MANY SERVER QUEUEING PROCESSES WITH POISSON INPUT AND EXPONENTIAL SERVICE TIMES

SAMUEL KARLIN and JAMES MCGREGOR

1. Introduction. A birth and death process is a stationary Markoff process whose state space is the non-negative integers and whose transition probability matrix

$$P_{ij}(t)=\Pr\{x(t)=j \mid x(0)=i\} \tag{1.1}$$

satisfies the conditions (as $t\to 0$)

$$p_{ij}(t)=\begin{cases} \lambda_i t+o(t) & \text{if } j=i+1\,, \\ \mu_i t+o(t) & \text{if } j=i-1\,, \\ 1-(\lambda_i+\mu_i)t+o(t) & \text{if } j=i\,, \end{cases} \tag{1.2}$$

where $\lambda_i>0$ for $i\geqq 0$, $\mu_i>0$ for $i\geqq 1$, and $\mu_0\geqq 0$. The process is called a queueing process if $\mu_0=0$ and $\lambda_i=\lambda$ for all i. The state of the system is then interpreted as the length of a queue for which the inter-arrival times have a negative exponential distribution with parameter λ, and for which the service times have a negative exponential distribution whose parameter μ_n depends on the length of the line. The classical case of a single server queue corresponds to $\mu_n=\mu$, $n\geqq 1$, and has been discussed by Reuter and Lederman [9] and Bailey [1].

The so-called telephone trunking problem (Feller [3]) arises from a queueing process with infinitely many servers, each of whose service time distribution has the same parameter μ, so that $\mu_n=n\mu$, $n\geqq 1$. Besides these two special cases, we discuss a queue with n servers, each of whose service time has a negative exponential distribution with the same parameter μ, so that $\mu_k=k\mu$ for $1\leqq k\leqq n$, $\mu_k=n\mu$ for $k\geqq n$. Our methods can also be used to study queueing processes with several servers whose service times have negative exponential distributions not all with the same parameter.

A sample of the type of problems treated is as follows:

(1) to obtain a usable formula for the transition probability $P_{ij}(t)$;

(2) to compute the distribution of the length of a busy period;

(3) to compute the distribution of the number of customers served during a busy period;

(4) to compute the distribution of the maximum length of the queue during a busy period; and similar questions.

Received May 29, 1957.