MANAGEMENT SCIENCE
Vol. 27, No. 8, August 1981
Printed in U.S.A.

A TWO SERVER QUEUE WITH NONWAITING CUSTOMERS RECEIVING SPECIALIZED SERVICE*

P. H. BRILL† AND M. J. M. POSNER‡

The stationary distributions of the waiting time and number in the system are obtained for a variant of an $M/M/2$ queue in which non-waiting customers receive a different rate of service from those who must wait in line. Characteristics of the solution, and applications of the model are discussed. The solution has also been generalized to any finite number of servers, using the same technique. The analysis utilizes the new System Point (SP) method for analyzing queues. This method employs the relationship between virtual waiting time sample function down- and up-crossings of a level, and the probability density function of the virtual wait at that level. Brief outlines of the SP approach, and relevant theory are included.
(QUEUES; MULTIPLE SERVER; WAITING TIME DEPENDENT SERVICE)

1. Introduction

The purpose of this paper is two-fold. Firstly, it introduces a relatively new class of multiple server queues with potentially wide application: multiple server queues with service time depending on waiting time. Secondly, it demonstrates the usefulness of the System Point (SP) method, which was discovered while one of the authors was solving this particular class of queueing models by conventional means.

Results for single server queues with service time depending on waiting time have been obtained by means of classical methods in J. A. Buzacott [8], J. R. Callahan [10], M. Libura [12], M. J. M. Posner [15], and P. D. Welch [17]. Little or no work in this area has been reported previously for multiple server queues. However, a criterion for ergodicity was obtained in S. Sugawara and M. Takahashi [16]. This paper treats a variant of the $M/M/2$ queue in which non-waiting customers receive a different rate of service from those who must wait in line. The analysis is by means of the SP method, and is based on Chapter 3 of [3]. The solution has also been generalized to any finite number of servers in Chapter 5 of [3]. It has also been solved for the case where the service time parameter is a general step function of the waiting time in Chapter 4 of [3]. In maintaining the stated purpose of this paper, only the two-server model with two levels of service is presented here. The SP method was originally introduced, developed, and applied by Brill during 1974 [3], and is further elucidated in [4] and [5].

Queues with service time depending on waiting time arise whenever the customer who starts the busy period of a server requires a setup time from that server. This situation may arise among bank tellers, machine repairmen, hospital emergency teams, copying systems in offices, computer terminals, automatic parking gates, etc.

Specifically, the initial customer to use a bank teller may incur a longer service time if the teller is performing another noninterruptable task when he arrives. Customers

*Accepted by Martin K. Starr, Special Editor; received September 15, 1978. This paper has been with the authors 17 months for 2 revisions.

†University of Illinois, Chicago Circle.
‡University of Toronto.

0025-1909/81/2708/0914$01.25

who arrive when other customers are present, consequently obtain a faster rate of service. Similarly, the initial machine requiring repair by a repairman may incur a setup time by the repairman, while those which arrive when the repairman is "tooled up" obtain a faster service rate. In hospital situations following a disaster, patients who initially require an emergency team, may incur a setup time (gathering the team from around the hospital, equipment setup time). Those who arrive while the team is already active obtain a faster service rate. In some offices the copying machine power supply is turned on by start-of-busy-period (nonwaiting) users. The machine warm up time must then be added to the workload for this type of customer. Users who arrive while the machine is in use obtain a faster rate of service, since the power supply is left on if there are people waiting. A similar remark applies to computer terminal rooms where users turn off the power supply to a terminal if no customers are waiting for that terminal.

Queues with service time depending on waiting also arise in supermarkets, steel plants, and in natural phenomena. In supermarket checkout lines, customers who wait in line may increase their checkout time by purchasing magazines, candy bars, cosmetics, etc., which are marketed next to the lines for this purpose. In a steel plant, hot ingots arrive at a mill to be rolled into sheets of steel, but must be stripped of their outer molds before the mill can process them. For this operation their outer shells are first required to cool down. Then the ingots must be reheated in soaking pits to the required "rolling" temperature. The longer they wait before stripping and/or reheating the more they cool, and then it takes a longer time for them to be reheated. This "nothing hot" delay has been dealt with in J. A. Buzacott and J. R. Callahan [9].

Two natural phenomena which generate these queueing models are forest fires and epidemics. Forest fire initial attack systems are modelled as ordinary multiple-server queues in J. H. Bookbinder and D. L. Martell [1]. The perimeter of a fire will depend on the waiting time between the first report of its occurrence and the moment it is "attacked" by a helicopter crew. Epidemics are similar to forest fires in that they spread while waiting to be "attacked" by a mass immunization program or other corrective measures.

