THE CAFETERIA PROCESS—TANDEM QUEUES WITH 0-1 DEPENDENT SERVICE TIMES AND THE BOWL SHAPE PHENOMENON

RICHARD R. WEBER

University of Cambridge, Cambridge, England

GIDEON WEISS

Georgia Institute of Technology, Atlanta, Georgia

(Received December 1991; revision received May 1993; accepted June 1993)

Customers move through a series of M service stations. Each customer, independent of all others, requires service from only one of the stations, for a duration of 1 time unit, this being station i with probability p_i. The customer has zero service at all the other stations, but there is no overtaking between the customers, and so queueing occurs. In the case where there is unlimited waiting room between the servers, we show that the system is interchangeable—permuting the order of the stations has no effect on the distribution of the output stream. When there is no waiting room between the stations we investigate optimal loads of the servers in terms of optimal p_i's for up to 10 stations, and observe that optimal loads exhibit the *bowl phenomenon*. We also obtain some bounds on the throughput for equal loads as a function of M.

A queue of people is moving along a cafeteria service counter. Each person wishes to pick up only one item, from its location along the line, and does not need to spend time anywhere else, except when blocked by preceding customers. Where along the line should we place the most popular items to increase its throughput?

A line to manufacture electronic circuit boards includes a series of automatic insertion machines. Each machine has a magazine containing a selection of parts it will insert. A mixed variety of boards of different types are assembled by the line, requiring different selections of parts to be inserted. The line operates in a synchronized fashion—all the parts move along the line, with no overtaking, until a number of them are positioned under appropriate insertion machines, and movement of the line stops. The appropriate machines then perform the insertion operations, within a constant synchronized amount of time, at the end of which the line is set in motion again to reposition the boards. What is the throughput of such a line?

These two generic scenarios can be modeled by a series of queues in tandem, with the following unusual feature—the service requirements of the customers are deterministic (at least as a first approximation), and the random element is the location along the line, or the station in the series of queues, at which the service is given.

We introduce the following model, which we name the cafeteria process: Customers move through a series of M service stations. All the services are deterministic, with a duration of 1 time unit. Each customer requires service from only one of the stations. Let S_n denote the station at which customer n is served; we assume S_n are independent identically distributed, $n = 1, 2, \ldots$, with $P(S_n = i) = p_i$, $i = 1, \ldots, M$. We call p_i the station loads, and refer to $p_i = 1/M$ as the *equal load* case. Customers can move through stations where they are not served with a delay of 0; however, no overtaking is allowed, and so customer n enters service in station i or moves through station i only after customer $n - 1$ leaves that station. Hence, queueing and congestion occur.

We consider two versions of the system: The first version has infinite waiting room between the stations (but no jockeying for position in the waiting room), and we consider it with a general stream of customer arrivals. In the second version there is no waiting room between the stations, and there is an infinite supply of customers in front of the first station.

Subject classifications: Production/scheduling, flexible manufacturing/line balancing: bowl shape phenomenon. Queues, optimization: cafeteria process. Queues, tandem: interchangeability, 0-1 dependent service times.
Area of review: STOCHASTIC PROCESSES AND THEIR APPLICATIONS.

0030-364X/94/4204-0895 $01.25
© 1994 Operations Research Society of America

Apart from its suitability to model the above and similarly applied scenarios, the cafeteria model has several interesting features, and sheds some new light on related research in tandem queues and flow lines.

It is well known that infinite buffer flow lines in which jobs have either all deterministic or all exponentially distributed processing times have the remarkable property of *interchangeability*—for a general input process, the output process of the line is unchanged if the stations are permuted. The cafeteria model *with infinite buffers* between the stations provides, as we show in Section 2, an unexpected third example.

A little understood property of flow lines with manufacturing blocking is the *bowl phenomenon*, which has become part of the folklore of the optimal design of flow lines, yet is based principally on simulation studies. The cafeteria model in which there is *no buffer space* between the service stations provides perhaps the simplest nontrivial model of a flow line with manufacturing blocking. We present an elegant Markovian description for it. This Markovian formulation enables the analysis of fairly long lines, and a direct demonstration of the bowl phenomenon. This is the content of Sections 3–5.

Calculating the throughput of a flow line with finite buffer space is a difficult problem in general. Of particular interest are symmetric systems where one studies the throughput as a function of the number of stations. In Sections 6 and 7 we study the throughput of the cafeteria process with zero buffers and equal loads, as M becomes large. We obtain an upper bound of $\sqrt{(2/\pi)M}$ on the throughput and conjecture that it is asymptotically equal to $\sqrt{(1/2)M}$. Section 8 concludes with a literature survey and a discussion that puts our results in context.

1. PRELIMINARY REMARKS

1.1. The Bowl Phenomenon

A flow line is a production line that is arranged as a series of stations in tandem. The throughput of the line is governed by the speed of the machines, the amount of work performed by each machine, and the buffer space between the machines. It is not necessarily the case that a uniform line, in which all the stations are identical and perform similar amounts of work, achieves the highest possible throughput. The bowl phenomena occurs if the throughput can be increased either by allocating resources such as processing capacity or buffer space unequally, with more resources in the center of the line and less at the two ends, or alternatively, by dividing the work to be done on the products so that more work is done by the machines at both ends of the line and less in the center.

The bowl phenomena was first noted by Hillier and Boling (1966), and has been much studied since. Some intuition as to why it occurs can be obtained by considering the effects of starvation and blocking. A station is blocked if there is no room in the buffer immediately downstream for a job that has completed; a station is starved if the machine at the station is free, but there is no upstream job waiting to start service. Notice that the last station in a flow line is never blocked. Also, if, as commonly occurs, there is always a queue of jobs in front of the line available to begin service, then the first station is never starved. In contrast, stations in the middle of the line can experience both blocking and starvation. This argues that stations in the middle of the line need a greater share of resources than those at the ends, or, alternatively, should be assigned a lighter workload.

Despite many years of research a theoretical demonstration of the bowl phenomenon has not yet been provided. The aim of this paper is to present a simple model of a flow line, for which it is possible to conduct a theoretical analysis that goes further than for any models previously considered. In our model, each customer requires service at exactly one station, and no service at the other stations. The bowl phenomenon here means placing the less frequently requested stations toward the middle of the line, and our results confirm its advantage for up to 10 stations. If a proof of the optimality of bowl-shaped allocations is to be accomplished, then it is likely that this simple model is a good place to start.

1.2. Dependent Service Times

In much of the work on tandem queueing systems it is assumed that each customer's service times at successive stations are independent. Yet in two of the major application areas of tandem queues the service times at successive stations will, as a rule, not be independent: In communication networks, messages travel through a series of relay nodes—the transmission time at each node is related to the message length, and so the service times of a given message in the successive nodes are positively dependent. Such systems were analyzed by Kelly (1982, 1984).

In manufacturing flow lines, one can think of chopping the total amount of work required by a part into work for the various machines. Often this chopping will introduce additional variability into the process; thus, if a typical part requires an amount X of processing in total, it may be divided into $X_1, \ldots, X_M$ so

that $Var(X) < Var(X_1) + \cdots + Var(X_M)$, and hence the series of processing times may be negatively correlated. The cafeteria process offers an extreme example of the latter kind: While $X = 1$ with zero variance the processing time on machine i is 1 or 0 with probabilities p_i and $1 - p_i$, respectively, and has variance $p_i(1 - p_i)$.

1.3. A Note on Scheduling of the Cafeteria Process

In the definition of the cafeteria process it has been assumed that station S_n at which customer n is served is independent of the stations at which other customers require service. Suppose instead that there is an infinite supply of waiting customers and we are allowed to choose which one to serve next, or in other words, we are allowed to schedule or sequence the customers. We will require from our schedule that the long-run average fraction of customers scheduled for service at each station equals the station load. Then the marginal distribution of S_n will be the same as if the customers were taken in random order, but the service requirements of successive customers will no longer be independent. Kelly (1984) discusses a similar scheduling problem.

Consider the equal load case $p_i = 1/M$. It is easy to determine schedules that maximize or minimize the throughput. To maximize throughput one should order the customers cyclically, so that each cycle of M customers requires their services at stations $M, M - 1, \ldots, 2, 1$. For this order, at each time period all the stations will be loaded by a full cycle, all the machines will process all the customers in the cycle simultaneously, and all the M customers will depart after one time unit. The throughput of this schedule is $\lambda = M$, it achieves full utilization of the service stations, and hence is maximal.

We conjecture that the worst schedule that can be based on cycles of length M is to order the customers in each cycle so that they need service at stations $1, 2, \ldots, M - 1, M$. (This is to be read as a sequence in which a customer needing service at station M is followed by one needing service at station $M - 1$, etc.) Now each cycle takes M time units to complete, because when a customer is served in station k the succeeding customer will be waiting behind it in station $k - 1$, for service at station $k + 1$, and no other customers are served. Cycles, however, overlap, and it is easy to see that cycles start at intervals of $\lceil M/2 \rceil$, so the throughput is $\lambda = 2$ if M is even, and $\lambda = 2M/(M + 1)$ if M is odd. One can, however, get arbitrarily close to the absolutely smallest possible throughput of 1 by using longer cycles: $1, 1, 1, \ldots, 1, 2, \ldots, 2, \ldots, M - 1, M - 1, \ldots, M - 1, M, \ldots, M$, in which each service station is scheduled r times. For $r \geq M - 1$ this has a throughput of $\lambda = rM/(rM - M + 1)$.

Our results so far lead us to conjecture that the random, independent case that we study has a throughput of $\lambda = O(\sqrt{M})$.

2. INFINITE BUFFER SPACE AND INTERCHANGEABILITY

In this section, we imagine that the buffer space between every two stations is infinite. The result of this section is the following.

