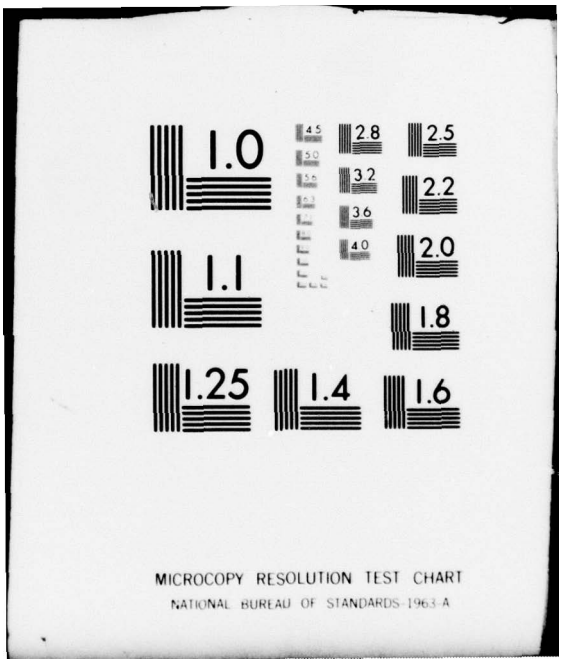
AD-A066 584 CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER F/G 12/1
ON THE OPTIMAL ASSIGNMENT OF SERVERS AND REPAIRMAN.(U)
NOV 78 C DERMAN, G J LIEBERMAN, S M ROSS N00014-77-C-0299
UNCLASSIFIED ORC-78-22 NL

| OF |
AD
A066584

END
DATE
FILMED
5-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

(12)

ORC 78-22
NOVEMBER 1978

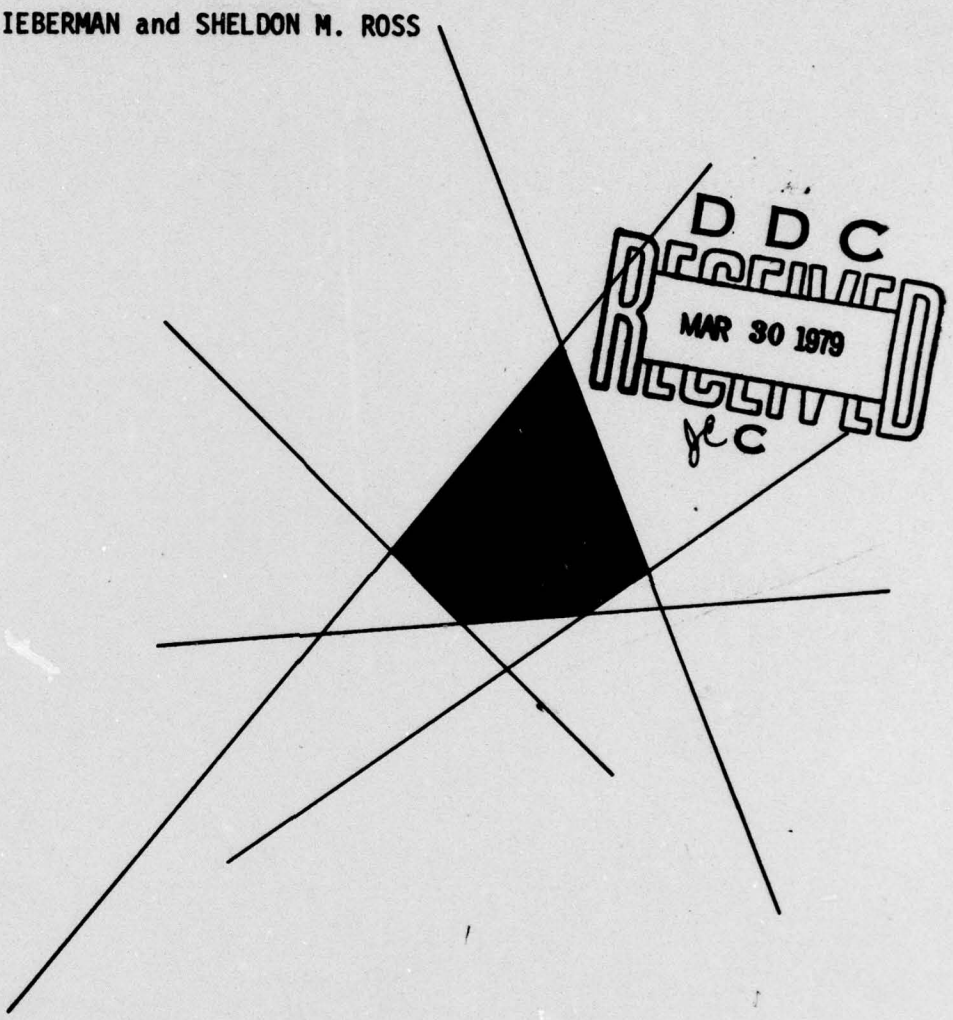
ON THE OPTIMAL ASSIGNMENT OF SERVERS AND REPAIRMAN

