QUEUING PROBLEMS WITH HETEROGENEOUS ARRIVALS AND SERVICE

U. Yechiali

New York University, New York, N.Y.

and

P. Naor

The Technion, Haifa, Israel

(Received September 29, 1969)

This paper studies a two-level modification of the $M/M/1$ queuing model where the rate of arrival and the service capacity are subject to Poisson alternations. The ensuing 'two-dimensional' problem is analyzed by using partial-generating-function techniques, which appear to be essential in the present context. The steady-state probabilities and the expected queue are evaluated, and numerous special and extreme cases are analyzed in detail.

IN THE literature on queuing theory, a great number of probabilistic models possessing a variety of properties have been discussed. Ordinarily, in these models the parameters describing arrival intensity and/or service capacity possess one of the following characteristics: (i) The parameters are homogeneous in time. (ii) The parameters are not constant but vary in time; however, their temporal dependence is a datum of the model. (iii) The parameters, if left by themselves, are homogeneous, but heterogeneity is introduced by control action, e.g., customers are refused admittance to the waiting line if the queue size exceeds a certain level, or, again, service capacity is reinforced for the same reason.

An additional set of queuing problems may be considered as possessing characteristics of service *heterogeneity*, to wit, when the service station is subject to breakdown (e.g., GAVER[4] and AVI-ITZHAK AND NAOR[2]). In these models, the service rate alternates in a random fashion between a fixed arbitrary positive level and zero. Another study (SCOTT[8]) attempts to generalize by considering random changes in either the arrival or the service parameters.

The purpose of this paper is to discuss a further generalization, that is, the model analyzed here is one where both arrival intensity and service capacity undergo Poissonian jumps between *two* levels. Processes with similar underlying structure have been treated by others (e.g., KEILSON

AND WISHART[5,6]), but the case where the underlying chain consists of two states is of special interest by virtue of the added simplicity, the potential practical importance and the explicit solvability of the model. A number of areas suggest themselves where such models may be of practical use. Thus, for instance, a computing facility may be retained by a number of clients, each emitting a steady Poisson stream of customers. The appearance and disappearance of such a client is associated with the simultaneous Poissonian increase and decrease of overall arrival rate and service capacity.

Again, some production processes are associated with product diversification that may bring about random intensity changes of input and output. Another approach is to view the model as a discrete analog of certain continuous-time storage processes (e.g., GANI[3] and MILLER[7]).

The problem under consideration is a two-dimensional generalization of the typically one-dimensional basic queuing systems. It possesses the following characteristics: A stream of Poisson-type customers arrives at a single service station. The arrival pattern is *not* homogeneous; rather there exist two arrival intensities at which the system is capable of operating. The time interval during which the system functions at level i ($i=1, 2$) is an exponentially distributed random variable possessing the expected value $1/\eta_i$. Furthermore, it is assumed that any realization of a time interval associated with uniform arrival rate λ_i is independent of previous history. Whatever has been said about arrival characteristics holds for the service pattern as well. Service time is assumed to be exponentially distributed; if the system is at level i, the service intensity possesses the value μ_i, and, as before, statistical independence between any two realizations is assumed. Let it be mentioned in passing that the notion of independence is to be understood in a *conditional* sense: given that the system is at level i, previous history is of no predictive value.

We have, then, a single-server queuing system that oscillates between two feasible levels denoted by 1 and 2. The persistence of the system at any level is governed by a random mechanism: if the system functions at level i (i.e., the arrival and service rates are λ_i and μ_i, respectively) it tends 'to jump' to the alternative level with Poisson intensity η_i. We note explicitly that, once they have joined the queue, customers do *not* wear labels 1 or 2; rather the service rendered to them possesses the instantaneous rate associated with the present level of the system. Hence some basic properties of the queuing process with which this study is concerned (e.g., state probabilities, expected queue size, etc.) do not depend on the specification of the queue discipline.

We sum up and restate the setting of this study in a more formal way: Let $X(t)$ denote a Markov process on the states $\{(i, m)\}(i=1, 2;\ m=$

$0, 1, 2, \cdots$) and let its transition probabilities $\{P_{(jn),(im)}(t)\}$ be stationary, that is, for $t>0$.

