A Queuing-Type Birth-and-Death Process Defined on a Continuous-Time Markov Chain

Uri Yechiali

Tel-Aviv University, Tel Aviv, Israel

(Received April 12, 1971)

This paper considers an n-phase generalization of the typical $M/M/1$ queuing model, where the queuing-type birth-and-death process is defined on a continuous-time n-state Markov chain. It shows that many models analyzed in the literature can be considered special cases of this framework. The paper focuses on the steady-state regime, and observes that, in general, closed-form results for the limiting probabilities are difficult to obtain, if at all possible. Hence, numerical methods should be employed. For an interesting special case, explicit results are obtained that are analogous to the classical solutions for the simple $M/M/1$ queue.

CONSIDER A 'phase-process' that is an n-state irreducible, aperiodic Markov chain with probability transition matrix $\|q_{ij}\|$, $i, j=1, 2, \cdots, n$, and $q_{ii}=0$ for all i. Consider also a birth-and-death process defined on the states $0, 1, 2, 3, \cdots$. Let $X(t)$ and $Y(t)$, respectively, denote the phase of the Markov chain and the state of the birth-and-death process at time t. Let $Y(t)$ and $X(t)$ be related such that whenever $X(t)=i$, $i=1, 2, \cdots, n$, and $Y(t)=m$, $m=0, 1, 2, 3, \cdots$, the birth-and-death rates are λ_{im} and μ_{im}, respectively. Moreover, assume that the phase process is a continuous-time Markov chain, where the sojourn time in phase i is an exponentially distributed random variable with mean $1/\eta_i$. We thus have a birth-and-death process on the nonnegative integers with rates depending on both the states $Y(t)$ and an extraneous phase process that is a continuous-time Markov chain.

In the context of queuing theory, we may think of $Y(t)$ as the number of customers in the system at time t, where the system is said to be in state E_{im} at time t if $X(t)=i$ and $Y(t)=m$. The arrival process is a heterogeneous Poisson process with arrival rate that assumes the value $\lambda_{im}\geqq 0$ whenever the system is in state E_{im}. The service process is of an exponential type with rate of service $\mu_{im}\geqq 0$ whenever $X(t)=i$, $Y(t)=m\neq 0$, and with $\mu_{i0}=0$ for all i.

Define the limiting probabilities of the system as follows:

$$P_{im}=\lim_{t\to\infty} P[X(t)=i,\ Y(t)=m]. \qquad (i=1, 2, \cdots, n;\ m=0, 1, 2, \cdots) \quad (1)$$

As is well known,[1] these limits always exist and are either all positive (steady-state case) or all vanish. In the sequel we will focus our interest on the steady-state regime.

Balance Equations

Writing Kolmogorov's forward differential equations,[1] and passing to the limit as t goes to infinity, we obtain the steady-state balance equations of the system:

$$(\lambda_{i0}+\eta_i \sum_{j=1, j\neq i}^{j=n} q_{ij}) p_{i0} = \mu_{i1} p_{i1} + \sum_{j=1, j\neq i}^{j=n} \eta_j q_{ji} p_{j0}, \qquad (i=1, 2, \cdots, n) \quad \text{(2a)}$$

$$(\lambda_{im}+\mu_{im}+\eta_i \sum_{j=1, j\neq i}^{j=n} q_{ij}) p_{im} = \lambda_{i,m-1} p_{i,m-1} + \mu_{i,m+1} p_{i,m+1} + \sum_{j=1, j\neq i}^{j=n} \eta_j q_{ji} p_{jm}. \qquad (i=1, 2, \cdots, n;\ m>0) \quad \text{(2b)}$$

Define $\eta_{ij}=\eta_i q_{ij}$ for $i, j=1, 2, \cdots, n$, and let $\lambda_{im}=0$ for $m<0$. On recalling that, for every i, $\mu_{i0}=0$ and $q_{ii}=0$, equation (2) may be rewritten as

$$(\lambda_{im}+\mu_{im}+\eta_i) p_{im} = \lambda_{i,m-1} p_{i,m-1} + \mu_{i,m+1} p_{i,m+1} + \sum_{j=1}^{j=n} \eta_{ji} p_{jm}. \qquad (i=1, 2, \cdots, n;\ m\geqq 0) \quad (3)$$

By examining the set of equations (3), it is readily seen that, once the n probabilities p_{i0} $(i=1, 2, \cdots, n)$ are known, all probabilities p_{im} may be calculated recursively. However, as will be apparent in Section II, it is difficult, if at all possible, to obtain explicit solutions for the probabilities p_{i0}. Moreover, the p_{i0}'s can not be found by using equation (3) alone. In order to obtain these probabilities we will have to use techniques employing generating functions. In a special interesting case, though, we will be able to derive some closed-form results that will resemble the well known results for the classical $M/M/1$ queuing model. These results will be developed in Section III.