by

CYRUS DERMAN, GERALD J. LIEBERMAN and SHELDON M. ROSS

AD A066584

DDC FILE COPY



DDC
RECEIVED
MAR 30 1979
DDC

OPERATIONS RESEARCH CENTER

This document has been approved for public release and sale; its distribution is unlimited.

UNIVERSITY OF CALIFORNIA · BERKELEY

79 03 27 117

(12)

ON THE OPTIMAL ASSIGNMENT OF SERVERS AND REPAIRMAN

by

Cyrus Derman
Department of Civil Engineering
and Engineering Mechanics
Columbia University
New York, New York

and

Gerald J. Lieberman
Department of Operations Research
Stanford University
Stanford, California

and

Sheldon M. Ross
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

DDC
MAR 30 1979
C

NOVEMBER 1978 ORC 78-22

This research has been partially supported by the Office of Naval Research under Contract N00014-77-C-0299 and the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-77-3213 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 14 ORC-78-22	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 6 ON THE OPTIMAL ASSIGNMENT OF SERVERS AND REPAIRMAN.		5. TYPE OF REPORT & PERIOD COVERED 9 Research Report.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) 10 Cyrus Derman, Gerald J. Lieberman ~~and~~ Sheldon M. Ross		8. CONTRACT OR GRANT NUMBER(s) 15 N00014-77-C-0299 AFOSR-77-3213
9. PERFORMING ORGANIZATION NAME AND ADDRESS Operations Research Center University of California Berkeley, California		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 16 2304/A5 17 A5
11. CONTROLLING OFFICE NAME AND ADDRESS United States Air Force Air Force Office of Scientific Research Bolling AFB, D.C. 20332		12. REPORT DATE 11 Nove~~mber 1~~978 13. NUMBER OF PAGES 12 12 13p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 61102F		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also supported by the Office of Naval Research under Contract N00014-77-C-0299.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Assignment
Queueing System
N Servers
Exponential Services
Repairman

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE ABSTRACT)

270 750 JOB

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

Consider an N server queuing system in which service times of server i are exponentially distributed random variables with rate λ_i. Customers arrive in accordance with some arbitrary arrival process. If a customer arrives when all servers are busy, then he is lost to the system; otherwise, he is assigned to one of the free servers according to some policy. Once a customer is assigned to a server he remains in that status until service is completed. We show that the policy that always assigns an arrival to that free server whose service rate is largest (smallest) stochastically minimizes (maximizes) the number in the system. The result is then used to show that in an N component system in which the ith component's up-time is exponential with rate λ_i and in which the repair times are exponential with rate μ, the policy of always repairing the failed components whose failure rate λ is smallest stochastically maximizes the number of working components.

