J. Appl. Prob. **13**, 195–199 (1976)
Printed in Israel

ON THE RELATIONSHIP BETWEEN THE DISTRIBUTION OF MAXIMAL QUEUE LENGTH IN THE $M/G/1$ QUEUE AND THE MEAN BUSY PERIOD IN THE $M/G/1/n$ QUEUE

ROBERT B. COOPER AND
BORGE TILT, *Georgia Institute of Technology*

Abstract

Takács has shown that, in the $M/G/1$ queue, the probability $P(k \mid i)$ that the maximum number of customers present simultaneously during a busy period that begins with i customers present is $P(k \mid i) = Q_{k-i}/Q_k$, where the Q's are easily calculated by recurrence in terms of an arbitrary $Q_0 \neq 0$. We augment Takács's theorem by showing that $P(k \mid i) = b_{k-i}/b_k$, where b_n is the mean busy period in the $M/G/1$ queue with finite waiting room of size n; that is, if we take Q_0 equal to the mean service time, then $Q_n = b_n$.

MAXIMAL QUEUE LENGTH DISTRIBUTION; QUEUES WITH FINITE WAITING ROOM

1. Introduction

We consider the $M/G/1/n$ queue: customers arrive according to a Poisson process, with rate λ, at a system composed of a single server and n waiting positions. Service times are identically distributed, positive random variables, independent of the arrival process and each other, with distribution function $H(x)$ and mean value $a = \int_0^\infty x dH(x)$. An arriving customer who finds the server busy and all n waiting positions occupied is cleared from the system; all others wait as long as necessary for service. The order of service is not specified.

This model, with explicit specification of the maximum allowable queue size, is important in applications because it allows one to examine the relationship between the number of waiting positions provided and the proportion of customers who will be denied service. The finiteness of the size of the waiting room makes the analysis of this model more difficult than that of its infinite-waiting room counterpart (see, for example, Cohen (1969), Cooper (1972) and Riordan (1962)). Of interest in this context are the mean duration of the busy period as a function of the number n of waiting positions, and the distribution of maximal queue length during a busy period in the corresponding system with an unlimited number of waiting positions. In this paper we show that these two quantities are intimately connected. In particular, we use this observation to

Received in revised form 6 June 1975.

augment a theorem of Takács, namely Theorem 3 of Takács (1969), which we now state. In what follows, we adopt the notation and terminology of Takács (1969).

Let $\xi(t)$ be the queue size at time t, that is, the total number of customers in the system at time t. $\xi(0)$ is the initial queue size, that is, the number of customers already waiting for service at time $t = 0$. Let θ_0 be the length of the initial busy period, and for $0 \leqq i \leqq k$ define

(1) $$P(k \mid i) = P\left\{ \sup_{0 \leqq t \leqq \theta_0} \xi(t) \leqq k \mid \xi(0) = i \right\}$$

as the probability that the maximal queue size during the initial busy period is $\leqq k$ given that the initial queue size is i. Let π_j be the probability that exactly j customers arrive during a service time; then

(2) $$\pi_j = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} \, dH(x) \qquad (j = 0, 1, 2, \cdots),$$

with generating function $\pi(z) = \sum_{j=0}^\infty \pi_j z^j$ given, for $|z| \leqq 1$, by

(3) $$\pi(z) = \psi(\lambda - \lambda z),$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service-time distribution function,

(4) $$\psi(s) = \int_0^\infty e^{-sx} \, dH(x).$$

Then Takács's theorem (Theorem 3 of Takács (1969)) is as follows.

Theorem. For $0 \leqq i \leqq k$ *we have*

(5) $$P(k \mid i) = \frac{Q_{k-i}}{Q_k}$$

where

(6) $$Q(z) = \sum_{k=0}^\infty Q_k z^k = \frac{Q_0 \pi(z)}{\pi(z) - z}$$

for $|z| < \delta$ *and* δ *is the smallest non-negative real root of*

(7) $$\pi(z) = z.$$

If $\lambda a \leqq 1$, *then* $\delta = 1$ *and if* $\lambda a > 1$, *then* $\delta < 1$. Q_0 *is an arbitrary non-null constant.*