Proposition 1. *The distribution of the departure stream from the final station is invariant under any permutation of stations.*

This result is similar to a result that is known to hold when service times at station i are independent and exponentially distributed with rate μ_i, $i = 1, \ldots, M$, and customers require service from all stations. Burke (1956) proved that if there are Poisson arrivals to the first station, then the output process is a Poisson process of the same rate in the stationary regime. A stronger result says: Given that the system starts empty and an arbitrary arrival process, the distribution of the transient output process is the same for all orders of the servers. This result, the *interchangeability of /M/1 queues*, has been proved by Weber (1979). Lehtonen (1986), Tsoucas and Walrand (1987), and Anantharam (1987) provided alternative proofs. An extension to the case $M = 2$, with a finite buffer at station two has been given by Chao, Pinedo and Sigman (1989). Kijima, Makimoto and Shirakawa (1990) proved a more general interchangeability result.

Proposition 1 can be seen as a special case of an even more general interchangeability result of Weber (1992). However, that result is more than is needed for the cafeteria model, and an interesting proof of Proposition 1 can be constructed using the result for interchangeability of /M/1 queues.

Proof of Proposition 1. The proof is in three parts: First, the case of two stations and all customers present at time 0, second, two machines and a general arrival stream, third, extension to any number of machines. The proof of the first part is the most interesting; the others are based essentially on inductive bookkeeping.

Part i. (Two stations, no arrivals) Suppose that there are just two stations and let $p = p_1$, $q = p_2 = 1 - p$.

Server i is assigned to station i, so each customer has its service time of 1 at the first station with probability p, or at the second station with probability q. Let $D_{n,0}$ be the arrival time of customer n. Let $D_{n,i}$ denote its departure time from station i and $X_{n,i}$ its service time at station i. Consider first the special case $D_{n,0} = 0$ for all n. That is, assume all customers are at station 1 at the start. Then

$$D_{n,1} = D_{n-1,1} + X_{n,1} \tag{1}$$

$$D_{n,2} = \max\{D_{n,1}, D_{n-1,2}\} + X_{n,2}. \tag{2}$$

Define $N_{n,i} = n - D_{n,i}$. Using the fact that $X_{n,1} + X_{n,2} = 1$, simple algebra gives

$$N_{n,1} = N_{n-1,1} + X_{n,2} \tag{3}$$

$$N_{n,2} = \min\{N_{n-1,1}, N_{n-1,2} + X_{n,1}\}. \tag{4}$$

Interestingly, (3) and (4) have a useful interpretation. Consider two tandem queues with an infinite intermediate buffer and an infinite number of customers present at station 1 at the start. Suppose that service times at the first and second stations are distributed as independent, exponential random variables with parameters q and p, respectively. Let $Z(t)$ be the number of customers that at time t have completed service at station 1 but not at station 2. In a uniformization of this queueing system potential transitions of $Z(t)$ occur as a Poisson process of rate 1. At the time of a potential transition either $Z(t)$ decreases by 1 (corresponding to a potential service completion at station 2 when $Z(t) > 0$), or $Z(t)$ increases by 1 (corresponding to a potential service completion at station 1), or $Z(t)$ does not change (if there is a potential service completion at station 2 when $Z(t) = 0$). It is clear that $N_{n,i}$, as defined by (3) and (4), can be interpreted as the number of customers that have completed service at station i following the nth such potential transition. Now the interchangeability result for tandem /M/1 servers implies that the output process $\{N_{n,2}, n = 1, \dots\}$ is stochastically unchanged by an exchange of p and q. Since $D_{n,2} = n - N_{n,2}$ it follows that $\{D_{n,2}, n = 1, \dots\}$ is also stochastically unchanged by exchanging p and q. This proves Proposition 1 for this special case.

The argument above made use of a novel identity: The *departure time of the nth customer* in one queueing system was identified as the *number of departures by the nth observation time* in a second queueing system. We are not aware of any previous use of this sort of argument.

Part ii. (Two stations, general arrivals) Consider the case in which customers are not all present at the start. This requires that (1) be modified. The equations which define the $D_{n,i}$'s are now (2) and

$$D_{n,1} = \max\{D_{n,0}, D_{n-1,1}\} + X_{n,1}. \tag{5}$$

Except for $D_{1,0} = 0$, we allow the arrival process to be arbitrary, with $D_{n,0}$ perhaps depending on $D_{1,2}, D_{2,2}, \dots, D_{n-1,2}$. Take as an inductive hypothesis that regardless of the arrival times of the first $n - 1$ customers the joint distribution of $(D_{1,2}, D_{2,2}, \dots, D_{n-1,2})$ is symmetric in p and q. This is true for $n = 2$. Fix numbers $(d_{1,2}, d_{2,2}, \dots, d_{n-1,2})$. Consider the event $A_{i,j}$ that $(D_{i,2}, D_{i+1,2}, \dots, D_{j,2}) = (d_{i,2}, d_{i+1,2}, \dots, d_{j,2})$. Let B_l be the event that $D_{l,0} \geq D_{l-1,2}$ and $D_{j,0} < D_{j-1,2}$ for all $j < l$. That is, customer l is the first customer who arrives to find the system empty. Since B_l is a function of $D_{1,0}, D_{2,0}, \dots, D_{l,0}$ and $D_{1,2}, D_{2,2}, \dots, D_{l-1,2}$, the inductive hypothesis implies that $P(A_{1,l-1} \cap B_l)$ is symmetric in p and q for $2 \leq l \leq n$. It is clear that the inductive hypothesis also implies that $P(A_{l,n} | A_{1,l-1} \cap B_l)$ is symmetric in p and q because the distribution of $(D_{l,2}, D_{l+1,2}, \dots, D_{n,2})$ conditional on B_l is the same as one would obtain for the first $n - l + 1$ departure times in a problem where customers $1, \dots, n + l - 1$ arrive at times $D_{l,0}, \dots, D_{n,0}$.

Now let C be the event that none of $B_2, \dots, B_n$ occurs: i.e., $D_{j,0} < D_{j-1,2}$ for all $2 \leq j \leq n$. Consider a realization of $\{(X_{1,1}, X_{1,2}), \dots, (X_{n,1}, X_{n,2})\}$ such that C occurs. This means that each of the first n customers arrives before the previous customer departs station 2. We claim that for this realization the departure times would be unchanged if all customers were present at the start. Suppose a subsidiary inductive hypothesis that $(D_{1,2}, D_{2,2}, \dots, D_{j-1,2})$ are unchanged if the arrival times of all the first $j - 1$ customers are reduced to 0. Recalling that $D_{1,0} = 0$, this is true for $j = 2$. Suppose this is the case and consider the arrival of customer j at time $D_{j,0}$.

Since $D_{j,0} < D_{j-1,2}$ and $D_{j-1,2}$ is unchanged by making customers $1, \dots, j - 1$ arrive earlier, customer $j - 1$ is yet to complete service at station 2 when customer j arrives. There are two cases to consider. On the one hand, if customer $j - 1$ is yet to complete service at the first station, then it is clear that the whole journey of customer j is unaltered by setting his arrival to 0. On the other hand, if customer $j - 1$ has completed service at the first station, and is yet to complete service at the second station, then customer j will start service as soon as he arrives at station 1 and be available for service at station 2 no later than $D_{j,0} + 1$. However, by hypothesis C and the inductive hypothesis that $D_{j-1,2}$ is unchanged if the arrival times of customers $1, \dots, j - 1$ are

reduced to 0, we have $D_{j-1,2} \geq D_{j,0} + 1$. So customer j cannot start service at station 2 any earlier than $D_{j,0} + 1$, even if it is available for service earlier. Thus, there can be no decrease in the time at which customer j starts service at station 2 even if $D_{j,0}$ is also reduced to 0, along with $D_{1,0}, \ldots, D_{j-1,0}$. This concludes an inductive step of the subsidiary inductive hypothesis and we see that on realizations for which C occurs the departure times are the same as if all customers had been present at the start. Applying the result of part i, we see that $P(A_{1,n} \cap C)$ is symmetric in p and q. Proposition 1 for two stations and any arrival process, follows from

$$P(A_{1,n}) = P(A_{1,n} \cap C) + \sum_{l=2}^{n} P(A_{l,n} | A_{1,l-1} \cap B_l) P(A_{1,l-1} \cap B_l)$$

and the above arguments, which show that every term on the right-hand side is symmetric in p and q.

Part iii. (Any number of stations) The generalization to more than two stations is clear. Suppose that server i is assigned to station i. Consider stations i and $i + 1$. The arrivals to these stations are the output of stations $1, 2, \ldots, i - 1$. Each customer requires service at station i, or service at station $i + 1$, or service at neither station, with probabilities p_i, p_{i+1} and $1 - p_i - p_{i+1}$, respectively. By the result stated in the previous paragraph the output process from station $i + 1$ is statistically unchanged if the servers at stations i and $i + 1$ are swapped. The output process from station $i + 1$ feeds into the downstream stations. Making use of the theorem for these stations, we see that the output process from station m is unchanged by a swap of the servers at stations i and $i + 1$. This proves Proposition 1.

In fact, if there are just two stations, then the result holds even if the intermediate buffer, for customers in addition to the one being served, is of finite size $b \geq 0$. This means that customer n may not begin service at station 1 until customer $n - b - 2$ has departed station 2. Then the appropriate equations defining the $D_{n,i}$'s are (2) and

$$D_{n,1} = \max\{D_{n,0}, D_{n-1,1}, D_{n-b-2,2}\} + X_{n,1}. \quad (6)$$

But consider a new arrival process for which $\bar{D}_{n,0} = \max\{D_{n,0}, D_{n-b-2,2}\}$, a nondecreasing function of n. Then (6) becomes

$$D_{n,1} = \max\{\bar{D}_{n,0}, D_{n-1,1}\} + X_{n,1}. \quad (7)$$

Recall that in the proof above, we allowed the arrival time of customer n to be a function of $D_{1,2}, D_{2,2}, \ldots, D_{n-1,2}$. The result follows immediately. This argument for the finite buffer case is substantially simpler than that given by Chao, Pinedo and Sigman and is equally valid for tandem /M/1 queues with a finite intermediate buffer, as considered in their paper.

