G. Schay, Jr.

Approximate Methods for a Multiqueueing Problem

Abstract: Two approximate methods are described and used to obtain the stationary distribution of the length of a queue which is a small part of a system of many queues. The methods are based on an analogy to statistical mechanics, and give simple approximate solutions of a problem whose exact handling would be extremely complex and would require much more information than is available.

Introduction

In this paper we discuss the following queueing problem: M queues are served by a single server. The number of items in each queue is unlimited, losses are not allowed. The server moves around from queue i to queue $i + 1$ mod M at a constant speed, regardless of whether it serves a queue or not. If the server is free when it arrives at a queue, and if that queue is not empty, then its first item will be picked up by the server, which holds it for some time, while moving along unhindered from queue to queue. When the service of an item terminates, the server takes its next item from the next non-empty queue it encounters. On each pass of the server, at most one item is taken from each queue.

The concrete setup in which this situation arose was part of a proposed TELE-PROCESSING® system[1] and only very little information was available about the input and output processes. We had to find approximations for the equilibrium distribution of any of the individual queues' length, with only minimal hypotheses about these processes.

Approximate techniques were, however, necessary not merely for this reason alone. Even in the simplest cases of well-specified input and output distribution, the exact calculations would require the determination of the joint distribution of the lengths of all queues, although one is usually (as in our case) interested only in the distributions describing any single queue. Since the number of parameters in the joint distribution is very large, it is practically impossible to find the exact solution. We shall fix our attention at a given queue and assume that the average arrival rate at this one is much less than that at all the M queues taken together. This assumption permits us to use approximate methods which are motivated by an analogy to Gibbs' method in statistical mechanics. In that theory a small system is considered as being in weak interaction with a large one, in such a way that the interaction is considered on the one hand strong enough to equalize the temperature, but on the other hand so weak as to permit consideration of the two systems as being mechanically independent in equilibrium.

Thus, the above assumption on the arrival rates will be replaced by certain independence assumptions. These make the mathematical problem fully determined and serve as approximations to the physical situation. Two different approximate methods will be described and used to treat the problem. In both of them a discrete time scale is used, the time between two successive arrivals of the server at a given queue being the unit of time. It is also assumed that either none or only one item can arrive at each queue in unit time. The existence of a stationary state of the whole system is also taken for granted with both methods. About the service times we assume only that each one has the same expected value.

In the case of the first method, the following further assumptions are made: First, the probability of a unit arriving at any of the queues at any time does not depend on that time. Second, the three random variables describing (a) whether at any time i a unit arrives at the given queue, (b) the number of items in the same queue at time $i - 1$, and (c) the state of the server (busy or not) at time i, are independent of each other.

It is important to point out that we do not require the independence of the inter-arrival times and of the service times. In this application the arrivals were certainly not independent.

In our second method the following hypotheses differ from those of the first method. First we take the arrival distribution at each queue as given by the probability $f(i)$ of having a time interval i between the arrival of any two consecutive units. This implies the assumption made in the first method, that the probability of a unit's arrival at any time does not depend on that time, but in addition to that, it requires somewhat more.

The inter-arrival times, although identically distributed, are still not required to be independent. The second assumption, differing from the corresponding one in the first method is that (a) any given inter-arrival time (b) the number of items in the selected queue at the beginning of that time, and (c) the state of the server at every later time are independent of each other. In the first method we compute the stationary distribution of the selected queue's length at any time, and in the second method, at the time of arrivals at that queue. Both these methods lead to equations that are standard in the theory of queues. The above conditions, however, are entirely different from those under which they have been deduced in the literature,[2] and give the solution of different problems.

A concrete model

The abstract queueing problem described in the Introduction can be illustrated by the following concrete example.

Let us say a shop manufactures M different products; for instance, screws of different sizes. They wish to pack them in boxes with L screws of the same size in each box. They have a revolving circular conveyor partitioned into M sections S_j, $j = 1, \cdots, M$. M is assumed to be a number of the order of 10 or 100. Each S_j contains boxes labelled j. There are $M + 1$ workers around the conveyor with M of them making screws. The j^{th} worker makes the j^{th} kind of screw, and puts each screw into a box labelled j, when S_j is next to him. He starts a new box when, and only when, one gets filled. (In the following we shall often use the term "units" to refer to the full boxes in the queue.) The $(M + 1)^{\text{th}}$ worker puts the full boxes into a machine A whenever a full box passes by him if machine A is not still occupied by a previous unit. If machine A is occupied, he does nothing.