The work in [2] utilizes the method of D. V. Lindley [13] which treats the sequence of customer waiting times as an embedded Markov chain. The appropriately simplified integral equations for the waiting time distribution are then derived after much algebraic manipulation, and the solution is tedious. In M. Eizenman and M. J. M. Posner [11], generating functions are used in a birth-and-death analysis of the two-server model. The complete probability distribution of the number in the system can be obtained only by extensive transform inversions, or numerical recursion. The resulting Laplace-Stieltjes transforms for the waiting time distribution are too complex to be inverted without a considerable expenditure of time and effort.

This model would seem to fit very simply into the context of Neuts' "matrix geometric method" [14], although care is required because the Markov chain derived from the matrix generator is reducible. For the two-server model presented in this paper, the system point method calculations are much easier than those of the matrix geometric method. However, this may not be true for more complex models. The system point method is an alternative approach, which is sometimes easier to implement, and often leads to insights into system behavior.

The SP method works for a very broad class of models, far from being restricted to queues with service time depending on waiting time. It has been successfully applied to

the solution of problems which involve: dams with general release rule [6], multiple server queues with heterogeneous servers and reneging depending on waiting time [7], queues depending on the number in the system at start of service epochs (Example 5 of [4]), queues with multiple Poisson inputs (Example 4 of [4], Chapter 8 of [3]), computers with several buffers having finite capacity (Example 1 of [5]).

This paper briefly outlines the concepts and theory of the SP method in the framework of the $M/M/2$ variant under consideration. The SP approach gives intuitive insight into the mathematical structure of the model not readily apparent using the aforementioned methods of analysis. It leads to a simpler, less tedious solution technique. A complete description of the SP method for a general exponential model is given in [5].

2. The Model and System Point Method

The variant of $M/M/2$ treated here has first-come, first-served queue discipline, and customer arrival rate λ. For the nth customer the waiting time before using the first available server is denoted by W_n, and the service time is denoted by $S_n(W_n)$. $S_n(W_n)$ is distributed exponentially with parameter $\mu(W_n)$. The service parameter depends on the value of W_n only, and not on n. Thus, for all n,

$$\Pr(S_n(W_n) \leqslant x \mid W_n = w) = 1 - \exp(-\mu(w)x), \qquad x \geqslant 0, w \geqslant 0. \tag{1}$$

Arriving customers who find at least one free server start service immediately, without waiting, and receive "special" service rate μ_0. All other customers receive service rate μ_1. This represents an extension of the single server model in Posner [15]. Customers whose service times have parameter μ_j are called type j customers, $j = 0, 1$. A sufficient condition for the stationary distributions of the waiting time and of the number in the system to exist is $\lambda < 2\mu_1$ [3], [5], [10].

The SP method, specific to this model, utilizes the following model description. Let $W(t)$ denote the virtual wait in the queue at time t, and let m_j denote the number of servers occupied by type j customers, $j = 0, 1$. A server configuration is defined to be any vector (m_0, m_1) such that $0 \leqslant m_0 + m_1 \leqslant$ (number of servers) $- 1 = 2 - 1 = 1$. This model has, therefore, three possible server configurations. Let the random variable $M(t)$ denote the *system* configuration *at time* t. Then $M(t)$ is said to be $m = (m_0, m_1)$ if a customer who arrives at time t would enter service when the server configuration is (m_0, m_1). Thus $M(t)$ depends on the server configuration at time $t + W(t)$. The occupancy numbers (m_0, m_1) are distributed among the servers *other* than those being entered. That is, (m_0, m_1) represents the type of the *other* customer in service, if any, that an arrival at time t would find when he enters service. Partition the set of possible configurations into two disjoint subsets $\mathfrak{M} = \{(1,0),(0,1)\}$ and $\mathfrak{M}_0 = \{(0,0)\}$. Hence $\mathfrak{M}$ contains the two possible system configurations which might occur at arrival epochs of customers who must either wait or find exactly one server free. $\mathfrak{M}_0$ contains that single configuration $(0,0)$ which occurs at arrival epochs of customers who find the system empty.

