J. Appl. Prob. **15**, 406–413 (1978)
Printed in Israel

ON THE OPTIMAL ASSIGNMENT OF CUSTOMERS TO PARALLEL SERVERS

RICHARD R. WEBER, *University of Cambridge*

Abstract

We consider a queuing system with several identical servers, each with its own queue. Identical customers arrive according to some stochastic process and as each customer arrives it must be assigned to some server's queue. No jockeying amongst the queues is allowed. We are interested in assigning the arriving customers so as to maximize the number of customers which complete their service by a certain time. If each customer's service time is a random variable with a non-decreasing hazard rate then the strategy which does this is one which assigns each arrival to the shortest queue.

QUEUING; SHORTEST LINE; STOCHASTIC ORDER; MULTI-SERVER

1. Introduction

Consider a queuing system having m identical servers, each with its own queue. Identical customers arrive according to some arbitrary stochastic process and upon arrival each customer must be assigned to some server's queue. No jockeying amongst the queues is allowed. Each server serves its queue according to a first-in first-out (FIFO) discipline and the time taken to serve any customer is a random variable with a non-decreasing hazard rate. This means that the longer a customer has been in service the more likely it is to complete service shortly. We are interested in assigning the arriving customers so as to maximize the number of customers that complete their service by a certain time. S^* is the strategy of assigning customers to queues which puts each arriving customer at the rear of the queue for which the expected time until it will begin service is least. It is the result of this paper that amongst all possible assignment strategies, S^* is optimal in the sense of stochastic order. It maximizes for all k and s the probability that k or more customers complete service by time s.

Winston (1977) suggested this problem and proved the optimality of S^* in the case where the servers are identical exponential and the customer arrival process is Poisson. We begin by giving a quick proof of the result for exponential servers

Received 11 August 1977; revision received 30 September 1977.

while making no assumption about the arrival process. We then treat the more general case of a service distribution which has a non-decreasing hazard rate. It is this last case which best models situations of parallel servers found in supermarket checkouts and highway toll stations.

2. Optimality of S^*: exponential servers

Theorem 1. Assume that the servers are identical exponential. Then S^* is optimal in that it maximizes for all initial states, k and s the probability that k or more customers complete service by time s.

Proof. The proof is by induction on k. Clearly the theorem is true for $k = 1$. Assume it is true for $k = 1, \cdots, k' - 1$. By the inductive hypothesis it must be optimal to assign arrivals according to S^* after the time of the first service completion. So we only need to show that S^* is optimal prior to that time. Note that there is no difference between queues with the same number of customers. S^* is the strategy which assigns each arriving customer to a queue with the least number of customers.

Suppose at the start that queue 1 contains more customers than queue 2. Suppose that during the time prior to the first service completion strategy S_1 assigns the first arrival to queue 1. Let S_2 be a strategy which instead assigns that first arrival to queue 2 and then up to the time of the first service completion assigns each arrival to the same queue as it would have been assigned had S_1 been followed, except that the first arrival, if any, which would have been assigned by S_1 to queue 2 is assigned instead by S_2 to queue 1. We also assume that up to the time of the first service completion we do not service queue 2 as long as queue 2 would be empty had S_1 been followed. Then just after the first service completion occurs, irrespective of when and in which queue it takes place, we find either that the states of the queues reached by S_1 and S_2 are identical, or that the state reached by S_2 has one less customer in queue 1 and one more customer in queue 2 than the state reached by S_1. So at the time of the first service completion the effects of S_1 and S_2 will have been either identical or so as to cause a difference in the placement of one customer. If the latter is the case then S_2 is at least as good as S_1. For imagine that the one customer differently placed had been an arrival just after the time of the first service completion. Using the inductive hypothesis for $k = k' - 1$ we see that to have placed this customer in queue 2 is at least as good as having placed it in queue 1, since prior to its placement queue 1 was no shorter than queue 2. So S_2 is a strategy at least as good as S_1. If we drop the assumption that the service of queue 2 is delayed as long as it would be empty had S_1 been followed, then S_2 is even better. Hence it is optimal to assign the first arrival (and by similar arguments all subsequent arrivals) to a queue with the least number of

customers. The theorem is thus true for $k = k'$ and the proof by induction is complete.

