ON PRE-EMPTIVE RESUME PRIORITY QUEUES[1]

By Peter D. Welch

IBM Corporation, Yorktown Heights, N.Y.

0. Summary. The following queueing problem is considered. Customers arrive at a service facility at r priority levels. At each priority level the input process is Poisson and these processes are mutually independent. The service times have an arbitrary distribution function which depends upon the priority level. A single server serves under a pre-emptive resume discipline. Results are obtained which characterize the transient and asymptotic distribution of the queue sizes and the waiting times. The analysis proceeds through reductions of the processes of interest to corresponding processes in a simple generalization of an M/G/1 queue.

1. Introduction. This paper discusses the problem of pre-emptive resume priority queues with Poisson inputs, general service times, and a single server. Such queues have previously been studied by Miller [5], Jaiswal [3], Gaver [2], and Keilson [4]. This treatment differs from previous analyses in that it is based upon discrete parameter imbedded Markov processes. These imbedded processes turn out to be equivalent to corresponding processes in simple generalizations of M/G/1 queues.

We assume that customers arrive at a service facility from r different sources which are assigned priority levels $1, \cdots, r$. Level 1 has the highest priority and level r, the lowest. Those customers arriving at priority level k will be called k-customers. From each source, or equivalently at each priority level, the arrivals form a Poisson process with parameter λ_k and these processes are mutually independent. Formally, let

$$\tau_{k,n} = \text{arrival time of the } n\text{th arriving } k\text{-customer},$$

$$\tau_{k,0} = 0, \quad \text{and} \quad \theta_{k,n} = \tau_{k,n} - \tau_{k,n-1}.$$

Then, for fixed k, $\{\theta_{k,n} : n = 1, 2, \cdots\}$ is a sequence of identically distributed, independent random variables with

$$P\{\theta_{k,n} \leqq x\} = 1 - e^{-\lambda_k x} \qquad x \geqq 0$$

$$= 0 \qquad x < 0$$

and the processes $\{\theta_{k,n}\}$ for $k = 1, \cdots, r$ are mutually independent.

We assume that the service times have an arbitrary distribution which is a function of the priority level. Formally, let

$$\chi_{k,n} = \text{service time of the } n\text{th departing } k\text{-customer}.$$

Received 26 March 1963.

[1] This paper is based on part of a Ph.D. thesis submitted to the Department of Mathematical Statistics, Columbia University.

Then, for fixed k, $\{\chi_{k,n} : n = 1, 2, \cdots\}$ is a sequence of identically distributed, independent, positive random variables. We let $G_k(x) = P\{\chi_{k,n} \leqq x\}$. Further we assume that the processes $\{\chi_{k,n}\}$ $k = 1, \cdots, r$ are mutually independent and are independent of the interarrival times. We define

$$\phi_k(s) = \int_0^\infty e^{-sx}\, dG_k(s), \text{ Re } (s) \geqq 0, \quad \text{and} \quad \alpha_k = \int_0^\infty x\, dG_k(x).$$

We will be considering a single server system subject to a pre-emptive resume priority service discipline. In the study of priority queues three service disciplines have received major attention: the head-of-the-line, pre-emptive resume, and pre-emptive repeat. In the head-of-the-line discipline, the initiation of service is determined by priority, but once the service of a customer has begun, it is continued to completion without interruption. In the pre-emptive resume discipline, the initiation of service is exactly as with the head-of-the line discipline; however, if a k-customer is being served, any customer arriving at priority levels 1 through $k - 1$ will pre-empt the server. That is, for the leading k-customer the absence of customers at priority levels 1 through $k - 1$ is a necessary and sufficient condition for both the initiation and continuation of service. The pre-emptive resume discipline is further characterized by the resumption of service at the point it was interrupted; that is, the total time the server is involved with the nth departing k-customer is his service time, $\chi_{k,n}$. The pre-emptive repeat discipline differs from the pre-emptive resume in that, if a customer's service is interrupted, it must be started anew. In all three disciplines the service within the priority level is on an order of arrival basis.