Define $\langle W(t), M(t)\rangle$ as the state at time t, with $W(t) \geqslant 0$ and $M(t) \in \mathfrak{M}_0 \cup \mathfrak{M}$. The stochastic process $\{\langle W(t), M(t)\rangle, t \geqslant 0\}$ is called the System Point (SP) Process, and it is assumed that the random variables $\langle W(t), M(t)\rangle$ converge weakly to the equilibrium random variables $\langle W, M\rangle$. That is, $\lim_{t\to\infty} \Pr(W(t) \leqslant w, M(t) = m) = \Pr(W \leqslant w, M = m)$. For $w > 0$ denotes the mixed partial densities of $\langle W, (1,0)\rangle$ and $\langle W, (0,1)\rangle$ at $W = w$ by $f_{10}(w)$ and $f_{01}(w)$ respectively. Let P_{00}, P_{10} and P_{01} denote the probabilities

that an arriving customer does not wait and the configuration is $(0,0),(1,0)$, and $(0,1)$ respectively, at the time of arrival. Let the density function of the waiting time of an arbitrary arrival be denoted by $g(w) = f_{10}(w) + f_{01}(w), w > 0$. For $w \geq 0$ and $m \in \mathfrak{M}$, consider customers who find the state to be $\langle w, m\rangle$ upon arriving. These customers will obtain service at rate

$$\mu(w,m) = \begin{cases} \mu_0 & \text{if } w = 0, m \in \mathfrak{M}_0 \cup \mathfrak{M}, \\ \mu_1 & \text{if } w > 0 \text{ (only if } m \in \mathfrak{M}). \end{cases} \tag{2}$$

The times from their start of service epochs until the first departures from the system have a common probability distribution which is exponential with parameter $m_0\mu_0 + m_1\mu_1 + \mu(w,m)$. Denote the random variable with this common distribution function by $S(w,m)$. This variable plays an important role in SP theory.

If the realized state of the system is $\langle w, m\rangle$ at time t, the state may be pictured as a point with coordinates (t,w) in a coordinate system corresponding to configuration m. (t,w) is called the System Point (SP).[1] If $m \in \mathfrak{M}$, then the range of values of w is $w \geq 0$, and (t,w) would be a point in the nonnegative quadrant of a two-dimensional Cartesian coordinate system. Since $\mathfrak{M}$ contains two states, there are two such coordinate systems, which are called "pages" in SP theory. For $m \in \mathfrak{M}_0$, the range of w is $w = 0$, and $(t,0)$ is a point on a "line" corresponding to m. The two pages may be thought of as being one behind the other, like the leaves of a book. The projection of these pages is called the "cover." Figure 1 depicts the pages, line and a possible sample function traced out by the SP over time, for the present model.

If the two pages and the "zero line" in the figure are superimposed, the resulting sample function will be piecewise continuous and resemble a graph of the usual virtual waiting time for a single server queue. In Figure 1, the first customer arrives at τ_1 when the state is $\langle 0,0\rangle$. The SP jumps to $(\tau_1^+, 0)$ on page 1; for any newly arriving customer would find the system in configuration $(1,0)$. The second customer arrives at τ_2 before customer one has ended service. The resulting jump is to point $(\tau_2^+, W(\tau_2^+))$ on page 1 where $W(\tau_2^+)$ is the sample value of a random variable $S(0,(1,0))$ which is exponentially distributed with parameter $2\mu_0$. This jump terminates on page 1 since any immediately arriving customer would enter service with the configuration being $(1,0)$. At τ_3 the jump happens to be to page 1 because for this realization, any immediately arriving customer would enter service with the system being in configuration $(1,0)$. This implies that of the two customers in service after customer 3 enters service, it is the type 1 customer (customer 3) who completes service first. At τ_4, the jump happens to be to page 2. Here any immediately arriving customer would start service with the system being in configuration $(0,1)$, because it is the type 0 customer who completes service first. Upon reflection it is seen that the length of any jump has an exponential distribution and: any jump from $(t,0)$ on page 1 has parameter $2\mu_0$; any jump from (t,w) on page 1, $w > 0$, has parameter $\mu_0 + \mu_1$; any jump from $(t,0)$ on page 2 has parameter $\mu_0 + \mu_1$, and any jump from (t,w), $w > 0$, on page 2 has parameter $2\mu_1$.

The SP theory makes an intimate connection between the down-crossing rates of levels in the state space by the SP and the steady state waiting time density function. General theorems, all definitions, and all proofs are given in [4]. Here is stated only the theory relevant to the subsequent analysis.

[1] SP is an abbreviation for System Point (point). It may be adjectival as in "SP process," or nominal as in "SP," which refers to the point (t,w). The meaning will always be clear from the context.