$$P_{(jn),(im)}(t)=\Pr\{X(t+s)=(i, m)|X(s)=(j, n)\} \qquad (i, j=1, 2;\ m, n=0, 1, 2, \cdots) \tag{1}$$

is independent of $s\geqq 0$. Furthermore, for $h\downarrow 0$, the $\{P_{(jn),(im)}(t)\}$ satisfy

$$P_{(im),(i,m+1)}(h)=\lambda_i h+o(h), \tag{2}$$

$$P_{(i,m+1),(im)}(h)=\mu_i h+o(h), \tag{3}$$

$$P_{(1m),(2m)}(h)=\eta_1 h+o(h), \tag{4}$$

$$P_{(2m),(1m)}(h)=\eta_2 h+o(h), \tag{5}$$

$$P_{(im),(im)}(h)=1-(\lambda_i+\mu_i+\eta_i)h+o(h), \qquad (m\neq 0) \tag{6}$$

$$P_{(i0),(i0)}(h)=1-(\lambda_i+\mu_i)h+o(h), \tag{7}$$

$$P_{(jn),(im)}(0)=\begin{cases}1, & \text{if} \quad (jn)=(im),\\ 0, & \text{otherwise.}\end{cases} \tag{8}$$

The rates η_i, λ_i, and μ_i are nonnegative though at least one of the $\eta-s$, one of the $\lambda-s$, and one of the $\mu-s$ must be positive. Ordinarily all rates will be assumed finite. Only in Section II, when some extreme cases will be studied, will we allow η_1 or η_2 or both to tend to infinity.

The set of transition probabilities $\{P_{(jn),(im)}(t)\}$ satisfies the backward Kolmogorov differential equations, and, from the theory of recurrent events, it is known that for all (i, m) the limits $\lim_{t\to\infty} P_{(jn),(im)}(t)=p_{im}$ exist and are independent of the initial state (j, n). The set $\{p_{im}\}$ satisfies

$$p_{10}(\lambda_1+\eta_1)=p_{11}\mu_1+p_{20}\eta_2, \tag{9a}$$

$$p_{20}(\lambda_2+\eta_2)=p_{21}\mu_2+p_{10}\eta_1, \tag{9b}$$

$$p_{1m}(\lambda_1+\eta_1+\mu_1)=p_{1,m-1}\lambda_1+p_{1,m+1}\mu_1+p_{2m}\eta_2, \qquad (m>0) \tag{9c}$$

$$p_{2m}(\lambda_2+\eta_2+\mu_2)=p_{2,m-1}\lambda_2+p_{2,m+1}\mu_2+p_{1m}\eta_1. \qquad (m>0) \tag{9d}$$

It is convenient to present the set (9) in diagrammatic form, as in Fig. 1.

The various equations appearing in the set (9) may be considered as a representation of a law relating to the steady-state regime: The average rate at which a point (i.e., a state) is entered equals the average rate at which a transition from the point occurs.

Again from the theory of recurrent events it can be deduced that (for positive η_1 and η_2) the probabilities $\{p_{im}\}$ are either all positive or, alternatively, all equal to zero. Indeed inspection of (9) shows immediately that all $p_{im}-s$ are positive if one of them is positive; and all $p_{im}-s$ vanish if one of them equals zero.