3. Optimality of S^*: servers of non-decreasing hazard rate

In considering the case where the service distribution has a non-decreasing hazard rate we work in discrete time using the following model. Time proceeds in discrete periods t ($t = 1, 2, 3, \cdots$). During period t all arriving customers are first assigned to queues and then the customers at the front of queues are each given one period of service. An *age* is associated with each customer. A customer which has received no service has age 0. A customer at the front of a queue with age x at the start of period t completes service that period with probability $p(x)$. If it does not complete service that period then its age at the start of period $t+1$ is $x+1$. Although in the usual course of events customers will have integer ages we shall allow starting states in which the ages take any values in $[0, \infty)$. We assume that $p(x)$ is non-decreasing, continuous and differentiable in x and that $p(x) = 1$ for all $x \geqq$ some N. S^* assigns each arriving customer to a queue which amongst those with the least number of customers has a front customer who has been in service the longest.

A state of the queues is specified by $(n; x) = (n_1, \cdots, n_m; x_1, \cdots, x_m)$, where n_i is the number of customers in queue i and x_i is the age of the customer at the front of queue i. We shall denote by $(n; x; i)$ the state which is derived from the state $(n; x)$ by the addition of exactly one more customer to queue i. Similarly $(n; x; i, j)$ denotes the state which is derived from the state $(n; x)$ by the addition of exactly one customer to each of the queues i and j. Given the customer arrival process we shall denote by $P^S_{k,t,s}(n; x)$ the probability that k or more customers complete service by the end of period $s-1$ given that we start at the beginning of period t in state $(n; x)$ and apply assignment strategy S. We shall write $P_{k,t,s}(n; x)$ for $P^{S^*}_{k,t,s}(n; x)$, and $P_{k,t,s}(n_1, n_2, \cdots, n_m; *, x_2, \cdots, x_m)$ for $P_{k-1,t,s}(n_1 - 1, n_2, \cdots, n_m; 0, x_2, \cdots, x_m)$.

Theorem 2. S^* is optimal in that for any other strategy S, $P_{k,t,s}(n; x) \geqq P^S_{k,t,s}(n; x)$ for all $(n; x)$, k, t and s.

Proof. The proof is by induction on t. Clearly the theorem is true for $t = s$. Assume it is true for $t = t'+1, \cdots, s$. By the inductive hypothesis it must be optimal to apply S^* from period $t'+1$ onwards. So we only need to show that it is optimal to apply S^* at period t'. We say that 1 is longer than 2 for $(n; x)$ when queue 1 has a greater expected service time than queue 2. This occurs if n_1 is greater than n_2 or if n_1 equals n_2 with x_1 less than x_2.

Consider the assignment at t' of the customers which arrive that period. Suppose that by assigning all but one of the arrivals we reach state $(n; x)$ for

which 1 is longer than 2, and that by assigning the one remaining arrival we can reach either of the states $(n; x; 1)$ or $(n; x; 2)$. Now for any state $(n; x)$ we shall define:

$$D_{k,t,s}(n; x) = P_{k,t,s}(n; x; 2) - P_{k,t,s}(n; x; 1).$$

It is the result of Theorem 3 that if 1 is longer than 2 for $(n; x)$ then $D_{k,t,s}(n; x) \geqq 0$ for all k, t and s. Applying this result for $t = t'$ to a problem in which there are no arrivals at t' but otherwise the same arrivals as in our original problem, we deduce that it is at least as good to assign the arrivals at t' so as to reach the state $(n; x; 2)$ as to assign them so as to reach the state $(n; x; 1)$. From this it is clear that it is optimal to assign arrivals according to S^* at period t' and the inductive proof is complete.

