Therefore, if A and$B$ are the two least valuable items, then $b_A^* + b_B^* < M$ where b_A^* and b_B^* are the winning bids on items A and B. In this case any bidder can obtain greater profits by bidding $b_A^* + \epsilon$ on A, $b_B^* + \epsilon$ on B, and 0 on all other items. This contradicts the Nash assumption, so the theorem is proved. Q.E.D.[8]

[8] The author gratefully acknowledges helpful comments on earier drafts by Kim Border, Ronald Braeutigam, Robert Forsythe and Roger Noll.

References

1. BLACKETT, D. W., "Some Blotto Games," *Naval Res. Logist. Quart.*, Vol. 1, No. 1 (1954), pp. 55–60.
2. COOK, W. D., KIRBY, M. J. L. AND MEHNDIRATTA, S. L., "Game Theoretic Approach to a Two Firm Bidding Problem," *Naval Res. Logist. Quart.*, Vol. 22, No. 4 (1975), pp. 721–739.
3. ENGELBRECHT-WIGGANS, R., "Auctions and Bidding Models: A Survey," *Management Sci.*, Vol. 26, No. 2 (1980), pp. 119–142.
4. ——— AND WEBER, R. J., "An Example of a Multi-Object Auction Game," *Management Sci.*, Vol. 25, No. 12 (1979), pp. 1272–1277.
5. GRIESMER, J. H. AND SHUBIK, M., "Toward a Study of Bidding Processes, Part II: Games with Capacity Limitations," *Naval Res. Logist. Quart.*, Vol. 10, No. 2 (1963), pp. 151–173.
6. LUCE, R. D. AND RAIFFA, H., *Games and Decisions*, Wiley, New York, 1957.
7. ROTHKOPF, M. H., "Bidding in Simultaneous Auctions With a Constraint on Exposure," *Operations Res.*, Vol. 25, No. 4 (1977), pp. 626–629.
8. SAKAGUCHI, M., "Pure Strategy Solutions to Blotto Games in Closed Auction Bidding," *Naval Res. Logist. Quart.*, Vol. 9, Nos. 3–4 (1962), pp. 253–264.
9. STARK, R. M. AND MAYER, R. H., JR., "Some Multi-Contract Decision Theoretic Competitive Bidding Models," *Operations Res.*, Vol. 19, No. 2 (1971), pp. 469–483.
10. ——— AND ROTHKOPF, M. H., "Competitive Bidding: A Comprehensive Bibliography," *Operations Res.*, Vol. 27, No. 2 (1979), pp. 364–390.
11. VICKREY, W., "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J. Finance*, Vol. 16, No. 1 (March 1861), pp. 8–37.

Notes* I

MANAGEMENT SCIENCE
Vol. 26, No. 9, September 1980
Printed in U.S.A.

ON THE MARGINAL BENEFIT OF ADDING SERVERS TO $G/GI/m$ QUEUES†

RICHARD R. WEBER‡

The mean queueing time in a $G/GI/m$ queue is shown to be a nonincreasing and convex function of the number of servers, m. This means that the marginal decrease in mean queueing time brought about by the addition of two extra servers is always less than twice the decrease brought about by the addition of one extra server. As a consequence, a method of marginal analysis is optimal for allocating a number of servers amongst several service facilities so as to minimize the sum of the mean queueing times at the facilities.
(QUEUES; MULTISERVER; DESIGN OF QUEUES)

* All Notes are refereed.

† Accepted by Marcel F. Neuts; received December 1, 1979. This paper has been with the author 2 weeks for 1 revision.

‡ Cambridge University, England.

0025-1909/80/2609/0946$01.25

Copyright © 1980, The Institute of Management Sciences

1. Marginal Allocation

As in a paper by Rolfe [3], we consider a service system made up of N facilities. Each facility behaves like a $G/GI/m$ queue; it consists of a number of servers which operate in parallel and serve customers in the order of their arrival. The arrival and queueing processes at the N facilities are independent. Customers arriving at the ith facility have independent and identically distributed interarrival times with mean $1/\lambda_i$. Their service times are independent and identically distributed with mean $1/\mu_i$. A customer's "queueing time" is the time it has to wait before beginning service, and a facility's "mean queueing time" is defined as the expected steady-state (stationary) queueing time at the facility. We suppose that each facility has enough servers to ensure that the mean queueing time there is finite.