The Probabilities $p_{i\cdot}$

Summation of (3) over m yields

$$\eta_i p_{i\cdot} = \sum_{j=1}^{j=n} \eta_{ji} p_{j\cdot}, \qquad (4)$$

where $p_{i\cdot} = \sum_{m=0}^{m=\infty} p_{im}$ is the (marginal) probability of the phase process being in phase i $(i=1, 2, \cdots, n)$. Clearly, result (4) could have been obtained directly by considering, independently, the phase process $X(t)$ by itself.

In contrast to our apparent inability to derive closed-form solutions for the probabilities p_{im}, results for the phase probabilities $p_{i\cdot}$ can be obtained explicitly. We recall[1] that, for the irreducible aperiodic finite Markov chain $\|q_{ij}\|$, the limiting probabilities $\{\Pi_j, j=1, 2, \cdots, n\}$ are all positive and uniquely satisfy

$$\sum_{j=1}^{j=n} \Pi_j = 1, \qquad \Pi_j = \sum_{i=1}^{i=n} \Pi_i q_{ij}. \qquad (j=1, 2, \cdots, n) \quad (5)$$

Since $\sum_{i=1}^{i=m} p_{i\cdot} = 1$, the unique solution of (4) is

$$p_{i\cdot} = [\Pi_i/\eta_i]/[\sum_{k=1}^{k=n} (\Pi_k/\eta_k)]. \qquad (i=1, 2, \cdots, n) \quad (6)$$

Clearly, $p_{i\cdot}$ is independent of the arrival and service rates.

Equation (6) may be interpreted as follows: Given that the system is in phase i, the mean sojourn time is $1/\eta_i$. Hence, the 'average' cycle length of the phase process is $\sum_{k=1}^{k=n} (\Pi_k/\eta_k)$, while the fraction of time the system spends in phase i is $p_{i\cdot} = (\Pi_i/\eta_i)/\sum_{k=1}^{k=n} (\Pi_k/\eta_k)$.

A Necessary and Sufficient Condition

Starting with $m=0$ and summing each of equations (3) over i we obtain the recurrence relations

$$\sum_{i=1}^{i=n} \lambda_{im} p_{im} = \sum_{i=1}^{i=n} \mu_{i,m+1} p_{i,m+1}. \qquad (m=0, 1, 2, \cdots) \quad (7)$$

By summing (7) over all m, we arrive at

$$\sum_{m=0}^{m=\infty} \sum_{i=1}^{i=n} \lambda_{im} p_{im} = \sum_{m=1}^{m=\infty} \sum_{i=1}^{i=n} \mu_{im} p_{im}. \tag{8}$$

We now restrict ourselves to the consideration of an n-phase generalization of the $M/M/1$ queuing process. That is, for $i=1, 2, \cdots, n$ we let $\lambda_{im}=\lambda_i$ for all $m\geqq 0$, $\mu_{im}=\mu_i$ for $m>0$, and $\mu_{i0}=0$. Equation (3) is now transformed into

$$(\lambda_i+\mu_i+\eta_i)p_{im}=\lambda_i p_{i,m-1}+\mu_i p_{i,m+1}+\sum_{j=1}^{j=n}\eta_{ji}p_{jm}, \quad (i=1, 2, \cdots, n;\ m\geqq 0) \tag{3'}$$

where $p_{im}=0$ for $m<0$. On defining $\hat{\mu}=\sum_{i=1}^{i=n}\mu_i p_{i\cdot}$ and $\hat{\lambda}=\sum_{i=1}^{i=n}\lambda_i p_{i\cdot}$, we can rewrite equation (8) as

$$\hat{\mu}-\hat{\lambda}=\sum_{i=1}^{i=n}\mu_i p_{i0}. \tag{9}$$

Since the steady-state regime exists if and only if $p_{i0}>0$ the necessary and sufficient condition for its existence is $\hat{\mu}-\hat{\lambda}>0$. That is, for the equilibrium condition, the average service capacity of the system must exceed the average arrival rate.