Theorem 3. For all k, t and s the derivatives in (1), (2) and (5) exist and

$$d^2/dx_1 dx_2\{P_{k,t,s}(n; x)\} \leqq 0, \tag{1}$$

$$d/dx_2\{P_{k,t,s}(n; *, \cdots, x_m) - P_{k,t,s}(n; N, \cdots, x_m)\} \leqq 0, \tag{2}$$

$$\begin{aligned}\{P_{k,t,s}(n; *, *, \cdots, x_m) - P_{k,t,s}(n; *, N, \cdots, x_m)\} \\ - \{P_{k,t,s}(n; N, *, \cdots, x_m) - P_{k,t,s}(n; N, N, \cdots, x_m)\} \leqq 0,\end{aligned} \tag{3}$$

and with 1 longer than 2 for $(n; x)$,

$$D_{k,t,s}(n; x) \geqq 0, \tag{4}$$

$$d/dx_1\{D_{k,t,s}(n; x)\} \leqq 0 \tag{5}$$

and

$$D_{k,t,s}(n; *, \cdots, x_m) - D_{k,t,s}(n; N, \cdots, x_m) \leqq 0. \tag{6}$$

Proof. All of (1)–(6) are proved by induction on t. Clearly they are true for $t = s$. Assume that they hold for $t = t' + 1, \cdots, s$. The steps by which we show that (1)–(6) are true for $t = t'$ are routine but lengthy. So we will explain the inductive step in detail for only (1) and (5) and just indicate the check of the inductive step for the others. Let $p_i = 1 - q_i$ and p_i' denote respectively $p(x_i)$ and $d/dx_i\{p(x_i)\}$.

I. Inductive step for (1). It is enough to consider the case $m = 2$ where there are no arrivals at t':

$$\begin{aligned}P_{k,t',s}(n; x) = {} & p_1 p_2 . P(*, *) + p_1 q_2 . P(*, x_2 + 1) \\ & + q_1 p_2 . P(x_1 + 1, *) + q_1 q_2 . P(x_1 + 1, x_2 + 1),\end{aligned}$$

where for simplicity we have omitted the suffices $k, t' + 1, s$ and the argument n

from the four P's on the right-hand side. By the inductive hypothesis for $t = t' + 1$ the right-hand side is differentiable and

$$d^2/dx_1dx_2\{P_{k,t',s}(n;x)\}$$

$$(7) \qquad = q_1q_2 \,.\, d^2/dx_1dx_2\{P(x_1+1, x_2+1)\}$$

$$(8) \qquad + p_1'q_2 \,.\, d/dx_2\{P(*, x_2+1) - P(x_1+1, x_2+1)\}$$

$$(9) \qquad + q_1p_2' \,.\, d/dx_1\{P(x_1+1, *) - P(x_1+1, x_2+1)\}$$

$$(10) \qquad + p_1'p_2' \,.\, [\{P(*,*) - P(*, x_2+1)\} - \{P(x_1+1, *) - P(x_1+1, x_2+1)\}].$$

Now by the inductive hypothesis for (1) we have $(7) \leqq 0$. By the inductive hypothesis for (1) and (2) we have,

$$(11) \qquad \begin{aligned} 0 &\geqq d/dx_2\{P(*, x_2+1) - P(N, x_2+1)\} \quad \text{by (2)} \\ &\geqq d/dx_2\{P(*, x_2+1) - P(x_1+1, x_2+1)\} \quad \text{by (1).} \end{aligned}$$

So noting that $p_i' \geqq 0$ we have $(8) \leqq 0$ and similarly $(9) \leqq 0$. By the inductive hypothesis for (1) and using (11) above we have,

$$(12) \qquad \begin{aligned} 0 &\geqq \{P(*,*) - P(*, N)\} - \{P(N, *) - P(N, N)\} \quad \text{by (3)} \\ &\geqq \{P(*,*) - P(*, N)\} - \{P(x_1+1, *) - P(x_1+1, N)\} \quad \text{by (11)} \\ &\geqq \{P(*,*) - P(*, x_2+1)\} - \{P(x_1+1, *) - P(x_1+1, x_2+1)\} \quad \text{by (11).} \end{aligned}$$

So we have $(10) \leqq 0$ and the inductive step showing that (1) is true for $t = t'$ is complete.