3. THREE OR FOUR MACHINES WITH NO BUFFER SPACE

Throughout the remainder of the paper we assume that there is no buffer space between stations and that there is an infinite supply of customers in front of the first station. In this section, we analyze cafeteria systems with three and four stations. We can obtain an explicit formula for the long-run throughput of the system, as a function of the station loads, and obtain the optimal loads.

We begin by considering the 3-station system. Number the stations as 1, 2, 3 with loads p_1, p_2, p_3, and let $\{S_n, n = 1, 2, \ldots\}$ be the sequence of stations at which customers $n = 1, 2, \ldots$ are served. Hence, the S_n's are i.i.d., taking the values 1, 2, 3 with probabilities p_1, p_2, p_3.

We will derive a formula for the throughput $\lambda(\underline{p})$ of the system as a function of $\underline{p} = (p_1, p_2, p_3)$. Note that any two successive customers either leave simultaneously, or with a difference of one time unit. Let $Q(\underline{p}) = P(\textit{interdeparture time is } 0)$ be the steady-state probability that the interdeparture time is 0, so that $1 - Q(\underline{p})$ is the steady-state probability that the interdeparture time is 1, and the steady-state throughput is $(\lambda(\underline{p}) = 1/(1 - Q(\underline{p}))$.

We study $Q(\underline{p})$ by considering an infinite sequence of customers $\ldots, -2, -1, 0, 1, 2, \ldots$, and looking at a fixed customer n. By considering customer n and some of its predecessors, $n - 1$, $n - 2$, etc., it is possible to determine if n departs simultaneously with $n - 1$, and so obtain $Q(\underline{p})$.

Note first that if customer n requires service at station i, and customer $n - 1$ requires service at station j, and $i < j$, then customer n will receive service at station i at the same time or earlier than the time at which customer $n - 1$ receives service at station j, and subsequently, the two customers will both be completed, will, henceforth, occupy adjacent stations, and leave the system together; this argument holds for any number of servers. For three servers this means that the interdeparture time is zero if (S_n, S_{n-1}) is (1, 2), (1, 3) or (2, 3). At the same time, we will have interdeparture time between $n - 1$ and n equal to 1 if $S_n = 3$, or if the pair (S_n, S_{n-1}) is (2, 2)

or (2, 1). There is one more possibility for an interdeparture time of 0, namely when $(S_n, S_{n-1}, S_{n-2}, S_{n-3}) = (1, 1, 3, 3)$. To see that the interdeparture time is indeed 0 in this case, note that when customer $n-3$ is served in station 3, customer $n-2$ is queueing behind him at station 2, and customer $n-1$ is served at station 1. At the next time unit, customer $n-3$ will have departed, and customer $n-2$ will be served at station 3. Simultaneously, customer $n-1$, who has been served, will be waiting at station 2, and customer n will be served at station 1. At the end of this time unit all three customers, $n, n-1, n-2$, will leave together. It is easy to see that we have covered all the possible cases, and so we have the following.

Proposition 2. *The probability of zero interdeparture time $Q(\underline{p})$ for a three-server system with zero buffer space is*:

$$Q(\underline{p}) = p_1p_2 + p_1p_3 + p_2p_3 + p_1^2p_3^2.$$

From this result we obtain:

Proposition 3. *The throughput of the 3-server zero buffer system is maximized by* $p_1 = p_3 = -1/2 + \sqrt{3}/2 = 0.3660$, $p_2 = 2 - \sqrt{3} = 0.2679$, *and has the value* $\lambda = 1.5339$.

Proof. Direct calculus shows that the optimal load has $p_1 = p_3 = \frac{1}{2}(1 - p_2)$, so the maximal throughput is obtained by maximizing $p_1^4 - 3p_1^2 + 2p_1$. Simple calculus shows this is maximized by $p_1 = -1/2 + \sqrt{3}/2$.

We now consider the throughput of a 4-station system, again with no buffers. As before, we list the complete set of possible sequences of service stations for customers $n, n-1, \ldots$ for which customers n and $n-1$ will depart the system together.

There is an infinite number of such sequences. The following example illustrates this fact: Consider the sequence 2 1 4 1 4 1 4 2 4 4, which are the stations at which customers $n, n-1, \ldots, n-9$, are served. At time t, customers $n-6, \ldots, n-9$ occupy the four stations, and customers $n-7, n-9$ are served in stations 2 and 4. At $t+1$ customers $n-5 \ldots n-8$ occupy the servers, and customers $n-5, n-8$ are served in stations 1 and 4. For each of the next two time periods, $t+2$ and $t+3$, two customers will be served, one each in stations 1 and 4. At period $t+4$ we finally have customers $n, n-1, n-2$ occupying machines 2 3 4; customers $n, n-2$ are served and customer $n-1$, who has had its service in period $t+3$, is queueing between them. At the end of period $t+4$ all three customers $n, n-1, n-2$ will leave together. It is clear from this example that one can obtain other sequences in which customers $n, n-1$ leave together by adding an arbitrary additional number of pairs of customers that are served at stations 1 4 into the middle of the sequence 2 1 4 1 4 1 4 2 4 4. We will use the notation 2 $\overline{1\,4}$ 2 4 4 to denote all the sequences having a run of one or more than 14 pairs in place of $\overline{14}$.

Proposition 4. *The probability of zero interdeparture time $Q(\underline{p})$ for a 4-server system with zero buffer space is*:

$$\begin{aligned} Q(\underline{p}) = {} & p_1p_2 + p_3p_4 + (p_1p_4)^2(p_2 + p_3 + p_1p_4) \\ & + \frac{1}{1 - p_1p_4}(p_1p_4 + p_2p_3 + p_1p_3 \\ & + p_2p_4 + (p_1p_3)^2 + (p_2p_4)^2 \\ & + (p_1p_4)^2(p_1p_3 + p_2p_4 + p_2p_3)). \end{aligned}$$

Proof. By looking at all the possible sequences of customers, lexicographically, we obtain the complete list of cases of the simultaneous departure of customers $n, n-1$:

3 4	1 4	1 1 3 3
2 4	1 3	1 1 3 2 4 4
2 3	1 2	1 1 3 $\overline{1\,4}$ 4
2 2 4 4	1 $\overline{1\,4}$ 4	1 1 3 $\overline{1\,4}$ 3
2 $\overline{1\,4}$ 4	1 $\overline{1\,4}$ 3	1 1 3 $\overline{1\,4}$ 2 4 4
2 $\overline{1\,4}$ 3	1 $\overline{1\,4}$ 2 4 4	1 1 2 4 4
2 $\overline{1\,4}$ 2 4 4	1 1 3 4 4	1 1 1 4 4 4

The proof is completed by adding the probabilities; for the groups with an infinite number of sequences, e.g., 2 $\overline{1\,4}$ 2 4 4 we sum a geometric series of probabilities.

The optimal load and throughput are given by:

Proposition 5. *The throughput of the 4-server zero buffer system is maximized by* $p_1 = p_4 = 0.3048$, $p_2 = p_3 = 0.1952$, *and has the value* $\lambda = 1.68939$.

Proof. Since it is straightforward but laborious calculus we give only a sketch of the proof. We used the Mathematica package to perform the algebraic steps, so the reader can verify them with this or any other algebraic manipulation package. Since the throughput is $1/(1 - Q(\underline{p}))$, we need to maximize $Q(\underline{p})$. We reparameterize p_1, p_2, p_3, p_4 via $2a = p_1 + p_4$, $2b = p_1 - p_4$, and $2c = p_3 - p_2$, where $0 \leq a \leq 1/2$, $-a \leq b \leq a$, $a - 1/2 \leq c \leq 1/2 - a$.

Writing $Q(\underline{p})$ in terms of a, b, c, note that it is quadratic in c, and c^2 has a negative coefficient. It is, therefore, immediate to obtain the value of the maximizing c as a function of a and b, and substitute it

into $Q(\underline{p})$ to obtain a function of a and b only, which is the value of $Q(\underline{p})$ maximized over c. The new expression is a rational function of b, which can be factored into a ratio of two linear terms divided by two linear terms, the whole multiplied by a 6th-order polynomial. It can now be shown that the whole expression is maximized over the range $-a \leq b \leq a$ by $b = 0$ for all $0 \leq a \leq 1/2$. Substituting into the maximizing c, one obtains likewise that $c = 0$. Thus, the maximal value of $Q(\underline{p})$ for a given a is obtained when $b = c = 0$, that is, when the loads are symmetric. Substituting $b = c = 0$ into $Q(\underline{p})$, one obtains an expression which is optimized over b and c, and is a rational function of a alone, of order 8 in the numerator, and 2 in the denominator. It is now straightforward to find the optimal value of a and of the throughput, as stated in the proposition.

4. A MARKOV CHAIN DESCRIPTION OF THE CAFETERIA PROCESS

While the approach of Section 3 gave us the throughput of a cafeteria of 3 and 4 stations, it is impractical to extend it to 5 or more stations. In this section, we present an alternative approach that is based on a Markov chain formulation. The Markov chain will have as its parameter not time, but customers; its states are defined in terms of a vector $W_n \in \{0, 1\}^{M-1}$ that describes the journey of customer n through the system. The components of W_n are 0s and 1s and they indicate in which of the stations $2, \ldots, M$ customer n spends 1 unit of time, and through which stations he passes with no delay (sojourn 0). We let

$$W_{n,i-1} = \begin{cases} 1 & \text{customer } n \text{ stays in station } i \\ 0 & \text{otherwise} \end{cases} \qquad i = 2, \ldots, M.$$

Note that W_n describes where customer n stops, without specifying at which of these stations he is actually served.

Given W_n, the journey of customer n through stations $2, \ldots, M$, we now consider the journey of customer $n + 1$ through stations $1, \ldots, M$. Assume first that customer $n + 1$ requires no service at all. Then if customer n stops at station i, customer $n + 1$ will stop simultaneously at station $i - 1$. We can therefore define

$$U'_{n+1,i} = \begin{cases} W_{n,i} & i = 1, \ldots, M-1 \\ 0 & i = M \end{cases}$$

to describe the journey of customer $n + 1$ through stations $1, \ldots, M$, if he requires no service.

