BULL. AUSTRAL. MATH. SOC. 60K25
VOL. 10 (1974), 219-244. (60K30)

Queuing and the provision of service facilities

Peter D. Finch

Results are obtained which enable one to assess the extent to which the waiting times of queuing customers depend on the number of servers which provide them with service.

1. Introduction

In many practical queueing situations one is interested in the extent to which the waiting times of customers depend on the number of servers used to provide them with service. Except for some very special cases, queuing theory does not answer such questions. Our purpose here is to pose these questions in a way which gives insight into practical problems rather than the mathematical problems associated with the many server queue of queuing theory.

Our method is based on the use of certain "standard" queuing situations as yardsticks against which one can compare actual queuing situations. Of course the usefulness of such a method involves the propriety of the yardstick in question and so it involves judgement of the extent to which propriety is achieved. Such judgements can seldom be made out of context and are relative to the practical situation under study. The standard queuing systems introduced below are judged to be appropriate in some practical situations but we do not claim they have universal propriety.

Received 6 November 1973.

2. Standard queuing systems

Let $C_1, C_2, \ldots, C_n$ be a finite number of customers each of whom is to be provided with service. For each $m = 1, 2, \ldots, n$ let C_m have service time s_m independently of which one of several available servers provides him with service. For the time being we make no assumptions about the stochastic nature of service times, we simply regard $(s_1, s_2, \ldots, s_n)$ as some given sequence of positive real numbers.

Our immediate purpose is the introduction of a standard queuing system for providing service to the customers $C_1, C_2, \ldots, C_n$ by means of k servers. The system in question is essentially one in which the servers do not become idle but to permit ease of comparison for different values of k it is convenient to arrange things so that corresponding initial situations are similar. For that purpose we introduce a sequence of fictitious customers $C_1', C_2', C_3', \ldots$ with respective service times $s_1', s_2', s_3', \ldots$. We impose no stochastic structure on these fictitious service times, for the time being we simply regard them as another given sequence of positive real numbers.

For each $k = 1, 2, \ldots$ define $Q_{1,1}^{(k)} \le Q_{1,2}^{(k)} \le \ldots \le Q_{1,k}^{(k)}$ to be the numbers $s_1', s_2', \ldots, s_k'$ arranged in non-decreasing order and for each $m = 2, 3, \ldots, n$ define $Q_{m,1}^{(k)} \le Q_{m,2}^{(k)} \le \ldots \le Q_{m,k}^{(k)}$ to be the numbers

$$Q_{m-1,1}^{(k)}+s_{m-1}, Q_{m-1,2}^{(k)}, \ldots, Q_{m-1,k}^{(k)},$$

arranged in non-decreasing order. The quantities $Q_{m,j}^{(k)}$ can be interpreted in the following way. Suppose that all the customers $C_1', C_2', C_k', C_1, C_2, \ldots, C_n$ are present at time 0 and are to be given service by k servers. At time 0 the customers $C_1', C_2', \ldots, C_k'$ start service simultaneously, one to each of the k servers, and thereafter customers are served in the order $C_1, C_2, \ldots, C_n$ none of the servers remaining idle whilst one of these customers has yet to start service.

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

Then $Q_{m,1}^{(k)}$ is the time of the mth departure from the system, in other words it is the time at which C_m starts service.

Suppose now, in contrast to the situation just considered, that only the customers $C_1', C_2', \ldots, C_k'$ are present at time 0 and that they then start service simultaneously, one to each of the k servers; thereafter the customers $C_1, C_2, \ldots, C_n$ arrive, in that order, at the times $A_1^{(k)} < A_2^{(k)} < \ldots < A_n^{(k)}$. It is supposed that the customers $C_1, C_2, \ldots, C_n$ are to be served in the order of their arrival and that no server is idle whilst a customer who has already arrived is waiting to start service. We say that the sequence of arrival times $\left(A_1^{(k)}, A_2^{(k)}, \ldots, A_n^{(k)}\right)$ generates a standard k-server queue relative to the sequence of service times $(s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n)$ when

$$(1) \qquad \forall\, m = 1, 2, \ldots, n : A_m^{(k)} \leq Q_{m,1}^{(k)} .$$

In a standard k-server queue no server is idle until $n + 1$ customers have received service and the waiting time of customer C_m is

$$(2) \qquad w_m^{(k)} = Q_{m,1}^{(k)} - Q_m^{(k)} , \quad 1 \leq m \leq n .$$

For each $m = 1, 2, \ldots, n$ one has

$$Q_{m,j}^{(k)} \leq A_{m,j}^{(k-1)} , \quad j = 1, 2, \ldots, k-1 , \quad k > 1 .$$

It follows that if $\left(A_1^{(k)}, A_2^{(k)}, \ldots, A_n^{(k)}\right)$ is a sequence of arrival times for $C_1, C_2, \ldots, C_n$ which generates a standard k-server queue relative to the sequence $(s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n)$ then it also generates a standard l-server queue relative to the sequence $(s_1', s_2', \ldots, s_l', s_1, s_2, \ldots, s_n)$ for each $l = 1, 2, \ldots, k-1$. Moreover in such an l-server queue the waiting time of customer C_m would be

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$w_m^{(l,k)} = Q_{m,1}^{(l)} - A_m^{(k)} , \quad 1 \le m \le n , \quad 1 \le l \le k , \tag{3}$$

so that

$$w_m^{(l,k)} - w_m^{(k)} = Q_{m,1}^{(l)} - Q_{m,1}^{(k)} , \quad 1 \le m \le n , \quad 1 \le l \le k , \tag{4}$$

would not depend on the arrival times $A_1^{(k)}, A_2^{(k)}, \ldots, A_n^{(k)}$. It is to be emphasised that in (2), (3) and (4) the waiting times refer to the same customers with the same respective service times and the same respective arrival times but for two different service arrangements, namely, k and l servers respectively with similar though slightly different initial conditions.

3. Practical motivation

The practical usefulness of (4) comes about in the following way. We often wish to reduce customer waiting times in an existing queuing situation where the servers are not idle over some period under consideration. In such cases one is dealing with a standard l-server queue and one wants to determine $k > l$ so that the use of k servers will reduce waiting times to some specified level. One cannot apply (4) directly to determine the improvement in waiting times obtained by using k rather than l servers because arrival times which generate a standard l-server queue relative to some sequence of service times may not generate a standard k-server queue relative to those service times when $k > l$. However one can apply (4) indirectly in the following way.