In his proof, Takács shows that

(8) $$Q_k = \sum_{j=0}^k \pi_j Q_{k+1-j} \qquad (k = 0, 1, 2, \cdots),$$

from which (6) follows immediately. (The proof given in Takács (1969) is more elementary than his earlier proofs — see the references in Takács (1969). Cohen (1967), (1969) has also studied the distribution of maximal queue length; his results (see pp. 252, 571–2 of Cohen (1969)) appear to be more complicated than those of Takács.) Note that the probability $P(k \mid i)$ is the same whether the waiting-room size n is finite or infinite, as long as $n+1>k$. It is worth remarking here that, with $P(0 \mid 0)=1$, (5) implies $P(k \mid i)= P(k \mid k)/P(k-i \mid k-i)$; thus Q_k may be given the interpretation $[P(k \mid k)]^{-1}$.

In this paper we show that

$$P(k \mid i) = \frac{b_{k-i}}{b_k}, \tag{9}$$

where b_n is the mean busy period in the $M/G/1/n$ queue; that is, if we take $Q_0 = b_0 = a$, then $Q_n = b_n$.

2. The mean busy period

We begin with the following recurrence on n for the $M/G/1/n$ queue:

$$b_n = a + \sum_{j=1}^{n-1} \pi_j \sum_{k=n-j+1}^{n} b_k + \sum_{j=n}^{\infty} \pi_j \sum_{k=1}^{n} b_k \qquad (n = 0, 1, 2, \cdots), \tag{10}$$

with the convention that any undefined sum is taken to equal zero. To prove (10) observe first that, clearly, $b_0 = a$. Now assume $n \geqq 1$. Observe that the busy period is composed of the service time of the first customer plus some additional time if there are any new arrivals during the first service time. Suppose that exactly $j (1 \leqq j \leqq n-1)$ arrivals occur during the first service time. Then, as the second service time begins, there will be $j-1$ customers waiting in the queue. Since the length of the busy period does not depend on the order of service of waiting customers, we can imagine that none of these $j-1$ waiting customers will enter service until any and all new customers are served who enter the waiting room after the start of the second service time. Thus the mean time until the next (if there is one) of the original j customers enters service is b_{n-j+1} (because only $n-j+1$ waiting positions were available to new arrivals during this time). Hence, using this queue discipline, the mean time required to serve all of the original j waiting customers is $b_{n-j+1} + \cdots + b_n$; this explains the second term on the right side of (10). Finally, if $j \geqq n$ customers arrive during the first service time, then the mean time until the completion of service of those n customers who enter the waiting room during the first service time is $b_1 + \cdots + b_n$.

Equation (10) can be written

$$b_n = a + \sum_{k=0}^{n-1} \Big(1 - \sum_{j=0}^{k} \pi_j\Big) b_{n-k} \qquad (n = 0, 1, \cdots). \tag{11}$$

If we subtract the $(n-1)$th equation from the nth in the above set, we get the following system, which appears most suitable for calculation of $b_1, b_2, \cdots$ by recurrence:

$$(12) \qquad b_n = \begin{cases} \pi_0^{-1} a & (n=1) \\ \pi_0^{-1}\left[(1-\pi_1)b_{n-1} - \sum_{j=1}^{n-2} \pi_{n-j} b_j\right] & (n=2,3,\cdots). \end{cases}$$

Observe that Equation (12) can be written

$$(13) \qquad b_k = \sum_{j=0}^{k} \pi_j b_{k+1-j} \qquad (k=0,1,\cdots;\ b_0 = a).$$

Equation (13) is identical to Equation (8); therefore, Equation (9) is true as asserted, and further, if we take $Q_0 = a$ then $Q_n = b_n$ $(n = 0,1,2,\cdots)$.

Similar results have been obtained by Tomko (1967), Cohen (1971) and Rosenlund (1973).

3. 'Direct' derivation of Equation (9)

We have shown that the distribution of maximal queue length is a ratio of mean busy periods, but this equality appeared as a consequence of the fact that the quantities b_n and Q_n $(n = 0,1,2,\cdots)$ satisfy the same recurrence (8). We now show that (9) can be obtained directly from arguments that relate only to the b_n $(n = 0,1,2,\cdots)$.

We begin by defining the i-busy period as the continuous busy time of a server that starts serving when i customers are in the system. (The 1-busy period is thus the ordinary busy period.) Let $B_k(i)$ be the duration of the i-busy period in the $M/G/1/k$ queue $(1 \leqq i \leqq k+1)$, and define $B_\infty(i) = B(i)$, $B_k(1) = B_k$, $E[B_k(i)] = b_k(i)$, and $b_k(1) = b_k$. (Note that, in Takács's notation, $\theta_0 = B_\infty(i)$ when $\xi(0) = i$.)