Additional servers are to be allocated amongst the facilities so as to minimize the sum of the mean queueing times at the N facilities. We show that the optimal allocation can be achieved by using an algorithm of "marginal allocation." This algorithm constructs the optimal allocations of $1, 2, \ldots, M$ additional servers one by one. The optimal allocation of M servers is obtained by optimally allocating $M - 1$ servers and then allocating the Mth server to that facility where its addition produces the greatest decrease in mean queueing time. Suppose that after optimally allocating $M - 1$ servers the number of servers at facility i is m_i $(i = 1, \ldots, N)$. Let $\overline{W}^i_{m_i}$ denote the mean queueing time at facility i. The method of marginal allocation is to allocate the Mth server to a facility j such that

$$\overline{W}^j_{m_j} - \overline{W}^j_{m_j+1} = \max_{1 \leqslant i \leqslant N} \left\{ \overline{W}^i_{m_i} - \overline{W}^i_{m_i+1} \right\}.$$

In some cases the quantities $\overline{W}^i_{m_i}$ can be calculated from an explicit formula; in other cases they can be estimated by simulation. If the marginal allocation algorithm were not optimal then it would be necessary to calculate or estimate as many as NM such quantities to find the optimal allocation. But because marginal allocation is optimal, only $N + M - 1$ such evaluations must be carried out.

It is easy to show (see Fox [2]) that the marginal allocation algorithm is optimal if the mean queueing time at each facility is a nonincreasing and convex function of the number of servers at the facility. Suppose we drop the reference to a particular facility. The statement that a facility's mean queueing time is a nonincreasing and convex function of the number of servers is the statement that for all m such that $\overline{W}_m$ is finite,

$$\overline{W}_m - \overline{W}_{m+1} \geqslant \overline{W}_{m+1} - \overline{W}_{m+2} \geqslant 0. \tag{1}$$

Rolfe proved (1) for a facility that behaves like a $M/D/m$ queue by direct calculation with an explicit formula for $\overline{W}_m$. He was unable to carry out a similar calculation for any other queue, but conjectured that (1) should hold for interarrival distributions other than exponential, and for generally distributed service times. Dyer and Proll [1] subsequently proved Rolfe's conjecture for a facility that behaves like a $M/M/m$ queue. The following theorem establishes Rolfe's conjecture for any arrival process and any independent, identically distributed service times. It thereby verifies the general applicability of marginal analysis to the allocation problem.

2. Convexity of the Mean Queueing Time

THEOREM. *The mean queueing time in a $G/GI/m$ queue is a nonincreasing and convex function of m.*

Proof. We shall prove the theorem in discrete time. It is merely a technicality to deduce the theorem in continuous time by approximating continuous distributions by appropriate discrete ones (Weber [5] gives an example of the sort of argument required). So without loss of generality we assume that the service and interarrival times take only integer values. Consider the first k customers to arrive. The theorem is proved by using induction on k to show that the total expected queueing time of these k customers is a nonincreasing and convex function of m. It follows that the k customers' average expected queueing time is also a nonincreasing and convex function of m. Provided that the queue is nonsaturating, the number of customers at the facility will return to zero infinitely often. The returns to zero are regeneration times, and standard results for regenerative processes (see Ross [4]) ensure that as k tends to infinity the limit of the k customers' average expected queueing time equals the mean (expected steady-state) queueing time.

As it happens, the proof does not depend on the nature of the arrival process. The customer interarrival times need be neither independent nor identically distributed. For convenience in notation and exposition we begin by assuming that the system initially contains exactly k customers, which are as yet unserviced, and that no additional customers arrive after the start. Later we shall see that we might as well have permitted some customers to arrive after the start.

