J. Appl. Prob. **51,** 880–884 (2014)
Printed in England

OPTIMAL SERVER SELECTION IN A QUEUEING LOSS MODEL WITH HETEROGENEOUS EXPONENTIAL SERVERS, DISCRIMINATING ARRIVALS, AND ARBITRARY ARRIVAL TIMES

SHELDON M. ROSS,* *University of Southern California*

Abstract

We consider a multiple server queueing loss system where the service times of server i are exponential with rate μ_i, where μ_i decreases in i. Arrivals have associated vectors $(X_1, \ldots, X_n)$ of binary variables, with $X_i = 1$ indicating that server i is eligible to serve that arrival. Arrivals finding no idle eligible servers are lost. Letting I_j be the indicator variable for the event that the jth arrival enters service, we show that, for any arrival process, the policy that assigns arrivals to the smallest numbered idle eligible server stochastically maximizes the vector $(I_1, \ldots, I_r)$ for every r if the eligibility vector of arrivals is either (a) exchangeable, or (b) a vector of independent variables for which $\mathbb{P}(X_i = 1)$ increases in i.

Keywords: Loss model; server eligibility; server selection; stochastic maximization

2010 Mathematics Subject Classification: Primary 60K25; 90B22
Secondary 90B36; 68M20

1. Introduction

Consider an n-server system such that each arrival has an associated vector of binary values $(x_1, \ldots, x_n)$ with the interpretation that server i is eligible to serve that arrival if $x_i = 1$ and is ineligible if $x_i = 0$, $i = 1, \ldots, n$. The binary vectors of successive arrivals are assumed to be independent and identically distributed having the distribution of $(X_1, \ldots, X_n)$. There is no queue allowed and an arrival that is not assigned to an idle server that is eligible for that arrival is lost. Moreover, we assume that, independent of all else, the time it takes server i to serve a customer is exponential with rate μ_i, $i = 1, \ldots, n$, where, without loss of generality, we suppose that $\mu_i \geq \mu_{i+1}$, $i < n$.

Optimization problems regarding the preceding loss model have previously been considered only in the special case where $\mathbb{P}(X_i = 1,\ i = 1, \ldots, n) = 1$, that is, only in the case where arrivals can always be served by any of the n servers. Yao [8] assumed that the arrival process is a renewal process and only considered 'priority list' policies, where a priority list policy is specified by a permutation $i_1, i_2, \ldots, i_n$ with the instruction to give an arrival to the idle server that appears earliest on this list. Yao showed that, among priority list policies, the policy $\pi^o \equiv (1, \ldots, n)$ minimized the rate of lost customers. Derman *et al.* [2] showed that, for any arrival process, the total amount of time up to the moment of the nth arrival that the number being served is less than k is, for every k and n, stochastically maximized by π^o. Because their arrival process is general, giving one the option to add an arrival at time t, they also showed that

Received 26 April 2013; revision received 24 September 2013.
* Postal address: Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA. Email address: smross@usc.edu
This material is based upon work supported by the National Science Foundation under contract/grant number CMMI1233337.

the total amount of the first t time units that the number being served is less than k is, for every k and t, stochastically maximized by π^o. Katehakis [5] used a continuous-time Markov decision process to show, in the case of Poisson arrivals, that π^o minimized the rate at which customers are lost (a result that also follows from [2] by the PASTA principle). Hordijk and Koole [4] showed that, for a general arrival process, π^o stochastically minimized the cost incurred at a fixed time t for a variety of cost functions, including that which incurs a cost 1 only when all servers are busy. Sobel [7] assumed that the arrival process was Poisson and showed that, for a variety of cost rules, including that which incurs a cost 1 whenever an arrival is lost, π^o minimized the long-run average cost per unit time. References to some papers earlier than those cited in the preceding are noted in the bibliography.

For our problem, in which not all servers are eligible to serve an arrival, we define the priority list policy $(1, \ldots, n)$ as the policy that gives an arrival to the smallest indexed server that is both arrival eligible and idle. Let I_j be the indicator variable for the event that the jth arrival enters service. For any arrival process that is independent of service times, we show that the priority list policy $(1, \ldots, n)$ stochastically maximizes the vector $(I_1, \ldots, I_r)$ for every r if the eligibility vector of arrivals is either (a) exchangeable, or (b) a vector of independent random variables for which $\mathbb{P}(X_i = 1)$ increases in i.

