A Finite Difference Equations Approach to a Priority Queue

ALEXANDRE BRANDWAJN

Amdahl Corporation, Sunnyvale, California

The main goal of this note is to demonstrate the use of a direct approach, as opposed to generating functions, for solving balance equations of a queueing system whose state is described by two independent variables. The basic idea is to consider a function of two variables as a set of functions of one independent variable. Hence the balance equations may be viewed as a set of simultaneous difference equations, and solved using the appropriate techniques. This method may allow the inclusion of queue-dependent service and arrival rates, and an easy treatment of finite queueing rooms. This is illustrated in the example of a two-level preemptive priority queue with exponentially distributed interarrival and service times.

THE STATIONARY BEHAVIOR, with respect to the numbers of customers at service facilities, of many queueing systems may be mathematically described by a set of finite difference equations. This is the case not only for systems with exponentially distributed interarrival and service times, but also for systems with general distributions if such distributions are approximated (as may be done with arbitrary accuracy) by a finite number of exponential "stages" (Cox [1955]). For finite difference equations with more than one independent variable no general solution is known (Jordan [1965]), even when the coefficients in the equations are constant, i.e. when the instantaneous service and arrival rates do not depend on current system state. Therefore, unless such a queueing system belongs to the "product-form" (or "separable") class of systems (Baskett et al. [1975]), its solution must be sought on an individual basis. The method that is used almost exclusively by researchers in such cases is the generating functions method (Jordan). This method, however, is but one of several possible approaches to the solution of finite difference equations (see Boole [1970], and Jordan), and despite its advantages, it has several drawbacks. Specifically, it is not well-suited to state-dependent systems (equations with variable coefficients); it is often more difficult to apply to finite population systems than to systems with infinite population; and, most importantly, the generating function, once

Subject classification: 698 priority queue.

obtained, must be inverted. This last may turn out to be very difficult. To quote Boole (p. 277), "It must be borne in mind that the discovery of the generating function is but a step toward the solution of the difference-equation, and that the next step, viz. the discovery of the general term of its development by some *independent* process, is usually far more difficult than the direct solution of the original difference-equation would be."

The main goal of this note is thus to demonstrate the use of an alternative approach whereby one attempts to solve the difference equations directly. The queueing system considered is a two-level priority queue, level 1 customers having preemptive priority over level 2 customers. The interarrival and service times are assumed to be exponentially distributed random variables, the arrival and service rates for level 2 customers depend upon that level's queue length. A preempted level 2 customer resumes service only when the higher priority queue becomes empty, and the service discipline for each level is first-come-first-served. If the service and arrival rates are constant, the system described is classic (Saaty [1961], Jaiswal [1968]). The treatment of queue-dependent rates is a new result.

Denote by n_1 and n_2 the respective current numbers of level 1, and level 2 customers in the system. We denote by λ_1 and μ_1 the constant arrival and service rates for level 1 customers, and by $\lambda_2(n_2)$ and $\mu_2(n_2)$ the queue-dependent arrival and service rates for customers of level 2. Let $p(n_1, n_2)$ be the stationary joint probability distribution (if it exists) for the numbers of customers at each priority level. The system balance equations are readily obtained as:

$$\mu_1 p(n_1 + 1, 0) - [\lambda_1 + \lambda_2(0) + \mu_1]p(n_1, 0) + \lambda_1 p(n_1 - 1, 0) = 0, \tag{1}$$

$$n_1 = 1, 2, \cdots ;$$

$$\mu_1 p(n_1 + 1, n_2) - [\lambda_1 + \lambda_2(n_2) + \mu_1]p(n_1, n_2) + \lambda_1 p(n_1 - 1, n_2) \tag{2}$$

$$+ \lambda_2(n_2 - 1)p(n_1, n_2 - 1) = 0, \quad n_1 = 1, 2, \cdots ; \quad n_2 = 1, 2, \cdots ;$$

$$\mu_1 p(1, 0) - [\lambda_1 + \lambda_2(0)]p(0, 0) + \mu_2(1)p(0, 1) = 0; \tag{3}$$

$$\mu_1 p(1, n_2) - [\lambda_1 + \lambda_2(n_2) + \mu_2(n_2)]p(0, n_2) + \lambda_2(n_2 - 1) \tag{4}$$

$$\cdot p(0, n_2 - 1) + \lambda_2(n_2 + 1)p(0, n_2 + 1) = 0, \quad n_2 = 1, 2, \cdots .$$

These equations must be complemented by the normalizing condition $\sum_{n_1, n_2} p(n_1, n_2) = 1$.