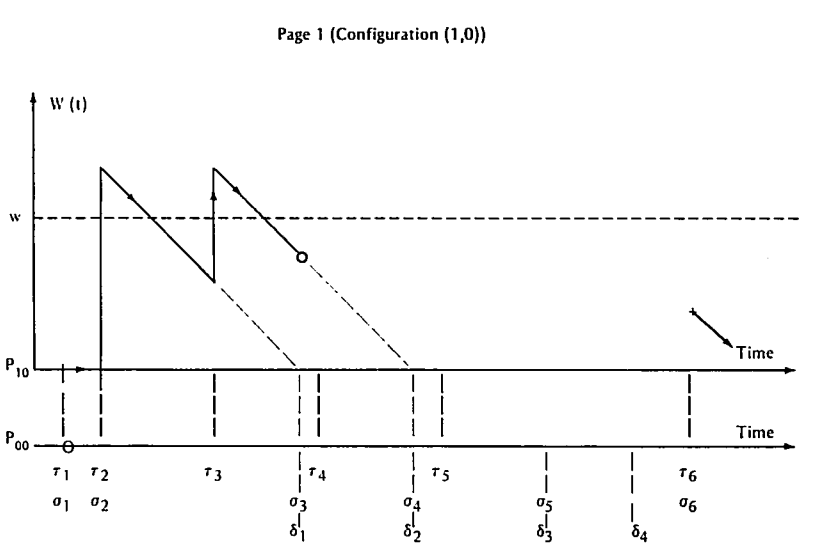

FIGURE 1. Motion of SP over Line and Pages for a Possible Realization.

Let $\mathfrak{D}_t(w,m)$ and $\mathfrak{I}_t(m)$ denote the number of down-crossings of level $w>0$, and impacts with level 0, on page $m \in \mathfrak{M}$ during $(0,t]$, respectively. Let $E[\cdot]$ denote expectation.

THEOREM 1.

$$\lim_{t\to\infty} E\big[\mathfrak{D}_t(w,m)\big]/t = f(w,m), \qquad w>0, \tag{3}$$

$$\lim_{t\to\infty} E\big[\mathfrak{I}_t(m)\big]/t = f(0^+,m). \tag{4}$$

Here

$$f(w,m) = \begin{cases} f_{10}(w) & m=(1,0) \\ f_{01}(w) & \text{if } m=(0,1). \end{cases} \tag{5}$$

Let $\mathfrak{U}_t(w,m,k)$ and $L_t(w,m,k)$ denote the number of up-crossings of level $w>0$ and leaps of a level $w\geqslant 0$ during $(0,t]$ which originate on page m and terminate on page k, for m and $k \in \mathfrak{M}$. Leaps are defined to originate from level zero.

THEOREM 2.

$$\lim_{t\to\infty} E\big[\mathfrak{U}_t(w,m,k)\big]/t = \lambda \int_{\alpha=0}^{w} p(\alpha,m,k)\exp(-C_{\alpha,m}(w-\alpha))F(d\alpha,m) \tag{6}$$

$$\lim_{t\to\infty} E\big[\mathfrak{L}_t(w,m,k)\big]/t = \lambda p(0,m,k)\exp(-C_{o,m}w)F(0,m), \tag{7}$$

where

$$C_{\alpha,m} = \sum_{j=0}^{1} m_j\mu_j + \mu(\alpha,m) \tag{8}$$

$$F(0,m) = \begin{cases} P_{10} & m=(1,0) \\ P_{01} & \text{if } m=(0,1) \end{cases} \tag{9}$$

$$\begin{aligned} F(d\alpha,m) &= f(\alpha,m)d\alpha \quad \text{if } \alpha>0 \\ &= F(0,m) \qquad \text{if } \alpha=0. \end{aligned} \tag{10}$$

$$p(\alpha,m,k) = \Pr(\textit{a jump from level } \alpha \textit{ on page } m \textit{ terminates on page } k). \tag{11}$$

In the present model

$$\begin{aligned} &p(w,(1,0),(0,1)) = \mu_0/(\mu_0+\mu_1), \qquad w>0 \\ &p(0,(1,0),(0,1)) = 0, \end{aligned} \tag{12}$$

$$\begin{aligned} &p(w,(0,1),(1,0)) = 0, \qquad w>0 \\ &p(0,(0,1),(1,0)) = \mu_1/(\mu_0+\mu_1). \end{aligned} \tag{13}$$

The first formula of (12) is the probability that a type 0 customer completes service before a type 1 customer when both are in service together. The second indicates that if a type 0 customer is in service alone, an arriving customer receives rate μ_0 so that

any additional newly arriving customer cannot then enter service with the other customer being type 1. The first formula of (13) indicates that if two type 1's are in service together an immediate arrival, upon entry into service can see only a type 1 in the other server. The last is the probability of a type 1 completing service before a type 0 when both types are in service together. Comparison with Figure 1 indicates the intuitive connotation associated with (12) and (13).