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. SPECIAL

ON THE OPTIMAL ASSIGNMENT OF SERVERS AND REPAIRMAN

by

C. Derman, G. J. Lieberman and S. M. Ross

0. INTRODUCTION AND SUMMARY

Consider an N server queuing system in which service times of server i are exponentially distributed random variables with rate λ_i. Customers arrive in accordance with some arbitrary arrival process. If a customer arrives when all servers are busy, then he is lost to the system; otherwise, he is assigned to one of the free servers according to some policy. Once a customer is assigned to a server he remains in that status until service is completed. We shall show that the policy that always assigns an arrival to that free server whose service rate is largest (smallest) stochastically minimizes (maximizes) the number in the system. This result generalizes a result of Seth [1] who obtained a similar result in the case where $N = 2$ and the customers arrive in accordance with a Poisson process. The result is then used to show that in an N component system in which the ith component's up-time is exponential with rate λ_i and in which the repair times are exponential with rate μ, the policy of always repairing the failed components whose failure rate λ is smallest stochastically maximizes the number of working components.

1. THE QUEUEING SYSTEM

Consider an N server queueing system in which the service times of server i are exponentially distributed random variables with rate λ_i, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Customers arrive in accordance to some arbitrary deterministic point process. If a customer arrives when all servers are busy, then he is lost to the system; otherwise, he is assigned to one of the free servers, according to some policy. We show that the policy that always assigns an arrival to that free server which service rate is largest (smallest) stochastically minimizes (maximizes) the number in the system.

Theorem:

The number of customers in the system is stochastically minimized (maximized) by the policy which always assigns an arrival to the free server having the largest (smallest) service rate.

Proof:

Let T_n denote the time of the nth arrival, and for a given constant c let $L(c)$ denote the amount of time in $[0,T_n]$ in which the number of customers in the system is no greater than c. Now consider the problem of choosing a policy which maximizes, under any initial conditions as to which servers are initially busy, the probability that $L(c)$ is no greater than d, for some d. We shall show, by induction on n, that for any c and d, the optimal policy has the property that it always assigns an arrival to server 1 (the fastest server) when that server is free. This is obvious for $n = 1$. So, for a given c and d, consider the probability where we are only interested up to the time of the

(n + 1)st arrival. If server 1 is busy immediately after the time of the first arrival, then the result follows by the induction hypothesis. So consider the case where the first arrival is assigned to some server - say server i - thus leaving server 1 free. Call the best policy of this type policy A. Now by the induction hypothesis it follows that the next arrival will be assigned to server 1. Thus, policy A assigns the initial arrival to server i and the next arrival to server 1. We contrast this, with policy B, which assigns the first arrival to server 1 and the next arrival to server 1 if he is free and to server i otherwise, and then continues optimally. Let S_1 and S_i denote the service times of the initial customer if he were to be assigned either to server 1 or server i. (Thus, S_1 and S_i are respectively exponential with rates λ_1 and λ_i). Letting t denote the time between the first and second arrivals, we compare policies A and B in the following 3 cases.

Case 1:

$$t < \min (S_1, S_i) .$$

In this case, both policies are identical in that one of servers i and 1 is busy until time t at which they both become busy.

Case 2:

$$t > \max (S_1, S_i) .$$

In this case, under policy A, one of the two servers is busy until time S_i and then both servers are free until time t when server 1 becomes busy. Under policy B, it is the same except that one of the servers

69 03 27 117

is busy until time S_1 . As the conditional distribution of S_1 given that $t > \max(S_1,S_i)$ is stochastically smaller than that of S_i given that $t > \max(S_1,S_i)$, it follows that policy B is preferable in this case.

Case 3:

$$\min(S_1,S_i) < t < \max(S_1,S_i) .$$

The situation under policy A can be described in this case as either

(a) one of the two servers is busy for a time S_i , then both are free, and then server 1 is busy at time t , or

(b) one of the two servers is busy until time t at which time both become busy.

Situation (a) occurs when $S_i < S_1$ and situation (b) otherwise. Under policy B , the possible situations are

(a') one of the two servers is busy for a time S_1 , then both are free, then server 1 is busy at time t ,

(b') same as situation (b).