In the next section we present a direct solution method, and outline its application to the set of difference equations (1)–(4). We also consider the case where the queueing room is finite.

Section 2 is devoted to a brief discussion of the classic system with constant arrival and service rates. We conclude by summarizing the

advantages and drawbacks of the direct solution method presented, and we indicate a few queueing problems to which it may be applied.

1. THE SOLUTION METHOD

The unknown probability distribution $p(n_1, n_2)$ is a function of two independent variables: n_1 and n_2. The basic idea of the direct solution method presented here is to consider such a function of two variables as a set of functions of one variable. In our case it is convienient, as will be seen shortly, to choose n_1 as this variable, and to use n_2 as an index to identify the function. We thus consider instead of $p(n_1, n_2)$ a set of functions $p_{n_2}(n_1)$. We have, of course,

$$p_{n_2}(n_1) = p(n_1, n_2), \quad n_1 = 0, 1, \cdots; \quad n_2 = 0, 1, \cdots. \tag{5}$$

Using (5) one may regard the difference equations for $p(n_1, n_2)$ as a set of simultaneous difference equations for the functions of one variable $p_{n_2}(n_1)$. Such simultaneous equations can often be solved by simply eliminating all the functions but one, and solving the resulting differences equation for that function (Jordan). In our case, things are even simpler. Equation 1 involves, without any elimination, only one function, viz. $p_0(n_1)$, and may easily be solved since with respect to our independent variable n_1 its coefficients are constant. Then, the $p_{n_2}(n_1)$ for consecutive values $n_2 = 1, 2, \cdots$ may be computed using (2) since it involves only the functions $p_{n_2}(n_1)$ and $p_{n_2-1}(n_1)$ which are already known.

If you consider difference equations for a function of two variables as a set of simultaneous difference equations for a set of functions of only one variable, this results in an approach similar to Boole's symbolic method for solving partial difference equations.

Let us now introduce the following operators, using the notations of Jordan. E, the operator of displacement $Ef(x) = f(x + 1)$; Δ, the operator of difference $\Delta f(x) = f(x + 1) - f(x)$; and Δ^{-1}, the operator of indefinite summation $\Delta^{-1}\phi(x) = f(x)$, if $\Delta f(x) = \phi(x)$. Equations 1 and 2 may now be rewritten in the following form

$$\psi_0(E) p_0(n_1) = 0, \quad n_1 = 0, 1, \cdots; \tag{6}$$

$$\psi_{n_2}(E) p_{n_2}(n_1) = -\lambda_2(n_2 - 1) E p_{n_2-1}(n_1),$$
$$n_1 = 0, 1, \cdots, \quad \text{and} \quad n_2 = 1, 2, \cdots; \tag{7}$$

where $\psi_{n_2}(E) = \mu_1 E^2 - [\lambda_1 + \lambda_2(n_2) + \mu_1]E + \lambda_1$, for $n_2 = 0, 1, \cdots$.

The general solution of a homogeneous difference equation with constant coefficients

$$\psi_{n_2}(E) p_{n_2}(n_1) = 0, \tag{8}$$

may be expressed in terms of the roots of the characteristic equation

$\psi_{n_2}(r) = 0$. The roots of this quadratic equation are $r_{1,2}(n_2) = (\lambda_1 + \lambda_2(n) + \mu_1 \pm \sqrt{\delta(n_2)})/(2\mu_1)$, where $\delta(n_2) = [\lambda_1 + \lambda_2(n_2) + \mu_1]^2 - 4\mu_1\lambda_1$. The general solution of (8) is of the form $\sum_{i=1}^{2} d_i r_i^{n_1}$, where the d_i are arbitrary constants. Since $p_0(n_1)$ must tend to zero as $n_1 \to \infty$, we reject the root $r_1(0)$ which is greater than unity, i.e. we set d_1 to zero. When the remaining root, $r_2(0)$, is less than unity, i.e.

$$(\lambda_1 + \lambda_2(0) + \mu_1 - \sqrt{[\lambda_1 + \lambda_2(0) + \mu_1]^2 - 4\mu_1\lambda_1})/(2\mu_1) < 1,$$

the solution of (6) may be expressed as

$$p_0(n_1) = C_0[r_2(0)]^{n_1}, \quad n_1 = 0, 1, \cdots. \tag{9}$$

C_0 is an "arbitrary" constant, to be determined from the left-out equations (Eqs. 3 and 4, and the normalizing condition).