Note that (2) and (3) are difference versions of (1) in which differentiation is replaced by differencing between x_i taking the values $*$ and N. Since the states $(n; *, x_2)$ and $(n; N, x_2)$ which appear in (2) are nearly the same we can choose S^* to assign arrivals to exactly the same queues in both states. A similar remark applies to the four states appearing in (3). So once again arrivals are irrelevant to the checking of the inductive step for (2) and (3) and the check follows similar lines as that for (1).

II. Inductive step for (5). We begin by showing that we can disregard arrivals at t'. The two states appearing in the definition of $D(n;x)$ are $(n;x;2)$ and $(n;x;1)$. As 1 is longer than 2 for $(n;x)$ an arriving customer will be assigned by S^* to the queues in these two states so that they result as one of the following pairs:

(a) $(n;x;2,1)$ $(n;x;1,2)$,

(b) $(n;x;2,2)$ $(n;x;1,2)$, where 1 is longer than 2 for $(n;x;2)$,

(c) $(n;x;2,j)$ $(n;x;1,j)$, some $j > 2$ where 1 is longer than 2 for $(n;x;j)$, or

(d) $(n;x;2,j)$ $(n;x;1,2)$, some $j > 2$ where 1 is longer than j for $(n;x;2)$.

If the assignments occur as in (a) then the two states become identical and $D = 0$. If they occur as in (b), (c) or (d) the pair of states which results is still of the type appearing in the definition of D, provided that in case (d) we interchange the labels on queues 2 and j. But since the derivative of interest is with respect to x_1 such an interchange of labels on 2 and j is of no consequence. So it is enough to check the inductive step when $m = 2$ and there are no arrivals at t'. We treat separately the cases of $n_2 > 0$ and $n_2 = 0$.

Suppose first that $n_2 > 0$. Then

$$d/dx_1\{D_{k,t',s}(n;x)\}$$

$$= q_1q_2 \,.\, d/dx_1\{D(x_1+1, x_2+1)\} \tag{13}$$

$$+ q_1p_2 \,.\, d/dx_1\{D(x_1+1, *)\} \tag{14}$$

$$+ p_1'q_2 \,.\, \{D(*, x_2+1) - D(x_1+1, x_2+1)\} \tag{15}$$

$$+ p_1'p_2 \,.\, \{D(*, *) - D(x_1+1, *)\}, \tag{16}$$

where we have abbreviated the notation on the right-hand side as in I. Now by the inductive hypothesis for (5) we have (13) $\leqq 0$ and (14) $\leqq 0$. When $n_1 \neq n_2$ then by the inductive hypothesis for (5) and (6) we have

$$0 \geqq D(*, x_2+1) - D(N, x_2+1) \quad \text{by (6)}$$

$$\geqq D(*, x_2+1) - D(x_1+1, x_2+1) \quad \text{by (5)}.$$

So we have (15) $\leqq 0$ and similarly (16) $\leqq 0$. When $n_1 = n_2$ then 1 will not be longer than 2 at $t'+1$ if the customer at the front of queue 1 completes service at t'. But then each of the terms in (15) and (16) is $\leqq 0$ and so (15) and (16) are $\leqq 0$.

Suppose now that $n_2 = 0$. Using the result (1) which is true for $t = t'$ by I above we have

$$d/dx_1\{D_{k,t',s}(n;x)\} = d/dx_1\{P_{k,t',s}(n_1, 1; x_1, 0) - P_{k,t',s}(n_1+1, 0; x_1, 0)\}$$

$$\leqq d/dx_1\{P_{k+1,t',s}(n_1, 2; x_1, N) - P_{k,t',s}(n_1+1, 0; x_1, 0)\} \quad \text{by (2)}$$

$$= q_1 \,.\, d/dx_1\{D(n_1, 0; x_1+1, 0)\} \tag{17}$$

$$+ p_1'\{D(n_1, 0; *, 0) - D(n_1, 0; x_1+1, 0)\}. \tag{18}$$

By the inductive hypothesis for (5) we have (17) $\leqq 0$ and by the inductive hypothesis for (6) we have (18) $\leqq 0$. This completes the inductive step which shows that (5) is true for $t = t'$. Note that (6) is a difference version of (5) and the inductive step for (6) can be checked similarly. We deduce (4) from (5) and (6) by observing that for $n_1 = n_2$ and $x_1 = x_2$ we have $D(n;x) = 0$ and that D is

increasing as the length of queue 1 increases. This concludes the proof of Theorem 3.