Now suppose customer $n + 1$ requires service at station k. The journey U'_{n+1} must be modified. First we need to make $U_{n+1,k} = 1$. Next, let

$$l = \begin{cases} \min(i : i \geq k,\ W_{n,i} = 1) & \text{if such exists} \\ 0 & \text{otherwise} \end{cases}$$

and note that if such a nonzero l exists, then the stay and service of a customer $n + 1$ in station k is simultaneous with the stay of customer n in station $l + 1$, and therefore customer $n + 1$ will not actually stay in station l. Define e_i as the unit vector in the ith coordinate direction of $\mathcal{R}^M$ and take $e_0 = 0$. We can write $U_{n+1} = U'_{n+1} + e_k - e_l$, to describe the journey of customer $n + 1$ through stations $1, \ldots, M$. Finally, let

$$W_{n+1,i-1} = U_{n+1,i} \quad i = 2, \ldots, M.$$

Then W_{n+1} is fully determined by W_n, the route of the previous customer, and by k, the station at which customer $n + 1$ is served. Hence, W_n is Markovian and has M possible transitions from W_n to W_{n+1}, occurring with probabilities p_k, $k = 1, \ldots, M$. Now let

$$X_n = \left(\sum_{i=2}^{M} 2^{i-2} W_{n,i-1} \right) + 1.$$

Then $X_n - 1$ has W_n as its binary representation, and so X_n is an isomorphic chain to W_n with state-space $1, \ldots, 2^{M-1}$. The following MATLAB algorithms calculate the transitions and the probability transition matrix for the Markov chain X_n:

```
function j = transition(i, k)
  % set j to the value of X_n + 1 when X_n = i and
customer n + 1 is served in station k
  q = fix((i - 1)/2^(k - 1)); r = rem((i - 1)/2^
(k - 1));
    if q == 0 q1 = 0;
    else
      x = q; l = 0;
       while rem(x, 2) == 0  x = x/2; l = l + 1;
end
      q1 = (x - 1)*2^l;
    end
      j1 = r + (q1 + 1)*2^(k - 1);
  j = fix(j1/2) + 1;

function A = mat(m, p)
  construct the transition matrix for an m station caf-
eteria with loads p
  n = 2^(m - 1); A = zeros(n);
    for i = 1:n
      for k = 1:m
        j = transition(i, k); A(i, j) = A(i, j) +
p(k);
      end
    end.
```

As an illustration here are the first few transition matrices:

$$A_2 = \begin{pmatrix} p_1 & p_2 \\ p_1 & p_2 \end{pmatrix} \quad A_3 = \begin{pmatrix} p_1 & p_2 & p_3 & 0 \\ p_1 & p_2 & p_3 & 0 \\ p_1 & p_2 & 0 & p_3 \\ 0 & p_1+p_2 & 0 & p_3 \end{pmatrix}$$

$$A_4 = \begin{bmatrix} p_1 & p_2 & p_3 & 0 & p_4 & 0 & 0 & 0 \\ p_1 & p_2 & p_3 & 0 & p_4 & 0 & 0 & 0 \\ p_1 & p_2 & 0 & p_3 & 0 & p_4 & 0 & 0 \\ 0 & p_1+p_2 & 0 & p_3 & 0 & p_4 & 0 & 0 \\ p_1 & p_2 & p_3 & 0 & 0 & 0 & p_4 & 0 \\ 0 & p_2 & p_1+p_3 & 0 & 0 & 0 & p_4 & 0 \\ 0 & 0 & p_1 & p_2+p_3 & 0 & 0 & 0 & p_4 \\ 0 & 0 & 0 & p_1+p_2+p_3 & 0 & 0 & 0 & p_4 \end{bmatrix}$$

In principle we can use these transition matrices to completely analyze the M station cafeteria model. However, the state space grows rapidly, being of size 2^{M-1}. We have not been able to derive any general properties of these Markov chains; instead, we have carried out numerical calculations for up to $M = 10$ stations by using the MATLAB package on an IBM PC and a MacIntosh IIci. The calculations involved matrices as large as 512×512.

Given the transition matrix A for the M station cafeteria, one can solve for π, the 2^{M-1} steady-state probability vector of X_n. It is clear that X_n is irreducible and aperiodic: If $M-1$ successive customers all require their service at station 1, then the state will become $X_{n+M-1} = 0$, and after one additional such customer the state will be $X_{n+M} = 0$. So state 0 can be reached from every other state, and it has period 1. To see that every state can be reached from state 0, consider any $1 \leq X_n \leq 2^{M-1}$, and let W_n be the corresponding vector of 0s and 1s. Let the number of 1s in W_n be r, and the location of the last 1 be l. Then W_n is reached with probability p_{l+1} from $X_{n-1} = 2(X_n - 1 - 2^{l-1}) + 1$, and the corresponding W_{n-1} has only $r-1$ 1s. Hence, by induction on r, X_n can be reached from 0.

Define the sojourn probability O_k, $k = 1, \ldots, M$ as the steady-state probability that customer n stays at station k. For $k = 2, \ldots, M$ one obtains O_k by summing π_i over all i for which W_n has 1 in position $k-1$, that is

$$O_k = \sum_{i \in \sigma_k} \pi_i, \quad \sigma_k = \{i : \text{rem}((i-1)/2^{k-2}, 2) > 0\}.$$

For $k = 1$, $O_1 = p_1 + (1-p_1)O_2$ because customer $n+1$ will stay in station 1 if it is served in station 1, or if customer n stays in station 2.

We now turn to calculating λ, the steady-state throughput of the M station cafeteria. One way to obtain λ is to apply Little's formula, $L = \lambda W$, to station 1. In this case $L = 1$, because station 1 is always occupied by a customer. Also, the sojourn of each customer is either 0 or 1, and it is 1 with probability O_1; so $W = O_1$. Hence, $\lambda = 1/O_1$.

Applying Little's formula to station k, $k = 2, \ldots, M$, we can also obtain the fraction of time or steady-state probability OC_k that the station is occupied. Now the expected number of customers in the station is $L = OC_k$, while $W = O_k$, hence $OC_k = \lambda O_k = O_k/O_1$.

In addition, the average number of customers in the system is $\sum_{k=1}^{M} OC_k$, and the average sojourn time of a customer in the system is $\sum_{k=1}^{M} O_k$.

There is another way to calculate the throughput. Consider the steady-state probability that customer $n+1$ departs later than customer n, say $P_1 = P(\textit{interdeparture} = 1)$. Let L_n denote the last station at which customer n stops. The interdeparture time between customers n and $n+1$ is 1 if customer $n+1$ is served in station k, and $L_n \leq k$. So $P_1 = \sum_{k=1}^{M} p_k P(L_n \leq k)$, or equivalently, $P_1 = \sum_{l=1}^{M} P(L_n = l) \sum_{k=l}^{M} p_k$. If $X_n = i$, then $L_n = l$ if $2^{l-2} < i \leq 2^{l-1}$. Define $w_i = \sum_{k=l}^{M} p_k$, where $2^{l-2} < i \leq 2^{l-1}$, $i = 1, \ldots, 2^{M-1}$. Then in steady state: $1/\lambda = P_1 = \sum_{i=1}^{2^{M-1}} w_i \pi_i$.

5. COMPUTATIONAL RESULTS FOR UP TO TEN MACHINES

The Markov chain formulation of the M station cafeteria model in Section 4 provides a complete picture of the process and all the quantities of interest. Unfortunately, the state space grows rapidly as 2^{M-1}, and we have been unable to derive general results for the steady-state behavior of this chain. Numerical calculations can be used to explore the throughput for $M > 4$. This section summarizes our results for up to $M = 10$ stations.

Calculations were carried out for equal loads, $p_1\, p_2 = \cdots = p_M = 1/M$. Calculations were also performed for loads $p_1 = p_M = 2/(M+2)$, $p_2 = p_3 = \cdots = p_{M-1} = 1/(M+2)$, the '2-1 loads' case, discussed by Yamazaki, Sakasegawa and Shanthikumar (1992). Finally we searched numerically for loads that maximize the throughput, and calculated quantities related to these "optimal" loads. The results of our search for the optimal loads are:

- Optimal loads seem to be symmetric: $p_k = p_{M-k+1}$, $k = 1, \ldots, M$.

- Optimal loads seem to be "bowl-shaped": $p_1 \geq \cdots \geq p_{\lfloor M+1/2 \rfloor} = p_{\lceil M+1/2 \rceil} \leq \cdots \leq p_M$.
- The bowl is very flat, so that approximately $p_2 \approx p_3 \approx \cdots \approx p_{M-1}$.
- To a lesser extent, the optimal loads for the first and last machines are approximately double the optimal loads of the other machines: $p_1 = p_M \approx 2p_k$, $k = 2, \ldots, M-1$.

The increase in the throughput due to optimal, as against other, loadings is very slight. It amounts to 2% relative to the equal loads, and only 0.07% relative to the 2-1 loads.

The results of our numerical search for the optimal loads can only be regarded as conjectured optimal loads. This is because we searched for a local optimum, and we have no guarantee that it is a global optimum. We do, however, conjecture that the throughput is a concave function of the loads, in which case any local optimum is global; we also conjecture that the optimal loads are symmetric. One numerical feature of the local maximum which we found is that the throughput is a flat function of the loads in the neighborhood of the maximum. Thus, even when we approached the maximal value of λ to within 10^{-15}, the values of the loads could only be pinpointed to within 10^{-7}. All the results in this section are accurate in all the digits displayed.

Table I lists the optimal loads obtained through numerical search for 2–10 machines. The value for two machines is trivial. The optimal loads for 3 and 4 machines are proven optimum values. For 5–10 machines these are conjectured optimal loads. The symmetric solution was obtained from a search over all possible loads for up to 7 machines. For 8–10 machines we searched only over symmetric loads.

Table II shows the throughput for equal loads, 2-1 loads, and optimal loads. Also included are upper and lower bounds on the throughput, as derived in Section 6.