It will be convenient to think of our process as a multiple queue process with the customers at each priority level waiting in separate queues. Hence, we define

$\xi_k(t)$ = size of the kth queue at time t

= number of k-customers waiting or being served at time t,

$\tau'_{k,n}$ = departure time of the nth departing k-customer,

$\xi_{k,n} = \xi_k(\tau'_{k,n} + 0)$.

We further define

$\eta_k(t)$ = virtual waiting time of the kth priority level

= amount of time a k-customer arriving at time t would have to wait before his service begins,

$\eta_{k,n} = \eta_k(\tau_{k,n} - 0)$ = waiting time of the nth arriving k-customer.

For the pre-emptive resume discipline under consideration the following results have been reported: Miller [5] characterized the transient and asymptotic behavior of $\{\eta_k(t)\}$ for $k = 1, \cdots, r$; Jaiswal [3] studied the case $r = 2$ and characterized the transient and asymptotic behavior of the bivariate process

$\{(\xi_1(t), \xi_2(t))\}$ (Keilson [4] also discussed this case), and Gaver [2] characterized the transient and asymptotic behavior of $\{\xi_k(t)\}$ for $k = 1, \cdots, r$. In this paper we determine the transient and asymptotic behavior of the sequences $\{\xi_{k,n}\}$ and $\{\eta_{k,n}\}$ for $k = 1, \cdots, r$ and point out a new argument yielding Miller's results on the virtual waiting times. Our arguments take the form of reductions of the above processes to comparable processes in simple generalizations of M/G/1 queues.

2. Some preliminary results and observations. We will find it useful to consider an alternative description of the input process. Consider the process of all arrivals. Assume this process is Poisson with parameter $\Lambda_r = \sum_{i=1}^{r} \lambda_i$. Now let $\{d_n : n = 1, 2, \cdots\}$ be a sequence of identically distributed, independent random variables taking on the values $\{1, 2, \cdots, r\}$ and let Prob $\{d_n = i\} = \lambda_i/\Lambda_r$, $i = 1, \cdots, r$. If we assign the nth arriving customer to a priority level according to the value assumed by the random variable d_n, we have an input process equivalent to that described earlier. That this is true follows from the fact that in both cases

$$\text{Prob}\{\text{exactly } j_1 \text{ 1-customers arrive, } \cdots, j_r \text{ } r\text{-customers arrive in an interval of length } t\} = \prod_{i=1}^{r} (\lambda_i t)^{j_i} e^{-\lambda_i t}/j_i!$$

and that for any set of nonoverlapping intervals these arrival events are independent. By the same token, we can consider the arrivals at the first k priority levels ($k = 2, \cdots, r-1$) as either the superposition of k independent Poisson processes with parameters $\lambda_1, \cdots, \lambda_k$, or as a single Poisson process with parameter $\Lambda_k = \sum_{i=1}^{k} \lambda_i$ and with priority determined by probabilities $\lambda_1/\Lambda_k, \cdots, \lambda_k/\Lambda_k$.

We will have occasion below to consider the sequence of service times of customers of the first k priority levels ($k = 1, \cdots, r$) taken in their order of arrival. That is, we will be interested in the quantities

$$\tilde{\chi}_{k,n} = \text{service time of the } n\text{th arriving customer of priority level} \leqq k.$$