The problem is to determine the probability distribution of the number of units in S_j as a function of some parameters describing the distribution of the arrival rates of the screws to each S_j and the speed of the machine A which handles the full boxes.

Method 1

In all that follows, we consider only one arbitrarily chosen section S_j, which we call simply S.

We also use the following notations:

i denotes the measure of time in units of revolutions of the conveyor, with increments occurring when section S leaves output A. $i = 0, 1, 2, \cdots$

$n(i)$ represents the number of full boxes in S at time i, or more precisely, the number immediately after the instant when S has left the output A for the i^{th} time. $n(i)$ is a random variable which may take the values $0, 1, 2, \cdots$

$N(i)$ is a random variable which is 0 if no unit arrives at S during the cycle which ends when S leaves A for the i^{th} time, and which is 1 if a unit does arrive in this cycle. (We assume that only one screw can enter S in a cycle, so no more than a single unit can arrive at a time.)

We consider finally a random variable $\zeta(i)$ which is 0 if A is busy and 1 if A is free when S arrives at A for the i^{th} time.

We assume a stationary distribution for $N(i)$, i.e., one independent of time, and use the following notation for the probabilities of the states of $N(i)$: $P[N(i) = 1] = p_N$ and $P[N(i) = 0] = 1 - p_N = q_N$. We assume that if a sufficient time has elapsed after the start of the operation, the process $\zeta(i)$ also becomes stationary, and its distribution may be given as $P[\zeta(i) = 1] = p_\zeta$ and $P[\zeta(i) = 0] = 1 - p_\zeta = q_\zeta$.

Finally, we assume that on the average, at least Mp_N units would arrive on the whole conveyor during a cycle, that is, the chosen S carries on the average at most $1/M$ of the total traffic. This, however, does not imply that all S_j's are equivalent; we only want to insure that S is a small part of the system.

Consider now the question how can $n(i + 1) = m \geq 1$ be realized from different states at time i. This can happen in four mutually exclusive ways:

1. $n(i) = m - 1$, and at time $i + 1$ one box gets filled and none leaves S.
2. $n(i) = m$, and at $i + 1$ nothing happens.
3. $n(i) = m$, and at $i + 1$ one box gets filled and immediately departs.
4. $n(i) = m + 1$, and at $i + 1$ nothing comes in, and one box departs.

This gives us the following relation:

$$\begin{aligned} P[n(i+1) = m] &= P[n(i) = m-1,\ N(i+1) \\ &\quad = 1,\ \zeta(i+1) = 0] \\ &+ P[n(i) = m,\ N(i+1) = 0,\ \zeta(i+1) = 0] \\ &+ P[n(i) = m,\ N(i+1) = 1,\ \zeta(i+1) = 1] \\ &+ P[n(i) = m+1,\ N(i+1) = 0,\ \zeta(i+1) = 1] \\ &\quad \text{for } m \geq 1\,. \end{aligned} \tag{1}$$

We now make the approximation that the variables $n(i)$, $N(i + 1)$ and $\zeta(i + 1)$ are independent. This may be justified by the following argument:

First, we have assumed that S carries, on the average, at most one M^{th} of the total traffic. Thus, if M is a large number, then $\zeta(i + 1)$ can depend only very weakly on the state of S, that is on $n(i)$ and $N(i + 1)$ since many of the S_i will be competing for the use of A.

Second, the fact whether something arrives at time $i + 1$ at S, that is, the value of $N(i + 1)$, is determined by causes external to the system. It can, however, depend on the time of the previous arrivals at S. These, on the other hand, partially determine $n(i)$. Thus, $N(i + 1)$ is related to $n(i)$ through the latter's dependence on the arrivals at S prior to $i + 1$. If $N(i + 1)$ is independent of the earlier arrivals, then it is also independent of $n(i)$. But even if $N(i + 1)$ does depend on the earlier arrivals, its dependence on $n(i)$ must be considerably weaker than that. This is so, because the previous arrivals do not determine $n(i)$ completely, as

$n(i)$ depends strongly on the previous values of ζ (this, of course, does not contradict the weak dependence of $n(i)$ on $\zeta(i+1)$, postulated earlier) determined primarily by the other queues of the system. If the mutual dependence of the arrivals is such that an arrival at time i makes other arrivals close to that time unlikely (as in the case of our model), then the assumption of independence attributes greater probabilities than the actual ones to the arrival of a unit if the queue in S is long. The reason for this is the following. In the case of such dependence, the conditional probability $P[N(i+1)=1|n(i)=m]$, of a unit arriving at S if there are m units there, has to decrease monotonically with increasing m, since the more units there are in S, the more likely it is that the last one has just arrived, and then for a considerable time no new unit will arrive. Now the assumption of the independence of $N(i+1)$ and $n(i)$ means that we replace each $P[N(i+1)=1|n(i)=m]$ for every m by their mean value

$$P[N(i+1)=1]=\sum_{m=0}^{\infty} P[N(i+1)=1|n(i)=m] \times P[n(i)=m].$$

Thus, by making this assumption we increase the probabilities of long queues, and this makes the design more conservative, that is, the error introduced by the approximation is in the admissible direction.

