} \beta^{k-n}(z) \alpha_k(z) H_2(z) + (-1)^{N+1-n} \beta^{N+1-n}(z) h_{N+1}(z), \quad \text{(A28)}$$

so that, in particular,

$$H_2(z) = (z-\lambda)^{-1} \sum_{k=0}^{N} (-1)^k \beta^k(z) \varepsilon_k(\lambda) - H_2(z) \sum_{k=0}^{N} (-1)^k \beta^k(z) \alpha_k(z) + (-1)^{N+1} \beta^{N+1}(z) h_{N+1}(z) \quad \text{(A29)}$$

follows from (A28) by simply noting that $H_2(z) = h_0(z)$ as indicated just before (55). Now, from (54) and (A24) with $n = N+1$ we can obtain

$$h_{N+1}(z) = (z-\lambda)^{-1}\varepsilon_{N+1}(\lambda) - H_1(z)\lambda p_1 e^{-zab_1} e^{z(b_2-b_1)} e^{-z(a-1)b_1(N+1)} - H_2(z)\lambda p_2 e^{-zab_2} e^{-z(a-1)b_1(N+1)} \quad \text{(A30)}$$

by using the fact that $h_n(z) = H_1(z)$ for $n \ge N+2$, as indicated by (55). Also, from (54), (A24), and (55), but with $n = N+2$ this time, we get

$$D_1(z) H_1(z) = \varepsilon_{N+2}(\lambda) - \lambda p_2 e^{-zab_2} e^{-z(b_2-b_1)} H_2(z), \quad \text{(A31)}$$

where we have let

$$D_1(z) = z - \lambda + \lambda p_1 e^{-zab_1} \quad \text{(A32)}$$

for notational convenience.

By using (A31) in (A30) we can obtain an expression for $h_{N+1}(z)$ in which the only unknown function is $H_2(z)$. This expression can then be used in turn in (A29) to establish that

$$H_2(z) = \frac{D_1(z)}{(z-\lambda) D_a(z)} \sum_{k=0}^{N+1} (-1)^k \beta^k(z) \varepsilon_k(\lambda) + \left[\frac{-\lambda p_1 e^{-zab_1}}{z-\lambda}\right]^{N+2} \frac{e^{z(b_2-b_1)}}{D_a(z)} \varepsilon_{N+2}(\lambda) \quad \text{(A33)}$$

after some algebraic manipulation, where

$$D_a(z) = z - \lambda + \lambda p_1 e^{-zab_1} + \lambda p_2 e^{-zab_2}, \quad \text{(A34)}$$

as in (26). We can now combine (A31) with (34) in order to get $\Omega(z)$ in terms of $H_2(z)$ only, and then employ (A33) to establish that

$$\Omega(z) = \Omega_a(z) \frac{Q_N(z)}{(z-\lambda)^{N+2} D_1(z)}, \quad \text{(A35)}$$

where

$$\Omega_a(z) = \frac{z(1-a\rho)}{D_a(z)} \quad \text{(A36)}$$

as in (30);

$$Q_N(z) = D_1(z)\phi(z) \sum_{k=0}^{N+1} (z-\lambda)^{N+1-k}(-\lambda p_1 e^{-zb_1})^k \frac{\varepsilon_k(\lambda)}{(1-a\rho)} + \frac{\varepsilon_{N+2}(\lambda)}{(1-a\rho)} \{p_1 D_a(z)(z-\lambda)^{N+2} + \phi(z)[-\lambda p_1 e^{-zab_1}]^{N+2} e^{z(b_2-b_1)}\} \quad \text{(A37)}$$

is the numerator function; and

$$\phi(z) = p_2[D_1(z) - \lambda p_1 e^{-zab_2} e^{-z(b_2-b_1)}] \quad \text{(A38)}$$

has been defined for notational convenience. We note that (A35)–(A38) establish the desired transform $\Omega(z)$, except for the set of positive real constants $\{\varepsilon_n(\lambda)\}_{n=0}^{N+2}$, which have yet to be determined.

