Note: On the Efficiency of Imbalance in Multi-Facility Multi-Server Service Systems

Linda V. Green • Debashis Guha
Columbia University, 423 Uris Hall, New York, New York 10027

We consider the problem of simultaneously allocating servers and demands in a service system with independent multiple facilities. We assume a fixed number of facilities and total servers which must service a given Poisson arrival stream. We also assume that service times are identically distributed and independent of the server or facility. The allocation decision is one of simultaneously determining the number of servers and the fraction of the total arrival stream for each facility in order to optimize a given performance measure. Several performance measures are considered including minimizing expected system delay and equalizing delays across facilities. Our findings demonstrate that the overall system performance improves as the individual facilities become more unbalanced in the number of allocated servers. More formally, we show that if there is a server allocation that is maximal under the partial order of majorization, then it is optimal.
(*Queues: Multi-Facility; Multiserver; Queues: Optimization; Service System Design*)

Introduction

It is well known from both experience and theory (Smith and Whitt 1981) that the performance of a service system can be improved when separate facilities serving distinct arrival streams are combined to serve all the streams together. However, there are many situations in which it is undesirable or infeasible to combine facilities. One case arises from the need to locate service facilities reasonably near their respective arrival sources. Examples include consumer businesses such as banks, retail stores, and fast food restaurants where geographic proximity is an important dimension of customer service. A variation of this is in systems in which servers must travel to customers, such as in emergency systems and many repair operations, so that travel time is part of the service time and thus cannot get too large. The combination of facilities may also be restricted by physical size and/or manageability limitations. A critical characteristic in designing such multi-facility systems is that performance will depend on both the allocation of servers and the allocation of work to the respective facilities. A related situation arises when the service needs of the arrivals and hence the types of servers needed to meet them may differ. This occurs, for example, in manufacturing and repair systems where individual workers and/or machines are trained or designed to deal with specific and different demand types. Thus certain arrivals must be directed to certain servers and resource combination is infeasible. An important issue in such a situation is the degree of flexibility that should be designed into certain servers so that the total workload of the system can be handled optimally with respect to a given performance criterion. Again, this must be decided in conjunction with the allocation of workload among the different server types. See Guha (1990) for a discussion of this case.

In this paper we will deal with the first class of multi-facility service systems. That is, we will assume that the demand stream is homogeneous and has the same service time distribution independent of the server or facility. Thus workload can be allocated in a continuous fashion rather than by distinct groupings. We will also assume that the total numbers of facilities and servers are fixed, but each facility can vary in the number of assigned servers and its associated arrival population. In such systems, the allocation decision is one of

0025-1909/95/4101/0179$01.25

simultaneously determining the number of servers and the fraction of the total arrival stream for each facility in order to achieve a given overall performance objective. Once these determinations are made, we assume that facilities operate independently of one another. It is important to note that these allocations of both servers and customers are *static* and not determined by dynamic control rules.

The purpose of this paper is to provide evidence based on simple approximations and numerical results that the overall efficiency of a multi-facility system with Poisson arrivals increases as the individual facilities become more unbalanced, i.e. the number of servers assigned to each becomes more "uneven." So, for example, in the two facility case (where each facility must have at least one server) with a total of s servers, the optimal allocation of servers would be 1 and $s - 1$ assuming, of course, an associated optimal allocation of demand. More formally, we show that for any given number of facilities and a fixed total number of servers, if there is a server allocation that is maximal under the partial order of majorization, then it is optimal. We show that this is true for several performance measures including, perhaps counterintuitively, equalizing delays across facilities.

To our knowledge, no previous work has addressed the central issue of this paper—the simultaneous assignment of servers and work in a service system with independent multiple facilities. The server allocation problem and the work allocation problem for such systems have been studied separately. Allocating a fixed budget of servers to a given number of facilities, each with given demand, was first considered by Rolfe (1971) who showed that a marginal allocation procedure is optimal assuming Poisson arrivals and constant service times. Dyer and Proll (1977) extended this to the case of exponential service. Lee and Cohen (1985) considered the problem of dynamically allocating incoming demands of different classes to facilities, each with a given number of servers.