The solution of (7) is obtained as a sum of the general solution of a homogeneous equation (Eq. 8) and a particular solution of the whole Eq. 7 (see Jordan). Such a particular solution may be obtained as a result of the operation (see Jordan, p. 558)

$$-((\lambda_2(n_2 - 1)E)/(\psi_{n_2}(E)))p_{n_2-1}(n_1). \tag{10}$$

When the function $p_{n_2-1}(n_1)$ is of the form $\sum_j g_j a_j^{n_1}$ where g_j and a_j are constants, (10) yields $\sum_j g_j[-((\lambda_2(n_2 - 1)a_j)/(\psi_{n_2}(a_j)))a_j^{n_1}]$ (see Jordan). Given (9) and the general solution of (8) it is clear that $p_{n_2-1}(n_1)$ is of the form considered.

Using the fact that $\psi_k(r_2(k)) = 0$, and hence

$$\psi_{n_2}(r_2(k)) = r_2(k)[\lambda_2(k) - \lambda_2(n_2)], \tag{11}$$

we obtain as a particular solution for $n_2 = 1$

$$\lambda_2(0)/(\lambda_2(1) - \lambda_2(0))C_0[r_2(0)]^{n_1}.$$

Leaving out again the root $r_1(1)$ we have the following general solution

$$p_1(n_1) = C_1[r_2(1)]^{n_1} + (\lambda_2(0)/(\lambda_2(1) - \lambda_2(0)))C_0[r_2(0)]^{n_1},$$

where C_1 is an "arbitrary constant," and $r_2(1)$ must be less than unity.

It is not difficult to see that in general we have

$$P_{n_2}(n_1) = \sum_{i=0}^{n_2} C_i[r_2(i)]^{n_1} \prod_{j=i}^{n_2-1} (\lambda_2(j)/(\lambda_2(j + 1) - \lambda_2(i))), \tag{12}$$

$$n_1 = 0, 1, \cdots, \quad n_2 = 0, 1, \cdots.$$

This solution exists only if

$$r_2(n_2) < 1, \quad \text{for all } n_2. \tag{13}$$

It is easy to show that (13) is equivalent to the simple condition $\lambda_1 < \mu_1$. This is in accordance with the preemptive priority of level 1 customers.

The constants C_{n_2}, $n_2 = 0, 1, \cdots$, have to be determined using Eqs. 3 and 4 and the normalization condition which may impose yet other conditions for the existence of the steady-state probabilities.

In our case the following approach to the determination of the C_{n_2} seems attractive. From the system balance equations the stationary marginal probability of the number of level 2 customers may be expressed as $p(n_2) = H \prod_{k=1}^{n_2} \lambda_2(k-1)/u(k)$, where $u(n_2)) = \mu_2(n_2)p(n_1 = 0 | n_2)$. For each value of n_2 the conditional probabilities $p(n_1 | n_2)$ must sum to unity. Define a set of new constants G_{n_2} with $C_{n_2} = p(n_2)G_{n_2}$. From (9) we obtain $G_0 = 1 - r_2(0)$. In general, using (12), we get $G_{n_2} = f_{n_2}(1 - r_2(n_2))$, where f_{n_2} is determined from the following recurrence relations $f_k = (1 + g_k(1 - r_2(k))/h_k)^{-1}$ with $g_k = \sum_{i=0}^{k-1} t_{k,i}$, $h_k = 1 - \sum_{i=0}^{k-1} (1 - r_2(i))t_{k,i}$ and $t_{k,i}$ given by $t_{k,i} = f_i \prod_{j=i}^{k-1} ((\mu_2(j+1))/(\lambda_2(j+1) - \lambda_2(i))) \prod_{j=i}^{k-2} p(n_1 = 0 | j+1)$. We have $p(n_1 = 0 | k) = G_k/h_k$ and $f_0 = 1$. The remaining constant H may be expressed as $H = \{1 + \sum_{n_2=1}^{\infty} \prod_{k=1}^{n_2} \lambda_2(k-1)/u(k)\}^{-1}$.

Let us now consider the case when there is a finite queueing room of, say, $N_2(N_2 \geq 1)$ positions for customers of level 2. This may be accounted for by simply setting $\lambda_2(n_2) = 0$ for $n_2 = N_2$ in (7). With this convention (12) remains valid for all $n_2 = 0, 1, \cdots , N_2$.