Table III lists some quantities that measure the "bowl phenomenon" in the cafeteria system. The first half of the table describes the actual shape of the bowl. The first column gives the "rim-to-center ratio" of the bowl, measured by the ratio of the load of station 1 (or station M) over the average of the loads of stations $2, \ldots, M-1$; the second column gives the "flatness of the inside of the bowl," measured by the standard deviation of the loads divided by their mean for the center stations $2, \ldots, M-1$. The second half of the table gives a comparison between the optimal throughput achieved by the "bowl," and the throughput of equal loads and of 2-1 loads.

6. AN UPPER AND A LOWER BOUND OF THE THROUGHPUT

It is clear that the throughput of a cafeteria system is an increasing function of the number of stations, because the amount of work per customer is constant, and adding stations increases the total service capacity. However, blocking also increases with the number of servers.

In this and the next section, we will assume that the stations are equally loaded with $p_1 = \cdots = p_M = 1/M$, and we study the behavior of the throughput λ as a function of the number of the stations M. As we saw in Section 5, the equal load case is quite close to the optimal.

In Section 4, we described the journey of a customer through the system by means of the Markov processes U_n, the binary M-vector indicating in the ith coordinate a stop at station i. Recall also L_n, the last station at which the customer stopped on his journey, and let $Y_n = M - L_n + 1$. From the equations at the end of Section 4 we have in the equal load case:

Table I
Optimal Loads for Cafeteria Stations

M	p_1	p_2	p_3	p_4	p_5
1	1				
2	0.5				
3	0.3660254	0.2679492			
4	0.3048205	0.1951795			
5	0.2678645	0.1554610	0.1533490		
6	0.2422009	0.1301452	0.1276539		
7	0.2229412	0.1124547	0.1099665	0.1092752	
8	0.2077598	0.0993096	0.0969503	0.0959803	
9	0.1953766	0.0891115	0.0869109	0.0858433	0.0855154
10	0.1850167	0.0809431	0.0788989	0.0778133	0.0773279

Table II
Throughput of Cafeteria Processes

M	Lower Bound	Equal Loads	2-1 Loads	Optimal Loads	Upper Bound
1	1	1	1	1	1
2	1.3333	1.3333	1.3333	1.3333	1.33333
3	1.5000	1.52830189	1.52811736	1.53392163	1.58824
4	1.6000	1.67498160	1.68554913	1.68938733	1.80282
5	1.6667	1.79792344	1.81988240	1.82149141	1.99171
6	1.7143	1.90643322	1.93822049	1.93854243	2.16240
7	1.7500	2.00496150	2.04475858	2.04478722	2.31931
8	1.7778	2.09601450	2.14221272	2.14276860	2.46532
9	1.8000	2.18117116	2.23244800	2.23415967	2.60242
10	1.8181	2.26150678	2.31680080	2.32013312	2.73208

Proposition 6. *The throughput of the M station stationary cafeteria process with equal loads is* $\lambda = M/E(Y_n)$.

It is instructive to think of U_n as representing a Markovian particle system on M locations: $U_{n,k} = 1$ if location k is occupied by a particle at time n. The transition of the particle system from U_n to U_{n+1} when $S_{n+1} = k$ (customer $n + 1$ is served in station k) is:

1. each particle moves one location to the left; a particle that was in position 1 is dropped, and location M is now unoccupied;
2. a "new" particle is added at location k (we may now have two particles in location k);
3. the first of the "original" particles that lies at or to the right of k is located, if such exists; this particle is dropped (we say it is "killed" by the new particle at k).

The transitions of the particle system can also be seen as a motion to the left: Each particle moves one location to the left, except for the particle that is "killed," which moves all the way to position k. In addition, a particle that occupies position 1 moves out on the left, while if k is $\geq$ the rightmost occupied position, a particle moves in from the right to position k.

Let $u = \max\{k : U_{n,k} = 1\}$ and $v = \max\{k : k < u, U_{n,k} = 1\}$ be the locations of the two rightmost particles at time n, and let $S_{n+1} = k$. If $k < v$ the leftmost particle at time $n + 1$ is unchanged, and its position becomes $u - 1$. If $k \geq v$ the new particle becomes leftmost, and its position at time $n + 1$ is k—if $v \leq k < u$ the particle at u is killed, if $u \leq k$ the previous leftmost particle at u now becomes the second leftmost particle, in position $u - 1$. In addition, the motion of the particle at v depends on the remaining particles of U_n.

Recall that $Y_n = M - u + 1$ and denote $Y_n = x$, $z = M - v + 1$. It is easy to see that $P(Y_{n+1} = x + 1) = (M - z + 1)/M$, and $P(Y_{n+1} = y) = 1/M$ for $y \leq z, y \neq x + 1$. Note that Y_n is not Markovian, but it is ergodic.

Upper and lower bounds on the throughput are obtained by defining two modified particle systems, in which the position of the last particle Y_n evolves as a Markov chain.

Table III
Describing the Bowl Phenomena for the Cafeteria System

	Bowl Shape Characteristics		Comparison of Throughput	
M	Rim-to-Center Ratio	Center Coefficient of Variation	Equal Loads % Suboptimal	2-1 Loads % Suboptimal
3	1.366		0.368	0.380
4	1.562		0.860	0.228
5	1.731	0.00788	1.311	0.0884
6	1.879	0.0112	1.684	0.0166
7	2.012	0.0137	1.986	0.0014
8	2.133	0.0157	2.231	0.0259
9	2.245	0.0174	2.429	0.0767
10	2.350	0.0204	2.527	0.1436

6.1. An Upper Bound on the Throughput

We begin with the upper bound. Denote the processes describing the modified system by U_n^U, Y_n^U. The transition of the modified system from n to $n + 1$ for $S_n = k$ follows the same rules as 1–3 above, for the unmodified system, except for step 3, *where if the particle that is to be killed is the rightmost, then it is not killed.*

Proposition 7. *Assume that* $U_0 = U_0^U$, *and assume that the processes* U_n *and* U_n^U *are coupled by the same sequence of newly arriving particles* S_n. *Then for all* $n \geq 0$, $U_{n,i}^U \geq U_{n,i}$, $i = 1, \ldots, M$.

Proof. The proposition trivially holds for 0, so we assume that it holds for n and prove it for $n + 1$. If $U_{n,i}^U \geq U_{n,i}$, then the inequalities clearly still hold after steps 1 (move to the left), and 2 (add particle in position $S_n = k$). Let $\tilde{U}_n$, $\tilde{U}_n^U$ denote the resulting vectors. In step 3, assume that a particle in location $l \geq k$ is killed in the modified process. Then $\tilde{U}_{n,l}^U = 1$, while for $k \leq i < l$, $\tilde{U}_{n,i}^U = 0$. Then by induction $\tilde{U}_{n,i} = 0$ for $k \leq i < l$, and so if $\tilde{U}_{n,l} = 1$, it will also be killed in the unmodified system. This is all that one needs to prove.

The conclusion from this is that if the unmodified cafeteria and the modified system start with $U_0 = U_0^U$ and have coupled inputs, then for all n, $Y_n^U \leq Y_n$. Hence, if $U_0 \cong U_0^U$, then for all n: $Y_n^U \leq_{ST} Y_n$. Hence, as $n \to \infty$, the same relation holds for the steady-state distributions, $Y_\infty^U \leq_{ST} Y_\infty$, and hence $\lambda = M/E(Y_\infty) \leq M/E(Y_\infty^U)$, which gives our upper bound.

We now analyze Y_n^U. Note that if the new particle is at or to the right of the original last particle, it becomes last, while if the new particle falls to the left of the original last particle, then in the modified system, because the original last particle cannot be killed, it will just move to the left by one location. Hence, for equal loads,

$$p_{kj} = P(Y_{n+1}^U = j | Y_n^U = k) = \begin{cases} 1/M & \text{if } j \leq k \\ (M-k)/M & \text{if } j = k+1. \end{cases}$$

Clearly this is a Markov chain, and the transition matrix for Y_n^U is

$$P = \begin{bmatrix} \frac{1}{M} & \frac{M-1}{M} & 0 & \cdots & 0 \\ \frac{1}{M} & \frac{1}{M} & \frac{M-2}{M} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{M} & & & & \frac{1}{M} \\ \frac{1}{M} & & & \cdots & \frac{1}{M} \end{bmatrix}$$

Proposition 8. *The steady-state probabilities for* Y_n^U *are*:

$$\alpha_k = \lim_{n\to\infty} P(Y_n^U = k) = \frac{k}{M^k} \frac{(M-1)!}{(M-k)!}.$$

Proof. We need to show that the α_k satisfy the steady-state equations and sum to one. It is easy to check by induction starting from M that for $1 \leq k \leq M$

$$\sum_{i=k}^{M} \alpha_i = \frac{(M-1)!}{(M-k)!} \frac{1}{M^{k-1}}.$$

In particular, putting $k = 1$, we get $\sum_{i=1}^{M} \alpha_i = 1$. The steady-state equation for α_k, $k = 2, \ldots, M$, is

$$\alpha_k = \frac{M-k+1}{M} \alpha_{k-1} + \frac{1}{M} \sum_{i=k}^{M} \alpha_i,$$

which is easy to verify. Finally, for $k = 1$ the steady-state equation reads

$$\alpha_1 = \frac{1}{M} \sum_{i=1}^{M} \alpha_i = \frac{1}{M},$$

as required.