Suppose that $W_1, W_2, \ldots, W_n$ is a sequence of desired upper bounds for the waiting times of the customers $C_1, C_2, \ldots, C_n$ and suppose further that for some $k > l$ one did have a standard k-server queue for which

$$w_m^{(k)} \le W_m , \quad m = 1, 2, \ldots, n . \tag{5}$$

Then from (4) one would have

$$w_m^{(l,k)} \le \left[Q_{m,1}^{(l)} - Q_{m,1}^{(k)}\right] + W_m , \quad 1 \le m \le n , \tag{6}$$

whatever the actual arrival times in the k-server queue provided only that

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

they did generate a standard queue in respect of the service times in question. If, however, it should happen that observations on the existing l-server queue indicated that for some m, $1 \leq m \leq n$, the inequality in (6) did not hold then it would seem that the k in question is not large enough to achieve the desired reduction in waiting times. In other words we want to find the smallest $k > l$ such that (6) holds when the $w_m^{(l,k)}$ are replaced by actual observed waiting times.

The preceding argument is admittedly heuristic but it corresponds to the following practical situation. Suppose that one is confronted by an l-server system in which waiting times are judged to be too long. As a result of a detailed analysis one recommends that the use of $k > l$ servers will meet prescribed standards bounding the waiting times of customers. However to justify this recommendation one is called upon to "explain" why fewer than k servers will not meet the standards in question. Such an explanation could take the following heuristic form. Consider the operation of the k-server queue over a period of time in which the waiting times of customers are small and do meet the prescribed standards but the servers do not become idle. Consider a particular history of this k-server process and suppose that one had in fact used k' servers; where $l \leq k' < k$, during the period in question. Then the use of k servers rather than k' would be "justified" if one could show that for such histories one would not meet the prescribed standard of service for any $k' < k$. It is this sort of justification which is attempted by the argument based on (4).

In practice, however, the situation is not quite as simple as that just envisaged. In the first place one is usually dealing with a relatively large number of customers and one is interested in the waiting times of the later arriving customers; in other words n is large and one is interested in the asymptotic value of $w_m^{(k)}$ when $m \leq n$ is also large. In the second place one is usually studying a queuing situation with its future operation in mind and the arrival times and service times under consideration will be hypothetical ones which are thought to typify the sort of variability to be expected in the situation at hand. One has then to consider not just one history but a set of histories which are taken to

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

typify the joint variability in arrival times and service times to be expected in the situation under study. In other words one has to formulate a stochastic model of the variability in question. However when one is dealing with standard queues one cannot characterise the variability in arrival times independently of the variability in service times because one has to preserve the inequalities in (1).

In the next section we formulate a general stochastic model for standard queues and show how a particular case of that model can be used to investigate the problem of reducing waiting time by providing additional servers.

4. Stochastic assumptions

At the most general level our stochastic model for the standard k-server queue consists in the assumption that the $2n + k$ quantities $A_1^{(k)}, A_2^{(k)}, \ldots, A_n^{(k)}, s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n$ have a joint distribution such that with probability one the inequalities in (1) hold simultaneously for each $m = 1, 2, \ldots, n$. This model will be called the general standard k-server queue and denoted by $GSQ^{(k)}$. However this model is so general that without further specification of the joint distribution in question it does not lead to informative results.

A particular case of the general model which does lead to informative results is the one in which the *marginal* distribution of $s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n$ is such that these $n + k$ quantities are independently and identically distributed with common distribution $F(x)$ having a finite mean μ and a finite variance σ^2. We call this case the general standard k-server queue with common independent service times and we denote it by $GSQ^{(k)}CI$. The interest of this model comes from the fact that there are many practical situations where it does not seem too unreasonable to suppose that service times are independently and identically distributed. In such cases $GSQ^{(k)}CI$ provides a suitable framework for discussing various arrival patterns which ensure that the servers do not become idle.

In $GSQ^{(k)}CI$ the quantity $Q_{m,1}^{(k)}$ is just the mth renewal point in the superposition of the k independent renewal process generated by the k servers. It is well-known from renewal theory that, for large m,

$$E\left(Q_{m,1}^{(k)}\right) \simeq \frac{m\mu}{k} + \frac{(k-1)(\mu^2-\sigma^2)}{2k\mu} . \tag{7}$$

5. Reducing mean waiting times

Suppose now that $k \geq 2$ and consider a particular realisation of $GSQ^{(k)}IC$, say

$$R^{(k)} = (A_1, A_2, \ldots, A_n, s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n) ,$$

with associated waiting times $w_m^{(k)}(R^{(k)})$, $1 \leq m \leq n$. If $1 \leq l < k$, then

$$R^{(l,k)} = (A_1, A_2, \ldots, A_n, s_1', s_2', \ldots, s_l', s_1, s_2, \ldots, s_n)$$

is a corresponding realisation for a $GSQ^{(l)}IC$ with waiting times $w_m^{(l,k)}(R^{(l,k)})$, $1 \leq m \leq n$. By (4),

$$w_m^{(l,k)}(R^{(l,k)}) - w_m^{(k)}(R^{(k)}) = Q_{m,1}^{(l)} - Q_{m,1}^{(k)} , \tag{8}$$

where the $Q_{m,1}^{(k)}$ are determined by $(s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n)$ in the way described in Section 2 and the $Q_{m,1}^{(l)}$ are determined, in like manner, by $(s_1', s_2', \ldots, s_l', s_1, s_2, \ldots, s_n)$. If we take expectations over all realisations $R^{(k)}$ of the $GSQ^{(k)}CI$ in question then equation (8) gives

$$E^{(k)}\left(w_m^{(l,k)}\right) - E^{(k)}\left(w_m^{(k)}\right) = E\left(Q_{m,1}^{(l)}\right) - E\left(Q_{m,1}^{(k)}\right) . \tag{9}$$

On the left-hand side of (9), $E^{(k)}$ denotes expectation over the realisations $R^{(k)}$ whereas on the right-hand side of (9), E denotes corresponding marginal expectations over service time distributions.

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

Making use of (7) we obtain

$$E^{(k)}\left[w_m^{(l,k)}\right] - E^{(k)}\left[w_m^{(k)}\right] \simeq \left\{\frac{1}{l} - \frac{1}{k}\right\}\left\{m\mu - \frac{\mu^2-\sigma^2}{2\mu}\right\} . \tag{10}$$

From expression (10) one can derive a crude rule of thumb to determine the number of servers required to achieve a desired reduction in mean customer waiting time. Thus suppose that the operation of an l-server queue has led to a situation where servers are not idle and the later arriving customers suffer excessively long delays. To shorten these delays one wants to replace the current system by one with more than l servers. How large should $k > l$ be to ensure that asymptotically $E\left[w_m^{(k)}\right] \leq W$?