2. Proof of optimality in the exchangeable case

In this section we suppose that the eligibility random vector $X_1, \ldots, X_n$ is exchangeable. That is, we suppose that there are probabilities α_j, $\sum_{j=0}^n \alpha_j = 1$, such that $\alpha_j = \mathbb{P}(\sum_{i=1}^n X_i = j)$ and, given that $\sum_{i=1}^n X_i = j$, all $\binom{n}{j}$ sets of j servers are equally likely to be the set of eligible servers.

Recall that I_j is the indicator of the event that arrival j is served, $j \geq 1$.

Theorem 1. *For an arbitrary arrival time process, if μ_i decreases in i then, for every $r \geq 1$, the priority list policy $(1, \ldots, n)$ stochastically maximizes the vector $(I_1, \ldots, I_r)$. That is, for any r and any increasing function $h(x_1, \ldots, x_r)$, the priority list policy $(1, \ldots, n)$ maximizes $\mathbb{E}[h(I_1, \ldots, I_r)]$.*

To prove the theorem, we will need a couple of lemmas.

Lemma 1. *Let X, Y, and I be independent, with X being exponential with rate μ, Y being exponential with rate $\lambda < \mu$, and $\mathbb{P}(I = 0) = \lambda/\mu = 1 - \mathbb{P}(I = 1)$. Then $W \equiv X + IY$ is exponential with rate λ.*

Proof. The moment generating function of $X + IY$ is

$$\mathbb{E}[e^{sW}] = \mathbb{E}[e^{sX}]\mathbb{E}[e^{sIY}] = \frac{\mu}{\mu - s}\left[\frac{\lambda}{\mu} + \frac{\mu - \lambda}{\mu}\frac{\lambda}{\lambda - s}\right] = \frac{\lambda}{\lambda - s}$$

which proves the result.

Note that Lemma 1 allows us to couple an exponential random variable X with rate μ and an exponential random variable W with rate $\lambda < \mu$ in such a manner that W is either equal to X or is larger than X by a random amount that is independent of X and is exponential with rate λ.

Definition. Say that the n-server system under consideration is in state S if S is the set of servers that are currently busy. Also, for states $S_2 = (i, i_1, \ldots, i_k)$ and $S_1 = (j, i_1, \ldots, i_{k'})$, say that S_2 dominates S_1 if $i \leq j$ and $k \leq k'$.

https://doi.org/10.1239/jap/1409932680 Published online by Cambridge University Press

Lemma 2. *If S_2 dominates S_1 then, for any r and any increasing function $h(x_1, \ldots, x_r)$,*

$$\sup_{\pi} \mathbb{E}_{\pi}[h(I_1, \ldots, I_r) \mid S_2] \geq \sup_{\pi} \mathbb{E}_{\pi}[h(I_1, \ldots, I_r) \mid S_1],$$

where π is a policy, and $\mathbb{E}_{\pi}[X \mid S]$ denotes the expected value of X given that policy π is employed and S is the state at time 0.

Proof. Suppose that $S_2 = (i, i_1, \ldots, i_k)$ dominates $S_1 = (j, i_1, \ldots, i_{k'})$. We now show by induction on r that, for any policy π_1, there is a policy π_2 such that

$$\mathbb{E}_{\pi_2}[h(I_1, \ldots, I_r) \mid S_2] \geq \mathbb{E}_{\pi_1}[h(I_1, \ldots, I_r) \mid S_1]$$

whenever h is an increasing function of r variables. To begin, consider two scenarios, scenario one having initial state S_1 and scenario two having initial state S_2, and let π_1 be a policy that is to be used in scenario one. Couple the arrival processes in the two scenarios so that arrivals are at identical times. Let $\boldsymbol{T} = (T_1, T_2, \ldots)$ be the sequence of interarrival times, and let $\boldsymbol{T}' = (T_2, T_3, \ldots)$.

Couple service times so that the initial service time of server i_m is the same in both scenarios for $m = 1, \ldots, k$. Couple the initial service times of server j in scenario one with that of server i in scenario two so that either they are equal or the service time of server j exceeds that of server i by an exponential random variable with rate μ_j. (That this is possible follows from Lemma 1.) Because the set of idle servers seen by the first arrival in scenario one is a subset of the set of idle servers seen by the first arrival in scenario two, the lemma follows when $r = 1$. So assume that it is true for r, and now let h be an increasing function of $r + 1$ variables. To complete the proof, we couple the eligibility vectors of the first arrival in the two scenarios, but how we do so depends on which of the following two cases results.