Let $\mathcal{D}_t(w) = \sum_{m\in\mathfrak{M}} \mathcal{D}_t(w,m)$ and $\mathcal{U}_t(w) = \sum_{m\in\mathfrak{M}}\sum_{k\in\mathfrak{M}} \mathcal{U}_t(w,m,k)$ denote the total numbers of upcrossings and downcrossings of level w respectively during $(0,t]$.

THEOREM 3.

$$\lim_{t\to\infty} E\big[\mathcal{D}_t(w)\big]/t = \lim_{t\to\infty} E\big[\mathcal{U}_t(w)\big]/t = g(w), \qquad w>0. \tag{14}$$

THEOREM 4.

$$g(w) = \lambda \sum_{m\in\mathfrak{M}} \int_0^w \exp(-C_{\alpha,m}(w-\alpha))F(d\alpha,m), \qquad w>0, \tag{15}$$

$$g(0^+) = \lambda \sum_{m\in\mathfrak{M}} F(0,m). \tag{16}$$

THEOREM 5. *For every* $m\in\mathfrak{M}$, *and* $w>0$

$$\begin{aligned} f(w,m) &+ \lambda \sum_{\substack{l\in\mathfrak{M}\\ l\neq m}} \int_{\alpha=0}^{w} p(\alpha,l,m)\big[1-\exp(C_{\alpha,l}(w-\alpha))\big]f(\alpha,l)d\alpha \\ &+ \lambda \sum_{l\in\mathfrak{M}} p(0,l,m)\big[1-\exp(-C_{0,l}w)\big]F(0,l) \\ &= f(0^+,m) + \lambda\int_0^w p(\alpha,m,m)\exp(-C_{\alpha,m}(w-\alpha))f(\alpha,m)d\alpha \\ &+ \lambda \sum_{\substack{l\in\mathfrak{M}\\ l\neq m}} \int_0^w p(\alpha,m,l)f(\alpha,m)d\alpha. \end{aligned} \tag{17}$$

The balance equations for the zero waiting time states $\langle 0;m\rangle$, $m\in\mathfrak{M}_0\cup\mathfrak{M}$ can also be obtained by usual stationary set balance yielding

$$\left(\lambda + \sum_{j=0}^{J} m_j\mu_j\right)F(0,m) = \begin{cases} f(0^+,m) + \lambda\sum_r p(0,r,m)F(0,m), & m\in\mathfrak{M}, \\ \lambda\sum_r p(0,r,m)F(0,r) + \sum_s \sum_{j=0}^{J} s_j\mu_j p(0,s,m)F(0,s), & m\in\mathfrak{M}_0, \end{cases} \tag{18}$$

where $r\in\mathfrak{M}_0$ with $\sum_0^J r_j = \sum_0^J m_j - 1$, and $s\in\mathfrak{M}_0\cup\mathfrak{M}$ with $\sum_0^J s_j = \sum_0^J m_j + 1$, and $J=1$ for the present model.

In (18) the LHS represents the long run average rate of exit by the SP from state $\langle 0;m\rangle$, $m\in\mathfrak{M}_0\cup\mathfrak{M}$, while the terms on the RHS are the corresponding rates of entry into $\langle 0,m\rangle$ for $m\in\mathfrak{M}$ and $m\in\mathfrak{M}_0$ respectively.

The SP method generates the model equations in a much simpler form than traditional techniques, based on the interconnections between: the motion of the SP,

the geometry of the sample function space, the theorems on level crossings and set balance, and the normalizing condition that all probabilities sum to 1. The ease of equation generation is due to the fact that these interconnections seem to be "easy" for the vast majority of potential users.

3. The System Point Analysis

Focussing attention on the motion of the SP over the "line" and "pages," and applying the theorems, leads to the model equations directly. In particular, (15), (17) and (18) facilitate the construction of these model equations. By this means, from (17), the following model equation for the partial density f_{10} is obtained:

$$f_{10}(w) + \lambda(1 - e^{-2\mu_0 w})P_{10} + (\lambda\mu_1/\nu)(1 - e^{-\nu w})P_{01}$$
$$= (\lambda\mu_1/\nu)\int_0^w e^{-\nu(w-z)} f_{10}(z)dz$$
$$+ (\lambda\mu_0/\nu)\int_0^w f_{10}(z)dz + f_{10}(0^+), \qquad w > 0, \tag{19}$$

where $\nu = \mu_0 + \mu_1$. Take the derivative with respect to w in (19) yielding

$$\langle D - \lambda\rangle f_{10}(w) = -2\lambda\mu_0 P_{10}e^{-2\mu_0 w} - \lambda\mu_1 P_{01}e^{-\nu w}$$
$$- \lambda\mu_1\int_{z=0}^w e^{-\nu(w-z)} f_{10}(z)dz \tag{20}$$

where $\langle D\rangle$ is the usual differential operator i.e., $\langle D\rangle f = f'$, $\langle D - \lambda\rangle f = f' - \lambda f$, etc. Operate on (20) with $\langle D + \nu\rangle$, resulting in

$$\langle D^2 + (\nu - \lambda)D - \lambda\mu_0\rangle f_{10}(w) = -2\lambda\mu_0 A P_{10}e^{-2\mu_0 w}. \tag{21}^2$$