The problems of server and workload allocation in the context of queueing networks has been addressed in several papers including Stecke and Solberg (1985), Shanthikumar and Yao (1988) and Dallery and Stecke (1990) which considered closed queueing networks. Of these, the results in Stecke and Solberg are most similar to those reported in this paper. They include findings for some special cases that unbalanced configurations of servers are superior to balanced ones and unbalanced workloads are better than unbalanced ones. Hillier and So (1991) examined the issue of simultaneous allocation of servers and workload in the context of a production line. Their major finding, based on numerical results, parallels ours. That is, they found that when simultaneously optimizing the server and work allocations, the optimal server allocation is one in which every station receives just a single server except for one of the two end stations which receives all the other servers. They call this the "L phenomenon." Most recently, Calabrese (1992) examined workload allocation in an open Jackson network and found that server pooling—combining servers into fewer but larger groups—always improves performance. In all of these network models, the measure of performance used is throughput.

In §1 we describe the model formulation and define the concept of majorization. Section 2 first examines the case of two facilities for which we present numerical results to support our conjectures. We then prove that if the result holds for two facilities, it holds for any fixed number of facilities. We end this paper in §3 with our conclusions.

1. Assumptions and Definitions

Let M equal the total number of facilities and N the total number of servers, where $N > M$. We assume that service times are identically distributed and are independent of each other and the facility at which the service is performed. Let the service rate of each server be 1, which corresponds to measuring time in the scale of mean service times. We assume that the total arrival stream is Poisson with rate λ.

Our problem can then be expressed as finding a server allocation vector $(s_1, \ldots, s_M)$ and an associated arrival rate vector $(\lambda_1, \ldots, \lambda_M)$ so that some total system cost which may be

$$F = \max_i [f(\lambda_i, s_i)]$$

or

$$F = \sum_{i=1}^{M} f(\lambda_i, s_i)$$

is minimized and

$$\sum_{i=1}^{M} \lambda_i = \lambda$$

and

$$\sum_{i=1}^{N} s_i = N$$

where $f(\lambda_i, s_i)$ is some performance measure or cost for facility i which depends on the arrival rate and server allocation to i. This function f may be, for instance, the mean queue size, the expected delay, the probability of delay, or the demand-weighted probability of delay. To avoid trivial solutions (e.g. all servers and customers are allocated to a single facility) and to allow for real size constraints, we assume that the number of servers at each facility i has both an upper and lower bound, u_i and l_i with $l_i > 0$ and $u_i < N$ for all i. Again, it is important to note that each λ_i represents a fraction of the total arrival stream that is sent to facility i with fixed probability independent of the state of the system. Thus, each new demand is immediately routed to its predetermined facility and is served in a first-come, first-served discipline.

The major result we want to establish is that if there is a server allocation that is maximal under the partial order of majorization, then there is an associated arrival stream allocation such that system performance is optimized.

We now define majorization. For any $(x_1, \cdots x_M) \in R^M$, let $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[M]}$.

DEFINITION 1. For $x, y \in R^M$, $x \prec y$ if

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \quad \text{for } k = 1, \ldots, M-1$$

and

$$\sum_{i=1}^{M} x_{[i]} = \sum_{i=1}^{M} y_{[i]}.$$

When $x \prec y$, x is said to be *majorized* by y.

The definition of majorization, introduced by Hardy, Littlewood, and Polya (1952) is a formalization of the idea of unevenness of the components of a vector. We will also need the following definitions.

DEFINITION 2. A real-valued function f defined on a set $A \in \mathbf{R}^m$ is said to be *Shur-concave* on A if

$$x \prec y \quad \text{on} \quad A \Rightarrow f(x) \geq f(y).$$

DEFINITION 3. If $x = (x_1, \ldots, x_M)$ and $x' = (x'_1, \cdots x'_M)$ are vectors with integer arguments, then if $x_i > x_j$ for given i, j and $x'_i = x_i - 1$, $x'_j = x_j + 1$ and $x'_k = x_k$ for all $k \neq i, j$, then the vector x' is said to be obtained from x by a *transfer*.

For any two vectors x, y with integer arguments, if x is majorized by y, then x can be obtained from y by a finite number of transfers (Marshall and Olkin 1979). This leads to the following:

LEMMA 1. *A function f of integer arguments is Schur-concave if and only if for any x, $f(x') \geq f(x)$ where x' is obtained from x by a transfer.*

2. Conjectures and Results

We first examine the case of two facilities. We will consider the measures of expected delay and probability of delay. We will also consider two system objectives: minimizing the average system performance and minimizing the maximum delay. Note that under our assumptions, the latter objective is equivalent to equalizing delays at each facility and may be more appropriate when issues of equity are important.