I. RELATED MODELS

A SIMILAR SYSTEM with $n=2$, and $\lambda_{im}=\lambda_i$, $\mu_{im}=\mu_i$ ($\mu_{i0}=0$) for $i=1$, 2 and $m=0$, 1, 2, $\cdots$ was considered by YECHIALI AND NAOR.[4] In that study the system was viewed as an $M/M/1$ queuing process defined on a continuous-time Markov chain with $q_{ij}=\begin{Vmatrix}0 & 1\\ 1 & 0\end{Vmatrix}$, and mean sojourn times $1/\eta_i$. MITRANI AND AVI-ITZHAK[2] studied an $M/M/n$ queuing system with service interruptions. In their case, $Y(t)$, the phase of the process, was defined as the number of operating servers. It is easy to transform the present general case into their case. To do this a 'zero' phase is added to $\|q_{ij}\|$ such that the Markov chain consists of $n+1$ phases $\{0, 1, 2, \cdots, n\}$. We then let $\lambda_{im}=\lambda$ for all i and m; $\mu_{0m}=0$ for all m; $\mu_{im}=m\mu$, if $i\geqq m$; and $\mu_{im}=i\mu$, if $i\leqq m$. In their notation the breakdown rate of an operative server was ξ, whereas the repair rate was η. Thus, after letting $\eta_i=i\xi+(n-i)\eta$ and

$$q_{ij}=\begin{cases} i\xi/\eta_i, & \text{if } j=i-1,\\ (n-i)\eta/\eta_i, & \text{if } j=i+1,\\ 0, & \text{otherwise,}\end{cases}$$

the transformation is complete.

A closely related model was treated by NEUTS,[3] who studied an $M/G/1$ queuing process defined on an extraneous continuous-time Markov chain (=the phase process). The assumptions he made are that $\lambda_{im}=\lambda_i$ for all (im), and that the service-time realizations are drawn from an arbitrary distribution function $H_i(\cdot)$ whenever the process is in phase i. In addition, his model differs from the present one and the one discussed in reference 4 by assuming that the service-time distribution of a customer depends only on the phase state at the *beginning* of his service, whereas here the rate of service of an individual customer may fluctuate with the changes in phase.

The models analyzed in references 4 and 2 were shown to be generalizations of many special cases studied previously in the literature. The present study is thus a further generalization of the theory. In addition, references 2, 3, and 4 contain discussions on the pragmatic aspects of the model together with indications of areas of applications. These studies also contain extensive lists of references.

II. GENERATING FUNCTIONS AND ZERO PROBABILITIES

LET THE PARTIAL generating functions $G_i(z)$, $|z|\leqq 1$, $i=1, 2, \cdots, n$, be defined by $G_i(z)=\sum_{m=0}^{m=\infty} z^m p_{im}$. Multiplying each equation of (3′) by z^m $(m=0, 1, 2, \cdots)$ and summing over all m, we obtain a set of n linear equations involving $G_i(z)$:

$$[\lambda_i z(1-z)+\mu_i(z-1)+\eta_i z]G_i(z)-\sum_{j=1}^{j=n}\eta_{ji}zG_j(z)=\mu_i p_{i0}(z-1). \qquad (i=1, 2, \cdots, n) \tag{10}$$

Following the notation and method introduced in reference 2, we let

$$f_i(z)=\lambda_i z(1-z)-\mu_i(1-z)+\eta_i z, \qquad (i=1, 2, \cdots, n)$$

$$A(z)=\left\|\begin{matrix} f_1(z) & -\eta_{21}z & -\eta_{31}z & \cdots & -\eta_{n1}z \\ -\eta_{12}z & f_2(z) & -\eta_{32}z & \cdots & -\eta_{n2}z \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -\eta_{1n}z & -\eta_{2n}z & -\eta_{3n}z & \cdots & f_n(z) \end{matrix}\right\|,$$

$$g(z)=\begin{bmatrix} G_1(z) \\ G_2(z) \\ \vdots \\ G_n(z) \end{bmatrix}, \qquad b=\begin{bmatrix} \mu_1 p_{10} \\ \mu_2 p_{20} \\ \vdots \\ \mu_n p_{n0} \end{bmatrix}.$$

In matrix form equation (10) is now written as

$$A(z)g(z)=(z-1)b. \tag{11}$$

For all values of z where $A(z)$ is nonsingular we have

$$|A(z)|G_i(z)=|A_i(z)|(z-1), \qquad (i=1, 2, \cdots, n) \tag{12}$$

where $|A|$ stands for the determinant of a matrix A, and the matrix $A_i(z)$ is obtained by replacing the ith column of $A(z)$ with b.