In the case when there is a finite queueing room N_1 for customers of level 1, Eqs. 6 and 7 are valid only up to $n_1 = N_1 - 1$. For $n_1 = N_1$ a boundary equation must be added:

$$-[\lambda_2(0) + \mu_1]p_0(N_1) + \lambda_1 p_0(N_1 - 1) = 0, \quad \text{for} \quad n_2 = 0, \quad (14)$$

and

$$-[\lambda_2(n_2) + \mu_1]p_{n_2}(N_1) + \lambda_1 p_{n_2}(N_1 - 1)$$
$$= -\lambda_2(n_2 - 1)p_{n_2-1}(N_1), \quad \text{for} \quad n_2 = 1, 2, \cdots . \quad (15)$$

Since the maximum value of n_1 is now finite, both roots of the characteristic equation, $r_1(n_2)$ and $r_2(n_2)$, have to be present in the general solution of the homogeneous equation (8). This yields for $n_2 = 0$, $p_0(n_1) = B_0[r_1(0)]^{n_1} + C_0[r_2(0)]^{n_1}$, $n_1 = 0, 1, \cdots , N_1$, and in general $p_0(n_1) = \sum_{i=0}^{n_2} \{B_i[r_1(i)]^{n_1} + C_i[r_2(i)^{n_1}\} \prod_{j=i}^{n_2-1} ((\lambda_2(j))/(\lambda_2(j+1) - \lambda_2(i)))$.

The new "arbitrary" constants B_i are disposed of so as to satisfy the boundary equations (14 and 15), and the remaining C_i's are determined, as previously, from Eqs. 3 and 4 and the normalization condition.

As a final point, note that the solution (12) can only be used directly if no two arrival rates to level 2, $\lambda_2(n_2)$, are identical. (This is evident from the very form of (12) and also from (10) with (11).)

In the next section we briefly discuss the application of the solution method presented to the case where $\lambda_2(n_2)$ is a constant.

2. SOLUTION WITH CONSTANT ARRIVAL RATES

When the arrival rate to level 2 is constant (λ_2) Eqs. 6 and 7 become $\psi(E)p_0(n_1) = 0$, $n_1 = 0, 1, \cdots$; and

$$\psi(E)p_{n_2}(n_1) = -\lambda_2 E p_{n_2-1}(n_1), \quad n_1 = 0, 1, \cdots, \tag{16}$$

where $\psi(E) = \mu_1 E^2 - (\lambda_1 + \lambda_2 + \mu_1) E + \lambda_1$.

The corresponding characteristic equation $\psi(r) = 0$ has two roots r_1 and r_2. $p_0(n_1)$ is obtained in the same manner as in Section 2 (rejecting r_1), $p_0(n_1) = C_0 r_2^{n_1}$, $n_1 = 0, 1, \cdots$. The following $p_{n_2}(n_1)$ have the form (from 16))

$$p_{n_2}(n_1) = C_{n_2} r_2^{n_1} + u_{n_2}(n_1), \quad n_1 = 0, 1, \cdots, \quad n_2 = 1, 2, \cdots,$$

where $u_{n_2}(n_1)$ is a particular solution of the nonhomogeneous equation (16).

As previously, the C_{n_2} have to be determined from the boundary equations and the normalization condition.

However, unlike in Section 2, formula (10) cannot be used directly to determine a particular solution $u_{n_2}(n_1)$ of (16) since its denominator vanishes ($\psi(r_2) = 0$). Instead of (10) we may use a partial fraction decomposition derived from it (see Jordan, p. 565)

$$u_{n_2}(n_1) = \sum_i b_i r_i^{n_1-1} \Delta^{-1}[r_i^{-n_1} p_{n_2-1}(n_1)],$$

where the r_i's are the roots of $\psi(r) = 0$, and

$$b_i = 1/[d\,\psi(r)/dr]_{r=r_i}. \tag{17}$$

Equation 17 yields $b_{1,2} = \pm[(\lambda_1 + \lambda_2 + \mu_1)^2 - 4\lambda_1\mu_1]^{-1/2} \equiv \pm b$, and we obtain for $n_2 = 1$, $p_1(n_1) = C_1 r_2^{n_1} + \lambda_2 b C_0 n_1 r_2^{n_1}$. Denoting by $n_1^{(m)} = n_1(n_1 - 1) \cdot \cdots \cdot (n_1 - m + 1)$ the factorial of degree m of n_1, with $n_1^{(0)} = 1$, and $n_1^{(1)} = n_1$, and using the fact that $\Delta^{-1} n_1^{(m)} = n_1^{(m+1)}/(m + 1)$, and

$$\Delta^{-1}[a^{n_1} n_1^{(m)}] = a^{n_1}/(a - 1) \sum_{k=0}^{m} (a/(a - 1))^k (m!/(m - k)!) n_1^{(m-k)},$$

one can derive the $p_{n_2}(n_1)$ for following values of n_2.