Since $\lim_{w\to\infty} f_{10}(w) = 0$, it follows that

$$f_{10}(w) = a_0 e^{Rw} + BP_{10}e^{-2\mu_0 w}, \qquad w > 0, \tag{22}$$

where a_0 is a constant to be determined, and R is the negative root of $x^2 + (\nu - \lambda)x - \lambda\mu_0 = 0$.

Similarly, the model equation for the total density g can be written using (15) as:

$$g(w) = \lambda P_{10}e^{-2\mu_0 w} + \lambda P_{01}e^{-\nu w} + \lambda\int_{z=0}^w e^{-\nu(w-z)} f_{10}(z)dz$$
$$+ \lambda\int_{z=0}^w e^{-2\mu_1(w-z)} f_{01}(z)dz. \tag{23}$$

Operating on (23) with $\langle D + 2\mu_1\rangle$, and using $g(w) = f_{10}(w) + f_{01}(w)$ yields

$$\langle D + 2\mu_1 - \lambda\rangle g(w) = \lambda A\Big[2P_{10}e^{-2\mu_0 w} + P_{01}e^{-\nu w}$$
$$+ \int_{z=0}^w e^{-\nu(w-z)} f_{10}(z)dz\Big]. \tag{24}$$

[2]The definitions of A, and of the other symbols used to simplify the expressions, are summarized in Table 1 along with the numbers of the equations in which they first occur.

Substitute for the integral in (24) from equation (20), and then substitute for $\langle D - \lambda \rangle f_{10}(w)$ using (22), resulting in

$$\langle D + 2\mu_1 - \lambda \rangle g(w) = Ca_0 e^{Rw} + EP_{10} e^{-2\mu_0 w} \tag{25}$$

from which we obtain

$$g(w) = a_1 e^{-(2\mu_1 - \lambda)w} + Ha_0 e^{Rw} + QP_{10} e^{-2\mu_0 w} \tag{26}$$

where a_1 is a constant to be determined. The mixed partial density $f_{01}(w)$ is then found by taking $g(w) - f_{10}(w)$ so that

$$f_{01}(w) = a_1 e^{-(2\mu_1 - \lambda)w} + (H - 1)a_0 e^{Rw} + (Q - B)P_{10} e^{-2\mu_0 w}, \qquad w > 0. \tag{27}$$

The model equations for the zero waiting time states are obtained by balancing SP rates into and out of the states $\langle 0,(1,0)\rangle$, $\langle 0,(0,1)\rangle$, and $\langle 0,(0,0)\rangle$ respectively, following the result in (18). This yields

$$f_{10}(0^+) + \lambda P_{00} = (\lambda + \mu_0)P_{10}, \tag{28}$$

$$g(0^+) = \lambda(P_{10} + P_{01}), \tag{29}$$

$$f_{01}(0^+) = (\lambda + \mu_1)P_{01}, \tag{30}$$

$$\lambda P_{00} = \mu_0 P_{10} + \mu_1 P_{01}. \tag{31}$$

Moreover, substitute into (20) for $f'_{10}(w)$ and $f_{10}(w)$ using (22), and then let $w \downarrow 0$, yielding

$$(R - \lambda)a_0 + (-2\mu_0 B - \lambda B + 2\lambda\mu_0)P_{10} + \lambda\mu_1 P_{01} = 0. \tag{32}$$

Finally, the normalizing condition is

$$P_{00} + P_{10} + P_{01} + \int_0^\infty g(z)dz = 1. \tag{33}$$

TABLE I

Symbols for Simplifying Expressions in the Equations

Symbol	Definition	Equation of First Occurrence
A	$\mu_1 - \mu_0$	(21)
R	$[\lambda - \mu_0 - \mu_1 - ((\mu_0 + \mu_1 - \lambda)^2 + 4\lambda\mu_0)^{1/2}]/2$	(22)
B	$-2\lambda A/(2\mu_0 - 2\mu_1 + \lambda)$	(22)
C	$(\lambda - R)A/\mu_1$	(25)
E	$2AB$	(25)
H	$C/(R + 2\mu_1 - \lambda)$	(26)
Q	$-B^2/\lambda$	(26)
S	$\lambda(R - 2\lambda)/[(\lambda + \mu_0)(R - \lambda) - B(R + 2\mu_0) + \lambda\mu_0]$	(27)
T	$(\lambda - \mu_0 S)/\mu_1$	(38)
U	$-\lambda + (\lambda + \mu_0 - B)S$	(39)
V	$(\lambda - Q)S + \lambda T - HU$	(40)