In equation (12), $G_i(z)$ is expressed in terms of the elements $\mu_i p_{i0}$ of b. Our problem is now to find the values of the n unknown 'zero' probabilities. Since $|A(z)|$ is a polynomial of degree $2n$, and $|A(1)|=0$, we may define a new polynomial $Q(z)$ of degree $2n-1$ by $|A(z)|=(z-1)Q(z)$. We thus have

$$Q(z)G_i(z)=|A_i(z)|. \qquad (i=1, 2, \cdots, n) \tag{13}$$

We recall that equation (9) gives us one linear relation in the n unknowns p_{i0}. In addition, from equation (13) it follows that $|A_i(z)|=0$ whenever $Q(z)=0$. This fact can be used to gain the additional $n-1$ linear relations (in the n unknown probabilities) in the following way.

THEOREM. *The polynomial* $Q(z)$ (*or, alternatively,* $|A(z)|$) *has exactly* $n-1$ *distinct real roots in the interval* $(0, 1)$.

The proof is a very elaborate adaptation of the proof given in reference 2 (pp. 632–634) and therefore will be omitted. We just indicate that, in principle, it is an inductive proof employing a sequence of polynomials $Q_k(z)$, $k=1, 2, \cdots, n$, which are the determinants of the main-diagonal minors of $A(z)$ starting from the lower right-hand corner of the matrix.

If we denote the $n-1$ distinct roots of $Q(z)$ in (0, 1) by z_j, $j=1, 2, \cdots, n-1$, equation (13) becomes

$$|A_i(z_j)|=0. \qquad (j=1, 2, \cdots, n-1; i=1, 2, \cdots, n) \quad (14)$$

However, from (13) we observe that, for each z_j and any pair $1 \leqq i, k \leqq n$, $|A_i(z_j)|/|A_k(z_j)|=G_i(z_j)/G_k(z_j)=$ constant. Thus, for each z_j we have n homogeneous linear equations that differ from each other only by a constant multiplier; that is, (14) yields only one independent equation for each z_j for $j=1, 2, \cdots, n-1$.

It is evident now how difficult, if at all possible, it is to obtain explicit expressions for the roots z_j, $j=1, 2, \cdots, n-1$. Even for the case $n=2$, it was pointed out in reference 4 that, in general, no closed-form relations are available for the probabilities p_{i0}, and, except for numerical results, no analytic comparison to the elegant results of the classical $M/M/1$ queue can be made. In one special case, however, such results are obtainable, and a comparison can be made. We analyze this case in the following section.

III. THE CASE $\lambda_i/\mu_i=\theta$ FOR ALL i

WE LET $\lambda_i/\mu_i=\theta$ for all $i=1, 2, \cdots, n$. As was shown in reference 4, this is the only case where we can derive results that are similar to the ones of the $M/M/1$ queue. It follows immediately that $\lambda_i/\mu_i=\theta$ for all i implies $\hat{\lambda}/\hat{\mu}=\theta$. Moreover, it was argued and shown in reference 4 that in this case (and in this case only) we have

$$p_{im}=p_{i\cdot}p_{\cdot m}, \qquad (i=1, 2, \cdots, n; m \geqq 0) \quad (15)$$

where $p_{\cdot m}=\sum_{i=1}^{i=n} p_{im}$ denotes the (marginal) probability of having m customers in the system, $m=0, 1, 2, \cdots$.

Equation (15) implies

$$p_{i0}/p_{i\cdot}=p_{\cdot 0}. \qquad (i=1, 2, \cdots, n) \quad (16)$$

Let the busy fraction ρ of the service station be represented by

$$\rho=1-\sum_{i=1}^{i=n} p_{i0}. \quad (17)$$

Substituting (16) in (17), we obtain

$$\rho=1-p_{\cdot 0}. \quad (18)$$

On combining (16) and (18), and utilizing (9) we arrive at

$$\hat{\lambda}/\hat{\mu}=\hat{\lambda}/\{\sum_{i=1}^{i=n} \mu_i[p_{i0}/(1-\rho)]\}=(1-\rho)[\hat{\lambda}/(\hat{\mu}-\hat{\lambda})],$$

from which it follows that

$$\rho=\hat{\lambda}/\hat{\mu}=\theta. \quad (19)$$

We can now show the following result.