From Proposition 8 we can now calculate the expected value of Y_n^U:

$$\begin{aligned} K_M = E(Y_n^U) &= \sum_{k=1}^{M} k\alpha_k = \sum_{k=1}^{M} \sum_{i=k}^{M} \alpha_i \\ &= 1 + \frac{M-1}{M} + \cdots + \frac{(M-1)\cdots(M-k)}{M^k} \\ &\quad + \cdots + \frac{(M-1)!}{M^{M-1}} \\ &= \sum_{k=0}^{M-1} (1 - 0/M)(1 - 1/M) \cdots (1 - k/M). \end{aligned}$$

Interestingly, K_M is also the solution to a problem that can be posed in terms of an urn of M balls. Suppose that balls are drawn at random from the urn one at a time, with replacement. It follows from the right-hand side above that K_M is the expected number of draws

until some ball is drawn for the second time. We now have an upper bound

$$\lambda \leq M/K_M \sim \sqrt{(2/\pi)M}.$$

The asymptotic form was discovered numerically and confirmed by some additional heuristic arguments. A direct justification is as follows. By the inequality that the geometric mean is less than the arithmetic mean,

$$(1 - 1/M)(1 - 2/M) \cdots (1 - k/M) \leq \left(1 - \frac{k+1}{2M}\right)^k$$

$$\leq e^{-k(k+1)/2M} \leq e^{-k^2/2M}.$$

Thus

$$K_M \leq 1 + \sum_{k=1}^{M-1} e^{-k^2/2M} \leq 1 + \int_0^\infty e^{-x^2/2M}\, dx$$

$$= 1 + \sqrt{(\pi/2)M}.$$

On the other hand,

$$\log(1 - j/M) = -j/M + O(j^2/M^2).$$

Hence

$$\prod_{j=0}^{k} (1 - j/M) = e^{-k(k+1)/2M + O(k^3/M^2)}.$$

Consider the lower bound obtained by summing (6.1) over $0 \leq k \leq M^\theta$ for $1/2 < \theta < 2/3$. Then

$$K_M/\sqrt{M} \geq \sum_{k=0}^{M^\theta} \frac{1}{\sqrt{M}} e^{-(k+1)^2/2M + O(M^{3\theta-2})}$$

$$\simeq \int_0^\infty \frac{1}{\sqrt{M}} e^{-x^2/2M}\, dx = \sqrt{\pi/2}.$$

6.2. A Lower Bound on the Throughput

To obtain a lower bound we define U_n^L, Y_n^L of a modified system whose evolution from time n to $n + 1$ for input $S_{n+1} = k$ follows the steps:

1. Each particle moves one location to the left.
2. A "new" particle is added in location k.
3. All the particles to the right of k are "killed."

The difference from the unmodified system is that more particles, all those to the right of the new particle and not just the leftmost one, are killed at every step.

Proposition 9. *Assume that $U_0 = U_0^L$, and assume that the processes U_n and U_n^L are coupled by the same sequence of newly arriving particles S_n. Then for all $n \geq 0$, $U_{n,i}^L \leq U_{n,i}$, $i = 1, \ldots, M$.*

Proof. This clearly holds for $n = 0$, so we assume that it holds for n and prove it for $n + 1$. By the induction hypothesis, $U_{n,i}^L \leq U_{n,i}$, $i = 1, \ldots, M$, and steps 1 and 2 certainly preserve this relationship. Let $\tilde{U}_n$, $\tilde{U}_n^L$ denote the resulting vectors. In step 3, assume that a particle in location $l \geq k$ is killed in the unmodified process. Then, if $\tilde{U}_{n,l}^L = 1$, the particle in location l of the modified process will also be killed. This is all that one needs to prove.

The conclusion is that if the unmodified cafeteria and the modified system start with $U_0 = U_0^L$, and have coupled inputs, then for all n, $Y_n^L \geq Y_n$. Hence, as in subsection 6.1, $M/E(Y_\infty^L)$, is a lower bound.

We now analyze Y_n^L. Note that in the modified system the new particle always becomes the rightmost particle. Hence, the analysis of Y_n^L is trivial, Y_n^L are simply independent identically distributed with $P(Y_n^L = i) = 1/M$, $i = 1, \ldots, M$. Hence, $E(Y_n^L) = (M + 1)/2$, and we have $\lambda \geq 2M/(M + 1) \sim 2$. Recall from subsection 1.3 that the value of 2 for the throughput is the lowest we could get for cyclic scheduling.

7. GRAPHIC DISPLAY, ASYMPTOTICS, AND SIMULATION RESULTS

In this section we present a graphic display of the realization of the cafeteria process. This display may be useful for gaining additional insight to the system. We then consider rescaling the system by rescaling both the stations and the customers by a factor of $\sqrt{M}$. We suggest that this rescaling may be useful in obtaining asymptotic properties of the cafeteria process as the number of stations grows. Simulation results support this scaling. We repeat our description of the journeys of successive customers given in Section 4. Consider, for example, an 11-station cafeteria, and a customer that stops at stations 1, 5, 6, 10. The stations at which he stops are indicated in the first line below:

```
1  0  0  0  1  1  0  0  0  1  0
   0  0  0  1  1  0  1  0  0  0  0
                     ↑  *  *
```

Here if the customer's journey starts at time t, he will be in station 1 at time t (that is, during the interval $(t, t + 1)$), in station 5 at time $t + 1$, etc., and leave the line after 4 time units, at time $t + 4$.

Consider the next customer, and assume first that he does not require any service. He will then enter the system when the first customer leaves station 1 at $t + 1$, and stop first in station 4, behind, and simultaneous with, the stay of the first customer at station 5. The

whole journey will include stops at stations 4, 5, 9 from time $t + 1$ to $t + 4$. The journey of the second customer is represented by the second line of the three above, where we have shifted the second customer's journey one position to the right—here 1s that are placed vertically above each other represent a simultaneous stop of the customers at two successive stations, for example, at time $t + 3$ (that is, during the interval $(t + 3, t + 4)$) the two customers occupy stations 10, 9.

Now assume that the second customer requires service at station 7. Then at time $t + 3$, instead of moving into station 9 the second customer will move into station 7, and stop there for service, while the first customer is in station 10. Note that during that time interval stations 8, 9 are unoccupied, as we have indicated by *s in the third row above.

Figure 1 shows the journey of 8 customers through a line of 11 stations (the customers discussed above are customers 3, 4 in the drawing and $t = 0$). Each horizontal bar represents the journey of a customer, the circles represent no stopping, the figures represent stops, filled figures are service periods, and hollow ones represent queueing behind (being blocked by) a previous customer. Successive customers are drawn vertically below each other, going down, so the vertical pointing down axis counts customers.

The 11 machines are represented by the top right to bottom left diagonals. For each machine one can read along the diagonal the customers that stopped at that machine, for example, machine 7 had stops of customers 1, 2, 4, of whom customer 4 received service, and the others were queueing (blocked). In addition, the diagonal line of machine 7 shows that at time 4 (during the time interval (4, 5)) it was unoccupied. Successive stations occupy successive diagonals from left to right, so the horizontal, left-to-right axis counts machines.

The heavy lines through the drawing represent time points. Time t (the interval $(t, t + 1)$ is represented by a heavy line that zigzags through the stations. The vertical portions represent stations that are occupied by successive customers (the first, the top one, is always being served, the others are blocked though some may be lucky enough to be served while being blocked). The horizontal portions of the line represent stations which are idle. The direction of the time axis is diagonal from top right to bottom left. The drawing keeps track of three different quantities—machines, customers and time—and is, therefore, almost a three-dimensional drawing.

To represent the evolution of the cafeteria process for many stations and customers, we use a slightly different drawing. Figure 2 represents a part of a simulated cafeteria process with 100 stations, which we ran for 500 customers. The drawing is obtained by rotating the previous drawing, as in Figure 1, through 45°, so that now each machine is represented by a

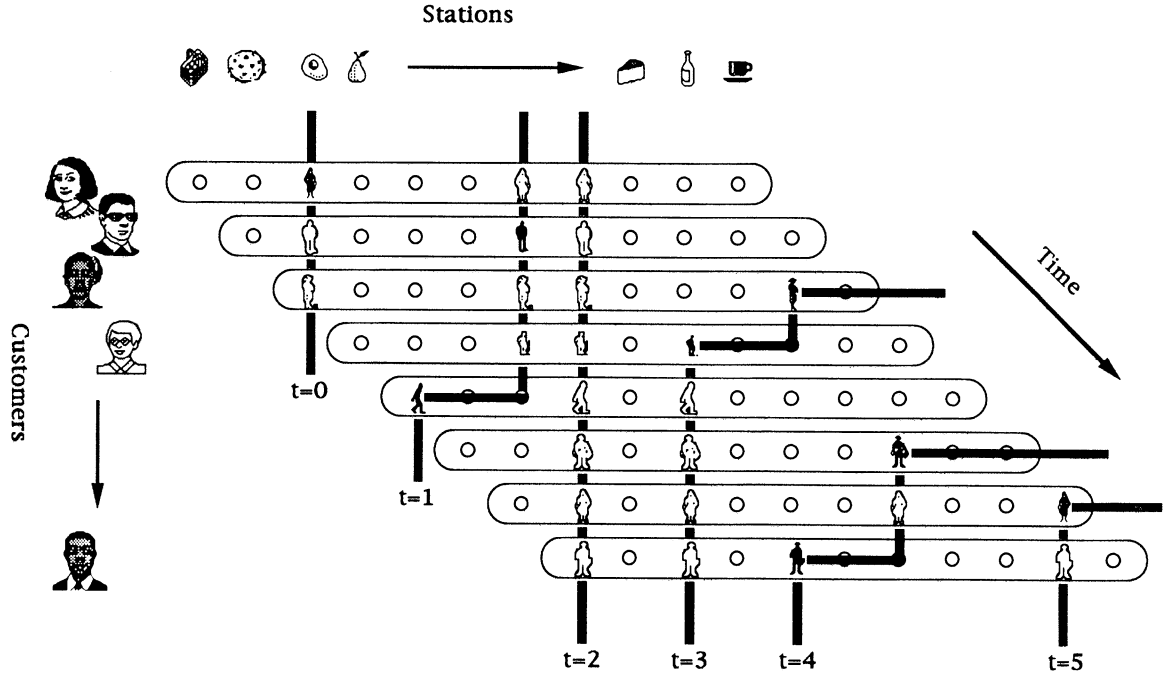

Figure 1. Eight customers moving through an 11-station cafeteria.