Now, it is clear that

$$(14) \qquad B_k(i) = B_{k+1-i} + B_k(i-1) \qquad (k=1,2,\cdots;\ i=2,\cdots,k+1),$$

from which it follows that

$$(15) \qquad b_k(i) = b_{k+1-i} + b_{k+2-i} + \cdots + b_k \qquad (k=0,1,\cdots;\ i=1,\cdots,k+1).$$

It is also true that

$$(16) \qquad b_i(i) = b_{k-1}(i) + \left[1 - P\left\{\sup_{0 \leqq t \leqq B(i)} \xi(t) \leqq k \mid \xi(0) = i\right\}\right] b_k \qquad (1 \leqq i \leqq k).$$

To prove (16) observe that during the i-busy period in the $M/G/1/k$ queue we can imagine that the customer C, if any, whose arrival causes the waiting room to be fully occupied for the first time, will not enter service until there are no other waiting customers. Then the mean time from the start of the i-busy period until the system is cleared of everyone but C (if he exists) is $b_{k-1}(i)$. If in fact no such customer C arrives during $[0, B(i)]$, then the i-busy period ends; if C does arrive, which occurs with probability $1 - P\{\sup_{0 \leqq t \leqq B(i)} \xi(t) \leqq k \mid \xi(0) = i\}$, then the additional mean time required to serve C and all his descendents is $b_k(1) = b_k$.

Equations (15) and (16) together imply (9). Thus, we have given a 'direct' proof of (9), as promised.

References

COHEN, J. W. (1967) Distribution of the maximum number of customers present simultaneously during a busy period for the queueing systems $M/G/1$ and $G/M/1$. *J. Appl. Prob.* **4**, 162–179.

COHEN J. W. (1969) *The Single Server Queue.* North Holland, Amsterdam.

COHEN, J. W. (1971) On the busy periods for the $M/G/1$ queue with finite and with infinite waiting room. *J. Appl. Prob.* **8**, 821–827.

COOPER, R. B. (1972) *Introduction to Queueing Theory.* Macmillan, New York.

RIORDAN, J. (1962) *Stochastic Service Systems.* Wiley, New York.

ROSENLUND, S. I. (1973) On the length and number of served customers of the busy period of a generalised $M/G/1$ queue with finite waiting room. *Adv. Appl. Prob.* **5**, 379–389.

TAKÁCS, L. (1969) On inverse queuing processes. *Zastosowania Matematyki* **10**, 213–224.

TOMKO, J. (1967) A limit theorem for a queue when the input rate increases indefinitely. (In Russian.) *Studia Sci. Math. Hung.* **2**, 447–454.

be followed by the first item, will not enter service until there are no other waiting customers. Then the mean time from the start of the C-busy period until the system is cleared of everyone but C (if he exists) is [illegible]. If in fact no such customer C arrives during [illegible], then the C-busy period ends; if C does arrive, which occurs with probability [illegible], then the additional mean time required to serve C and all his descendants is [illegible].

Equations ([illegible]) and ([illegible]) together imply (9). Thus, we have given a 'direct' proof of (9), as promised.

References

[illegible] (19[illegible]) Distribution of the maximum number of customers present simultaneously during a busy period for the queueing systems M/G/1 and GI/M/1. J. Appl. Prob. 4, [illegible].
Cohen, J. W. (19[illegible]) The Single Server Queue. North-Holland, Amsterdam.
[illegible] (19[illegible]) [illegible] the busy periods for the M/G/1 queue with finite and [illegible] waiting room. J. Appl. Prob. [illegible].
[illegible] (19[illegible]) [illegible] Introduction to Queueing Theory. Macmillan, New York.
[illegible] (19[illegible]) [illegible] Stochastic Service Systems. Wiley, New York.
[illegible] (19[illegible]) On the [illegible] number of early descendants in a [illegible] process [illegible] queueing [illegible] waiting room. Adv. Appl. Prob. [illegible].
[illegible] (19[illegible]) [illegible] Zastosowania Matematyki 10, [illegible].
[illegible] [illegible] when the input [illegible]. [illegible]