From the above remarks on the input process, it follows that, for fixed k, $\{\tilde{\chi}_{k,n} : n = 1, 2, \cdots\}$ is a sequence of identically distributed, independent random variables with $P\{\tilde{\chi}_{k,n} \leqq x\} = \sum_{i=1}^{k} \lambda_i G_i(x)/\Lambda_k$. We define

$$H_k(x) = P\{\tilde{\chi}_{k,n} \leqq x\} = \sum_{i=1}^{k} \lambda_i G_i(x)/\Lambda_k,$$

$$\Phi_k(s) = \int_0^\infty e^{-sx}\, dH_k(s) = \sum_{i=1}^{k} \lambda_i \phi_i(s)/\Lambda_k,$$

$$\bar{\alpha}_k = \int_0^\infty x\, dH_k(x) = \sum_{i=1}^{k} \lambda_i \alpha_i/\Lambda_k.$$

We now make an observation, in the form of a lemma, on the occupation time of the server and the busy periods of a G/G/1 queueing system. The oc-

cupation time of the server at time t is the sum of the remaining service time of the customer being served at time t and the service times of all customers in the queue at time t. It is identical to the virtual waiting time if the service is in order of arrival. The busy periods are the periods during which the server is continuously occupied.

LEMMA 1. *Consider a* G/G/1 *queueing process subject to the following conditions*:

(a) *the server is busy whenever there are customers in the queue,*

(b) *the total time any customer is in contact with the server is his service time. For such a process, the occupation time of the server at time* $t(t \geqq 0)$ *and the lengths of the busy periods are independent of the service discipline. In fact, they are the same even if customers are not served continuously but have their service broken up into parts.*

PROOF. For $n = 1, 2, \cdots$ let τ_n = arrival time of the nth arriving customer, χ_n = service time of the nth arriving customer, and $\eta(t)$ = occupation time of the server at time t. The lemma is obvious if we consider a sample time function of the process $\{\eta(t): t \geqq 0\}$. Regardless of the service discipline, $\eta(t)$ has jumps at the points τ_n of size χ_n and in between the jumps it either decreases with slope -1 if it is greater than zero or remains unchanged if it equals zero. The lengths of the busy periods are determined by $\eta(t)$; hence, their independence of the service discipline follows immediately.

Finally we note that, because of the pre-emptive nature of the service discipline, the behavior of quantities concerning the kth queue depends only on the first k queues and not in any way upon the $(k + 1)$st through rth queues. With respect to the kth queue, the situation is the same as if it were the last or rth queue. It is only with respect to the server that there is a difference, for in the one case he would be idle when the 1st through kth queues were empty, while in the other case, he might be busy serving a $(k + 1)$st through rth customer.

3. The queue sizes. We will assume initial departure points $\tau'_{k,0}$ for $k = 1, \cdots, r$ and study the sequences $\{\xi_{k,n} : n = 0, 1, 2, \cdots\}$. We first observe that, because of the last remark of Section 2, $\{\xi_{1,n} : n = 0, 1, \cdots\}$ behaves exactly as the comparable process of an M/G/1 queueing system with input parameter λ_1 and with a service time distribution function $G_1(x)$. Its behavior has been extensively studied; see, for example, Takács [6].

Let us consider then $\{\xi_{k,n} : n = 0, 1, 2, \cdots\}$ for $k \geqq 2$. We will show that this process is a homogeneous Markov chain identical to the comparable process in a simple generalization of an M/G/1 queue. Consider some fixed pair k, n and suppose that $\xi_{k,n} > 0$; that is, suppose that the nth departing k-customer leaves someone in the kth queue. Since a k-customer is departing, because of the pre-emptive nature of the discipline, it must be that the 1st through $(k - 1)$st queues are empty. Hence, the service of the next or $(n + 1)$st k-customer will begin immediately. The time from the beginning of his service until he departs we will call his service-plus-interruption time (this duration was called the "completion time" by Gaver [2]). This service-plus-interruption time is equal