For example suppose that a queuing system with 4 servers is causing excessive delays. It is observed that during periods immediately after starting up, when about 100 customers are served, none of the servers is idle and the later arriving customers are waiting on average from 8 to 10 mean service times. It is desired to reduce the waiting time of such customers to achieve an expected value of no more than 2 mean service times. Consider first a $GSQ^{(5)}IC$ system in which $E^{(5)}\left[w_{100}^{(5)}\right] \leq 2\mu$ and suppose that σ^2/μ is small compared to 100μ . From (10) we have

$$E^{(5)}\left[w_{100}^{(4,5)}\right] \simeq E^{(5)}\left[w_{100}^{(5)}\right] + \left[\frac{1}{4} - \frac{1}{5}\right]100\mu \leq 7\mu .$$

Thus if the arrival pattern in a standard 5-server queue was such that the mean waiting time of the 100th customer was no more than 2 mean service times we would expect that 4 servers used with the same arrival pattern would result in a mean waiting time for the 100th customer of no more than 7 mean service times. But in the observed 4 server queue the mean waiting time in question does exceed 7 mean service times. This suggests that the use of 5 servers in the practical situation under study would not achieve the desired asymptotic mean waiting time of no more than 2 mean service times. On the other hand consider a $GSQ^{(6)}IC$ system in which $E^{(6)}\left[w_{100}^{(6)}\right] \leq 2\mu$ and suppose again that σ^2/μ is small compared to 100μ . From (10),

$$E^{(6)}\left[w_{100}^{(4,6)}\right] \simeq E^{(6)}\left[w_{100}^{(6)}\right] + \left[\frac{1}{4} - \frac{1}{6}\right]100\mu \leq \left[10\ \frac{1}{3}\right]\mu$$

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

and since the observed mean waiting time does not exceed 10μ we conclude that the use of 6 servers in the practical situation under study will achieve the desired reduction in mean waiting time.

The general form of this argument is easily presented. Consider a $GSQ^{(k)}IC$ system in which asymptotically $E^{(k)}\left[w_m^{(k)}\right] \leq W$. If $l < k$ and

$$\frac{1}{k} > \frac{1}{l} - (W'-W)\left[m\mu - \frac{\mu^2-\sigma^2}{2\mu}\right]^{-1} \tag{11}$$

then (10) shows that asymptotically

$$E^{(k)}\left[w_m^{(l,k)}\right] < W' .$$

Thus (11) must be false if $W' \leq E^{(k)}\left[w_m^{(l,k)}\right]$, in other words k must satisfy

$$\frac{1}{k} \leq \frac{1}{l} - \left\{E^{(k)}\left[w_m^{(l,k)}\right]-W\right\}\left[m\mu - \frac{\mu^2-\sigma^2}{2\mu}\right]^{-1} . \tag{12}$$

In practice, of course, $E^{(k)}\left[w_m^{(l,k)}\right]$ in (12) might well be replaced by an estimate derived from observations of the current l-server situation. Moreover since m is assumed to be large there is often little to be lost by replacing (12) by

$$\frac{1}{k} \leq \frac{1}{l} - \frac{(W^*-W)}{m\mu} \tag{13}$$

where W^* is the estimate of $E^{(k)}\left[w_m^{(l,k)}\right]$.

Suppose then that in a $GSQ^{(l)}IC$ system customer C_m has mean waiting time W^*. The rule of thumb for finding the number of servers required to ensure that customer C_m has an expected waiting time of no more than $W < W^*$ is to take $k > l$ to be the smallest integer satisfying (13). It should be noted that this rule of thumb depends on the nature of the arrival process only to the extent that we are dealing with standard queues.

There are two aspects of the preceding discussion which need further investigation. In the first place we have used (7) to derive the

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

asymptotic result (10) but we have given no indication of how large m must be before the asymptotic relation (7) is applicable in practice. We discuss this question in Section 9. Secondly there are practical queuing situations where the rule of thumb derived above cannot be applied simply because there is no current l-server queuing in operation, in other words W^* and l in (13) are not available. In such cases one needs to be able to calculate the $E\left(w_m^{(k)}\right)$ directly. We proceed now to discuss a class of standard queues for which such calculations can be carried out.

6. Simple standard queues

Consider the standard k-server queue instroducted in Section 2. Write $A_0^{(k)} = 0$ and $Q_m^{(k)}$ for $Q_{m,1}^{(k)}$. Since

$$A_{m-1}^{(k)} < A_m^{(k)} \leq Q_m^{(k)} , \quad 1 \leq m \leq n , \tag{14}$$

there exists $\alpha_m^{(k)}$, $0 \leq \alpha_m^{(k)} < 1$, $1 \leq m \leq n$, such that

$$A_m^{(k)} = \alpha_m^{(k)} A_{m-1}^{(k)} + \left(1-\alpha_m^{(k)}\right) Q_m^{(k)} , \quad 1 \leq m \leq n , \tag{15}$$

namely

$$\alpha_m^{(k)} = \left(Q_m^{(k)} - A_m^{(k)}\right) \Big/ \left(Q_m^{(k)} - A_{m-1}^{(k)}\right) , \quad 1 \leq m \leq n . \tag{16}$$

Write

$$I_m^{(k)} = Q_m^{(k)} - Q_{m-1}^{(k)} , \quad 1 \leq m \leq n ,$$

and $Q_0^{(k)} = 0$. From (2) and (15) we obtain

$$w_m^{(k)} = \alpha_m^{(k)} \left(I_m^{(k)} + w_{m-1}^{(k)}\right) , \quad 1 \leq m \leq n , \tag{17}$$

where $w_0^{(k)} = 0$ and $I_1^{(k)} = Q_1^{(k)}$. It follows that

$$w_m^{(k)} = \sum_{r=1}^{m} \alpha_m^{(k)} \alpha_{m-1}^{(k)} \cdots \alpha_r^{(k)} I_r^{(k)} , \quad 1 \leq m \leq n . \tag{18}$$

The numbers $\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_n^{(k)}$ play the role of n parameters

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

which characterise the relationship between arrival times and service times in a standard k-server queue. If one had such a sequence of numbers one could define $A_1^{(k)}, A_2^{(k)}, \ldots, A_n^{(k)}$ recurrently from equations (15). The conditions $0 \le \alpha_m^{(k)} < 1$, $m = 1, 2, \ldots, n$ would ensure that (1) held and so one would obtain a sequence of arrival times which generated a standard k-server queue relative to the service times in question.

