THE ERGODICITY OF SERIES QUEUES WITH GENERAL PRIORITIES[1]

By Bruce McK. Johnson

Northwestern University

1. Introduction. In this paper we study the structure of and obtain certain limit theorems for a rather general and complex queue process. Specifically, the model considered is the following: Let t_j, $j = 1, \cdots$, denote arrival epochs and set $\tau_j = t_j - t_{j-1}$. The $\{\tau_j\}$ are assumed independent and identically distributed with finite expectation. At each arrival epoch the item type $(i = 1, \cdots, r)$ of the current arrival is determined by a multinomial experiment. That is, α_i, $i = 1, \cdots, r$, is the probability that an arriving item is of type i. Each arriving item is to be serviced first by a facility 1, secondly by a facility 2, $\cdots$, and finally by a facility q (q queues in series). The service time of type i items at facility k, $S_{i,k}$, are assumed nonnegative and independent and identically distributed with finite expectation. The service times of the jth arrival, $S_1^j, \cdots, S_q^j$, are taken to be independent. Also, all service times and interarrival times, the τ, are assumed independent.

The physical operation of the system is restricted as follows. No item is allowed to wait in a queue if the corresponding facility is idle. Also, after the servicing of an item is started it may not be displaced at the facility by another item. At those epochs when a facility becomes idle (completion epochs) a priority structure is employed to make a decision as to which of the items waiting in the corresponding queue, if any, is to be serviced next. The admissible priority structures (APS) are general in that they may depend on the state of the system at the time of such a decision. Although there are other possibilities, the APS are assumed to depend on: the elapsed time since the last arrival (V), the composition of the queues ($(L_1, \cdots, L_q)$, where $L_j = (L_{1,j}, \cdots, L_{q,j})$ and $L_{i,j}$ is the number of type i items waiting in queue j and being serviced by facility j), the types of items being serviced at the q queues ($K_1, \cdots, K_q$), and the elapsed service time of the items being serviced ($U_1, \cdots, U_q$). For simplicity we assume that items of the same type waiting in the same queue are not distinguishable to the priority structure. Thus, the decision made at a completion epoch is the type of item to be processed next. The order of servicing items of the same type at a facility is taken to be the order in which they joined the queue. Finally, we assume that the decision made at facility j, $j = 1, \cdots, q$, is a measurable function of the above variables with range, in $1, \cdots, r$.

Models of this generality have not evidently been studied. Sacks [10] (see also Loynes [9]) gives necessary and sufficient conditions for the existence of a bona fide limit distribution of the total time an item waits in q series queues. However,

Received 17 September 1964; revised 10 August 1965.

[1] Work supported in part by the National Science Foundation, GP-87, at Northwestern University. Based on a section of the author's Ph.D. dissertation written at Cornell University.