to his service time, plus the service times of all 1-through $(k-1)$-customers arriving while he is being served, plus the service times of all 1-through $(k-1)$-customers arriving while they are being served, etc. Hence, it is a busy period of the combined first $k-1$ queues with his service time as the initial waiting time. Now from the remarks of Section 2, we know that the input to the first $k-1$ queues is Poisson with parameter Λ_{k-1}. We also know that if the 1-through $(k-1)$-customers are served in order of arrival, they have a service time distribution function $H_{k-1}(x)$. Now, Lemma 1 of Section 2 states that the busy period of an M/G/1 system is independent of the order in which the customers are served and also is unchanged if a particular service time is broken up into parts. Hence, his service-plus-interruption time is equal to the initial busy period of an M/G/1 system with input parameter Λ_{k-1}, service time distribution function $H_{k-1}(x)$, and an initial waiting time with distribution function $G_k(x)$. Thus, from Takács [6] (remark 4, page 63) we have that this service-plus-interruption time has a distribution function with Laplace-Stieltjes transform $\phi_k\{s+\Lambda_{k-1}(1-\gamma_{k-1}(s))\}$ where $\gamma_k(s)$ is the unique root with minimum absolute value of the equation $z=\Phi_k(s+\Lambda_k(1-z))$. Hence we have, for $\xi_{k,n}>0$,

$$\begin{aligned}\xi_{k,n+1} = \xi_{k,n} - 1 + [&\text{number of } k\text{-customers arriving during a}\\ &\text{random length of time whose distribution function has}\\ &\text{Laplace-Stieltjes transform } \phi_k\{s+\Lambda_{k-1}(1-\gamma_{k-1}(s))\}].\end{aligned} \tag{3.1}$$

This is the same relationship as that for an M/G/1 process with input parameter λ_k and with a service time distribution whose Laplace-Stieltjes transform is $\phi_k(s+\Lambda_{k-1}(1-\gamma_{k-1}(s)))$.

Next, suppose that $\xi_{k,n}=0$. This means that the nth departing k-customer leaves the first k queues empty. Now, during the time from $\tau'_{k,n}$ to $\tau_{k,n+1}$, the time of the next arrival of a k-customer, the occupation time of the server with respect to the first $k-1$ queues will build up. (By the occupation time of the server with respect to the first $k-1$ queues we mean the length of time required to complete the service of all the customers in the first $k-1$ queues.) In other words, when the next k-customer arrives, the server will have a certain occupation time with respect to the first $k-1$ queues. We will define

$$\begin{aligned}\rho_{k,n+1} = &\text{ occupation time of the server with respect to the first}\\ &k-1 \text{ queues at the time } \tau_{k,n+1}, \text{ given } \xi_{k,n}=0.\end{aligned}$$

$$T_k(x) = P\{\rho_{k,n+1} \leqq x\}, \qquad \Xi_k(s) = \int_0^\infty e^{-sx}\, dT_k(x), \qquad \text{Re}\,(s) \geqq 0.$$

$T_k(x)$ is independent of n because of the Poisson nature of the arrivals and the fact that at time $\tau'_{k,n}$ the first $k-1$ queues are empty. Now the $(n+1)$st k-customer will have to wait until the first $k-1$ queues are empty before his service can begin. Consequently, by an argument exactly parallel to one given earlier, he will have to wait a length of time equal to the initial busy period of an M/G/1 system with input parameter Λ_{k-1}, service time distribution function

$H_{k-1}(x)$, and an initial waiting time distribution function $T_k(x)$. From Takács [6], (remark 4, page 63) we have that this "waiting" time has a distribution function with Laplace-Stieltjes transform $\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$. We will call this "waiting" time his delay time.