THEOREM. *If, for all* i, $\lambda_i/\mu_i=\theta$ (*which implies that* $\theta=\rho=\hat{\lambda}/\hat{\mu}$), *then* (i) *for* $\theta<1$,

$$p_{im}=p_{i\cdot}(1-\rho)\rho^m, \qquad (i=1, 2, \cdots, n; m \geqq 0) \quad (20)$$

and (ii) *for* $\theta \geqq 1$, $p_{im}=0$ *for all* m.

Proof. Part (ii) is a consequence of the condition $\hat{\mu}-\hat{\lambda}>0$. For (i) the proof will be by induction. By (16) the theorem is valid for $m=0$. Assuming that it holds up to $m>0$ (which implies that $p_{im}=p_{i,m-1}\rho$), and using (3′), we derive

$$p_{i,m+1}=p_{im}(\rho+1+\eta_i/\mu_i)-p_{i,m-1}\rho-\sum_{j=1}^{j=n} p_{jm}(\eta_{ji}/\mu_i)$$
$$=p_{im}\rho+(1/\mu_i)(p_{im}\eta_i-\sum_{j=1}^{j=n} p_{jm}\eta_{ji}).$$

But $p_{im}\eta_i=\sum_{j=1}^{j=n} p_{jm}\eta_{ji}$. This follows since, by using equation (6), we have

$$p_{im}\eta_i=p_{i\cdot}(1-\rho)\rho^m\eta_i=(1-\rho)\rho^m\pi_i/[\sum_{k=1}^{k=n}(\pi_k/\eta_k)],$$

and the use of equation (5) gives

$$\sum_{j=1}^{j=n}\eta_{jm}p_{jm}=(1-\rho)\rho^m\sum_{j=1}^{j=n}\eta_j q_{ji}(\pi_j/\eta_j)/[\sum_{k=1}^{k=n}(\pi_k/\eta_k)]$$
$$=(1-\rho)\rho^m\pi_i/[\sum_{k=1}^{k=n}(\pi_k/\eta_k)].$$

Hence, $p_{i,m+1}=p_{im}\rho=p_{i\cdot}(1-\rho)\rho^{m+1}$, which completes the proof.

The partial generating functions are now derived explicitly as

$$G_i(z)=p_{i\cdot}(1-\rho)\sum_{m=0}^{m=\infty}(z\rho)^m=p_{i\cdot}(1-\rho)/(1-\rho z)=p_{i\cdot}(\hat{\mu}-\hat{\lambda})/(\hat{\mu}-\hat{\lambda}z), \quad (i=1, 2, \cdots, n)$$

from which the average number of customers L, in the system is

$$L=\hat{\lambda}/(\hat{\mu}-\hat{\lambda}). \tag{21}$$

Alternatively, result (21) may be obtained directly from equation (20):

$$L=\sum_{i=1}^{i=n}\sum_{m=0}^{m=\infty} mp_{im}=\sum_{m=0}^{m=\infty} m(1-\rho)\rho^m=\rho/(1-\rho).$$

To summarize, the n-phase generalization of the (steady-state) $M/M/1$ queue will not yield, in general, closed-form solutions. Except for one interesting case where $\lambda_i/\mu_i=\theta$ for all i and simple results are obtained, numerical methods should be employed to solve any specific case. As was argued in reference 3, a library of computer routines may prove useful in such circumstances.

ACKNOWLEDGMENT

MUCH OF THIS work was done while the author was at New York University.

REFERENCES

1. S. KARLIN, *A First Course in Stochastic Processes*, Academic Press, New York, 1966.
2. I. L. MITRANI AND B. AVI-ITZHAK, "A Many-Server Queue with Service Interruptions," *Opns. Res.* **16,** 628–638(1968).
3. M. F. NEUTS, "A Queue Subject to Extraneous Phase Changes," *Adv. in Appl. Prob.* **3,** 78–119 (1971).
4. U. YECHIALI AND P. NAOR, "Queuing Problems with Heterogeneous Arrivals and Service," *Opns. Res.* **19,** 722–734(1971).