Thus the stochastic model $GSQ^{(k)}$ could be specified by starting with the quantities $\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_n^{(k)}, s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n$ and supposing that these have a joint distribution such that $0 \le \alpha_m^{(k)} < 1$, $1 \le m \le n$, with probability one. When the $\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_n^{(k)}$ are jointly distributed independently of the $n + k$ quantities $s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n$ we refer to the stochastic model as the simple standard k-server queue and we denote it by $SSQ^{(k)}$.

Some general results can be written down at once. Thus in $SSQ^{(k)}$ equation (18) yields

$$E\left(w_m^{(k)}\right) = \sum_{r=1}^{m} E\left(\alpha_m^{(k)}\alpha_{m-1}^{(k)} \ldots \alpha_r^{(k)}\right)E\left(I_r^{(k)}\right), \quad 1 \le m \le n, \tag{19}$$

for the expected waiting time of customer C_m. In particular, writing $s_1' = s_0$, we obtain

$$E\left(w_m^{(1)}\right) = \sum_{r=1}^{m} E\left(\alpha_m^{(1)}\alpha_{m-1}^{(1)} \ldots \alpha_r^{(1)}\right)E\left(s_{r-1}\right), \quad 1 \le m \le n. \tag{20}$$

In like manner one can write down an expression for the variance of $w_m^{(k)}$ in terms of the variances and covariances of the quantities $\alpha_m^{(k)}\alpha_{m-1}^{(k)} \ldots \alpha_r^{(k)}$ and those of the quantities $I_m^{(k)}$.

7. Asymptotic results in $SSQ^{(k)}{}_{CI}$

From (7) the independence and common distribution of service times

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

gives

$$E\left(I_m^{(k)}\right) \simeq \frac{\mu}{k} . \tag{21}$$

Thus, provided

$$\sum_{r=1}^{m} E\left(\alpha_m^{(k)}\alpha_{m-1}^{(k)} \ldots \alpha_r^{(k)}\right)$$

converges as $m \to \infty$, we obtain the following asymptotic expression for mean waiting time in $SSQ^{(k)}CI$:

$$E\left(w_m^{(k)}\right) \simeq \frac{\mu}{k} \sum_{r=1}^{m} E\left(\alpha_m^{(k)}\alpha_{m-1}^{(k)} \ldots \alpha_r^{(k)}\right) . \tag{22}$$

Since the expectations on the right of (22) do not exceed unity $E\left(w_m^{(k)}\right)$ is aymptotically bounded above by $m\mu/k$. When $k = 1$ the expressions (7), (21) and (22) are exact, in particular

$$E\left(w_m^{(1)}\right) \leq m\mu . \tag{23}$$

An interesting special case of $SSQ^{(k)}CI$ occurs when $\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_n^{(k)}$ are mutually independent. It is then possible to achieve any permitted sequence of asymptotic mean waiting times by an appropriate choice of $E\left(\alpha_1^{(k)}\right), E\left(\alpha_2^{(k)}\right), \ldots, E\left(\alpha_n^{(k)}\right)$. For suppose that F is a function defined on the non-negative integers with $F(0) = 0$ and

$$0 < F(m) < 1 + F(m-1) , \quad m \geq 1 . \tag{24}$$

Suppose that we wish to consider the model $SSQ^{(k)}CI$ when

$$\frac{k}{\mu} E\left(w_m^{(k)}\right) \simeq F(m) . \tag{25}$$

Write $\lambda_m = E\left(\alpha_m^{(k)}\right)$, $m \geq 1$. Then (22) gives

$$\frac{k}{\mu} E\left(w_m^{(k)}\right) \simeq \sum_{r=1}^{m} \lambda_m\lambda_{m-1} \ldots \lambda_r . \tag{26}$$

Thus we could achieve (25) by choosing $\lambda_1, \lambda_2, \ldots, \lambda_n$ successively so

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

that

$$\sum_{r=1}^{m} \lambda_m \lambda_{m-1} \cdots \lambda_r = F(m) \ .$$

In other words we have (25) when

$$\lambda_m = F(m)[1+F(m-1)]^{-1} \ , \quad 1 \le m \le n \ . \tag{27}$$

The inequalities (24) ensure that $0 < \lambda_m < 1$. It should be noted that a function F which does not satisfy the inequalities (24), at least asymptotically, cannot lead to a permissible sequence of asymptotic mean waiting times. To see this observe that (17) gives

$$E\left[w_m^{(k)}\right] \simeq \lambda_m \left[\frac{\mu}{k} + E\left[w_{m-1}^{(k)}\right]\right] \ .$$

Thus if (25) holds and $0 < \lambda_m < 1$, we must have

$$F(m) \simeq \lambda_m[1+F(m-1)] < 1 + F(m-1) \ .$$

It is convenient to have a symbol for the simple standard k-server queue with common independently distributed service times. When the $\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_n^{(k)}$ are mutually independent and have a common distribution with mean λ and variance τ^2 we denote it by $SCISQ^{(k)}{}_{CI}$. For this queuing system one obtains especially simple formulae. Thus (22) gives

$$E\left[w_m^{(k)}\right] \simeq \frac{\mu}{k}\left[\frac{\lambda-\lambda^{m+1}}{1-\lambda}\right] \ , \tag{28}$$

this formula being exact when $k = 1$. When λ^m is small there is little to be lost by replacing (28) by

$$E\left[w_m^{(k)}\right] \simeq \frac{\mu}{k}\left[\frac{\lambda}{1-\lambda}\right] \ . \tag{29}$$

It follows from (28) that for $SCISQ^{(k)}{}_{CI}$ the asymptotic mean waiting times are approximately constant, being small when λ is small and large when λ is close to 1 . This is plausible because λ close to 0 means that the arrival time $A_m^{(k)}$ is expected to be close to $Q_m^{(k)}$ whereas λ

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

near to unity means that the arrival time $A_m^{(k)}$ is expected to be near to $A_{m-1}^{(k)}$. On the other hand if m, though large, is fixed and we let $\lambda \to 1$ we have

$$\lim_{\lambda \to 1} E\left[w_m^{(1)}\right] = m\mu$$

and

$$\lim_{\lambda \to 1} E\left[w_m^{(k)}\right] \simeq \frac{m\mu}{k}, \quad k > 1 .$$

Of course, as noted above, $E\left[w_m^{(k)}\right]$ is asymptotically bounded above by $m\mu/k$.