Any three, of equations (28)–(31) together with (32) and (33), can be solved to obtain a_0, a_1, P_{00}, P_{10} and P_{01}, uniquely. First, substitute from (22) into (28) to yield

$$a_0 = -\lambda P_{00} + (\lambda + \mu_0 - B)P_{10}. \tag{34}$$

Then from (26) and (29) we obtain

$$a_1 + Ha_0 = (\lambda - Q)P_{10} + \lambda P_{01}. \tag{35}$$

From (31) and (32) it follows that

$$(R - \lambda)a_0 + (-2\mu_0 B - \lambda B + \lambda\mu_0)P_{10} + \lambda^2 P_{00} = 0. \tag{36}$$

Now, eliminate a_0 from (34) and (36); substitute for P_{10} in (31), and then in (34); finally substitute for P_{10}, P_{01} and a_0 in (35). These operations respectively yield

$$P_{10} = SP_{00}, \tag{37}$$

$$P_{01} = TP_{00}, \tag{38}$$

$$a_0 = UP_{00}, \tag{39}$$

and

$$a_1 = VP_{00}. \tag{40}$$

Substitute from (37)–(40) into (33), and use (26) to obtain

$$P_{00} = \left[1 + S + T + V/(2\mu_1 - \lambda) - U \cdot H/R + QS/2\mu_0\right]^{-1}. \tag{41}$$

The solution for the steady state distribution of the virtual waiting time is then given by (22), (26), (27), and (37)–(41). The probability of a zero wait is

$$G(0) = P_{00} + P_{10} + P_{01}. \tag{42}$$

Let P_n denote the stationary probability of n customers in the system at an instant at which service begins, $n \geq 0$. Then we obtain in the usual way, by conditioning on the waiting time, that

$$P_n = \int_{z=0}^{\infty} \frac{e^{-\lambda z}(\lambda z)^{n-2}}{(n-2)!}\, g(z)dz$$

$$= \left[a_1(\lambda/2\mu_1)^{n-1} + Ha_0(\lambda/(\lambda - R))^{n-1} + QP_{10}(\lambda/(\lambda + 2\mu_0))^{n-1}\right]/\lambda \tag{43}$$

for $n \geq 2$, with $P_0 = P_{00}$ and $P_1 = P_{10} + P_{01}$.

Table 1 is essentially a computer program for evaluating P_{00} using (41). Once P_{00} is obtained, the values of P_{10}, P_{01}, a_0 and a_1 can be evaluated from (37)–(40). These values then yield the total waiting time pdf g, and the partial waiting time pdf's f_{10} and f_{01} from (26), (22), and (27) respectively. The pdf of the number in the system is calculated from (43). Since the solution is in a closed form, all questions of sensitivity of waiting time, number-in-the-system, or throughput characteristics with respect to changes in the parameters λ, μ_0 and μ_1 can be answered. These sensitivity results can be obtained analytically, or conveniently by means of a computer program. Comparison of the solution with the simple $M/M/2$ queue having the same arrival rate, and

service parameter $(\mu_0 + \mu_1)/2$ in each server, is of interest. In the two cases

$$\text{(a)} \quad \mu_0 < \frac{\mu_0 + \mu_1}{2} < \mu_1, \qquad \text{(b)} \quad \mu_1 < \frac{\mu_0 + \mu_1}{2} < \mu_0 \tag{44}$$

the anticipated results occur. The values of $G(0)$ and p_0 are smaller in case (a), and in case (b) these values are larger than in the $M/M/2$ model. Recall that for $M/M/2$ the pdf of the waiting time in the queue is

$$g(w) = a\exp(-(2\mu - \lambda))w, \qquad w > 0,$$

where

$$\begin{aligned} a &= (\lambda^2/\mu)P_0, \\ P_1 &= (a/\lambda)P_0, \\ P_0 &= \left[1 + \lambda/\mu + \lambda^2/(\mu(2\mu - \lambda))\right]^{-1}, \\ \mu &= (\mu_0 + \mu_1)/2. \end{aligned} \tag{45}$$