We will now determine $\Xi_k(s)$. Applying Lemma 1 of Section 2, we know that the occupation time of the server with respect to the first $k - 1$ queues at time $\tau'_{k,n} + t$ is the same as the occupation time of server at time t in an M/G/1 queueing process with input parameter Λ_{k-1} and a service time distribution function $H_{k-1}(x)$ when at time zero the occupation time of the server is identically zero. Hence, we have from Takács [6] (Theorem 1, page 51) that

$$E\{e^{-\rho_{k,n+1}s} \mid \tau_{k,n+1} - \tau'_{k,n} = t\} = \exp\{st - [1 - \Phi_{k-1}(s)]\Lambda_{k-1}t\} \cdot \{1 - s\int_0^t \exp\{-su + [1 - \Phi_{k-1}(s)]\Lambda_{k-1}u\}P_0(u)\,du\} \tag{3.2}$$

where $\int_0^\infty e^{-st}P_0(t)\,dt = 1/\{s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\}$. Removing the condition in (3.2), $\Xi_k(s)$ is given by

$$\begin{aligned}\Xi_k(s) &= E\{e^{-\rho_{k,n+1}s}\} \\ &= \lambda_k\int_0^\infty E\{e^{-\rho_{k,n+1}s} \mid \tau_{k,n+1} - \tau'_{k,n} = t\}e^{-\lambda_k t}\,dt.\end{aligned} \tag{3.3}$$

Now, (3.2) and (3.3) determine $\Xi_k(s)$, but in a very awkward manner. Fortunately, through (3.3), $\Xi_k(s)$ can be determined directly. In fact, Takács [6], in his determination of (3.2), obtained and used the following result

$$\begin{aligned}&\int_0^\infty E\{e^{-\rho_{k,n+1}\zeta} \mid \tau_{k,n+1} - \tau'_{k,n} = t\}e^{-st}\,dt \\ &= \frac{1 - \zeta\int_0^\infty e^{-st}P_0(t)\,dt}{s - \zeta + \Lambda_{k-1}[1 - \Phi_{k-1}(\zeta)]} = \frac{1 - \zeta/\{s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\}}{s - \zeta + \Lambda_{k-1}[1 - \Phi_{k-1}(\zeta)]}.\end{aligned} \tag{3.4}$$

If we make the substitutions $s = \lambda_k$ and $\zeta = s$ in (3.4), we have from (3.3) and (3.4) that

$$\begin{aligned}\Xi_k(s) &= \frac{\lambda_k - \lambda_k s/\{\lambda_k + \Lambda_{k-1}[1 - \gamma_{k-1}(\lambda_k)]\}}{\lambda_k - s + \Lambda_{k-1}[1 - \Phi_{k-1}(s)]} \\ &= \frac{\lambda_k[\Lambda_k - s - \Lambda_{k-1}\gamma_{k-1}(\lambda_k)]}{[\Lambda_k - \Lambda_{k-1}\gamma_{k-1}(\lambda_k)][\Lambda_k - s - \Lambda_{k-1}\Phi_{k-1}(s)]}\end{aligned}$$

and hence that

$$\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\} = \frac{\lambda_k[\lambda_k - s + \Lambda_{k-1}\{\gamma_{k-1}(s) - \gamma_{k-1}(\lambda_k)\}]}{[\Lambda_k - \Lambda_{k-1}\gamma_{k-1}(\lambda_k)][\lambda_k - s]}. \tag{3.5}$$

Now we are assuming that $\xi_{k,n} = 0$. In this case we have shown that the $(n + 1)$st arriving k-customer will have to wait a delay time until his service

begins and that the distribution function of this delay time has Laplace-Stieltjes transform $\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$. Once his service begins, his service-plus-interruption time is the same as that obtained in the case $\xi_{k,n} > 0$ for again at the time his service begins, the first $k - 1$ queues are empty. Further, the waiting time and service-plus-interruption time are independent. Hence, we have, for $\xi_{k,n} = 0$,

$$\xi_{k,n+1} = \text{number of } k\text{-customers arriving during a random length of time whose distribution function has Laplace-Stieltjes transform } \Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}. \tag{3.6}$$