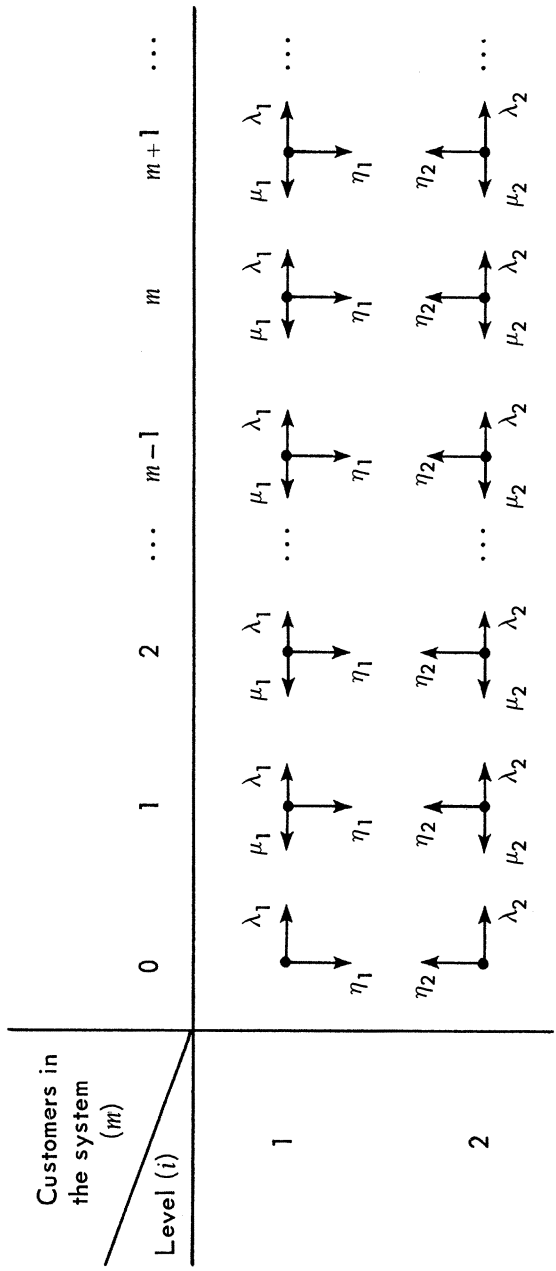

Figure 1

Simple algebraic operations on (9) yield

$$p_{1m}\lambda_1+p_{2m}\lambda_2=p_{1,m+1}\mu_1+p_{2,m+1}\mu_2. \qquad (m=0, 1, 2, \cdots) \quad (10)$$

Summation of (10) over all m yields.

$$p_{1\cdot}\lambda_1+p_{2\cdot}\lambda_2=(p_{1\cdot}-p_{10})\mu_1+(p_{2\cdot}-p_{20})\mu_2, \qquad (11)$$

where

$$p_{i\cdot}=\sum_{m=0}^{m=\infty} p_{im}. \qquad (12)$$

The quantity $p_{i\cdot}$ is the probability of the system being at level i.

Let two quantities $\hat{\lambda}$ and $\hat{\mu}$ be defined as

$$\hat{\lambda}=p_{1\cdot}\lambda_1+p_{2\cdot}\lambda_2 \qquad (13)$$

and

$$\hat{\mu}=p_{1\cdot}\mu_1+p_{2\cdot}\mu_2. \qquad (14)$$

The physical interpretation of these quantities is straightforward: $\hat{\lambda}$ is the average rate of customer arrivals; $\hat{\mu}$ is the average capacity of the system to render service.

Relation (11) may be written as

$$p_{10}\mu_1+p_{20}\mu_2=\hat{\mu}-\hat{\lambda}. \qquad (15)$$

For the case of interest—the steady-state regime—in which the set $\{p_{im}\}$ is positive throughout, we can deduce from equation (15) that the following relation must hold:

$$\hat{\mu}-\hat{\lambda}>0; \qquad (16)$$

that is, for steady-state conditions, the average service capacity of the system must exceed the average arrival rate.

On viewing the underlying two-state Markov process, we immediately obtain

$$p_{1\cdot}=\eta_2/(\eta_1+\eta_2) \qquad (17a)$$

and

$$p_{2\cdot}=\eta_1/(\eta_1+\eta_2), \qquad (17b)$$

which is, of course, consistent with set (10).

As a solution, we desire to express the state probabilities in their functional dependence on the parameter set $\{\lambda_1, \lambda_2, \mu_1, \mu_2, \eta_1, \eta_2\}$. Apparently there is no simple way of solving (9) in a straightforward recursive manner. The generating-function techniques proposed here go beyond the typical application. They are not just a compact and convenient way of presentation; rather, they appear to be *essential* for the analysis of the model under study.

I. GENERATING FUNCTIONS AND MODEL CHARACTERISTICS

WE DEFINE THE partial generating functions of the system as

$$G_i(z) = \sum_{m=0}^{m=\infty} z^m p_{im}, \qquad |z| \leqq 1,\ i = 1,\ 2. \tag{18}$$

Let the set of equations (9a) and (9c) be slightly modified and rewritten as

$$p_{10}(\lambda_1+\eta_1+\mu_1) = p_{20}\eta_2 + p_{11}\mu_1 + p_{10}\mu_1, \tag{19a}$$

$$p_{1m}(\lambda_1+\eta_1+\mu_1) = p_{1,m-1}\lambda_1 + p_{2m}\eta_2 + p_{1,m+1}\mu_1. \qquad (m>0) \tag{19b}$$

We multiply each equation of the set (19) by $z^m (m = 0,\ 1,\ \cdots)$ appropriately and sum over all m. This process results in

$$(\lambda_1+\eta_1+\mu_1)G_1(z) = \lambda_1 z G_1(z) + \eta_2 G_2(z) + (\mu_1/z)[G_1(z) - p_{10}] + p_{10}\mu_1. \tag{20}$$

In an analogous fashion, we obtain

$$(\lambda_2+\eta_2+\mu_2)G_2(z) = \lambda_2 z G_2(z) + \eta_1 G_1(z) + (\mu_2/z)[G_2(z) - p_{20}] + p_{20}\mu_2. \tag{21}$$