These $(N+3)$ unknowns can be obtained by appealing to the properties of the Laplace-Stieltjes transform $\Omega(z)$. From their respective definitions as transforms of equilibrium distribution functions, it follows that $\Omega(0) = \Omega_a(0) = 1$, which then implies that

$$Q_N(0) = (-\lambda)^{N+2}(-\lambda p_2) \quad \text{(A39)}$$

from (A35). Also, since the denominator of (A35) has a zero of order $(N+2)$ at the point $z = \lambda$, it follows that

$$\lim_{z\to\lambda} \frac{d^\nu}{dz^\nu} Q_N(z) = 0 \qquad (\nu = 0, 1, 2, \cdots, N+1) \quad \text{(A40)}$$

in order for $\Omega(z)$ to remain finite at that point. The above, (A40) along with (A39), then represents a system of $(N+3)$ independent, linear equations in the unknowns $\{\varepsilon_n(\lambda)\}_{n=0}^{N+2}$ and can therefore theoretically be solved straightforwardly by classical methods. Pragmatically, the calculations involved become increasingly tedious as N increases. We note that (A35) has an additional zero at some value of z that we shall designate as w_1, where $D_1(w_1) = 0$ ($D_1(z)$ can be shown to have a unique real root for $Re(z) \geq 0$ as in Calo [2]). A direct calculation readily yields $Q_N(w_1) = 0$ as required.

For the simplest case, $N = 0$, the transform of interest, as given by the appropriate expansions of (A35) and (A37), becomes

$$\Omega(z) = \Omega_a(z) \left\{ \frac{\phi(z)}{(z-\lambda)} \frac{\varepsilon_0(\lambda)}{(1-a\rho)} + \frac{\phi(z)}{(z-\lambda)^2}(-\lambda p_1 e^{-zb_1}) \frac{\varepsilon_1(\lambda)}{(1-a\rho)} + \left[p_1 \frac{D_a(z)}{D_1(z)} + \frac{\phi(z)e^{z(b_2-b_1)}}{(z-\lambda)^2 D_1(z)} \cdot (-\lambda p_1 e^{-zab_1})^2 \right] \frac{\varepsilon_2(\lambda)}{(1-a\rho)} \right\}. \quad \text{(A41)}$$

As indicated by (A39) and (A40), we now have three equations to solve for the constants $\varepsilon_0(\lambda)$, $\varepsilon_1(\lambda)$, and $\varepsilon_2(\lambda)$: namely, $Q_0(0) = -\lambda^3 p_2$, $Q_0(\lambda) = 0$, and $Q_0'(\lambda) = 0$. The first gives

$$p_2\varepsilon_0(\lambda) + p_1 p_2 \varepsilon_1(\lambda) + p_1^2 \varepsilon_2(\lambda) = (1-a\rho); \quad \text{(A42)}$$

the second yields

$$\varepsilon_1(\lambda) = \varepsilon_2(\lambda) e^{\lambda(b_2-ab_1)}; \quad \text{(A43)}$$

and the third provides

$$\varepsilon_0(\lambda) = \varepsilon_1(\lambda) e^{\lambda b_1(a-1)} \left[D_1(\lambda) \frac{\phi'(\lambda)}{\phi(\lambda)} + D_1'(\lambda) - b_1 D_1(\lambda) \right] - \varepsilon_2(\lambda)\lambda p_1 e^{-\lambda[ab_1+b_2-b_1]} \cdot \left[\frac{\phi'(\lambda)}{\phi(\lambda)} - (2ab_1 + b_2 - b_1) \right]. \quad \text{(A44)}$$