Equation (3.1) for $\xi_{k,n} > 0$ and Equation (3.6) for $\xi_{k,n} = 0$ plus independence conditions which follow trivially from the Poisson nature of the arrivals show that $\{\xi_{k,n} : n = 0, 1, 2, \cdots\}$ for $k \geqq 2$ behaves as the queue size of the following generalized M/G/1 system. Customers arriving when the server is busy are served immediately after the departure of the customer ahead of them. However, customers arriving when the server is idle must wait a random "delay time" until their service begins. The transient and asymptotic behavior of the queue size and the waiting time for this system is characterized in Appendix B of Welch [7], and the results are summarized in the Appendix of this paper. The asymptotic results were obtained earlier by Finch [1]. The specific equivalence to $\{\xi_{k,n}\}$ occurs when the service time and delay time distribution functions have Laplace-Stieltjes transforms $\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$ and $\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$, respectively.

Now by implicit differentiation of $\gamma_k(s) = \Phi_k(s + \Lambda_k(1 - \gamma_k(s)))$, one can easily show that

$$\lim_{s\to 0} \frac{d\gamma_k(s)}{ds} = -\left(\sum_{i=1}^{k} \frac{\lambda_i \alpha_i}{\Lambda_k}\right)\left(1 - \sum_{i=1}^{k} \lambda_i \alpha_i\right)^{-1}. \tag{3.7}$$

Using (3.7) we have that the expectation of the service-plus-interruption time is given by

$$-\lim_{s\to 0} d\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}/ds = \alpha_k\left(1 - \sum_{i=1}^{k-1} \lambda_i \alpha_i\right)^{-1},$$

and, using (3.5) and (3.7), that the expectation of the delay time is given by

$$-\lim_{s\to 0} d\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}/ds = \left\{[\Lambda_k - \Lambda_{k-1}\gamma_{k-1}(\lambda_k)]\left[1 - \sum_{i=1}^{k-1} \lambda_i \alpha_i\right]\right\}^{-1} - \lambda_k^{-1}.$$

Further, if we let $\phi(s) = \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$ and let $g(w)$ for $|w| < 1$ be the unique root with minimum absolute value of the equation $z = w\phi(\lambda_k(1 - z))$, which in this case is the equation

$$z = w\phi_k\{\lambda_k(1 - z) + \Lambda_{k-1}[1 - \gamma_{k-1}(\lambda_k(1 - z))]\},$$

then it is shown in Welch [7] (page 105) that $g(w) = w\phi_k\{\Lambda_k(1 - \theta_{k,k}(w))\}$, where $\theta_{k,k}(w)$ is the unique root with minimum absolute value of the equation

$$z = (w\lambda_k/\Lambda_k)\phi_k(\Lambda_k(1 - z)) + \sum_{i=1}^{k-1} (\lambda_i/\Lambda_k)\phi_i(\Lambda_k(1 - z)).$$

A series expansion for $\theta_{k,k}(w)$ is also given in Welch [7].

If we apply the above remarks, we obtain the following two theorems by substitution into Lemmas 2 and 3 of the Appendix.

THEOREM 1. *For $k \geqq 2$, if $\sum_{j=1}^{k-1} \lambda_j\alpha_j \leqq 1$, then $\{\xi_{k,n} : n = 0, 1, 2, \cdots\}$ is a homogeneous Markov chain. The generating functions $L_{k,n}(z) = \sum_{j=0}^{\infty} P\{\xi_{k,n} = j\}z^j$, $|z| \leqq 1$, are given by*

$$\sum_{n=0}^{\infty} L_{k,n}(z)w^n = \frac{zL_{k,0}(z)}{z - w\phi(\lambda_k(1 - z))} + \frac{wL_{k,0}(g(w))\phi(\lambda_k(1 - z))[z\mu(\lambda_k(1 - z)) - 1]}{[z - w\phi(\lambda_k(1 - z))][1 - g(w)\mu\{\lambda_k(1 - g(w))\}]} \quad |z| \leqq 1, \quad |w| < 1$$