Case 1. If, in scenario two, server i completes service before the first arrival then let the eligibility vectors for the first arrival in the two scenarios be identical. Also, if this case results then whoever policy π_1 assigns to the first arrival in scenario one should also be the server assigned by policy π_2 in scenario two. If there is no eligible idle server for the initial arrival in scenario one and there is at least one eligible idle server in scenario two then let π_2 assign the arrival to any of the idle eligible servers. Noting that the set of idle servers in scenario one at the moment of the first arrival is a subset of the set of idle servers in scenario two at that moment, it follows that if $N(S_1)$ and $N(S_2)$ denote the sets of busy servers after the first arrival in scenarios one and two, respectively, then $N(S_2)$ dominates $N(S_1)$. Moreover, $K_2 \geq K_1$, where K_1 and K_2 are respectively equal to the indicators of the events that the first arrival is served in scenarios one and two. (That is, K_1 and K_2 are respectively the values of I_1 in scenarios 1 and 2.) Now, for any sets of servers A_1 and A_2 such that A_2 dominates A_1, and binary values $k_2 \geq k_1$,

$$\begin{aligned}
&\mathbb{E}_{\pi_1}[h(I_1, \ldots, I_{r+1}) \mid S_1, \boldsymbol{T}, N(S_1) = A_1,\ K_1 = k_1] \\
&\quad \leq \sup_{\pi} \mathbb{E}_{\pi}[h(k_1, I_1, \ldots, I_r) \mid A_1, \boldsymbol{T}'] \\
&\quad \leq \sup_{\pi} \mathbb{E}_{\pi}[h(k_2, I_1, \ldots, I_r) \mid A_1, \boldsymbol{T}'] \\
&\quad \leq \sup_{\pi} \mathbb{E}_{\pi}[h(k_2, I_1, \ldots, I_r) \mid A_2, \boldsymbol{T}'] \\
&\quad = \sup_{\pi} \mathbb{E}_{\pi}[h(I_1, \ldots, I_{r+1}) \mid S_2, \boldsymbol{T}, N(S_2) = A_2,\ K_2 = k_2].
\end{aligned}$$

https://doi.org/10.1239/jap/1409932680 Published online by Cambridge University Press

In the first inequality above we used the fact that if, in scenario one, server j is busy at the moment of the first arrival then, whether or not server i had completed service in scenario two, the remaining service time of server j is exponential with rate μ_j, and so the situation after the first arrival is the same as if the problem began with the set of busy servers A_1 and we are interested in maximizing the quantity $\mathbb{E}_\pi[h(k_1, I_1, \ldots, I_r)]$ when the set of interarrival times are $\boldsymbol{T}'$. The second inequality follows because h is an increasing function, and the next inequality follows from the induction hypothesis. Hence, there is a policy π_2 such that

$$\mathbb{E}_{\pi_1}[h(I_1, \ldots, I_{r+1}) \mid S_1, \boldsymbol{T}, N(S_1), K_1] \leq \mathbb{E}_{\pi_2}[h(I_1, \ldots, I_{r+1}) \mid S_2, \boldsymbol{T}, N(S_2), K_2].$$

Taking expectations it follows that, for any policy π_1, there is a policy π_2 such that

$$\mathbb{E}_{\pi_1}[h(I_1, \ldots, I_{r+1}) \mid S_1] \leq \mathbb{E}_{\pi_2}[h(I_1, \ldots, I_{r+1}) \mid S_2],$$

which completes the induction and establishes the result in this case.

Case 2. If, in scenario two, server i has not yet completed service before the first arrival then with $X_1^{(1)}, \ldots, X_n^{(1)}$ equal to the eligibility vector for the first arrival in scenario one, let

$$X_s^{(2)} = X_s^{(1)}, \qquad s \neq i,\ s \neq j,$$
$$X_i^{(2)} = X_j^{(1)}, \qquad X_j^{(2)} = X_i^{(1)},$$

and take $X_1^{(2)}, \ldots, X_n^{(2)}$ as the eligibility vector for the first arrival in scenario two. (Because an eligibility random vector $X_1, \ldots, X_n$ is exhangeable, $X_1^{(2)}, \ldots, X_n^{(2)}$ has the appropriate distribution.) Now if the first arrival is assigned to server i in scenario one then assign it to server j in scenario two (which can be done because $X_j^{(2)} = X_i^{(1)}$); otherwise, make the same assignment in scenario two as is made in scenario one, with the exception that if the arrival is not eligible for any idle server in scenario one and there is an idle eligible server for it in scenario two, then in scenario two it should be assigned to one of its idle eligible servers. Now, as in case 1, for $i = 1, 2$, let $N(S_i)$ be the set of busy servers in scenario i after the first arrival. Because $N(S_2)$ dominates $N(S_1)$, and the number of lost customers after the first arrival will be no greater in scenario two than in scenario one, we can, as in case 1, apply the induction hypothesis to complete the proof.