For $SCISQ^{(1)}CI$ it is possible to derive a simple expression for the asymptotic variance of waiting times. Thus using (17) one finds that if λ^m is small then

$$\text{(30)} \qquad \operatorname{var}\left[w_m^{(1)}\right] \simeq \left\{\frac{\lambda^2+\tau^2}{1-\lambda^2-\tau^2}\right\}\left\{\frac{\mu^2(1+\lambda)}{1-\lambda} + \sigma^2\right\} - \mu^2 .$$

For example, if $\alpha_1^{(1)}, \alpha_2^{(1)}, \ldots, \alpha_n^{(1)}$ are uniformly and independently distributed on $(0, 1)$, then asymptotically $w_m^{(1)}$ has mean μ and variance $(\mu^2+\sigma^2)/2$.

The variance of waiting times in the case of more than one server can be very crudely estimated by observing that the process resulting from the superposition of a large number of independent renewal processes is approximately a Poisson process. Thus if k is large one could treat $I_1^{(k)}, I_2^{(k)}, \ldots, I_n^{(k)}$ as independent exponentially distributed random variables with common mean μ/k. Then (30) would apply with both μ and σ replaced by μ/k; we have then in $SCISQ^{(k)}CI$ when k is large

$$\text{(31)} \qquad \operatorname{var}\left[w_m^{(k)}\right] \simeq \frac{\mu^2}{k^2}\left[\frac{2}{1-\lambda} \cdot \left\{\frac{\lambda^2+\tau^2}{1-\lambda^2-\tau^2}\right\} - 1\right] .$$

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

8. The number of service facilities

We now show how the results of the last section can be used in a rough estimation of the number of servers required to achieve asymptotic mean times which do not exceed some preassigned upper bound. The problem in question is related to that of reducing mean waiting time as considered in Section 5, the difference being that we do not now have an existing multi-server queue to provide W^* and l in (13).

Suppose it is desired to set up a multi-server queuing system which will be able to deal with about n customers without the servers becoming idle and yet ensure that the asymptotic mean waiting times of the later arriving customers do not exceed $W = \Delta\mu$, where μ is the mean service time and it is assumed that the service times are independently and identically distributed. How many servers should we use?

Within the context of classical queuing theory this question is posed very imprecisely. There are no specific assumptions about the nature of the arrival process of the customers in question. Indeed even within the context of the standard queues of the last section it would seem that the arrival process should be taken into account because, for a fixed k, there are $SCISQ^{(k)}CI$ systems in which the asymptotic mean waiting times behave in a preassigned but arbitrary way.

Nonetheless the question as posed can be answered in a meaningful way by the use of standard queues as appropriate yardsticks. Thus consider a $SQ^{(k+1)}CI$ system in which the asymptotic mean waiting times are so small that they are negligible. Such systems exist, for instance by (28) an $SCISQ^{(k+1)}CI$ system with very small λ is such a queuing system. If the arrival process to the $SQ^{(k+1)}CI$ system were used with only k servers the asymptotic mean waiting time $E^{(k+1)}\left[w_m^{(k,k+1)}\right]$ could be obtained from the expression (10), namely

$$E^{(k+1)}\left[w_m^{(k,k+1)}\right] \simeq \frac{1}{k(k+1)}\left[m\mu - \frac{\mu^2-\sigma^2}{2\mu}\right] . \tag{32}$$

Moreover if $(\mu^2-\sigma^2)/2\mu$ is small compared to $m\mu$ (and for simplicity in exposition we will suppose that this is so) we have

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$E^{(k+1)}\left(w_m^{(k,k+1)}\right) \simeq \frac{m\mu}{k(k+1)} \quad . \tag{33}$$

Now for some values of k the expected waiting time in (32) would be bounded by $W = \Delta\mu$ whereas for other values of k it would not. There is, in fact, a smallest k for which the right-hand side of the expression (33) does not exceed $W = \Delta\mu$, namely the smallest positive integer k such that

$$k(k+1) \geq m/\Delta \quad . \tag{34}$$

We obtain therefore the following rule of thumb: to ensure that the asymptotic mean waiting time of customer C_m does not exceed $\Delta\mu$ use k servers where k is the smallest positive integer satisfying (34) or, more roughly,

$$k \geq [m/\Delta]^{\frac{1}{2}} \quad . \tag{35}$$

For example suppose we are dealing with 100 customers and wish the expected waiting time of C_{100} to be no more than 2 mean service times. Then (34) with $m = 100$ and $\Delta = 2$ gives $k = 7$, a result close to that of Section 5 where it was found that six servers would reduce the asymptotic mean waiting time to a similar level.

The logic behind this rule of thumb seems foreign to the type of argument used in applications of queuing theory because there is no explicit reference to the process whereby customers arrive. It might be argued, for instance, that the rule is misleading because there are *single* server queues in which the limiting distribution of waiting time has an expectation which does not exceed 2 mean service times whereas the rule in question would prescribe the use of 7 servers, surely a gross overestimate of the required number of servers. On the other hand it follows from (29), with $\lambda > 14/15$ and $k = 7$, that there are $SCISQ^{(7)}CI$ systems in which the asymptotic mean waiting time does exceed 2 mean service times. It would seem that in such situations the suggested rule would underestimate the required number of servers.

The resolution of these objections lies in the recognition of the sort of practical problem to which the rule is applicable. In other words the propriety of considering a standard $(k+1)$-server queue with negligible

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

waiting times comes, if at all, from the practical context of the practical situation under study and not from the mathematics *per se*. This is because many, though not all, practical queuing situations operate with customers leaving before service if their anticipated waiting times are too long. The effect of this is to produce an arrival process in the desired situation in which the addition of another server would substantially reduce customer waiting times. However in the derivation of a rule of thumb it is not necessary to suppose that the standard $(k+1)$-server queue does have negligible waiting times. Indeed if we assume that asymptotic mean waiting times in the $(k+1)$-server queue are bounded above by $\delta\mu$, with $\delta < \Delta$, then (35) is replaced by

$$k \geq [m/(\Delta-\delta)]^{\frac{1}{2}} . \tag{36}$$

In some practical queuing problems where one wants to provide service to customers one has not only some idea of the number of customers in question but also some idea of the length of the period during which they must be provided with that service. Suppose, for instance, that we require customer C_n to start service at an expected time no later than T. Then we have the additional constraint

$$E\left[Q_n^{(k)}\right] \leq T .$$

Thus from (7) we have the approximate constraint

$$k \geq \frac{1}{T}\left[n\mu + \frac{(\mu^2-\sigma^2)}{2\mu}\right] ,$$

and if we continue to assume that $(\mu^2-\sigma^2)/2\mu$ is small compared to $n\mu$ this inequality can be replaced by

$$k \geq n\mu/T . \tag{37}$$

For instance if we not only wished to achieve a mean waiting time for C_{100} of no more than 2 mean service times but also to ensure that C_{100} started service at an expected time no later then $T = 10\mu$ we would require $k = 10$ rather than $k = 7$ as given by (34). On the other hand if $T = 20\mu$ then $k = 5$ would meet the constraint on the expected starting time of C_{100} but the use of only 5 servers is suspect because