where

$$\phi(s) = \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\},$$

$$\mu(s) = \frac{\lambda_k[\lambda_k - s + \Lambda_{k-1}(\gamma_{k-1}(s) - \gamma_{k-1}(\lambda_k))]}{[\Lambda_k - \Lambda_{k-1}\gamma_{k-1}(\lambda_k)][\lambda_k - s]},$$

$$g(w) = w\phi_k\{\Lambda_k(1 - \theta_{k,k}(w))\},$$

$\theta_{k,k}(w)$ is the root with minimum absolute value of the equation

$$z = (\lambda_k w/\Lambda_k)\phi_k(\Lambda_k(1 - z)) + \sum_{i=1}^{k-1} (\lambda_i/\Lambda_k)\phi_i(\Lambda_k(1 - z)),$$

and $\gamma_k(s)$ is the root with minimum absolute value of the equation $z = \Phi_k(s + \Lambda_k(1 - z))$.

If $\sum_{j=1}^{k-1} \lambda_j\alpha_j > 1$, then there is a nonzero probability that the service-plus-interruption times and the delay times never terminate.

THEOREM 2. *For $k \geqq 2$, if $\sum_{j=1}^{k} \lambda_j\alpha_j < 1$, then the Markov chain $\{\xi_{k,n} : n = 0, 1, 2, \cdots\}$ is ergodic and, independent of the initial distribution, we have $\lim_{n\to\infty} P\{\xi_{k,n} = j\} = P_k(j)$ where $\{P_k(j)\}$ is a probability distribution whose generating function is given by*

$$\sum_{j=0}^{\infty} P_k(j)z^j = \frac{\phi(\lambda_k(1 - z))[\lambda_k z - \Lambda_k + \Lambda_{k-1}\gamma_{k-1}(\lambda_k(1 - z))][1 - \sum_{i=1}^{k} \lambda_i\alpha_i]}{\lambda_k[z - \phi(\lambda_k(1 - z))]}$$

where $\phi(s) = \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$ and $\gamma_k(s)$ is the root with minimum absolute value of the equation $z = \Phi_k(s + \Lambda_k(1 - z))$.

If $\sum_{j=1}^{k} \lambda_j\alpha_j \geqq 1$ and $\sum_{j=1}^{k-1} \lambda_j\alpha_j \leqq 1$, then independent of the initial distribution $\lim_{n\to\infty} P\{\xi_{k,n} = j\} = 0$, for all j.

If $\sum_{j=1}^{k-1} \lambda_j\alpha_j > 1$, then $\lim_{n\to\infty} P\{n\text{th } k \text{ departure occurs}\} = 0$.

It is interesting to note that the limiting distributions $\{P_k(j): j = 0, 1, \cdots\}$ given by Theorem 2 are the same as the limiting distributions $\{\lim_{t\to\infty} P(\xi_k(t) = j): j = 0, 1, 2, \cdots\}$ obtained by Gaver [2]. The generating functions of the limiting distributions $\{\lim_{t\to\infty} P(\xi_k(t) = j): j = 0, 1, \cdots\}$ are given by *his* equation (8.4) with the substitutions $\rho = \lambda_k\alpha_k(1 - \sum_{i=1}^{k-1}\lambda_i\alpha_i)^{-1}$, $\nu = \Lambda_{k-1}$, $\lambda = \lambda_k$, $E(D) = (\sum_{i=1}^{k-1}\lambda_i\alpha_i)\Lambda_{k-1}^{-1}(1 - \sum_{i=1}^{k-1}\lambda_i\alpha_i)^{-1}$, $b(x) = x$, $\hat{U}(s) = \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$ and $\hat{L}(s) = \gamma_{k-1}(s)$.