Next, we define a polynomial of the third degree, $g(z)$, as follows

$$\begin{aligned} g(z) = \lambda_1\lambda_2 z^3 - (\eta_1\lambda_2 + \eta_2\lambda_1 + \lambda_1\lambda_2 + \lambda_1\mu_2 + \lambda_2\mu_1)z^2 \\ + (\eta_1\mu_2 + \eta_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2 + \lambda_2\mu_1)z - \mu_1\mu_2. \end{aligned} \tag{22}$$

On utilizing (20), (21), and (22) we arrive at

$$g(z)G_1(z) = p_{20}\eta_2\mu_2 z + p_{10}\mu_1[\eta_2 z + \lambda_2 z(1-z) - \mu_2(1-z)]. \tag{23}$$

THEOREM. *For positive μ_1 and μ_2 and finite η_1 and η_2, the polynomial $g(z)$ possesses a unique root z_0 in the open interval* (0, 1).

Proof. (i) Let $z=0$; then $g(0) = -\mu_1\mu_2 < 0$.

(ii) Let $z=1$; since $G_1(1) = p_{1\cdot} > 0$ and $g(1)G_1(1) = \eta_2(p_{10}\mu_1 + p_{20}\mu_2) > 0$, it follows that $g(1) > 0$. Thus, the number of roots in the interval (0, 1) is odd (either *one* or *three*).

(iii) Assume—without loss of generality—that

$$\mu_2/\lambda_2 \geqq \mu_1/\lambda_1. \tag{24}$$

Since $\hat{\mu} > \hat{\lambda}$, it follows, then, that $\mu_2/\lambda_2 > 1$. But

$$g(\mu_2/\lambda_2) = (\mu_2/\lambda_2)\eta_2[\mu_1 - (\mu_2/\lambda_2)\lambda_1] \leqq 0.$$

Hence, there exists a root of $g(z)$ in the interval $(1,\ \mu_2/\lambda_2]$, and, therefore, the number of roots in the interval (0, 1) equals *one*. This completes the proof.

Typically, Cardano's solution of $g(z)$ will yield no simple algebraic expression for z_0, though—as will be shown later—in some limiting cases simplification is possible.

The probabilities p_{10} and p_{20} can now be obtained in the following manner: Setting $z=z_0$ in equation (23), we have

$$p_{20}\eta_2\mu_2 z_0+p_{10}\mu_1[\eta_2 z_0+\lambda_2 z_0(1-z_0)-\mu_2(1-z_0)]=0. \tag{25}$$

Inserting (15) in (25), we get

$$p_{10}=\eta_2(\hat{\mu}-\hat{\lambda})z_0/\mu_1(1-z_0)(\mu_2-\lambda_2 z_0), \tag{26}$$

and, similarly,

$$p_{20}=\eta_1(\hat{\mu}-\hat{\lambda})z_0/\mu_2(1-z_0)(\mu_1-\lambda_1 z_0). \tag{27}$$

We recollect that the busy fraction ρ of the service station is represented by

$$\rho=1-(p_{10}+p_{20}). \tag{28}$$

We may mention, in passing, that—contrary to intuition—typically, the busy fraction ρ does *not* equal the ratio of the average arrival intensity $\hat{\lambda}$ to the average service capacity $\hat{\mu}$; that is, $\rho\neq\hat{\lambda}/\hat{\mu}$, generally speaking, though in one special case the equality does hold. Indeed, in Section II we shall show that $\rho=\hat{\lambda}/\hat{\mu}$ if and only if $p_{10}/p_{20}=p_{1\cdot}/p_{2\cdot}$, and this equality occurs if and only if $\lambda_1/\mu_1=\lambda_2/\mu_2$.

While there seems to be no simple and compact formula relating $\{p_{im}\}$ to p_{10} and p_{20}, there are no serious computational difficulties. After some manipulation of equations (9) we arrive at a computationally convenient set of recursive expressions

$$p_{1m}=p_{1,m-1}(\lambda_1/\mu_1)+(\textstyle\sum_{j=0}^{m-1}p_{1j})(\eta_1/\mu_1)-(\sum_{j=0}^{m-1}p_{2j})(\eta_2/\mu_2), \quad (m>0) \tag{29a}$$

and

$$p_{2m}=p_{2,m-1}(\lambda_2/\mu_2)+(\textstyle\sum_{j=0}^{m-1}p_{2j})(\eta_2/\mu_2)-(\sum_{j=0}^{m-1}p_{1j})(\eta_1/\mu_2). \quad (m>0) \tag{29b}$$

The partial generating functions are completely known once the values of p_{10} and p_{20} have been established, as shown above. Hence, the expected queue size (as well as higher moments) can be determined by standard procedures.