4. Extensions and related results

The results of Section 3 can be extended by letting $N \to \infty$. We can also deduce that S^* is the optimal strategy in a continuous-time formulation. In continuous time the probability that a customer at the front of a queue with age x completes service in the next small interval δt is $h(x).\delta t$, where the hazard rate $h(x)$ is non-decreasing, continuous and differentiable in x. For a distribution on service times with density $f(x)$ and distribution function $F(x)$ we have that $h(x) = f(x)/\{1-F(x)\}$. Suppose we wish to maximize the probability that k or more customers complete service by time s. It is enough to consider the case where no queue initially contains more than k customers and where the total number of future arrivals and customers initially present is exactly mk. Assume customer i requires a service time s_i which has distribution function $F_i(s_i) = \{F(a_i + s_i) - F(a_i)\}/\{1-F(a_i)\}$ if the customer begins its service with age a_i. Unless the customer is one which is initially at the front of a queue it begins service with age 0. Customer i arrives at time t_i where t_i is 0 if the customer is present initially and where the t_i's have a joint distribution function $G(t_1, \cdots, t_{mk})$. Let S be a strategy which assigns arrivals to queues knowing only the F_i's, G, and the past history of the system. Define $A^S(n; x; s_1, \cdots, s_{mk}; t_1, \cdots, t_{mk})$ to be 1 or 0 as S does or does not achieve k service completions by time s when the initial state is $(n; x)$ and the service and arrival times are $s_1, \cdots, s_{mk}$ and $t_1, \cdots, t_{mk}$. Under S the probability of k or more service completions by time s is given by the multiple integral

$$\int_{t_1=0}^{\infty} \cdots \int_{t_{mk}=0}^{\infty} \int_{s_1=0}^{\infty} \cdots \int_{s_{mk}=0}^{\infty} A^S(\cdot) dF_1(s_1) \cdots dF_{mk}(s_{mk}) dG(t_1, \cdots, t_{mk}),$$

where $A^S(\cdot)dF_1 \cdots dF_{mk}dG$ is Riemann integrable. This integral may be approximated by a Riemann sum in which $[0, s)$ is divided into M segments of length s/M. Such a Riemann sum is just the probability of k or more service completions by the end of interval M in a discrete-time formulation of the problem and is maximized for all M by the strategy S^*. The multiple integral is the limit of Riemann sums as $M \to \infty$. So S^* must be optimal for the continuous-time problem.

There are several further claims which are true for S^*.

(i) If we must assign arrivals without knowing the ages of the customers at the fronts of the queues, so that we must assign at random amongst queues with the same number of customers, then S^* is still optimal.

(ii) If a reward 1 is earned every time a customer completes service and

rewards are discounted over time at discount rate α, then S^* maximizes the total discounted reward.

(iii) S^* minimizes the expected value of the mean customer waiting time.

(iv) S^* maximizes the probability that no server is idle at time t.

The method of proof used in Section 3 can also be applied to other stochastic multi-server assignment problems. In Weber and Nash (1978) we find the optimal strategy for the problem considered by Glazebrook and Nash (1976) of scheduling stochastically failing components in a piece of machinery so as to minimize the number of components which fail by a certain time.

References

GLAZEBROOK, K. D. AND NASH, P. (1976) On multi-server stochastic scheduling. *J. R. Statist. Soc.* B **38**, 67–72.

WEBER, R. R. AND NASH, P. (1978) An optimal strategy in multi-server stochastic scheduling. To appear.

WINSTON, W. (1977) Optimality of the shortest line discipline. *J. Appl. Prob.* **14**, 181–189.