Situation (a') occurs when $S_1 < S_i$ and (b) otherwise. As situation (a') is better than situation (a) (since the conditional distribution of S_1 given that we are in Case 3 is stochastically smaller than that of S_i), and as both are obviously better than situation (b), it follows since it is conditionally (on the event that Case 3 is in effect) more likely that S_1 is smaller than S_i than the reverse, that policy B is better than policy A . Hence, in all cases, policy B is better than policy A and

so by the induction hypothesis it follows that an arrival should be assigned to server 1 when that server is free. Now by conditioning on the set of service times for server 1 we are left with the same type of problem except that only servers 2 through N are available for assignment (it is automatic that an arrival is assigned to server 1 if he is free). Hence, by the same reasoning, it follows that if server 1 is busy and server 2 free, an arrival should be assigned to server 2. Repetition of this argument completes the proof that always assigning to the fastest server stochastically minimizes the number in the system.

The result for maximizing the number in the system is proved similarly.

Remarks:

(i) As the same policy is optimal for every deterministic arrival process, it follows that it is optimal for any stochastic arrival process which is assumed independent of the service times. (To see this, just condition on the set of arrival times.)

(ii) It is not essential that the system considered was a loss system - the result holds true for any finite or infinite capacity system.

(iii) The proof also shows that the percentage of lost customers is minimized by the policy of always assigning an arrival to the fastest free server. This generalizes a result of Seth [1] presented for $N = 2$ servers and Poisson arrivals. It is also interesting to note that Seth presented a counter-example to the (obvious modification of the above) when

the service distributions are assumed stochastically ordered but not necessarily exponential.

(iv) The foregoing server assignment model does not allow switching of servers when a server completes a service (i.e., no switching between arrivals). For example, if server 1 completes service at some instant, under the assumption of exponential service times, it would be reasonable to immediately reassign one of the customers already in service, but whose service is incomplete, to server 1.

Suppose such switching is allowable. The policy π which always assigns customer arrivals or reassigns customers to the available server with the largest service rate has the property that if, at any time, M customers are in the system, then the M fastest servers are the busy ones. Consequently, the customer departure rate is always largest under π.

It is thus clear (and not difficult to establish) that for this modified model, the assertion of Theorem 1 holds for π.

2. THE REPAIRMAN MODEL

Suppose we have a single repairman tending an N component system in which the ith component, when up, functions for an exponentially distributed length of time with rate λ_i where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. If a component fails when the repairman is idle, then repair is begun. The time to complete a repair (on any component) is exponentially distributed with rate μ. When more than one component is failed, a decision has to be made as to which one will be repaired. It is assumed to be allowable to reassign the repairman from one (failed) component to another. That is, for instance, if component i is being repaired and component j fails, then the repairman is allowed to switch from j to i (or any other failed component).

Theorem 2:

The length of time during which there are k or more components working in any interval of time $(0,t)$ is stochastically maximized by the strategy which always assigns the repairman to the failed component whose failure rate λ_i is the smallest.

Proof:

By imagining that the repairman is always working (even when all components are up) and by using the lack of memory property of the exponential distribution, it is easy to see that the above model is mathematically equal to one in which repair completion epochs occur in accordance with a Poisson process having rate μ, and at these epochs we are given the option of starting up any one of the components that are down at the time. (If no component is failed at a repair completion

epoch, then this opportunity is not used). Hence, the model is mathematically equivalent to the loss model of Section 1 under the assumption of Poisson arrivals in the sense that we interpret component i to be working whenever server i is busy, and we interpret the decision to put component i up at a repair completion epoch as the decision to assign an arrival to server i. The result now follows from Theorem 1.

Remarks:

(i) We have been informed by Smith that he has also obtained Theorem 2 by different methods.

(ii) Smith [2] was concerned with a series system in which repair times were exponential but were allowed to depend on the component under repair. He conjectured that the (asymptotic) probability that all components are functioning is maximized by the policy which always repairs the failed component having the smallest failure rate (no matter what the repair rates are). Thus, our results prove Smith's conjecture when the repair rates are the same. (Of course, we've proven optimality not only for a series but also for a k-of-N system.)

(iii) A remaining component-repair problem which has not been treated herein and which shall be treated in a future paper is the model that imposes the constraint that a repairman assigned to repair component i must complete his repair before being assigned to repair another failed component.