Combination of (15) and (23) results in

$$G_1(z)=[\eta_2(\hat{\mu}-\hat{\lambda})z+p_{10}\mu_1(1-z)(\lambda_2 z-\mu_2)]/g(z), \tag{30a}$$

$$G_2(z)=[\eta_1(\hat{\mu}-\hat{\lambda})z+p_{20}\mu_2(1-z)(\lambda_1 z-\mu_1)]/g(z). \tag{30b}$$

Let auxiliary quantities M_i be defined as

$$M_i=\textstyle\sum_{m=0}^{m=\infty} mp_{im}. \qquad (i=1,2) \tag{31}$$

Clearly, we have

$$(d/dz)G_i(z)|_{z=1}=M_i. \tag{32}$$

The quantity M_i may be considered as the contribution of level i to the mean queue size; it is the product of the probability of the system being at level i and the conditional mean queue size, given that the system is at level i.

Hence, the (unconditional) expected queue size Eq, after some development, is found to be

$$\begin{aligned} Eq &= M_1+M_2 \\ &= \hat{\lambda}/(\hat{\mu}-\hat{\lambda})+[\mu_1(\mu_2-\lambda_2)p_{10}+\mu_2(\mu_1-\lambda_1)p_{20}-(\mu_1-\lambda_1)(\mu_2-\lambda_2)]/ \\ &\quad (\eta_1+\eta_2)(\hat{\mu}-\hat{\lambda}). \end{aligned} \tag{33}$$

II. SPECIAL AND EXTREME CASES

IN THE PRECEDING section, the model was treated in rather general terms. We can envisage many cases where it is not necessary to assume that $\lambda_1 \neq \lambda_2$, or, alternately, that $\mu_1 \neq \mu_2$. Apparently there is no significant simplification in the expressions of the previous section if the more specialized assumptions $\lambda_1=\lambda_2$ or $\mu_1=\mu_2$ are introduced. However, there is one case where a specialized assumption causes the final expressions to be of extreme simplicity. This is the case where the traffic intensities λ_1/μ_1 and λ_2/μ_2 are equal, though arrival intensities and service capacities need not be equal.

First we investigate the properties of the ratio p_{10}/p_{20}. Use of the relations (26) and (27) leads to

$$p_{10}/p_{20}=(\eta_2/\eta_1)[1-(\lambda_1/\mu_1)z_0]/[1-(\lambda_2/\mu_2)z_0]. \tag{34}$$

Since $\eta_2/\eta_1=p_{1\cdot}/p_{2\cdot}$, it follows that, whenever $z_0>0$, $p_{10}/p_{20}=p_{1\cdot}/p_{2\cdot}$ if and only if $\mu_1/\lambda_1=\mu_2/\lambda_2$. In such a situation, let this ratio be defined as θ; that is,

$$\mu_1/\lambda_1=\mu_2/\lambda_2=\theta. \tag{35}$$

We have then, immediately,

$$\begin{aligned} \hat{\mu}/\hat{\lambda} &= (\mu_1 p_{1\cdot}+\mu_2 p_{2\cdot})/(\lambda_1 p_{1\cdot}+\lambda_2 p_{2\cdot}) \\ &= (\theta\lambda_1 p_{1\cdot}+\theta\lambda_2 p_{2\cdot})/(\lambda_1 p_{1\cdot}+\lambda_2 p_{2\cdot})=\theta. \end{aligned} \tag{36}$$

We recall (28), where the traffic intensity was expressed by $\rho=1-(p_{10}+p_{20})$. This intensity is usually not equal to $\hat{\lambda}/\hat{\mu}$. However, in the particular case under consideration (and in this case *only*) the equality $\rho=\hat{\lambda}/\hat{\mu}$ *does* hold. The proof is rather elementary: If the equality is assumed, then $p_{10}+p_{20}=(\hat{\mu}-\hat{\lambda})/\hat{\mu}$. But, on using (15), we obtain

$$\begin{aligned} p_{10}/p_{20} &= (\hat{\mu}-\mu_2)/(\mu_1-\hat{\mu}) \\ &= (\mu_1 p_{1\cdot}+\mu_2 p_{2\cdot}-\mu_2)/(\mu_1-\mu_1 p_{1\cdot}-\mu_2 p_{2\cdot})=p_{1\cdot}/p_{2\cdot}. \end{aligned} \tag{37}$$

A closely related argument can be presented to prove that the assumption $p_{10}/p_{20}=p_{1\cdot}/p_{2\cdot}$ implies the result $\rho=\hat{\lambda}/\hat{\mu}$. However, it will be more instructive to use the properties of the polynomial $g(z)$.