We first state our two major conjectures.

CONJECTURE 1. For any $s_1 > s_2$ and λ_1, λ_2 such that $f(\lambda_1, s_1 - 1) = f(\lambda_2, s_2 + 1)$ with $\lambda_1 + \lambda_2 = \lambda < s_1 + s_2$ there is a λ'_1, λ'_2 with $\lambda'_1 + \lambda'_2 = \lambda$ such that

$$f(\lambda'_1, s_1) = f(\lambda'_2, s_2) < f(\lambda_1, s_1 - 1) = f(\lambda_2, s_2 + 1).$$

This conjecture then states that when the objective is equalizing probability of delay or expected delay, a more uneven server allocation will result in lower delays.

For the case of minimizing total system performance, we have the following conjecture.

CONJECTURE 2. For any $s_1 > s_2$ and $\lambda < s_1 + s_2$, for every λ_1, λ_2 with $\lambda_1 + \lambda_2 = \lambda$ there is a λ'_1, λ'_2 with $\lambda'_1 + \lambda'_2 = \lambda$ such that

$$f(\lambda'_1, s_1) + f(\lambda'_2, s_2) < f(\lambda_1, s_1 - 1) + f(\lambda_2, s_2 + 1).$$

This implies that

$$\min_{\lambda_1'} [f(\lambda_1', s_1) + f(\lambda_2', s_2)] < \min_{\lambda_1} [f(\lambda_1, s_1 - 1) + f(\lambda_2, s_2 + 1)].$$

The left hand side of this inequality is the minimum cost of serving the total demand λ at two facilities with s_1, s_2 servers and the right hand side is the minimum cost of serving the same demand with $s_1 - 1$ and $s_2 + 1$ servers. If we assume that f is the expected queue length, then this conjecture states that the average system expected queue length or, equivalently by Little's formula, the expected system delay will be reduced by a more uneven server allocation. Similarly, f may be the demand-weighted probability of delay.

Tables 1 and 2 present the results of numerical experiments which support these conjectures where f is the expected queue length or the weighted probability of delay. We consider two-facility systems with a total number of servers ranging from 5 to 14. We assume that service times are exponential. For each we solve for the demand allocations which optimize the system performance measure for all possible server allocations and a range of feasible total demands corresponding to various total traffic intensities (rho) λ/N. In each case, as the imbalance (i.e. the difference between servers in the two facilities) increases, the system performance improves.

Table 1 System Measures for Objective of Equalizing Delays (Imbalance = Difference in No. of Servers Between Facilities)

Total Number of Servers = 5

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
3	0.0018	0.0626	0.2312	0.4823	0.6346
1	0.0070	0.0937	0.2743	0.5260	0.8289
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
3	0.0003	0.0408	0.3923	2.0261	13.1887
1	0.0015	0.0785	0.5410	2.4329	14.8113

Total Number of Servers = 8

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
6	0.0000	0.0120	0.1119	0.3506	0.4745
4	0.0002	0.0244	0.1474	0.3961	0.5756
2	0.0006	0.0337	0.1674	0.4208	0.7799
0	0.0008	0.0370	0.1739	0.4287	0.7878
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
6	0.0000	0.0065	0.1638	1.3445	11.2306
4	0.0000	0.0177	0.2671	1.7349	13.0115
2	0.0001	0.0279	0.3279	1.9368	13.9044
0	0.0002	0.0317	0.3478	2.0004	14.1824

Table 1 *Continued*

Total Number of Servers = 11

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
9	0.0000	0.0023	0.0561	0.2649	0.3906
7	0.0000	0.0057	0.0794	0.3042	0.4805
5	0.0000	0.0097	0.0964	0.3305	0.7233
3	0.0000	0.0130	0.1074	0.3470	0.7412
1	0.0001	0.0147	0.1128	0.3549	0.7498
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
9	0.0000	0.0011	0.0740	0.9477	9.9176
7	0.0000	0.0036	0.1323	1.2647	11.5716
5	0.0000	0.0073	0.1786	1.4684	12.5820
3	0.0000	0.0107	0.2096	1.5935	13.1886
1	0.0000	0.0126	0.2251	1.6535	13.4763