$$p_2(n_1) = r_2^{n_1}\{C_2 + C_1(\lambda_2/b)n_1^{(1)} + C_0(\lambda_2/b)^2[(n_1 + 1)^{(1)}/(r_1/r_2 - 1)$$
$$+ (n + 1)^{(2)}/2]\},$$

$$p_3(n_1) = r_2^{n_1}\{C_3 + C_2(\lambda_2/b)n_1^{(1)} + C_1(\lambda_2/b)^2[(n_1 + 1)^{(1)}/(r_1/r_2 - 1)$$
$$+ (n_1 + 1)^{(2)}/2] + C_2(\lambda_2/b)^3[(n_1 + 2)^{(2)}/(r_1/r_2 - 1)$$
$$+ (n_1 + 2)^{(3)}/(2 \cdot 3)]\}, \quad \text{etc.}$$

It is interesting to note that $p_{n_2}(n_1)$ has the form of a polynomial of

degree n_2 in n_1, multiplied by the exponential factor $r_2^{n_1}$, and that, curiously enough, the solution turns out to be more complex when the arrival rate λ_2 is constant than when it is variable.

It is clear that finite queueing rooms may be taken into account in a manner similar to that of Section 2, and do not constitute a particular problem for the method employed.

As a final point, let us note that the generating function for $p_0(n_1)$ is $\sum_{n_1=0}^{\infty} p_0(n_1) z^{n_1} = C_0(1 - r_2 z)^{-1}$ and is thus consistent with formula (11–37) of Saaty which gives the generating function of steady-state probabilities.

3. CONCLUSION

Using the example of a two-level preemptive priority queue we have presented a method which allows a direct solution of the system balance equations. In this way the often difficult problem of "inverting" the generating function is avoided. The method, close to Boole's symbolical method, is mainly suited to difference equations with two independent variables. It is based on the idea to consider a function of two (in principle, also more) independent variables as a set of functions of one variable, and hence to consider the given partial difference equations as a set of simultaneous equations for this set of functions. The method often allows the inclusion of queue-dependent service (and/or arrival) rates, and an easy inclusion of finite queueing rooms. (The treatment of a priority queue with queue-dependent arrival and service rates for lower priority customers presented in this note is a new result.)

However, the method presented also has drawbacks. Since it yields the state probabilities, and these may have complex expressions, the determination of expected values for the numbers of customers may be more complex than the generating functions approach where differentiation is involved. A similar remark holds with respect to the normalizing constant. Also, as the reader has probably noticed, the explicit determination of the "arbitrary" constants introduced in the process of solution may not be easy. Therefore, it may be a good idea in some cases to use this method and the generating functions method conjunctly. For instance, when the arrival and departure rates for level 2 customers are constant, the method of generating functions readily yields the normalizing constant C_0 but not the form of the state probabilities. The method discussed herein, on the other hand, yields the general form of the latter but the determination of C_0 is much more involved.

Nonetheless, the method presented is, in the author's opinion, worthwhile, as it allows an easy, elementary attack of otherwise difficult problems. These include, for instance, the $G/M/1$ and $M/G/1$ queues with correlated arrival and service processes. The basic underlying idea

is to use the Coxian representation of a general (well, almost) distribution by a finite number of exponential "stages" (Cox). The couple composed of the number of the current stage and of the total number of customers in the system is then an adequate description of the state of the queue. The correlation between arrival and service processes is simply expressed by letting the instantaneous rate of the exponential process depend on the current stage in the general process. The balance equations for the state description defined above are then treated as a finite set (as many as there are stages) of simultaneous difference equations with one variable, viz. the total number of customers in the system. The interested reader will find in Brandwajn [1979] a discussion of the application of this method to the $G/M/1$ and $M/G/1$ queues with both finite and infinite queueing rooms.

REFERENCES

BASKETT, F., K. M. CHANDY, R. R. MUNTZ AND F. PALACIOS. 1975. Open, Closed and Mixed Networks of Queues with Several Classes of Customers. *J. Assoc. Comput. Mach.* **22,** 248–260.

BOOLE, G. 1970. *Calculus of Finite Differences*, Ed. 5. Chelsea Publishing Co., New York.

BRANDWAJN, A. 1979. $G/M/1$ and $M/G/1$ Queue with Correlated Arrival and Service Processes—a Simple Approach, Research Report CS-1979-8, Duke University, Computer Science Dept. (April).

COX, D. R. 1955. A Use of Complex Probabilities in the Theory of Stochastic Processes. *Proc. Camb. Phil. Soc.* **51,** 313–319.

JAISWAL, N. K. 1968. *Priority Queues*. Academic Press, New York.

JORDAN, C. 1965. *Calculus of Finite Differences*. Chelsea Publishing Co., New York.

SAATY, T. L. 1961. *Elements of Queueing Theory*. McGraw-Hill, New York.