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

they could not meet the constraint on expected waiting time for an arrival process which generated a standard queue with 6 servers and negligible waiting times. In general then the rule of thumb can be modified to meet a constraint on the expected starting time of C_n by replacing (35) by

$$k \geq \max\left[\left(\frac{n}{\Delta}\right)^{\frac{1}{2}}, \frac{n}{N}\right] , \tag{38}$$

where $T = N\mu$ is the prescribed expected starting time of customer C_n .

The use of a constraint on the expected starting time of customer C_n is a rough and indirect way of taking the arrival process into account, it being implicitly assumed that the n customers arrive during a period whose expected length does not exceed $(N-\Delta)\mu$.

9. The asymptotic approximation

The preceding results have been based on the asymptotic formula (7) and so their usefulness depends on how large an m is required to make (7) an approximation which is useful in practice. We examine this question in detail for the case of an Erlang service time distribution; however it is convenient to start with some general results which hold for a general service time distribution.

Let $Q_{m,j}^{(k)}$, $j = 1, 2, \ldots, k$, $m = 1, 2, \ldots, n$ be defined as in Section 2 and suppose that $s_1', s_2', \ldots, s_k', s_1, s_2, \ldots, s_n$ are independently and identically distributed non-negative random variables with common distribution function $F(x)$ having finite mean μ and finite variance σ^2 . Write

$$G(x) = \mu^{-1} \int_0^x \{1-F(u)\}du , \quad x \geq 0 . \tag{39}$$

We recall that $G(x)$ is the limiting distribution function of forward recurrence time in a renewal process with renewal distribution $F(x)$.

Since

$$\sum_{j=1}^{k} Q_{m,j}^{(k)} = \sum_{j=1}^{k} Q_{m-1,j}^{(k)} + s_{m-1} , \quad m > 1 ,$$

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

we have

$$\sum_{j=1}^{k} Q_{m,j}^{(k)} = S_m^{(k)} \quad , \quad m \geq 1 \ , \tag{40}$$

where

$$S_m^{(k)} = s_1' + s_2' + \ldots + s_k' + s_1 + s_2 + \ldots + s_{m-1} \quad , \quad m \geq 1 \ ,$$

and we interpret s_0 as 0 . Rearrangement of (40) gives

$$\frac{1}{k} S_m^{(k)} - Q_{m,1}^{(k)} = \frac{1}{k} \sum_{j=1}^{k} \left\{ Q_{m,j}^{(k)} - Q_{m,1}^{(k)} \right\} \ . \tag{41}$$

Consider the pooled process formed by the k independent renewal processes generated by each of the k servers. The quantities $Q_{m,j}^{(k)} - Q_{m,1}^{(k)}$, $j = 2, 3, \ldots, k$ are, in order of magnitude, just the forward recurrence times measured from the mth renewal point of the pooled process to the next renewals on the other $k - 1$ component renewal processes. Since the component renewal processes are independent one would conjecture that the unordered forward recurrence times would be asymptotically independent and that each would have the asymptotic distribution function $G(x)$. If this conjecture is true then

$$\sum_{j=2}^{k} \left[Q_{m,j}^{(k)} - Q_{m,1}^{(k)} \right]$$

is asymptotically the sum of $k - 1$ independent random variables each of which has the distribution G . Thus from (41) we would have

$$Pr\left\{ \frac{1}{k} S_m^{(k)} - Q_{m,1}^{(k)} \leq x \right\} \simeq G^{(k-1)}(kx) \quad , \quad x \geq 0 \ , \tag{42}$$

where $G^{(k-1)}$ is the $(k-1)$-fold convolution of G with itself. Moreover, since the mean of the distribution G is $(\mu^2+\sigma^2)/2\mu$, we obtain (7) by formally taking means on each side of the asymptotic expression (42).

Both the conjecture formulated above and the asymptotic result (42) are true but we will not establish them here. What we do is to examine the rate of convergence to the asymptotic distribution in (42) for the special

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

case of Erlang service times. In the process of doing this we do, in fact, establish the truth of the conjecture in the special case of Erlang service times. However before specializing to the Erlang distribution we obtain a few general results.

Write

$$P_0(t, x) = F(t+x) - F(x) , \quad t > 0 , \quad x > 0$$

and, for $m = 1, 2, \ldots, t > 0$ and $x > 0$,

$$P_m(t, x) = \int_0^t [F(t+x-u)-F(t-u)]F^{(m)}(du)$$

where $F^{(m)}$ is the m-fold convolution of F with itself. Then for any one of the component renewal processes under consideration $P_m(t, x)$ is the joint probability that exactly m of its renewal points occur in $(0, t)$ and that its $(m+1)$th renewal point occurs before $t + x$.

Consider again the pooled process from the k servers and for $t > 0$, $x_2 > 0, \ldots, x_k > 0$ denote the quantity

$$(43) \qquad k \sum_{m=0}^{n-1} F^{(n-m)}(dt) \sum_{m(2)+\ldots+m(k)=m} \prod_{j=2}^{k} P_{m(j)}(t, x_j)$$

by $Q_n(dt; x_2, \ldots, x_k)$. The quantity in (43) is just the joint probability that the nth renewal point of the pooled process occurs in $(t, t+dt)$ and that of the forward recurrence times to the next renewals on the other $k - 1$ component processes one does not exceed x_2 , another x_3 and so on. Thus

$$(44) \qquad Q_n(x_2, x_3, \ldots, x_k) = \int_0^\infty Q_n(dt; x_2, \ldots, x_k)$$

is just the joint distribution function of the unordered quantities whose arrangement, in order of magnitude, is $Q_{n,j}^{(k)} - Q_{n,1}^{(k)}$, $j = 2, 3, \ldots, k$. Note that $Q_n(x_2, x_3, \ldots, x_k)$ is a symmetric function of $x_2, x_3, \ldots, x_k$ and that if we formally put $x_2 = x_3 = \ldots = x_k = \infty$ we obtain