Total Number of Servers = 14

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
12	0.0000	0.0004	0.0286	0.2048	0.1789
10	0.0000	0.0013	0.0431	0.2380	0.2059
8	0.0000	0.0025	0.0551	0.2621	0.2254
6	0.0000	0.0039	0.0644	0.2795	0.2396
4	0.0000	0.0052	0.0710	0.2914	0.2492
2	0.0000	0.0060	0.0749	0.2984	0.2549
0	0.0000	0.0062	0.0762	0.3007	0.2567
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
12	0.0000	0.0002	0.0352	0.6930	0.8252
10	0.0000	0.0007	0.0666	0.9453	1.1084
8	0.0000	0.0017	0.0964	1.1235	1.3066
6	0.0000	0.0030	0.1206	1.2506	1.4473
4	0.0000	0.0042	0.1382	1.3367	1.5423
2	0.0000	0.0051	0.1489	1.3867	1.5975
0	0.0000	0.0054	0.1524	1.4031	1.6156

These results were obtained by discretizing the total arrival rate λ by taking $\Delta = \lambda/10{,}000$ and using marginal allocation. For Table 1, each unit Δ was allocated to the facility with the smaller cost. For Table 2, this meant iteratively allocating each unit Δ to the facility where it would result in the smallest increase in cost. Since the performance measures we are considering are convex in λ, Lee and Cohen (1983) and Fox (1966)

have shown that the optimal solution can be found in this way.

Though these results were obtained assuming an exponential service time distribution, it is reasonable to expect that they will hold for more general service distributions since the widely known and used approximations for $M/G/c$ queues (see, e.g. Hokstad 1978) indicate that for a given service distribution, the expected queue size, etc. is a constant times the same measure for the comparable $M/M/c$ system.

Now we consider a service system with M independent facilities. First consider the case of minimizing the total system cost. Then the total cost incurred in the system is

$$F = \sum_{i=1}^{M} f(\lambda_i, s_i).$$

For a given feasible server allocation vector, $s = (s_1, \ldots, s_M)$, let $\Phi_M(s)$ be the cost of the system where the demand has been optimally allocated. That is,

$$\Phi_M(s) = \min_{(\lambda_1, \ldots \lambda_M)} \sum_{i=1}^{M} f(\lambda_i, s_i)$$

with $\lambda_i < s_i$ for all i. For the case $M = 2$ we write $\Phi_2(s)$ in the form $\Phi(s_1, s_2)$ with $s_1 \geq s_2$.

We wish to show that if the cost function f has the property that the total cost for a two facility system is

Table 2 System Measures for Objective of Minimizing Overall Delay (Imbalance = Difference in No. of Servers Between Facilities)

Total Number of Servers = 5

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
3	0.0018	0.0652	0.2438	0.5084	0.8290
1	0.0070	0.0952	0.2784	0.5322	0.8370
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
3	0.0003	0.0419	0.4172	2.2146	14.8292
1	0.0015	0.0799	0.5521	2.4841	15.2806

Total Number of Servers = 8

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
6	0.0000	0.0122	0.1192	0.3802	0.7713
4	0.0002	0.0255	0.1542	0.4132	0.7838
2	0.0006	0.0345	0.1705	0.4268	0.7890
0	0.0008	0.0376	0.1752	0.4308	0.7906
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
6	0.0000	0.0065	0.1704	1.4836	13.2858
4	0.0000	0.0184	0.2809	1.8452	14.0484
2	0.0001	0.0286	0.3354	1.9791	14.2995
0	0.0002	0.0322	0.3511	2.0167	14.3693

Table 2 *Continued*

Total Number of Servers = 11

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
9	0.0000	0.0023	0.0591	0.2905	0.7241
7	0.0000	0.0060	0.0846	0.3251	0.7393
5	0.0000	0.0101	0.1005	0.3427	0.7469
3	0.0000	0.0133	0.1095	0.3525	0.7511
1	0.0001	0.0149	0.1137	0.3569	0.7530
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
9	0.0000	0.0011	0.0755	1.0274	11.9878
7	0.0000	0.0037	0.1401	1.3720	12.9742
5	0.0000	0.0076	0.1867	1.5403	13.3631
3	0.0000	0.0110	0.2144	1.6282	13.5585
1	0.0000	0.0128	0.2273	1.6673	13.6431