The joint probabilities in Eq. (1) may now be written as products of the individual probabilities. Moreover, in the stationary limit these probabilities do not depend on i. Thus, writing $P[n(i)=k]=P_k$, Eq. (1) becomes

$$P_m = p_N q_\zeta P_{m-1} + (q_N q_\zeta + p_N p_\zeta)P_m + q_N p_\zeta P_{m+1} \quad \text{for} \quad m \geq 1. \tag{2}$$

Similarly for $m=0$,

$$P_0 = (q_N + p_N p_\zeta)P_0 + q_N p_\zeta P_1. \tag{3}$$

We also have the condition

$$\sum_{k=0}^{\infty} P_k = 1. \tag{4}$$

The solution of the infinite system (2), (3), and (4) is

$$P_m = (1-x)x^m, \tag{5}$$

where

$$x = \frac{p_N q_\zeta}{q_N p_\zeta}. \tag{6}$$

Let us see now what conclusions we can draw for the design of the system. First of all, in view of (5), the series in (4) converges only if $x<1$. With (6) this gives

$$p_\zeta > p_N. \tag{7}$$

This is, of course, intuitively rather obvious: The probability of finding the output A empty when S arrives there in any cycle has to be greater than the probability of the arrival of a unit at S during a cycle, if the queue is not to grow without limit. The inequality (7) does not tell, however, how long the queues will grow. Evidently, the nearer p_ζ is to p_N, the more probable long queues will become.

From (5) it follows that

$$P[n(i) \geq k] = \sum_{m=k}^{\infty} P_m = x^k. \tag{8}$$

That is, the probability of obtaining queues longer than or equal to k units is the k^{th} power of the parameter x, which is given by Eq. (6).

The probability q_ζ of A being busy can be expressed in terms of the average number $\bar{N}$ of units arriving per cycle at each of the M sections of the conveyor ($\bar{N}$ is now taken equal for every section), and the average time $\bar{T}$ (in units of number of cycles) spent by each box in A, as

$$q_\zeta = \begin{cases} 1 & \text{if} \quad M\bar{N}\bar{T} \geq 1 \\ M\bar{N}\bar{T} & \text{if} \quad M\bar{N}\bar{T} < 1. \end{cases} \tag{9}$$

On the other hand,

$$\bar{N} = 1p_N + 0q_N = p_N. \tag{10}$$

Therefore, x can also be expressed for $M\bar{N}\bar{T} \leq 1$ as

$$x = \frac{Mp_N{}^2\bar{T}}{(1-p_N)(1-Mp_N\bar{T})}. \tag{11}$$

Substituting this back into Eq. (8), we can see how extremely sensitive that probability is to changes in p_N. For example, p_N need be only very slightly under the blowup value of $1/M\bar{T}$ to make long queues almost impossible.

Method 2

We now turn to another method of solving the same problem. It is applicable only to those cases in which the input process is such that the probability of having a time interval i between the arrival of any two consecutive units at S is known and is the same. We denote it by $f(i)$. However, the independence of these time intervals is not required.

Stationary operation of the system is assumed again. Furthermore, we make the following independence assumptions. If a unit arrives at S at time i_0, then the variables $\zeta(i)$ for all $i \geq i_0$, $n(i_0)$, and the arrival time of the next unit at S are independent of each other. This assumption is reasonable either if, due to the total traffic from the M sections, the probability p_ζ that A is empty is very small, or if $\bar{T}$ is much less than a cycle time. For in the first case regardless whether at a time i A is free or busy, there is a good chance that, if it gets empty, it will pick up a new item from one of the queues before the time $i+1$, since these queues will be very likely non-empty, because of the high probability that A was busy at the time of its previous arrivals

at these queues. In the second case $\overline{T} \ll 1$ means that regardless whether A is free or busy at time i, it will be free very soon after i. Then the states of the queues encountered between that time and $i + 1$ will determine $\zeta(i + 1)$. These arguments justify the independence assumption for the $\zeta(i)$'s. Their independence for $i > i_0$ of $n(i_0)$ and the next arrival at S is based again on the admissible neglect of the influence of S on A, as compared to that of the other queues. The argument for the independence of $n(i_0)$ of the next arrival's time is exactly the same as the corresponding argument in the first method.