It is not difficult to verify that

$$g(\theta)=0. \tag{38}$$

Hence, we have the decomposition

$$g(z)=\lambda_1\lambda_2(z^2-kz+\theta)(z-\theta), \tag{39}$$

where

$$k=\eta_1/\lambda_1+\eta_2/\lambda_2+1+\theta. \tag{40}$$

The root of interest z_0, which is located in the interval (0, 1), is equal to

$$z_0=(k-\sqrt{k^2-4\theta})/2. \tag{41}$$

In formulas (26) and (27) the ratio $z_0/(1-z_0)(1-z_0/\theta)$ makes its appearance. Algebraic manipulation yields

$$z_0/(1-z_0)(1-z_0/\theta)=\theta\lambda_1\lambda_2/(\eta_1+\eta_2)\hat{\lambda}. \tag{42}$$

Substituting this result in (26) and (27), we obtain

$$p_{i0}=p_{i\cdot}(1-1/\theta). \qquad (i=1, 2) \tag{43}$$

Using (28), we finally get

$$\rho=1-(p_{10}+p_{20})=1/\theta=\hat{\lambda}/\hat{\mu}. \tag{44}$$

THEOREM. *If relation* (39) *holds, then*

$$p_{im}=p_{i\cdot}(1-\rho)\rho^m. \qquad (i=1, 2;\ m=0, 1, \cdots) \tag{45}$$

The proof will be by induction. By (43) the theorem is valid for $m=0$. By using (9a) and (9b), it is immediate that it holds for $m=1$. Assume now that it holds up to some $m>0$; then it holds for $m+1$ as well, since by (9c) we derive

$$\begin{aligned} p_{1,m+1}&=p_{1m}(\rho+\eta_1/\mu_1+1)-p_{2m}(\eta_2/\mu_1)-p_{1,m-1}\rho \\ &=[1/(\eta_1+\eta_2)](1-\rho)\rho^m[\eta_2(\rho+\eta_1/\mu_1+1)-\eta_1(\eta_2/\mu_1)-\eta_2] \\ &=p_{1\cdot}(1-\rho)\rho^{m+1}, \end{aligned} \tag{46}$$

and similarly for $p_{2,m+1}$. This completes the proof.

It is interesting to obtain the probability of having m customers in the queue regardless of level:

$$p_{\cdot m}=p_{1m}+p_{2m}=(1-\rho)\rho^m. \tag{47}$$

The partial generating functions are derived as

$$G_i(z) = p_{i\cdot}(1-\rho)\sum_{m=0}^{\infty}(z\rho)^m = p_{i\cdot}(\hat{\mu}-\hat{\lambda})/(\hat{\mu}-\hat{\lambda}z), \qquad (i=1, 2) \quad (48)$$

from which it is obtained that

$$Eq = \hat{\lambda}/(\hat{\mu}-\hat{\lambda}). \qquad (49)$$

The set of relations that were derived in this case is closely related to the single-server queue with Poisson input and exponential service ($M/M/1$). As pointed out before, this is the only case where such a simple extension of the $M/M/1$ formulas exist.

It is possible to obtain these relations by a slightly different avenue of approach. Let us assume that steady-state conditions have been attained and, furthermore, that at the present moment a transition from one level to the second level has taken place. Now this transition will carry no influence on the random variable 'number of customers present in the queue,' since the traffic intensity $\lambda_i/\mu_i(=\rho)$ has not changed (what has changed is the average number of transitions per unit time, which is different for the two levels). Hence, the conditional distributions of this random variable are identical for both levels 1 and 2 and the state probabilities are given simply by

$$p_{im} = p_{i\cdot}p_{\cdot m}, \qquad (50)$$

a formula equivalent to an appropriate combination of (45) and (47).

A special case of a model discussed by White and Christie,[9] by Avi-Itzhak and Naor,[1,2] and by Gaver[4] may also be regarded as a special case of the present model. In particular, it was assumed in these other studies that the service station is incapacitated from time to time and resumes its operation after a random time. In the notation of the present study, this is equivalent to assuming that $\lambda_1=\lambda_2=\lambda$ and $\mu_2=0$. Using (15), we obtain

$$p_{10} = p_{1\cdot} - \lambda/\mu_1. \qquad (51)$$

Substituting this value in (33), we derive

$$Eq = \{\lambda+[\lambda\mu_1/(\eta_1+\eta_1)]p_{2\cdot}\}/(\mu_1 p_{1\cdot}-\lambda), \qquad (52)$$

which is equivalent to queuing formulas obtained by the above authors.