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$(45) \qquad Pr\left\{Q_{n,1}^{(k)} \le t\right\} = \int_0^t Q_n(dt, \infty, \infty, \ldots, \infty) \ .$$

Specialising now to Erlang service times we introduce the following notation: $r > 0$ is a positive integer, $\lambda > 0$ is a positive real number,

$$e_{r,\lambda}(x) = \frac{\lambda^r x^{r-1}}{(r-1)!} e^{-\lambda x} \ , \quad x \ge 0 \ ,$$

$$E_{r,\lambda}(x) = \int_0^x e_{r,\lambda}(u)du \ , \quad x \ge 0 \ .$$

We recall that

$$1 - E_{r,\lambda}(x) = e^{-\lambda x} \sum_{j=0}^{r-1} \frac{(\lambda x)^j}{j!} \ , \quad x \ge 0 \ .$$

For the remainder of this paper we assume that $F = E_{r,\lambda}$, that is service times have an Erlang distribution. When $F = E_{r,\lambda}$ one has $\mu = r/\lambda$, $\sigma^2 = r/\lambda^2$ and G in (39) is given by

$$(46) \qquad G_{r,\lambda}(x) = r^{-1} \sum_{j=1}^{r} E_{r,\lambda}(x) \ , \quad x \ge 0 \ ,$$

with corresponding mean $(r+1)/2\lambda$. One finds that

$$(47) \qquad P_m(t, x) = e^{-\lambda t} \sum_{j=0}^{r-1} \frac{(\lambda t)^{mr+j}}{(mr+j)!} E_{r-j,\lambda}(x) \ , \quad m \ge 0 \ .$$

Equation (47) follows at once from the formula defining $P_m(t, x)$ with $F = E_{r,\lambda}$ but it may also be derived by the following probability argument. The renewal interval is the sum of r independent phases with identical exponential distributions. The points at which phases terminate form a Poisson process with parameter λ. Exactly m renewal points occur in $(0, t)$ if and only if $mr + j$ phase points occur in that interval for some $j = 0, 1, \ldots, r-1$. If exactly $mr + j$ phase points do occur there remain $r - j$ phase points until the next renewal point; thus the time to the next renewal point has an $E_{r-j,\lambda}$ distribution. This gives (47). In like manner when $F = E_{r,\lambda}$ one has

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$F^{(n-m)}(dt) = E_{(n-m)r,}(dt) \ . \tag{48}$$

Then substitution from (47) and (48) into (43) and integration with respect to t yields, from (44), the following expression for $Q_n(x_2, x_3, \ldots, x_k)$:

$$\sum_{j(2),\ldots,j(k)} k^{-(nr+j(2)+\ldots+j(k)-1)} B_{n,j(2),\ldots,j(k)} E_{j(2),\ldots,j(k)} \tag{49}$$

where

$$E_{j(2),\ldots,j(k)} = \prod_{l=2}^{k} E_{r-j(l),\lambda}(x_l)$$

and

$$(nr-mr-1)![(nr + j(2) + \ldots + j(k) - 1)!]^{-1} B_{n,j(2),\ldots,j(k)}$$

is

$$\sum_{m=0}^{n-1} \sum_{m(2)+\ldots+m(k)=m} \prod_{l=2}^{k} [(m(l)r+j(l))!]^{-1} \ .$$

Writing $\omega = \exp(2\pi i/r)$ for an rth root of unity one finds that $r^{k-1} B_{n,j(2),\ldots,j(k)}$ is

$$\sum_{s_2,\ldots,s_k} \omega^{-(j(2)s_2+\ldots+j(k)s_k)} \times \left\{1 + \omega^{s_2} + \ldots + \omega^{s_k}\right\}^{nr+j(2)+\ldots+j(k)-1} \tag{50}$$

where the summation is over all integers s_l , $l = 2, \ldots, k$ with $1 \le s_l \le r$.

Write

$$\theta_{r,k} = \max\left[\frac{1}{k}\left|\left(1 + \omega^{s_2} + \ldots + \omega^{s_k}\right)\right|\right]$$

where the maximum is taken over all integral $s_2, s_3, \ldots, s_k$ with $1 \le s_l \le r$ and $s_2 + \ldots + s_k < (k-1)r$. It is not difficult to see that

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

this maximum occurs when $s_2 = s_3 = \ldots = s_k = 1$ so that

$$\theta_{r,k} = \frac{1}{k}\left[1+2(k-1)\cos\frac{2\pi}{r} + (k-1)^2\right]^{\frac{1}{2}} \tag{51}$$

and $0 < \theta_{r,k} < 1$. Since the term on the right of (50) corresponding to $s_2 = s_3 = \ldots = s_k = r$ is just $k^{nr+j(2)+\ldots+j(k)-1}$ we obtain

$$\left|k^{-(nr+j(2)+\ldots+j(k)-1)}B^r_{n,j(2),\ldots,j(k)} - r^{-(k-1)}\right| \leq \frac{(r^{i-1}-1)}{r^{k-1}}\left[\theta_{r,k}\right]^{nr+j(2)+\ldots+j(k)-1} . \tag{52}$$

Now write

$$G_{r,\lambda,k}(x_2, \ldots, x_k) = \prod_{j=2}^{k} G_{r,\lambda}(x_j)$$

and for any multivariate distribution function H and integrable α,

$$E(\alpha|H) = \int \ldots \int \alpha(y_2, \ldots, y_k)H(dy_2, \ldots, dy_k) .$$

Using (52) in (49) we obtain

$$|E(\alpha|Q_n) - E(\alpha|G_{r,\lambda,k})| \leq (r^{k-1}-1)\theta^{nr-1}_{r,k}E(\alpha|G_{r,\lambda,k}) , \tag{53}$$

for any non-negative α. In particular, taking

$$\alpha(y_2, \ldots, y_k) = \begin{cases} 1 , & y_2 \leq x_2, \ldots, y_k \leq x_k , \\ 0 & \text{otherwise,} \end{cases}$$

we obtain

$$\lim_{n\to\infty} Q_n(x_2, \ldots, x_k) = \prod_{j=2}^{k} G_{r,\lambda}(x_j) ,$$

verifying, in the case of Erlang service times, that the forward recurrence times measured from $Q^{(k)}_{n,1}$ are asymptotically independent with common distribution function G. In like manner taking

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$\alpha(y_2, \ldots, y_k) = \begin{cases} 1 \text{ , if } y_2 + \ldots + y_k \leq kx \text{ ,} \\ 0 \quad \text{otherwise,} \end{cases}$$

we obtain

$$\lim_{n\to\infty} Pr\left\{\frac{1}{k} S_m^{(k)} - Q_{m,1}^{(i)} \leq x\right\} = G^{(k-1)}(kx)$$

verifying, for Erlang service times, the asymptotic result given in (42).