We are now ready to prove Theorem 1.

Proof of Theorem 1. Fix h, and consider any given policy π. Now, if π ever deviates from the priority list policy $(1, \ldots, n)$, it will enter a state that is dominated by the state that would have been entered if $(1, \ldots, n)$ were employed. By Lemma 2, it follows that there is a policy that uses $(1, \ldots, n)$ for the initial period that is at least as good as π with respect to the criterion $\mathbb{E}[h(I_1, \ldots, I_r)]$. But continuing this argument for each subsequent period shows that always using $(1, \ldots, n)$ minimizes $\mathbb{E}[h(I_1, \ldots, I_r)]$.

3. An extension to eligibility vectors of independent random variables

Suppose now that the eligibility vector $(X_1, \ldots, X_n)$ is not exchangeable but rather that it is a vector of independent random variables, and let $p_i = \mathbb{P}(X_i = 1)$. Then we can prove the following result.

Theorem 2. *if $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ and $p_1 \leq p_2 \leq \cdots \leq p_n$ then the priority list policy $(1, \ldots, n)$ stochastically maximizes $(I_1, \ldots, I_r)$ for every r.*

https://doi.org/10.1239/jap/1409932680 Published online by Cambridge University Press

Theorem 2 can be proven in the same manner as Theorem 1, with the exception that in case 2 (which holds when server i has not yet completed service in scenario two before the first arrival), we couple the eligibility vectors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ for the first arrival in the two scenarios in the following manner. For independent uniform (0, 1) random variables $U_1, \ldots, U_n$, set

$$
\begin{aligned}
X_m^{(1)} &= I\{U_m < p_m\}, \qquad m = 1, \ldots, n,\\
X_m^{(2)} &= I\{U_m < p_m\}, \qquad m \neq i, j,\\
X_j^{(2)} &= I\{U_i < p_j\},\\
X_i^{(2)} &= I\{U_j < p_i\},
\end{aligned}
$$

where $I\{A\}$ is the indicator of the event A. Note that, because $p_i < p_j$, if the first arrival is eligible for server i in scenario one then it is also eligible for server j in scenario two. The proof then continues as in the exchangeable case.

Remark. Yao [8] showed that, when μ_i is decreasing in i, the rate of lost customers under the priority list policy $\boldsymbol{L}_1 = (i_1, \ldots, i_k, i_{k+1}, i_{k+2}, \ldots, i_n)$ was less than when the priority list policy $\boldsymbol{L}_2 = (i_1, \ldots, i_k, i_{k+2}, i_{k+1}, \ldots, i_n)$ is employed, whenever $i_{k+1} < i_{k+2}$. We can prove the stronger result that $(I_1, \ldots, I_r)$ is stochastically larger when using $\boldsymbol{L}_1$ than it is when using $\boldsymbol{L}_2$ under either the conditions of Theorem 1 or Theorem 2. This is done by restricting attention to those policies that give highest priority to servers $i_1, \ldots, i_k$ in that order, and lowest priority to those servers $i_n, \ldots, i_{k+3}$ in that order, and then showing by the same approach used to establish Theorem 1 that the best policy under this restriction is that which gives server i_{k+1} priority over server i_{k+2}.

References

[1] Cooper, R. B. (1976). Queues with ordered servers that work at different rates. *Opsearch* **13,** 69–78.

[2] Derman, C., Lieberman, G. J. and Ross, S. M. (1980). On the optimal assignment of servers and a repairman. *J. Appl. Prob.* **17,** 577–581.

[3] Gregory, G. and Litton, C. D. (1975). A conveyer model with exponential service times. *Internat. J. Production Res.* **13,** 1–7.

[4] Hordijk, A. and Koole, G. (1992). On the assignment of customers to parallel queues. *Prob. Eng. Inf. Sci.* **6,** 495–511.

[5] Katehakis, M. N. (1985). A note on the hypercube model. *Operat. Res. Lett.* **3,** 319–322.

[6] Matsui, M. and Fukuta, J. (1977). On a multichannel queueing system with ordered entry and heterogeneous servers. *AIIE Trans.* **9,** 209–214.

[7] Sobel, M. J. (1990). Throughput maximization in a loss queueing system with heterogeneous servers. *J. Appl. Prob.* **27,** 693–700.

[8] Yao, D. D. (1987). The arrangement of servers in an ordered-entry system. *Operat. Res.* **35,** 759–763.