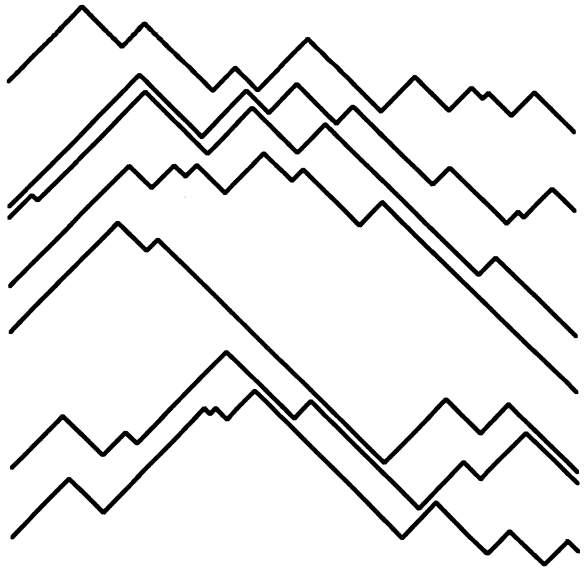

Figure 2. Simulation of a 100-station cafeteria over 7 time periods.

narrow vertical strip, and the machine axis is left to right. Each customer is represented by a narrow diagonal strip pointing from top left to bottom right, and the customer's axis runs vertically down. Each of the zigzag lines describes one time unit; the time axis also points vertically down. In each line, the diagonally ascending portions represent customers queueing at successive machines, and the descending portions represent unoccupied stations.

In the zigzag line of time t, a breakpoint from ascending to descending indicates service, because it represents the head of a line of customers queueing in successive stations, with a number of empty stations in front of them, and the customer at the head of the line, at the breakpoint, is being served. In Figure 2, the 7 successive time periods include 8, 7, 5, 6, 4, 6, 7 breakpoints, indicating at least that number of services (some services may occur without breakpoints, when a queueing customer is served).

Figure 2 suggests a rescaling and a possible asymptotic process that will describe the behavior of the cafeteria process for a large number of stations. Consider a lozenge-shaped quadrangle, of $\sqrt{M}$ successive machines and $\sqrt{M}$ successive customers. Fix a customer, who will require service from each machine with probability $1/M$ (exclusive, not independent, events), and the probability that he will require service from one of the $\sqrt{M}$ machines is therefore $1/\sqrt{M}$. For all of the $\sqrt{M}$ customers, the number of services on the $\sqrt{M}$ machines will be a binomial random variable, with $\sqrt{M}$ independent trials, $1/\sqrt{M}$ probability of success, and with expectation 1. For large M (and, hence, large $\sqrt{M}$), the number of services in the lozenge will converge to a Poisson random variable with rate 1. Furthermore, for large M the number of services in disjoint lozenges is nearly independent. Finally, as M becomes large the overwhelming majority of services will be at breakpoints of the zigzag lines. Rescaling the process of Figure 2 into units of $\sqrt{M}$ machines and $\sqrt{M}$ customers (time is not rescaled), and letting $M \to \infty$, the ascending-descending breakpoints of the line process will form a Poisson process of rate 1 in the plane.

We were unable to pursue the analysis of this process any further. However, we performed some simulation runs which confirm our conjecture that $\sqrt{M}$ is the correct scaling for this process. In these runs we simulated cafeteria processes with various numbers of stations, in the range $3 \leq M \leq 1{,}000$; starting from empty systems, we ran $15M$ customers for each value of M and discarded the initial 20%. Table IV summarizes the results, giving approximate 95% confidence limits for the throughputs. The values which were obtained are fitted extremely well by a linear function of $\sqrt{M}$. In the table we show the values of $\sqrt{2/\pi} + 1/2\sqrt{M}$ for comparison. We also include the upper and lower bounds of Section 6. Based on these results, we conjecture that the throughput of the cafeteria process is asymptotically $\sim 1/2\sqrt{M}$.

8. DISCUSSION

Four types of problems are discussed in the literature with regard to the optimal design of a flow line: For flow lines with infinite buffer space the objective is to minimize the average waiting time of a customer (the flow time). In the case of finite or zero buffer space the objective is to maximize the throughput (or minimize the time required to process a batch of customers—the makespan). In both cases, one is interested in the optimal order for given service stations, or one has a total amount of resources for the whole line, and is searching for the optimal allocation of these to the individual service stations.

In tandem queues with infinite buffer space the throughput of the line is essentially not affected by permuting the service stations: If the system is fed by a stationary input stream with an input rate lower than the service rate of all the stations, then the throughput will equal the input rate, each station will have finite queues in front of it, and the system will stabilize. If the input rate is higher than the service rates of some of the stations, then the throughput rate will equal the service rate of the slowest, or most heavily loaded,

Table IV
Estimated Throughput of Cafeteria Processes

M	Lower Bound	Simulation Estimates	Conjectured Asymptotics	Upper Bound
12	1.846	2.483 ± 0.336	2.530	2.973
15	1.875	2.727 ± 0.315	2.734	3.300
20	1.905	2.927 ± 0.462	3.034	3.778
25	1.923	3.333 ± 0.307	3.298	4.200
30	1.936	3.396 ± 0.530	3.536	4.581
40	1.951	4.000 ± 0.447	3.960	5.257
50	1.961	4.317 ± 0.457	4.333	5.853
75	1.974	5.085 ± 0.231	5.128	7.121
100	1.980	5.581 ± 0.310	5.798	8.190
200	1.990	7.818 ± 0.520	7.869	11.495
300	1.993	9.375 ± 0.550	9.458	14.031
500	1.996	11.858 ± 0.521	11.978	18.053
1,000	1.998	16.238 ± 0.656	16.609	25.443

station. Infinite queues will form in front of all the stations that are slower than all their predecessors, and all other stations will have finite queues; the system will be unstable. Even though permuting the stations has no effect on the throughput, different permutations may produce different waiting times for the jobs.

If the queues are interchangeable, as discussed in Section 2, then the total waiting time in the system is the same for all permutations (though the waiting times and queue sizes in front of the individual stations will be affected). The property of interchangeability remains a rare property, shared by only some special flow lines, including deterministic service times (when station i has a constant service time x_i which is the same for all customers (see Friedman 1965), independent exponential services, and the cafeteria process.

In the absence of interchangeability, one may wish to find the order of stations that will minimize the total expected waiting time of a customer for a stable system with a stationary input stream. This is a difficult question, which has received much attention in the literature. See, for example, Tembe and Wolff (1974), Pinedo (1982), Greenberg and Wolff (1988), as well as Whitt (1985) and Wein (1988).

If the problem is to design the optimal line by dividing a certain amount of resources, then for interchangeable lines, symmetry and convexity imply that equal allocation to all the stations is optimal, hence there is no "bowl shape." For general noninterchangeable lines that have infinite buffers between stations, it is a reasonable conjecture that the throughput is maximized by allocating equal service rates. It is possible that further gains in the average waiting times may be obtained from a bowl shape allocation of the variability.

Research on lines with zero or finite buffers presents many additional problems beyond the usual tandem queues. Early research in this area includes the pioneering papers of Avi-Itzhak and Yadin (1965) and of Avi-Itzhak (1965), as well as the more practice-oriented paper of Buzacott (1967). For further results, including discussions of dependent service times, see Boxma (1979), Wolff (1982), and Pinedo and Wolff (1982).

Work on optimal scheduling of customers through flow shops includes Foley and Suresh (1984, 1986), Wie and Pinedo (1986), and McCormick et al. (1989). When the buffer space between the stations is limited, and, in particular, in the case of zero buffer space, not only the waiting time but also the throughput may be affected by permuting the stations, and it makes sense to look for the order which maximizes the throughput. Interchangeability is even rarer in lines with finite buffers. In fact, the only known cases are those discussed in Section 2, namely deterministic, exponential, or cafeteria processes with two service stations and a finite buffer between them.

An important property of most flow lines, with zero, finite, or infinite buffers, which is much weaker than interchangeability, is the property of reversibility: If the order of the stations is reversed, the throughput remains unchanged. Reversibility of flow lines was proved by Muth (1979), Dattatreya (1978), and Yamazaki and Sakasegawa (1975). Yamakazi, Kawashima and Sakasegawa (1985) proved that if the system starts empty, then the distribution of the nth departure is the same if the line is reversed. Chao and Pinedo (1992) show that for three exponential stations

the output process (and not just the throughput) is the same if the order of the servers is reversed. They conjectured that this is true for any number of exponential stations in tandem.

The problem of obtaining the optimal order of stations among all possible permutations is hard, and has received much attention. It appears that a bowl shape permutation is optimal for 3 or 4 machines, but for more than 4 machines, a "saw tooth" solution may be optimal (Pinedo). Yamazaki phrased the following heuristic rule: "the optimal permutation is to keep the slow stations as far apart as possible." There are results to the effect that the first and last machines ought to be slower than the second and the one before last, in other words, the first and the last two machines should be arranged in a bowl shape (Ding and Greenberg 1991, Huang and Weiss 1991, Shanthikumar, Yamazaki and Sakasegawa 1991).

Finally, the problem that we address principally in this paper is that of designing the best flow line, by allocating resources to the stations to maximize throughput. Attempts to learn more about the bowl phenomenon for this problem have been extensive. Makino (1964), extending Hunt's (1956) early work for the same problem as Hillier and Boling, obtained improvements by assigning lower mean service time to the middle station. Patterson (1964) suggested alternating the mean service times between high and low along the line, while Davis (1966) conjectured that a low-medium-high pattern might be best. El-Rayad (1979a) conducted statistical tests for exponential, lognormal and normal distributions and found that only bowl-shaped designs performed consistently better than balanced lines. Muth (1984) showed that the optimality of the bowl shape was due to unbalancing the mean service times, not the variances (which might have been another interpretation of Hillier and Boling's results for exponential distributions). Hillier and Boling (1979) extended their early work to longer lines and service times drawn from Erlang distributions. In all cases, the optimal allocations were bowl shaped. They examined the effect of increasing buffer space and the shape parameter of the Erlang distribution, finding that curvature of the optimal bowl became less as the buffer size increased. They concluded that the optimal allocations are robust, in the sense that the throughput function has a flat maximum. El-Rayad (1979b), Chow (1987) and Kijima, Makimoto and Shirakawa (1989) discuss the allocation of buffers along the line.

