J. Appl. Prob. **16**, 690–695 (1979)
Printed in Israel

THE INTERCHANGEABILITY OF $\cdot/M/1$ QUEUES IN SERIES

RICHARD R. WEBER,* *University of Cambridge*

Abstract

A series of queues consists of a number of $\cdot/M/1$ queues arranged in a series order. Each queue has an infinite waiting room and a single exponential server. The rates of the servers may differ. Initially the system is empty. Customers enter the first queue according to an arbitrary stochastic input process and then pass through the queues in order: a customer leaving the first queue immediately enters the second queue, and so on. We are concerned with the stochastic output process of customer departures from the final queue. We show that the queues are interchangeable, in the sense that the output process has the same distribution for all series arrangements of the queues. The 'output theorem' for the $M/M/1$ queue is a corollary of this result.

OUTPUT PROCESSES; DEPARTURE PROCESSES; SERIES OF QUEUES; TANDEM QUEUES

1. Introduction

Series of queues are interesting to study because they form one of the simplest of queueing systems and are a fundamental element in many more complicated queueing networks. A system of queues in series (or queues in tandem), denoted by $\cdot/M_1/1 \to /M_2/1 \to \cdots \to /M_N/1$, consists of N queues of $\cdot/M/1$ type arranged in a series order. Each queue has an infinite waiting room and a single exponential (memoryless) server. The rates of the servers may differ. Initially the system is empty. Customers enter the first queue according to an arbitrary stochastic *input process* and then pass through the queues in order: a customer leaving the first queue immediately enters the second queue, and so on. The number of customers to enter the first queue in the interval $[0, t)$ is given by the random function $A(t)$. We are concerned with the stochastic *output process* of customer departures from the final queue, and the proof that its distribution is the same for all series arrangements of the queues. We say that the queues are *interchangeable*, meaning that the order of the queues does not affect the distribution of any statistic of the output process (such as the time of the nth customer departure).

Results concerning the outputs of series of queues have been surveyed by

Received 6 June 1978; revision received 4 July 1978.
*Postal address: Queens' College, Cambridge CB3 9ET, U.K.

Reich (1965) and Boxma (1977). If the servers are deterministic rather than exponential a result of Friedman (1965) shows that the queues are interchangeable. This is a purely combinatoric result. For exponential servers, the 'output theorem' of Burke (1956) states that the stationary distribution of the output of an unsaturated $M/M/1$ queue is a Poisson process with the same rate as the input. This means that if the input process is Poisson, with a rate less than that of any server, then the stationary output of the $\cdot/M_1/1 \rightarrow /M_2/1 \rightarrow \cdots \rightarrow /M_N/1$ system is the same for all orderings of the queues, being a Poisson process with the same rate as the input. But this tells us nothing about how the ordering of the queues might affect the transient output process (for example, the departure time of the nth customer); neither does it tell us what happens if the input process is not Poisson. Our interchangeability theorem holds for any input process and shows that the full output process is independent of the order of the queues.

2. The interchangeability theorem

It is the result of this section that, with an arbitrary input process, a number of $\cdot/M/1$ queues in series are interchangeable. We directly prove interchangeability for just two queues in series; but this suffices to prove it for any number. For suppose two queues in series are interchangeable. Notice that the input to a series of two queues might itself be the output of an 'upstream' series of queues. So two orderings of N queues in series have identically distributed outputs if they differ in the interchange of one adjacent pair of queues. By the interchange of adjacent pairs of queues we see that N queues in series are interchangeable. It now suffices to show that two queues in series are interchangeable.

Theorem. Two $\cdot/M/1$ queues in series are interchangeable: for any common input process, $A(t)$, the output processes of $\cdot/M_1/1 \rightarrow /M_2/1$ and $\cdot/M_2/1 \rightarrow /M_1/1$ have the same distribution.

Proof. It is sufficient to prove the theorem for arbitrary non-stochastic input processes. For if the theorem is true for every non-stochastic realization of the stochastic input process $A(t)$, then it is true for that stochastic input process. So suppose that no customers are present before time 0 and that the non-stochastic input process is such that the number of customers entering the first queue in the interval $[0, t)$ is $a(t)$. We say that two queues in series are in *state* (t, m, n), when the time is t, the first queue contains $a(t) - m - n$ customers, the second queue contains m customers, and n customers have already departed (customers which arrive at time t are considered to be 'just arriving'). Throughout what follows the phrase, 'for all (t, m, n)', is taken as meaning, 'for all states (t, m, n) which can be reached from the state $(0, 0, 0)$'. We consider the departure times of the first n^*

customers and show that their joint distribution is the same for both orderings of the queues. Consider first the system $\cdot/M_1/1 \rightarrow /M_2/1$, and suppose that the first server has rate λ and the second server has rate μ, where without loss of generality $\lambda + \mu = 1$. Suppose the system is in state (t, m, n) with $n < n^*$. Starting from this state, let $D_j^{\lambda}(t, m, n)$ be the departure time of the jth customer $(j = n+1, \cdots, n^*)$. Define the joint Laplace transform of the departure times of these $n^* - n$ customers as

$$F^{\lambda}(t, m, n) = E[\exp\{-\theta_{n+1}D_{n+1}^{\lambda}(t, m, n) - \cdots - \theta_{n^*}D_{n^*}^{\lambda}(t, m, n)\}],$$

$\theta_{n+1}, \cdots, \theta_{n^*}$ real and positive.