Total Number of Servers = 14

		Overall Probability of Delay			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
12	0.0000	0.0004	0.0297	0.2248	0.6827
10	0.0000	0.0013	0.0461	0.2573	0.6989
8	0.0000	0.0027	0.0581	0.2759	0.7080
6	0.0000	0.0041	0.0664	0.2878	0.7138
4	0.0000	0.0052	0.0720	0.2952	0.7174
2	0.0000	0.0060	0.0752	0.2994	0.7194
0	0.0000	0.0062	0.0762	0.3007	0.7200
		Expected Total Queue Size			
Imbalance	RHO = 0.1	RHO = 0.3	RHO = 0.5	RHO = 0.7	RHO = 0.9
12	0.0000	0.0002	0.0355	0.7365	10.8396
10	0.0000	0.0007	0.0701	1.0272	11.9219
8	0.0000	0.0018	0.1013	1.1927	12.4099
6	0.0000	0.0031	0.1246	1.2945	12.6835
4	0.0000	0.0043	0.1403	1.3575	12.8493
2	0.0000	0.0051	0.1495	1.3920	12.9345
0	0.0000	0.0054	0.1524	1.4031	12.9613

minimized by the most uneven server allocation (in the sense of majorization) then the cost of an M facility system for any fixed $M > 2$ is also minimized by the most uneven server allocation.

PROPOSITION 1. *Assume that f is such that $\Phi(s_1, s_2) < \Phi(s_1 - 1, s_2 + 1)$ with $s_1 > s_2$. Then for any $M > 2$ and feasible server allocation vectors s' and s such that s' is majorized by s, we have $\Phi_M(s) < \Phi_M(s')$.*

PROOF. Assume that for any real valued function Φ defined on integer vectors, if for any integer vectors a, a' such that a' is obtained from a by a transfer, the inequality $\Phi(a) < \Phi(a')$ holds. Then for any integer vectors b, b' such that b' is majorized by b, the inequality $\Phi(b) < \Phi(b')$ must hold since b' can be obtained from b by a finite number of transfers and the inequality holds for each. Hence it suffices to assume that s' is obtained from s by a single transfer. Suppose that

$$\Phi_M(s) = \sum_{k=1}^{M} f(\lambda_k, s_k)$$

and

$$\Phi_M(s') = \sum_{k=1}^{M} f(\lambda'_k, s'_k)$$

with

$$\sum_{i=1}^{M} \lambda_i = \sum_{i=1}^{M} \lambda'_i = \lambda$$

where $s'_i = s_i - 1$, $s'_j = s_j + 1$ and $s'_k = s_k$ for $k \neq i, j$.

Consider the terms $f(\lambda_i, s_i) + f(\lambda_j, s_j)$ in the first sum and $f(\lambda'_i, s_i - 1) + f(\lambda'_j, s_j + 1)$ in the second sum. By assumption, for the system consisting of only facilities i and j, $\Phi_2(s_i, s_j) < \Phi_2(s_i - 1, s_j + 1)$ for any feasible total arrival rate. This implies that there exists a y such that

$$f(y, s_i) + f(\lambda'_i + \lambda'_j - y, s_j) < f(\lambda'_i, s'_i) + f(\lambda'_j, s'_j).$$

Consider the feasible demand allocation vector α defined as follows:

$$\alpha_k = \lambda'_k \quad \text{for} \quad k \neq i, j$$

$$\alpha_i = y, \quad \alpha_j = \lambda'_i + \lambda'_j - y.$$

Then $f(\alpha_k, s_k) = f(\lambda'_k, s'_k)$ for $k \neq i, j$ and $f(\alpha_i, s_i) + f(\alpha_j, s_j) < f(\lambda'_i, s'_i) + f(\lambda'_j, s'_j)$. Adding, we have

$$\sum_{k=1}^{M} f(\alpha_k, s_k) < \sum_{k=1}^{M} f(\lambda'_k, s'_k).$$

The right-hand side is $\Phi_M(s')$ and since $\Phi_M(s)$ is the cost of the optimal demand allocation for s we have

$$\Phi_M(s) < \sum_{k=1}^{M} f(\alpha_k, s_k)$$

so that $\Phi_M(s) < \Phi_M(s')$. From Lemma 1, we get

PROPOSITION 2. *Given the assumption in Proposition 1, then $\Phi_M(s)$ is Schur-concave and hence if there is a server allocation that is maximal under the partial order of majorization then it is optimal.*