Results for $G(0)$, P_0 and $\lambda E[W]$, where $E[W]$ is the expected wait in the queue, are summarized in Table 2 for various values of $\rho_j = \lambda/\mu_j, j = 0, 1$. In Table 2 the values of $\rho_0 = \rho_1 = 3/2$ correspond to the $M/M/2$ queue with $\lambda = 3$ and $\mu = (\mu_0 + \mu_1)/2 = (2 + 2)/2 = 2$. It is clear that the waiting time distribution in cases (a) and (b) envelope that for the $M/M/2$ model. Observe also that the expected wait is quite sensitive to deviations in ρ_j: a 5% deviation of each μ_j from $(\mu_0, \mu_1) = (2, 2)$ to $(1.9, 2.1)$ leads to a change of 14%, and to a change of 24% for $(\mu_0, \mu_1) = (2.1, 1.9)$.[3]

TABLE 2

Probabilities of No Wait in Queue, Zero in System, and Expected Queue Size for Various $\rho_j, j = 0, 1$

ρ_0	ρ_1	$G(0)$	P_0	$\lambda E[W]$
3/1.90	3/2.10	0.370063	0.145717	1.646934
3/1.99	3/2.01	0.358784	0.1432970	1.893396
3/2	3/2	0.357143	0.14285	1.928571
3/2.01	3/1.99	0.355415	0.142378	1.965671
3/2.10	3/1.90	0.335399	0.136090	2.412708

[3]Financial assistance for this work was provided by National Research Council of Canada grant no. A4374 and in part by the University of Manitoba grant no. 431-2013-20. We acknowledge the contribution of Mr. M. Eizenman, and thank the referees for their pertinent comments.

References

1. Bookbinder, J. H. and Martell, D. L., "Time-Dependent Queueing Approach to Helicopter Allocation for Forest Fire Initial Attack," *INFOR*, Vol. 17 (1979), pp. 58–70.
2. Brill, P. H. and Posner, M. J. M., "Two Server Queues With Service Time Depending on Waiting Time," Working Paper #74-005, Department of Industrial Engineering, University of Toronto, 1974.
3. ———. "System Point Theory in Exponential Queues," Ph.D. Dissertation, University of Toronto, 1975.

4. BRILL, P. H. AND POSNER, M. J. M., "Level Crossings in Point Processes Applied to Queues: Single Server Case," *Operations Res.*, Vol. 25 (1977), pp. 662–674.
5. ——— AND ———, "The System Point Method in Exponential Queues," *Math. Operations Res.*, Vol. 6 (1981), pp. 31–49.
6. ——— AND MOON, R. E., "An Application of Queueing Theory to Pharmacokinetics," *J. Pharmaceutical Sci.*, Vol. 69 (1980), pp. 558–560.
7. ———, "A Queue with Heterogeneous Servers and Reneging Depending on Waiting Time" (submitted).
8. BUZACOTT, J. A., "The Effect of Queue Discipline on the Capacity of Queues with Service Time Dependent on Waiting Time," *INFOR*, Vol. 12 (1974), pp. 174–185.
9. ——— AND CALLAHAN, J. R., "The Capacity of the Soaking Pit—Rolling Mill Complex in Steel Production," *INFOR*, Vol. 9 (1971), pp. 87–95.
10. CALLAHAN, J. R., "A Queue With Waiting Time Dependent Service Times," *Naval Res. Logist. Quart.*, Vol. 20 (1973), pp. 321–324.
11. EIZENMAN, M. AND POSNER, M. J. M., "A Birth and Death Analysis of Two Server Queue With Non-Waiting Customer Receiving Specialized Service," Working Paper #78-002, Department of Industrial Engineering, University of Toronto, 1978.
12. LIBURA, M., "On a One-Dimensional Queueing System With Service Time Depending on Waiting Time," *Arch. Automatyki Telemechaniki*, Vol. 16 (1971), pp. 279–286.
13. LINDLEY, D. V., "The Theory of Queues with a Single Server," *Proc. Cambridge Philos. Soc.*, Vol. 48 (1952), pp. 277–289.
14. NEUTS, M. F., "Markov Chains with Applications in Queueing Theory, Which Have a Matrix-Geometric Invariant Probability Vector," *Advances Appl. Probability*, Vol. 10 (1978), pp. 185–212.
15. POSNER, M. J. M., "Single Server Queues With Service Time Depending on Waiting Time," *Operations Res.*, Vol. 21 (1973), pp. 610–616.
16. SUGAWARA, S. AND TAKAHASHI, M., "On Some Queues Occurring in an Integrated Iron and Steel Works," *J. Operations Res. Soc. Japan*, Vol. 8 (1965), pp. 16–23.
17. WELCH, P. D., "On a Generalized $M/G/1$ Queueing Process in which the First Customer of Each Busy Period Receives Exceptional Service," *Operations Res.*, Vol. 12 (1964), pp. 736–752.