Next, we examine a number of extreme cases.

Case A. It is assumed here that one of the level transition intensities, η_1, say, vanishes. It is immediately clear that, under such circumstances, we deal with an $M/M/1$ queuing system with λ_1 and μ_1 as arrival and service intensities. Indeed, factorization of $g(z)$ and further manipulation yield

$$p_{10} = (\hat{\mu}-\hat{\lambda})/\mu_1 = 1-\lambda_1/\mu_1. \qquad (53)$$

Further, the probabilities $\{p_{1m}\}$ are geometrically distributed, whereas all $\{p_{2m}\}$ vanish.

Case B. This is another extreme situation of some simplicity. Here we let $\eta_2 \to \infty$, whereas η_1 is positive and finite. Again it is clear that this case converges to an $M/M/1$ queue on level 1.

Case C. Here we assume that very rapid oscillations occur between the two levels 1 and 2. More specifically, we let η_1 and η_2 tend simultaneously to infinity with the proviso that the ratio η_1/η_2 tends to a positive and finite constant C.

We note that the probabilities associated with the levels may be presented as

$$p_{1\cdot} = 1/(1+C), \tag{54a}$$

$$p_{2\cdot} = C/(1+C). \tag{54b}$$

On utilizing (19), we get

$$Cp_{1m} = p_{2m}, \qquad (m = 0, 1, \cdots) \tag{55}$$

or, equivalently,

$$CG_1(z) = G_2(z). \tag{56}$$

Insertion of (55) in (15) results in

$$p_{10} = (\hat{\mu} - \hat{\lambda})/(\mu_1 + \mu_2 C) = (1 - \hat{\lambda}/\hat{\mu})p_{1\cdot}. \tag{57}$$

On using induction arguments, we eventually derive

$$p_{im} = p_{i\cdot}(1 - \hat{\lambda}/\hat{\mu})(\hat{\lambda}/\hat{\mu})^m. \tag{58}$$

In other words, we have again obtained a geometric distribution over the states m with parameter $\hat{\lambda}/\hat{\mu}$. The physical interpretation of (58) is simply that, in the case of extremely rapid oscillations between levels 1 and 2, the arrival becomes homogeneously Poissonian with weighted intensity $\hat{\lambda}$; an analogous statement holds true for weighted service capacity $\hat{\mu}$.

Case D. Under this heading we shall deal with a situation where transitions between levels are very sluggish; that is, oscillations occur infrequently. In more formal terms, we assume that the transition intensities η_1 and η_2 are arbitrarily close to zero, while the ratio η_1/η_2 equals a finite, nonzero constant C. We shall have to distinguish between two subcases:

D1. Here it is assumed that both arrival rates fall short of their corresponding service capacities, i.e., $\lambda_1 < \mu_1$, $\lambda_2 < \mu_2$. Combination of (22) and (23) (and letting $\eta_1 \downarrow 0^+$, $\eta_2 \downarrow 0^+$) yields

$$(\mu_1 - \lambda_1 z)G_1(z) \cong \mu_1 p_{10}. \tag{59}$$

Letting $z=1$ in (59) and further use of (9) result in

$$p_{im} \cong p_{i\cdot}(1-\lambda_i/\mu_i)(\lambda_i/\mu_i)^m. \tag{60}$$

This can be interpreted in the following manner: If oscillations between levels occur very infrequently and if, on both levels, arrival rate falls short of service capacity, the system settles in two distinct quasi-equilibria (one at a time). Each quasi-equilibrium is of $M/M/1$ type).

D2. In this subcase, we shall assume that one of the service capacities, μ_1, say, does not exceed its corresponding arrival rate λ_1, that is, $\lambda_1 \geqq \mu_1$. However, we recollect the condition that the weighted service capacity $\hat{\mu}$ has to exceed the weighted arrival rate, or, equivalently,

$$\mu_2 > \lambda_2 + (\lambda_1 - \mu_1)\eta_2/\eta_1. \tag{61}$$

Under such circumstances, there can be no quasi-equilibrium at level 1. For arbitrarily small η_i we are able—as will be shown below—to accumulate an appropriately large average queue whenever the system is at level 1. This large queue is decreased to quasi-equilibrium size of the type discussed in subcase D1 whenever the system functions at level 2.