Although $F^{\lambda}(t, m, n)$ is a function of the variables $\theta_{n+1}, \cdots, \theta_{n^*}$, we omit their explicit mention for notational convenience. For $n \geqq n^*$ define $F^{\lambda}(t, m, n) = 1$. Similarly define $F^{\mu}(t, m, n)$ for the system $\cdot/M_2/1 \rightarrow /M_1/1$.

To prove the theorem it is sufficient to show that $F^{\lambda}(0, 0, 0) = F^{\mu}(0, 0, 0)$, since if these two Laplace transforms are equal the joint distribution of the departure times of the first n^* customers are the same for both orderings of the queues. The equality of the Laplace transforms is proved by constructing, for all (t, m, n), $F_k^{\lambda}(t, m, n)$ and $F_k^{\mu}(t, m, n)$, which tend to $F^{\lambda}(t, m, n)$ and $F^{\mu}(t, m, n)$ respectively as k tends to ∞. We then show by induction on k that $F_k^{\lambda}(t, m, n) = F_k^{\mu}(t, m, n)$ for all k.

Consider the system $\cdot/M_1/1 \rightarrow /M_2/1$. When both servers are busy the time until a customer completes service in one or the other of the queues is distributed as an exponential random variable with mean 1 $(\lambda + \mu = 1)$. This suggests the following device. We suppose that in addition to the 'real' customers already considered the system contains an infinite number of 'virtual' customers in each queue. A server serves virtual customers whenever there are no real customers in its queue, but as soon as a real customer enters the queue the server attends to that real customer. So the presence of virtual customers makes no difference to the real ones. But by this device the servers are always busy and the time until the completion of a customer (real or virtual) in one or the other of the queues is always distributed as an exponential random variable with mean 1. The completion occurs in the first queue with probability λ and in the second queue with probability μ. By considering this first service completion, we have for all (t, m, n) with $n < n^*$,

$$(1) \qquad F^{\lambda}(t, m, n) = \int_t^{\infty} \{\lambda F^{\lambda}(s, m+1, n) + \mu F^{\lambda}(s, m-1, n+1)e^{-\theta_{n+1}s}\}e^{-(s-t)}ds,$$

with the provisions that

(i) if $m + n = a(s)$, then $F^{\lambda}(s, m+1, n)$ is replaced by $F^{\lambda}(s, m, n)$, and
(ii) if $m = 0$, then $F^{\lambda}(s, m-1, n+1)e^{-\theta_{n+1}s}$ is replaced by $F^{\lambda}(s, 0, n)$.

The provisions (i) and (ii) apply to cases where the first service completion is of a virtual customer in the first or second queue. A similar formula holds for $F^{\mu}(t, m, n)$.

For all (t, m, n) with $n \geqq n^*$ define $F_k^{\lambda}(t, m, n) = 1$ $(k = 0, 1, 2, \cdots)$. For all (t, m, n) with $n < n^*$ define

$$\phi_n = \theta_{n+1} + \cdots + \theta_{n^*},$$

$$F_0^{\lambda}(t, m, n) = e^{-\phi_n t}, \text{ and for } k = 0, 1, 2, \cdots,$$

$$F_{k+1}^{\lambda}(t, m, n) =$$

(2)
$$\int_t^{\infty} \{\lambda F_k^{\lambda}(s, m+1, n) + \mu F_k^{\lambda}(t, m-1, n+1) e^{-\theta_{n+1} s}\} e^{-(s-t)} ds,$$

with similar provisions on the terms within the integral as in (1). Similarly define $F_k^{\mu}(t, m, n)$. Clearly $0 \leqq F_0^{\lambda}(t, m, n) - F^{\lambda}(t, m, n) \leqq \exp(-\phi_n t)$. By subtracting (1) from (2) it is easy to show inductively that

$$|F_k^{\lambda}(t, m, n) - F^{\lambda}(t, m, n)| \leqq \frac{e^{-\phi_n t}}{(1+\theta_{n^*})^k} \to 0 \quad \text{as} \quad k \to \infty.$$

The proof of the theorem will be complete if we can show that $F_k^{\lambda}(0, 0, 0) = F_k^{\mu}(0, 0, 0)$. In fact, we shall show that for all (t, m, n) and k,

(3)
$$S_k^{\lambda}(t, m, n) = S_k^{\mu}(t, m, n),$$

where $S_k^{\lambda}(t, m, n)$ is defined by

$$S_k^{\lambda}(t, m, n) = \lambda^m F_k^{\lambda}(t, m, n) + \sum_{i=0}^{m-1} \mu \lambda^i F_k^{\lambda}(t, i, n).$$