Suppose that the service times of the k customers are given by the independent, identically distributed and integer-valued random variables $Y_1, \dots, Y_k$. As the system evolves in time we denote its state by $(X_1, \dots, X_m; n)$. The random variable X_i is the remaining service time of the customer currently assigned to the ith server, and n is the total number of customers in the system. Starting from this state, and conditional on $X_1, \dots, X_m$ taking the values $x_1, \dots, x_m$, we denote the expected value of the total remaining queueing time by $W_m(x_1, \dots, x_m; n)$.

Now let $x_1, \dots, x_m$ be any nonnegative integers (which we shall permit to be arbitrarily large, even if the Y_i's are bounded). We shall show that for all nonnegative integers, $x_1', x_2', x_1, \dots, x_m$, with $x_1' > x_1$, $x_2' > x_2$, and $m \geqslant 0$,

$$W_{m+2}(x_1', x_2, \dots, x_{m+2}; n) - W_{m+2}(x_1, x_2, \dots, x_{m+2}; n) \geqslant 0 \tag{2}$$

and

$$\begin{aligned}\{ W_{m+2}(x_1', x_2', \dots, x_{m+2}; n) - W_{m+2}(x_1, x_2', \dots, x_{m+2}; n)\} \\ - \{ W_{m+2}(x_1', x_2, \dots, x_{m+2}; n) - W_{m+2}(x_1, x_2, \dots, x_{m+2}; n)\} \geqslant 0.\end{aligned} \tag{3}$$

Suppose for the moment that (2) and (3) are true and $k \geqslant m + 2$. Let $Y_i = 0$ for $i > k$. From (2) we can deduce that

$$W_{m+2}(x_1', y_1, \dots, y_{m+1}; k+1) - W_{m+2}(x_1, Y_1, \dots, Y_{m+1}; k+1) \geqslant 0. \tag{4}$$

Note that if $k < m + 2$ then (4) is equal to 0, since there are then no customers waiting for sevice. We now let $x_1' \to \infty$ and $x_1 \to 0$ in (4) to get

$$\begin{aligned}\lim_{x_1' \to \infty} W_{m+2}(x_1', Y_1, \dots, Y_{m+1}; k+1) - \lim_{x_1 \to 0} W_{m+2}(x_1, Y_1, \dots, Y_{m+1}; k+1) \\ = W_{m+1}(Y_1, \dots, Y_{m+1}; k) - E_{Y_{m+2}}\left[W_{m+2}(Y_1, \dots, Y_{m+2}; k)\right].\end{aligned} \tag{5}$$

When the system has m servers we denote the expected total queueing time of the first k customers by $W_m(k)$. Taking an expectation over $Y_1, \dots, Y_{m+1}$ in (5) we conclude that

$$W_{m+1}(k) - W_{m+2}(k) \geqslant 0. \tag{6}$$

This shows that W_m is nonincreasing in m. Similarly we can put $(Y_1, \ldots, Y_m)$ in place of $(x_3, \ldots, x_{m+2})$ in (3), put $n = k + 2$, and then let $x_1', x_2' \to \infty$ and $x_1, x_2 \to 0$ to deduce that

$$\{W_m(k) - W_{m+1}(k)\} - \{W_{m+1}(k) - W_{m+2}(k)\} \geqslant 0. \tag{7}$$

This shows that W_m is convex in m.

It is helpful to think of x_1, x_1', x_2 and x_2' as lengths of time for which servers 1 and 2 are prevented or "blocked" from serving customers in the queue, because of the customers they are currently serving. Rather than directly compute the differences in queueing times amongst queues with m, $m + 1$ or $m + 2$ servers, we have considered queues that have m servers for a certain length of time, and then $m + 1$ or $m + 2$ servers at later times, as queues 1 or 2 become unblocked. By letting the lengths of the blocked times tend to 0 or ∞, queues with different numbers of servers are compared.