In a more formal way, let us rewrite (22) as

$$g(z) = (z-1)(\lambda_1 z - \mu_1)(\lambda_2 z - \mu_2) + (\eta_1+\eta_2)z(\hat{\mu} - \hat{\lambda}z). \tag{62}$$

Now $\eta_1, \eta_2 > 0$ (however small), and it follows that $z_0 < \mu_1/\lambda_1$ and $(\mu_1 - \lambda_1 z_0)$ is of the order $(\eta_1+\eta_2)$, since

$$(\mu_1 - \lambda_1 z_0)/(\eta_1+\eta_2) = z_0(\hat{\mu} - \hat{\lambda}z_0)/(1-z_0)(\mu_2 - \lambda_2 z_0). \tag{63}$$

From inspection of (26) it becomes immediately obvious that p_{10} is small of the order η_2. This is *not* the case for p_{20}. Formula (27) is not immediately applicable here, since both numerator and denominator are close to zero. If (63) is combined with (27) and (17), we obtain

$$p_{20} = p_{2\cdot}\{(\hat{\mu} - \hat{\lambda})[1 - (\lambda_2/\mu_2)z_0]\}/(\hat{\mu} - \hat{\lambda}z_0). \tag{64}$$

Since for very small η_i the quantity z_0 is close to μ_1/λ_1, we get, after substitututing $z_0 \cong \mu_1/\lambda_1$, the approximation

$$p_{20} \cong (\hat{\mu} - \hat{\lambda})/\mu_2. \tag{65}$$

The interpretation to be attached to this result is this: For very small η_i and $\lambda_1 \geqq \mu_1$ and $\lambda_2 < \mu_2$ such that $\hat{\lambda} < \hat{\mu}$, the idle capacity at level 1 has disappeared. Hence, by (15) the overall idle capacity equals $p_{20}\mu_2$, which is another way of expressing relation (65).

It remains to find an expression for the expected queue size. As was pointed out earlier [and can be observed from (33)], Eq can be made

arbitrarily large by an appropriate choice of η_i. However, the product of Eq and $(\eta_1+\eta_2)$ tends to a finite limit:

$$\begin{aligned}(\eta_1+\eta_2)Eq &\cong(\lambda_1-\mu_1)(\mu_2-\lambda_2-\mu_2 p_{20})/(\hat{\mu}-\hat{\lambda}) \\ &\cong(\lambda_1-\mu_1)[(\mu_2-\lambda_2)-(\hat{\mu}-\hat{\lambda})]/\hat{\mu}-\hat{\lambda}.\end{aligned} \tag{66}$$

This limit equals zero when $\lambda_1=\mu_1$ and is positive for the case $\lambda_1>\mu_1$.

ACKNOWLEDGMENTS

THIS RESEARCH was sponsored in part by the Logistics and Mathematical Statistics Branch, Office of Naval Research, Washington, D.C.

The authors wish to express their appreciation to I. MITRANY for a number of fruitful discussions.

REFERENCES

1. B. AVI-ITZHAK AND P. NAOR, "On a Problem of Preemptive Priority Queuing," *Opns. Res.* **9,** 664–672 (1961).
2. ——— AND ———, "Some Queuing Problems with the Service Station Subject to Breakdown," *Opns. Res.* **11,** 303–320 (1963).
3. J. GANI, "The Time-Dependent Solution for a Dam with Ordered Poisson Inputs," pp. 101–109 in *Studies in Applied Probability and Management Science,* K. J. ARROW, S. KARLIN, AND H. SCARF (eds.), Stanford University Press, (1962).
4. D. P. GAVER, JR., "A Waiting Line with Interrupted Service, Including Priorities," *J. Roy. Stat. Soc. B* **24,** 73–90 (1962).
5. J. KEILSON AND D. M. G. WISHART, "A Central Limit Theorem for Processes Defined on a Finite Markov Chain," *Proc. Cambridge Philos. Soc.* (GB) **60,** 547–67 (1964).
6. ——— AND ———, "Boundary Problems for Additive Processes Defined on a Finite Markov Chain," *Proc. Cambridge Philos. Soc.* (GB) **61,** 173–190 (1965).
7. R. G. MILLER, JR., "Continuous Time Stochastic Storage Processes with Random Linear Inputs and Outputs," *J. Math. and Mech.* **12,** 275–291 (1963).
8. MECKINLEY SCOTT, "A Study of Some Single-Counter Queuing Processes," Doctoral Thesis submitted to the University of North Carolina at Chapel Hill (1964).
9. HARRISON WHITE AND LEE S. CHRISTIE, "Queuing with Preemptive Priorities or with Breakdown," *Opns. Res.* **6,** 79–95 (1958).