Now consider the objective of equalizing delays. We have

PROPOSITION 3. *Given the objective Φ is to minimize the maximum delay, then under the assumption of Proposition 1, $\Phi_M(s)$ is Schur-concave.*

PROOF. Define

$$\Phi_M(s) = \min_{(\lambda_1, \ldots, \lambda_M)} \max_i [f(\lambda_i, s_i)],$$

and our proof is the same by replacing each sum by a maximum operation.

4. Conclusions

We have studied a service system with independent multiple facilities where the problem is to simultaneously allocate servers and demands in order to optimize a system performance measure. Under the assumption of homogeneous demands that can be continuously allocated, we have shown that if there is a server allocation that is maximal under the partial order of majorization, then it is optimal. This result is consistent with some previous observations and results for queueing networks and with the concept that queueing systems become more efficient as the number of servers increases (see, e.g., Whitt 1992). From an applications perspective, it implies that service territories should be designed to achieve unequal population sizes, particularly if geographic areas can be kept about the same.

We have assumed that service times are identically distributed and independent of the facility and the allocated demand at that facility. An important and interesting extension of this work would be to examine the problem of simultaneous allocation of servers and demands when service times increase as the fraction of the customer population assigned to a facility increases. This occurs, for example, when enlarging the arrival allocation to a facility means increasing travel times to customers as in many emergency and repair systems. For some discussion on this, see Guha (1990).

References

Calabrese, J. M., "Optimal Workload Allocation in Open Networks of Multiserver Queues," *Management Sci.*, 38 (1992), 1792–1802.

Dallery, Y. and K. E. Stecke, "On the Optimal Allocation of Servers and Workloads in Closed Queueing Networks," *Oper. Res.*, 38 (1990), 694–703.

Dyer, M. E. and L. G. Proll, "On the Validity of Marginal Analysis for Allocating Servers in $M/M/c$ Queues," *Management Sci.*, 24 (1977), 1019–1022.

Fox, B., "Discrete Optimization via Marginal Analysis," *Management Sci.*, 13 (1966), 210–216.

Guha, D., "On the Design of Multi-Facility Multiserver Service Systems," Ph.D. Dissertation, Columbia University New York, NY, 1990.

Hardy, G. H., J. E. Littlewood and G. Polya, *Inequalities* (2nd Ed.), Cambridge University Press, London and New York, 1952.

Hillier, F. S. and K. C. So, "On the Simultaneous Optimization of Server and Work Allocations in Production Line Systems with Variable Processing Times," presented at ORSA/TIMS meeting, Anaheim, CA, 1991.

Hokstad, P., "Approximations for the $M/G/m$ Queue," *Oper. Res.*, 26 (1978), 510–523.

Lee, H. L. and M. A. Cohen, "A Note on Marginal Allocation in Multiple Server Service Systems," *J. Appl. Prob.*, 20 (1983), 920–923.

—— and ——, "Multi-Agent Customer Allocation in a Stochastic Service System," *Management Sci.*, 21 (1985), 752–763.

Marshall, A. W. and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, NY, 1979.

Rolfe, A. J., "A Note on Marginal Allocation in Multiple Server Service Systems," *Management Sci.*, 17 (1971), 656–658.

Shanthikumar, J. G. and D. D. Yao, "Optimal Server Allocation in a System of Multi-Server Stations," *Management Sci.*, 33 (1987), 1173–1180.

Smith, D. R. and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," *Bell System Tech J.*, 60 (1981), 39–55.

Stecke, K. E. and J. J. Solberg, "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multiserver Queues," *Oper. Res.*, 33 (1985), 882–910.

Whitt, W., "Understanding the Efficiency of Multi-Server Service Systems," *Management Sci.*, 38 (1992), 708–723.

Accepted by Gabriel R. Bitran; received June 1992. *This paper has been with the authors* 8 *months for* 2 *revisions.*