It only remains to prove statements (2) and (3). The proof is by induction on n, and will be sufficiently illustrated by carrying it through for the case in which $m = 1$ (3 servers). In this case (2) and (3) are clearly true for $n \leqslant 4$, since there is at most one customer waiting in the queue. Assuming that (2) and (3) are true for $1, \ldots, n - 1$, we shall show that they are true for n, where $n > 4$. Consider a state $(x_1, x_2, x_3; n)$, and let $x = \min(x_1, x_2, x_3)$ be the time until the next customer in the queue begins service. Suppose that the service times of the next two customers to begin service are given by the random variables Y' and Y''. Notice that it is sufficient to prove (2) and (3) for $x_1' = x_1 + 1$ and $x_2' = x_2 + 1$. This is because for any $x_1' > x_1$ and $x_2' > x_2$ we can write (2) and (3) as (8) and (9) respectively.

$$\sum_{i=x_1}^{x_1'-1} \{W_3(i + 1, x_2, x_3; n) - W_3(i, x_2, x_3; n)\} \geqslant 0, \tag{8}$$

$$\begin{aligned}\sum_{j=x_2}^{x_2'-1} \Bigg[&\sum_{i=x_1}^{x_1'-1} \{W_3(i + 1, j + 1, x_3; n) - W_3(i, j + 1, x_3; n)\} \\ &- \sum_{i=x_1}^{x_1'-1} \{W_{m+3}(i + 1, j, x_3; n) - W_{m+3}(i, j, x_3; n)\} \Bigg] \geqslant 0.\end{aligned} \tag{9}$$

We therefore suppose that $x_1' = x_1 + 1$ and $x_2' = x_2 + 1$. The expectation operators that appear below are with respect to Y' and Y''. There are essentially three cases to consider, depending on whether x is the time until there is a customer service completion (i) in queue 3, (ii) in just one of queues 1 or 2, or (iii) in both queues 1 and 2 (but not in 3).

(i) $x = x_3$. In this case (2) and (3) become (10) and (11) respectively.

$$E[\{x_3 + W_3(x_1', x_2, x_3 + Y'; n - 1)\} - \{x_3 + W_3(x_1, x_2, x_3 + Y'; n - 1)\}]. \tag{10}$$

$$\begin{aligned}E\big[&[\{x_3 + W_3(x_1', x_2', x_3 + Y'; n - 1)\} - \{x_3 + W_3(x_1, x_2', x_3 + Y'; n - 1)\}] \\ &-[\{x_3 + W_3(x_1', x_2, x_3 + Y'; n - 1)\} - \{x_3 + W_3(x_1, x_2, x_3 + Y'; n - 1)\}]\big].\end{aligned} \tag{11}$$

By applying the inductive hypotheses that (2) and (3) are true for $n-1$, (10) and (11) are nonnegative, and the inductive step is established for this case.

(ii) $x = x_1 < x_2, x_3$, or $x = x_2 < x_1, x_3$. Suppose $x = x_1 < x_2$. In this case (2) and (3) become (12) and (13) respectively.

$$E\big[\{x_1' + W_3(x_1' + Y', x_2, x_3; n-1)\} - \{x_1 + W_3(x_1 + Y', x_2, x_3; n-1)\}\big]. \tag{12}$$

$$\begin{aligned} E\big[\big[&\{x_1' + W_3(x_1' + Y', x_2', x_3; n-1)\} - \{x_1 + W_3(x_1 + Y', x_2', x_3; n-1)\}\big] \\ &-\big[\{x_1' + W_3(x_1' + Y', x_2, x_3; n-1)\} - \{x_1 + W_3(x_1 + Y', x_2', x_3; n-1)\}\big]\big] \end{aligned} \tag{13}$$

By applying the inductive hypotheses that (2) and (3) are true for $n-1$, (12) and (13) are nonnegative, and the inductive step is established for this case.

(iii) $x = x_1 = x_2 < x_3$. In this case (2) and (3) become (14) and (15) respectively.

$$\begin{aligned} E\big[&\{x_2 + W_3(x_1', x_2 + Y', x_3; n-1)\} \\ &- \{x_2 + W_3(x_1, x_2 + Y', x_3; n-1)\}\big]. \end{aligned} \tag{14}$$

$$\begin{aligned} E\big[\big[&\{x_1' + x_2' + W_3(x_1' + Y', x_2' + Y'', x_3; n-2)\} \\ &- \{x_1 + x_2' + W_3(x_1 + Y', x_2' + Y'', x_3; n-2)\}\big] \\ &-\big[\{x_1' + x_2 + W_3(x_1' + Y'', x_2 + Y', x_3; n-2)\} \\ &- \{x_1 + x_2 + W_3(x_1 + Y', x_2 + Y'', x_3; n-2)\}\big]\big]. \end{aligned} \tag{15}$$

At this point we use the fact that Y' and Y'' are identically distributed. By interchanging their appearance in the third term of (15) and applying the inductive hypotheses that (2) and (3) are true for $n-1$ and $n-2$, (14) and (15) are nonnegative and the inductive step is established. This completes the proof of the theorem when there are no arrivals.