A simultaneous solution of (A42)–(A44) then explicitly establishes the three constants. These turn out to be

$$\varepsilon_0(\lambda) = e^{\lambda(b_2-b_1)}[1 - \lambda p_1 e^{-\lambda ab_1}(b_2 - ab_1)] \frac{(1-a\rho)}{d(\lambda)}$$

$$\varepsilon_1(\lambda) = e^{\lambda(b_2-ab_1)} \frac{(1-a\rho)}{d(\lambda)}$$

$$\varepsilon_2(\lambda) = \frac{(1-a\rho)}{d(\lambda)}, \quad \text{(A45)}$$

where we have defined

$$d(\lambda) = p_1^2 + p_1 p_2 e^{\lambda(b_2-ab_1)} + p_2 e^{\lambda(b_2-b_1)}[1 - \lambda p_1 e^{-\lambda ab_1}(b_2 - ab_1)] \quad \text{(A46)}$$

for notational convenience. These equations along with (A41) establish the desired transform. The equilibrium mean cumulative waiting time then follows from

$$E\{A_\infty\} = -\Omega'(0),$$

which in this case becomes

$$E\{A_\infty\} = \frac{\lambda a^2 E\{S^2\}}{2(1-a\rho)} + \frac{p_1 p_2}{\pi_0} \{(a+1)(b_2 - b_1) I_0 + I_1\}, \quad \text{(A47)}$$

where

$$I_0 = e^{\lambda(b_2-b_2)}[1 - \lambda p_1 e^{-\lambda ab_1}(b_2 - ab_1)] + p_1[e^{\lambda(b_2-ab_1)} - 1] - 1,$$

$$I_1 = p_1 \left\{ \left[\frac{2}{\lambda} - (b_1 + ab_2) \right] [e^{\lambda(b_2-ab_1)} - 1] - (b_2 - ab_1) \right\},$$

and

$$\pi_0 = p_2 e^{\lambda(b_2-b_1)}[1 - \lambda p_1 e^{-\lambda ab_1}(b_2 - ab_1)] + p_1[p_1 + p_2 e^{\lambda(b_2-ab_1)}].$$

We note that the first term of (A47) is just the mean waiting time in a standard M/G/1 queue with the same service time distribution as our second server, as in (50).

References

1. S. B. Calo, "Delay Properties of Message Channels," *Proceedings of the 1979 International Conference on Communications*, Boston, MA, June 1979, Institute of Electrical and Electronics Engineers, New York, pp. 43.5.1–43.5.4.
2. S. B. Calo, "Message Delays in Repeated-Service Tandem Connections," *IEEE Trans. Commun.* **COM-29,** 670–678 (1981).
3. R. M. Loynes, "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times," *Proc. Camb. Phil. Soc.* **58,** 497–520 (1961).
4. S. V. Tembe and R. W. Wolff, "The Optimal Order of Service in Tandem Queues," *Oper. Res.* **22,** 824–832 (1974).
5. P. J. Burke, "The Output of a Queueing System," *Oper. Res.* **4,** 699–704 (1956).
6. H. D. Friedman, "Reduction Methods for Tandem Queueing Systems," *Oper. Res.* **13,** 121–131 (1965).
7. I. Rubin, "Message Path Delays in Packet-Switching Communication Networks," *IEEE Trans. Commun.* **COM-23,** 186–192 (1975).
8. I. Rubin, "Communication Networks: Message Path Delays," *IEEE Trans. Info. Theory* **IT-20,** 738–745 (1974).
9. O. J. Boxma, "On a Tandem Queueing Model with Identical Service Times at Both Counters, I & II," *Report Mathematical Institute*, University of Utrecht, The Netherlands, 1978.
10. S. B. Calo, "The Message Channel, A Tandem Interconnection of Queues: Part II—Waiting Time Properties Under General Arrivals," *Research Report RC7170*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1978.
11. S. B. Calo, "The Message Channel, A Tandem Interconnection of Queues: Part I—Waiting Time Preserving Families of Queues and Related Realizations of the Message Channel," *Research Report RC6868*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1977.
12. L. Kleinrock, *Queueing Systems—Volume I: Theory*, John Wiley & Sons, Inc., New York, 1975.

Received January 12, 1981; revised June 19, 1981

The author is located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.