Shanthikumar and Yao (1991) defined a property of a collection of parametrized random variables that they call *strong stochastic convexity* and discussed applications to tandem queues. For example, in a model with exponentially distributed service times, finite buffers and an infinite number of customers in front of the first station, the expected departure time of the nth customer is a convex decreasing function of the service rates. This fact, combined with the fact that the reversed line has the same throughput, implies that the allocations of service rates should be symmetric about the middle. Similarly, for the same model, Meester and Shanthikumar (1990) established that the throughput is an increasing and concave function of the buffer sizes. It follows that an optimal buffer allocation should be nearly symmetric about the middle to within ± 1 that is due to the discrete nature of buffer sizes. All these results point to, but still leave open, the challenge to say something more precise about the optimality of bowl-shaped allocations.

In Sections 5, 6, 7, we analyze the throughput of the equal load zero buffer cafeteria process, as a function of M the number of stations. The throughput of balanced flow lines with blocking is another topic which is far from being well understood. Massey (1991a, b) analyzes a flow line of M exponential rate 1 servers with communications blocking. Other references are Kelly (1982, 1984), Srinivasan (1992), and Glynn and Whitt (1991). Note the throughput of our cafeteria process could be increased from $\leq O(\sqrt{M})$ to $O(M)$ by using a scheduling buffer of size $O(M)$ in front of the system, as mentioned in subsection 1.3. The idea of using buffers for rescheduling, and increasing the throughput appears in Kelly (1984).

ACKNOWLEDGMENT

This research was supported in part by NSF grants DDM-8914863 and DDM-9215233.

REFERENCES

ANANTHARAM, V. 1987. Probabilistic Proof of Interchangeability of $/M/1$ Queues in Series. *Queueing Syst. Theory and Applic.* **2,** 387–392.

AVI-ITZHAK, B. 1965. A Sequence of Service Stations With Arbitrary Input and Regular Service Times. *Mgmt. Sci.* **11,** 565–571.

AVI-ITZHAK, B., AND M. YADIN. 1965. A Sequence of Two Servers With No Intermediate Queue. *Mgmt. Sci.* **11,** 553–564.

BOXMA, O. 1979. On the Random Queueing Model With Identical Service Times at Both Counters, Parts I and II. *Adv. Appl. Prob.* **11,** 616–659.

BURKE, P. J. 1956. The Output of a Queueing System. *Opns. Res.* **4,** 699–704.

BUZACOTT, J. A. 1967. Automatic Transfer Lines With Buffer Stocks. *Int. J. Prod. Res.* **5,** 183–200.

CHAO, X., AND M. PINEDO. 1992. On Reversibility of Tandem Queues With Blocking. *Naval Res. Logist.* **39,** 957–974.

CHAO, X., M. PINEDO AND K. SIGMAN. 1989. On the Interchangeability and Stochastic Ordering of Exponential Queues in Tandem With Blocking. *Prob. Engin. and Info. Sci.* **3,** 223–236.

CHOW, W. 1987. Buffer Capacity Analysis for Sequential Production Lines With Variable Processing Times. *Int. J. Prod. Res.* **25,** 1183–1196.

DATTATREYA, E. S. 1978. Tandem Queueing Systems With Blocking. Ph.D. Dissertation, University of California, Berkeley.

DING, J., AND B. S. GREENBERG. 1991. Bowl Shapes are Better With Buffers—Sometimes. *Prob. Engin. and Infor. Sci.* **5,** 159–169.

EL-RAYAD, T. E. 1979a. The Efficiency of Balanced and Unbalanced Production Lines. *Int. J. Prod. Res.* **17,** 61–75a.

EL-RAYAD, T. E. 1979b. The Effect of Inequality of Interstage Buffer Capacity and Operation Time Variability on the Efficiency of Production Systems. *Int. J. Prod. Res.* **17,** 77–89.

FOLEY, R. D., AND S. SURESH. 1984. Stochastically Minimizing the Makespan in Flowshops. *Naval Res. Logist. Quart.* **31,** 551–557.

FOLEY, R. D., AND S. SURESH. 1986. Scheduling n Nonoverlapping Jobs and Two Stochastic Jobs in a Flow Shop. *Naval Res. Logist. Quart.* **33,** 123–128.

FRIEDMAN, H. D. 1965. Reduction Methods for Tandem Queueing Systems. *Opns. Res.* **13,** 121–131.

GLYNN, P. W., AND W. WHITT. 1991. Departures From Many Queues in Series. *Ann. Appl. Prob.* (to appear).

GREENBERG, B. S., AND R. W. WOLFF. 1988. Optimal Order of Servers for Tandem Queues in Light Traffic. *Mgmt. Sci.* **34,** 500–508.

HILLIER, F. S., AND R. M. BOLING. 1966. The Effect of Some Design Factors on the Efficiency of Production Lines With Variable Operation Times. *J. Indus. Eng.* **17,** 651–658.

HILLIER, F. S., AND R. M. BOLING. 1979. On the Optimal Allocation of Work in Symmetric Balanced Production Line Systems With Variable Operation Times. *Mgmt. Sci.* **25,** 721–728.

HUANG, C. C., AND G. WEISS. 1990. On the Optimal Order of M Machines in Tandem. *O. R. Letts.* **9,** 299–303.

HUNT, G. C. 1957. Sequential Arrays of Waiting Lines. *Opns. Res.* **4,** 674–683.

KELLY, F. P. 1982. The Throughput of a Series of Buffers. *Adv. Appl. Prob.* **14,** 663–653.

KELLY, F. P. 1984. Blocking, Reordering, and the Throughput of a Series of Servers. *Stoch. Proc. Applic.* **17,** 327–336.

KIJIMA, M., N. MAKIMOTO AND H. SHIRAKAWA. 1990. Stochastic Minimization of the Makespan in Flow Shops With Identical Machines and Buffers of Arbitrary Size. *Opns. Res.* **38,** 924–928.

LEHTONEN, T. 1986. On the Ordering of Tandem Queues With Exponential Servers. *J. Appl. Prob.* **23,** 115–129.

MCCORMICK, S. T., M. L. PINEDO, S. SHENKER AND B. WOLF. 1989. Sequencing in an Assembly Line With Blocking to Minimize Cycle Time. *Opns. Res.* **37,** 925–935.

MAKINO, T. 1964. On the Mean Passage Time Concerning Some Queueing Problems of the Tandem Type. *J. Opns. Res. Soc. Japan* **7,** 17.

MASSEY, W. A. 1991a. Balanced Queues in Series With Communication Blocking. *Math. O. R.* (to appear).

MASSEY, W. A. 1991b. Continuous Node Limits for Series Networks With Blocking (to appear).

MEESTER, L. E., AND J. G. SHANTHIKUMAR. 1990. Concavity of the Throughput of Tandem Queueing Systems With Finite Buffer Storage Space. *Adv. Appl. Prob.* **22,** 764–767.

MUTH, E. J. 1979. The Reversibility Property of Production Lines. *Mgmt. Sci.* **25,** 152–158.

PINEDO, M. 1982. On the Optimal Order of Stations in Tandem Queues. In *Applied Probability—Computer Science: The Interface*, R. Disney and T. Ott (eds.). Birkhauser, Boston, Mass., 307–326.

PINEDO, M., AND R. W. WOLFF. 1982. A Comparison Between Independent and Dependent Service Times in Tandem Queues. *Opns. Res.* **30,** 464–479.

SHANTHIKUMAR, J. G., AND D. D. YAO. 1991. Strong Stochastic Convexity: Closure Properties and Applications. *J. Appl. Prob.* **28,** 131–145.

SHANTHIKUMAR, J. G., G. YAMAZAKI AND H. SAKASEGAWA. 1991. Characterization of Optimal Order of Servers in a Tandem Queue With Blocking. *O. R. Letts.* **10,** 17–22.

SRINIVASAN, R. Queues in Series Via Interacting Particle Systems. *Math. O. R.* (to appear).

TEMBE, S. V., AND R. W. WOLFF. 1974. The Optimal Order of Service in Tandem Queues. *Opns. Res.* **30,** 148–162.

TSOUCAS, P., AND J. WALRAND. 1987. On the Interchangeability and Stochastic Ordering of $/M/1$ Queues in Tandem. *Adv. Appl. Prob.* **16,** 515–520.

WEBER, R. R. 1979. The Interchangeability of $\cdot/M/1$ Queues in Series. *J. Appl. Prob.* **16,** 690–695.

WEBER, R. R. 1992. The Interchangeability of Tandem Queues With Heterogeneous Customers and Dependent Service Times. *J. Appl. Prob.* (to appear).

WEIN, L. M. 1988. Ordering Tandem Queues in Heavy Traffic. Technical Report, Sloan School of Management, MIT, Cambridge, Mass.

WHITT, W. 1985. The Best Order for Queues in Series. *Mgmt. Sci.* **31,** 475–487.

WIE, S. H., AND M. L. PINEDO. 1986. On the Minimization of Expected Makespan and Flowtime in Stochastic Flowshops With Blocking. *Math. O. R.* **11,** 336–342.

WOLFF, R. W. 1982. Tandem Queues With Dependent Service Times in Light Traffic. *Opns. Res.* **30,** 619–635.

YAMAKAZI, G., AND H. SAKASEGAWA. 1975. Properties of Duality in Queueing Systems. *Ann. Instit. Statist. Math.* **27,** 201–212.

YAMAKAZI, G., H. SAKASEGAWA AND J. G. SHANTHIKUMAR. 1992. On Optimal Arrangement of Stations in a Tandem Queue System With Blocking. *Mgmt. Sci.* **38,** 137.

YAMAKAZI, G., T. KAWASHIMA AND H. SAKASEGAWA. 1985. Reversibility of Tandem Blocking Queueing Systems. *Mgmt. Sci.* **31,** 78–83.