A system in which some customers arrive after the start can be treated in the following manner. The proof is really the same except that the notation for a state of the system becomes more complicated, and we have to check what happens when a customer arrives. Suppose that the interarrival times take only integer values, and recall that customers are served in the order of their arrival. We are interested in showing that the expected queueing time of the first k arrivals is nonincreasing and convex in m. So suppose that the state $(x_1, x_2, x_3; t, n, i)$ denotes the fact that at time t there are n of the first k arrivals yet to complete service and that i of these n customers are yet to arrive ($i \leqslant n \leqslant k$). Let $T(n', i')$ denote the statement that the result is true for all $i \leqslant n < n'$, and also for $n = n'$ with $i \leqslant i'$. To establish $T(n, i)$ it suffices to show that $T(n', i')$ implies $T(n', i' + 1)$ and $T(n', n')$ implies $T(n' + 1, 0)$. Let the integer z be a realization of the time until the next arrival and let $x = \min(x_1, x_2, x_3, z)$. The implications are proved by considering cases in which x is equal to x_1, x_2, x_3 or z, along the same lines as above.[1]

[1] The author would like to thank the referees for a number of comments that have helped to improve the clarity of this paper.

References

DYER, M. E. AND PROLL, L. G., "On the Validity of Marginal Analysis for Allocating Servers in $M/M/m$ queues," *Management Sci.*, Vol. 23, No. 9 (May 1977), pp. 1019–1022.

FOX, B. L., "Discrete Optimization Via Marginal Analysis," *Management Sci.*, Vol. 13, No. 3 (November 1966), pp. 210–216.

ROLFE, A. J., "A Note on Marginal Allocation in Multiple Service Systems," *Management Sci.*, Vol. 17, No. 9 (May 1971), pp. 656–658.

4. ROSS, S. M., *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, Calif., 1970.

WEBER, R. R., "On the Optimal Assignment of Customers to Parallel Servers," *J. Appl. Probability*, Vol. 15 (1978), pp. 406–413.

Notes*
II

MANAGEMENT SCIENCE
Vol. 26, No. 9, September 1980
Printed in U.S.A.

SIMPLE INEQUALITIES FOR MULTISERVER QUEUES†

MATTHEW J. SOBEL‡

Simple inequalities are obtained for some operating characteristics of multiserver queueing models. "Loss system" and "delay system" results are presented.
(QUEUES; MULTISERVER; INEQUALITIES)

1. Introduction

It is difficult to obtain explicit formulae for operating characteristics of queueing models with more than one server. When formulae can be obtained, often they are complicated and depend on particular probability distributions. The results below are nonparametric and simple in form. Simplicity typically implies that an inequality is not sharp (cf. Kingman [6]). However, one of the principal inequalities below is simple *and* sharp. See Brumelle [2], [3] and his references for other inequalities for multiserver queueing models.

In a "loss system" with a Poisson input process, i.e., an $M/G/c/N$ model (so arriving customers are refused entry when there are N customers already inside the facility), let ρ denote the "traffic intensity" and let B denote the long-run probability that the facility is full. A principal result below is

$$(1 - 1/\rho)^{+} \leqslant B \leqslant 1 - 1/(\rho + 1)$$

for all c and N. For $c \geqslant 2$ and $\rho \geqslant 1.5$, $1 - 1/\rho$ seems very close to B.

*All Notes are refereed.

†Accepted by Marcel F. Neuts; received October 29, 1979. This paper has been with the author 3 weeks for 1 revision.

‡Georgia Institute of Technology.

0025-1909/80/2609/0951$01.25

Copyright © 1980, The Institute of Management Sciences