The stationary distribution of the queue lengths in S at the times of arrivals can be computed on the basis of these assumptions as follows:

Let R_{mn} denote the conditional probability of having n units in S when one arrives, if there were m there at the arrival of the previous unit. Let $k = m - n + 1$. It is easy to see that if $n > 0$, R_{mn} depends on $m - n$ only. Thus, writing $R_{mn} = R_{k-1}$, we get for $n > 0$

$$R_{k-1} = \sum_{i=k}^{\infty} \binom{i}{k} p_\zeta^{\,k} q_\zeta^{\,i-k} f(i) \quad \text{for} \quad k = 0, 1, 2, \cdots . \tag{12}$$

This is so because the general term in the sum is the probability of A being empty at exactly k occasions out of i independent trials, multiplied by the probability of having i trials. Since at each trial only one unit may be removed from S, the probability of removing k units is given by the sum of these terms with i running from k to infinity.

Let P_n denote the stationary probability of having n units in S when a new one arrives. This situation can arise from and only from the states of S with $n + k - 1$ units at the time of the previous arrival, with $k = 0, 1, \cdots$. With the transition probabilities computed above, we have thus

$$P_n = \sum_{k=0}^{\infty} R_{k-1} P_{k+n-1} \qquad n = 1, 2, \cdots . \tag{13}$$

Substituting from (12) into (13) we obtain

$$P_n = \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} \binom{i}{k} p_\zeta^{\,k} q_\zeta^{\,i-k} f(i) P_{k+n-1} . \tag{14}$$

We may seek the solution of Eq. (14) in the form $P_k = cx^k$. Substituting this into (14), and interchanging the order of the summations we get

$$x = \sum_{i=0}^{\infty} \sum_{k=0}^{i} \binom{i}{k} p_\zeta^{\,k} q_\zeta^{\,i-k} x^k f(i) . \tag{15}$$

Summing over k, we have

$$x = \sum_{i=0}^{\infty} (q_\zeta + p_\zeta x)^i f(i) . \tag{16}$$

If there exists a positive solution of this equation which is less than 1, then that x generates the P_k's. It is not difficult to see that such a solution exists whenever $p_\zeta \sum_{i=0}^{\infty} f(i) > 1$. From the normalization condition, it follows that $c = 1 - x$. Thus

$$P_k = (1 - x)x^k , \tag{17}$$

but x is now given by (16) instead of (6).

As an example, consider the particular input distribution

$$f(i) = \begin{cases} 0 & \text{if} \quad i < j \\ pq^{(i-j)} & \text{if} \quad i \geq j , \end{cases} \tag{18}$$

where $p = 1 - q$ is the probability of a unit's arrival at S if the previous one has arrived at least j cycles earlier. Then (16) becomes

$$x = \frac{p(q_\zeta + p_\zeta x)^j}{1 - (q_\zeta + p_\zeta x)q} . \tag{19}$$

In order to compare the two methods, we observe that the probability p_N, for the distribution (18) of the arrival of a unit during a cycle, is given by

$$p_N = \frac{1}{j + q/p} . \tag{20}$$

We give now a numerical example to illustrate the relation of the two methods and the remark about the sensitivity of the solution to changes in p_N, which also holds in the case of the second method.

Let $M = 32$, $\overline{T} = 3.125$. Then if $p_N = 0.01$, we get from (9) and (10) that $q_\zeta = 1$, and according to Eqs. (6) and (8), the queue blows up. Taking $p_N = 0.0097$, we get from (9) and (10) (which are valid for both methods), that $q_\zeta = 0.97$. Then the first method gives from Eq. (6) $x = 0.31$. The second method, for the distribution (18) with $j = 65$ and p expressed from (20), gives $x = 0.1$ as the solution of (19). Then, for example, the probability of finding queues longer than five units in S is from Method 1 by Eq. (8) equal to 8.8×10^{-4}, and is equal to 10^{-6} from Method 2.

Acknowledgment

I wish to express my thanks to Dr. J. M. Berger for his careful reading of the manuscript and for his many valuable suggestions. I am also indebted to Professor Mark Kac for appraising the manuscript and to Dr. M. A. Leibowitz for many illuminating discussions.

References and footnotes

1. Communicated to us by Dr. W. G. Spruth.
2. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1960.

Received February 8, 1961