The inequality (53) gives insight into the rate of convergence to the asymptotic formulae. For instance $E(\alpha|Q_n)$ is approximated by its asymptotic value $E(\alpha|G_{r,\lambda,k})$ with an absolute error not exceeding $100b\%$ of the approximating quantity provided

$$(r^{k-1}-1)\theta_{r,k}^{nr-1} \leq b \text{ .}$$

To indicate how this approximation works in practice we give, in the table below, for each r, $k = 2, 3, \ldots, 10$, first the value of $\theta_{r,k}$ and then the smallest integer n which ensures that

$$(r^{i-1}-1)\theta_{r,k}^{nr-1} \leq 0.01, 0.05$$

respectively. Thus the smaller of the two integer entries refers to the 5% level of approximation.

The table on the opposite page indicates that for small r and k, the approximation by the asymptotic distribution of forward recurrence times is quite close after a relatively small number of departures. For instance when $r = 4$ and $k = 5$ one attains 1% accuracy after the fourteenth departure from the system, that is after about 3 departures per server. Even in the extreme case of the table, namely $r = k = 10$, one attains 1% accuracy after about 15 departures per server. The table shows that there are many practical circumstances in which the asymptotic approximation will be good enough for practical purposes.

However (53) does not apply directly to the asymptotic expression for $E\left(Q_{n,1}^{(k)}\right)$. To consider this quantity we return to (45) which, in the case of Erlang service times, yields $Pr\left(Q_{n,1}^{(k)} \leq x\right)$ to be

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

1% and 5% approximation

r \ k	2	3	4	5	6	7	8	9	10
2	0.000 0 0	0.333 4 3	0.500 6 4	0.600 8 7	0.667 11 9	0.714 14 12	0.750 17 15	0.778 21 18	0.800 25 22
3	0.500 3 3	0.577 5 4	0.661 7 6	0.721 10 8	0.764 13 11	0.795 17 15	0.820 21 19	0.839 26 23	0.854 31 28
4	0.707 5 4	0.745 7 6	0.791 10 8	0.825 14 12	0.850 18 16	0.869 24 21	0.884 30 26	0.896 36 33	0.906 44 40
5	0.809 7 5	0.834 9 7	0.861 13 11	0.883 18 16	0.899 24 21	0.915 33 29	0.921 38 35	0.929 48 44	0.936 58 54
6	0.866 8 6	0.882 12 9	0.901 17 14	0.912 23 20	0.927 31 27	0.937 40 36	0.944 50 46	0.949 61 56	0.954 74 68
7	0.901 10 7	0.913 14 11	0.927 20 17	0.938 28 25	0.946 38 33	0.952 48 43	0.958 62 56	0.962 76 73	0.966 93 85
8	0.923 11 8	0.933 16 13	0.943 24 20	0.952 33 29	0.958 44 40	0.963 57 52	0.967 72 66	0.971 91 83	0.973 107 99
9	0.940 13 10	0.947 19 16	0.955 28 24	0.962 39 34	0.967 52 47	0.971 68 61	0.974 86 78	0.977 105 97	0.979 128 118
10	0.951 15 11	0.957 21 18	0.964 32 28	0.969 44 38	0.973 60 54	0.976 77 71	0.979 97 90	0.981 121 112	0.983 145 136

$$(54)\qquad \sum_{j(2),\ldots,j(k)=0}^{r-1} k^{-\left(nr+j(2)+\ldots+j(k)-1\right)} B_{n,j(2),\ldots,j(k)} \times \times E_{nr+j(2)+\ldots+j(k),k\lambda}(x) \ .$$

Thus

https://doi.org/10.1017/S0004972700040843 Published online by Cambridge University Press

$$(55)\quad E\left[Q_{n,1}^{(k)}\right] = \frac{1}{k\lambda} \sum_{j(2),\ldots,j(k)=0}^{r-1} \{nr+j(2) + \ldots + j(k)\} \times \times k^{-\left(nr+j(2)+\ldots+j(h)\right)} B_{n,j(2),\ldots,j(h)} .$$

We find then that (7) takes the form

$$(56)\quad E\left[Q_{n,1}^{(k)}\right] \simeq \frac{nr}{k\lambda} + \frac{(r-1)(k-1)}{2k\lambda}$$

and, using (52), that

$$(57)\quad \left|E\left[Q_{n,1}^{(k)}\right] - \frac{nr}{k\lambda} - \frac{(r-1)(k-1)}{2k\lambda}\right| \leq \left(r^{k-1}-1\right)\theta_{r,k}^{nr-1}\left[\frac{nr}{k\lambda} + \frac{(r-1)(k-1)}{2k\lambda}\right] .$$

Thus the above table again enables one to assess the percentage accuracy in using the asymptotic expression (7). However it should be noted that for all $n \geq 1$,

$$(58)\quad \frac{nr}{k\lambda} \leq E\left[Q_{n,1}^{(k)}\right] \leq \frac{nr}{k\lambda} + \frac{(r-1)(k-1)}{k\lambda} ,$$

these inequalities are often good enough for practical purposes. Note that (56) always gives a value mid-way between the bounds in (58). The difference between the upper and lower bounds in (58) does not exceed the mean service time and, for small r and k , the difference is even smaller. For example with $k = 3$ and $r = 5$ one finds that the expected time of the 9th departure lies between 3 and 3.53 mean service times, whereas the expected time of the 99th departure lies between 33 and 33.53 mean service times. Note that the asymptotic expression yields 3.26 mean service times as the expected time of the 9th departure; from the table we are locating the expectation in question to within 1% of the approximation, thus (57) locates the expected time of the 9th departure between 3.2274 and 3.2926 mean service times.

To establish (58) we return to (54) and observe that for any $j(2), \ldots, j(k)$, $0 \leq j(l) < r$, one has

$$E_{nr+(k-1)(r-1),k\lambda}(x) \leq E_{nr+j(2)+\ldots+j(k),k\lambda}(x) \leq E_{nr,k\lambda}(x) .$$

Thus

$$E_{nr+(k-1)(r-1),k\lambda}(x) \leq Pr\left[Q_{n,1}^{(k)} \leq x\right] \leq E_{nr,k\lambda}(x)$$

and (58) follows at once.

Monash University.