Note that $S_k^{\lambda}(t, 0, n) = F_k^{\lambda}(t, 0, n)$. We similarly define $S_k^{\mu}(t, m, n)$, and then prove (3) by induction on k. For $k = 0$, (3) is true since $S_0^{\lambda}(t, m, n) = S_0^{\mu}(t, m, n) = \exp(-\phi_n t)$. Suppose that (3) is true for k; we shall show that it is true for $k+1$. Summing (2) we have

$$\begin{aligned} S_{k+1}^{\lambda}(t, m, n) = \int_t^{\infty} \Bigg\{ & \lambda^{m+1} F_k^{\lambda}(s, m+1, n) + \mu \lambda^m F_k^{\lambda}(s, m-1, n+1) e^{-\theta_{n+1} s} \\ & + \sum_{i=0}^{m-1} \mu \lambda^{i+1} F_k^{\lambda}(s, i+1, n) \\ & + \mu^2 F_k^{\lambda}(s, 0, n) + \sum_{i=1}^{m-1} \mu^2 \lambda^i F_k^{\lambda}(s, i-1, n+1) e^{-\theta_{n+1} s} \Bigg\} e^{-(s-t)} ds. \end{aligned}$$

By writing $\mu^2 F_k^{\lambda}(s, 0, n)$ as $\mu F_k^{\lambda}(s, 0, n) - \mu \lambda S_k^{\lambda}(s, 0, n)$ and appropriately combining the terms within the integral we have

$$S^{\lambda}_{k+1}(t,m,n) = \int_t^{\infty} \{S^{\lambda}_k(s,m+1,n) - \mu\lambda S^{\lambda}_k(s,0,n) + \mu\lambda S^{\lambda}_k(s,m-1,n+1)e^{-\theta_{n+1}s}\}e^{-(s-t)}ds, \tag{4}$$

with the provisions that:

(i) if $m+n = a(s)$, then $S^{\lambda}_k(s,m+1,n)$ is replaced by $S^{\lambda}_k(s,m,n)$, and
(ii) if $m = 0$, then the last two terms in the integral are deleted.

By the inductive hypothesis for (3) every term within the integral of (4) is unchanged in value when the superscript λ is replaced by μ. Hence $S^{\lambda}_{k+1}(t,m,n) = S^{\mu}_{k+1}(t,m,n)$ for all (t,m,n), and the induction is complete. In particular. $S^{\lambda}_k(0,0,0) = S^{\mu}_k(0,0,0)$ for all k. This is just $F^{\lambda}_k(0,0,0) = F^{\mu}_k(0,0,0)$. Letting $k \to \infty$ we deduce that $F^{\lambda}(0,0,0) = F^{\mu}(0,0,0)$. This completes the proof of the theorem.

3. Applications

Burke's 'output theorem' for the $M/M/1$ queue may be easily deduced from our theorem. We do this by noticing that an $M/M/1$ queue can be viewed as two $\cdot/M/1$ queues in series, where the first queue contains an infinite number of customers. Suppose that the queue $M_1/M_2/1$ has an arrival rate λ less than the service rate μ. Our interchangeability theorem tells us that the output processes of $M_1/M_2/1$ and $M_2/M_1/1$ have the same distribution. But $M_2/M_1/1$ saturates and its stationary output is just a Poisson process of rate λ. So the stationary output of $M_1/M_2/1$ is also Poisson of rate λ.

It is clear that the interchangeability theorem continues to hold even when the input process is a function of the output process. A special case of this is a *cycle of queues*, formed by taking a series of queues and sending the output back into the first queue as input. In order that the cycle not saturate, we suppose that the process of external input, $A(t)$, terminates at some fixed time. We deduce the interesting result that if a fixed number of customers circle unendingly around a cycle of exponential server queues, then the outputs from all the queues have the same stationary distribution; moreover, this distribution is independent of the order of the queues.

We might reverse our model by thinking of the N servers as passing through n customers in series. This enables us to consider the sequencing of N non-identical 'jobs' through a flowshop of n identical 'machines'. We find that the time for all the jobs to pass through the n machines is independent of the order in which they are served at each machine. The mean finish time of the N jobs is minimized by passing them through the machines in 'shortest first' order.

A number of $\cdot/D/1$ (deterministic) queues in series are interchangeable. But the queues in a series of mixed $\cdot/D/1$ and $\cdot/M/1$ queues are not interchangeable.

If a $\cdot/D/1$ queue is last in such a series, then there is a minimum interval between one departure and the next. This is not the case if a $\cdot/M/1$ queue is last. Burke's output theorem is true for the $M/M/s$ queue. But a number of $\cdot/M/s$ queues in series are not interchangeable, as any simple example will show. For a discussion of the optimal ordering of non-interchangeable queues see Tembe and Wolff (1974).

References

BOXMA, O. J. (1977) Analysis of Models for Tandem Queues. Ph.D. Thesis, University of Utrecht.

BURKE, P. J. (1956) The output of a queueing system. *Opns Res.* **4,** 699–704.

FRIEDMAN, H. D. (1965) Reduction methods for tandem queueing systems. *Opns Res.* **13,** 121–131.

REICH, E. (1965) Departure processes. In *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, 439–457.

TEMBE, S. V. AND WOLFF, R. W. (1974) The optimal order of service in tandem queues. *Opns Res.* **22,** 824–832.