4. Waiting times. We define

$$W_{k,n}(x) = \text{Prob}\,\{\eta_{k,n} \leqq x\}, \qquad \Omega_{k,n}(s) = \int_0^\infty e^{-sx}\,dW_{k,n}(x) \qquad \text{Re}\,(s) \geqq 0.$$

Because of the pre-emptive nature of the service discipline, the process $\{\eta_{1,n} : n = 1, 2, \cdots\}$ behaves exactly as the waiting time process of an M/G/1 queueing system with input parameter λ_1 and with a service time distribution function $G_1(x)$. The transient and asymptotic behavior of such processes have been extensively studied; see, e.g., Takács [6].

For $k = 2, 3, \cdots, r$ from the remarks of Section 3, we know the following. The service-plus-interruption times (that is, the times from the beginning of service to departure) are independent, identically distributed random variables whose distribution function has Laplace-Stieltjes transform $\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$. If an arriving k-customer finds the kth queue nonempty, his service begins immediately after the departure of the k-customer ahead of him. If an arriving k-customer finds the kth queue empty, he must wait a delay time until his service begins. This sequence of delay times is a sequence of identically distributed, independent random variables whose distribution function has Laplace-Stieltjes transform $\Xi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$, defined by Equation (3.5). Further, these delay times are independent of the service-plus-interruption times.

Hence, the waiting time processes $\{\eta_{k,n} : n = 1, 2, \cdots\}$, for $k \geqq 2$, are special cases of the situation discussed in the Appendix, and with the proper substitutions, we obtain the following two theorems from Lemmas 4 and 5.

THEOREM 3. *For $k \geqq 2$, if $\sum_{i=1}^{k-1}\lambda_i\alpha_i \leqq 1$, then the $\Omega_{k,n}(s)$ are given by*

$$\sum_{n=1}^{\infty}\Omega_{k,n}(s)w^n = \frac{w(\lambda_k - s)\Omega_{k,1}(s)}{\lambda_k - s - w\lambda_k\phi(s)} + \frac{wg(w)\Omega_{k,1}\{\lambda_k(1 - g(w))\}[(\lambda_k - s)\mu(s) - \lambda_k]}{[\lambda_k - s - w\lambda_k\phi(s)][1 - g(w)\mu\{\lambda_k(1 - g(w))\}]}$$

$$\text{Re}\,(s) \geqq 0, \qquad |w| < 1$$

where $\phi(s)$, $\mu(s)$, and $g(w)$ are as defined in Theorem 1.

If $\sum_{i=1}^{k-1}\lambda_i\alpha_i > 1$, then there is a nonzero probability that the delay and service-plus-interruption times never terminate.

THEOREM 4. *For $k \geqq 2$, if $\sum_{i=1}^{k}\lambda_i\alpha_i < 1$, then $W_{k,n}(x)$ converges as $n \to \infty$,*

to a distribution function $W_k(x)$ *which has Laplace-Stieltjes transform*

$$\Omega_k(s) = \frac{[s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))][1 - \sum_{i=1}^{k} \lambda_i \alpha_i]}{s - \lambda_k + \lambda_k \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}}$$

where $\gamma_k(s)$ *is the root with minimum absolute value of the equation* $z = \Phi_k(s + \Lambda_k(1 - z))$.

If $\sum_{i=1}^{k} \lambda_i \alpha_i \geqq 1$ *and* $\sum_{i=1}^{k-1} \lambda_i \alpha_i \leqq 1$, *then* $\lim_{n\to\infty} W_{k,n}(x) = 0$, *for all* x.

If $\sum_{i=1}^{k-1} \lambda_i \alpha_i > 1$, *then* $\lim_{n\to\infty} P\{n\text{th } k \text{ departure occurs}\} = 0$.

It should be remembered that the above theorems concern the time a k-customer must wait until his service begins and that the time from the beginning of service until he departs is a service-plus-interruption time with a distribution function whose Laplace-Stieltjes transform is $\phi_k(s + \Lambda_{k-1}(1 - \gamma_{k-1}(s)))$.

It is interesting to note that the limiting distribution $W_k(x)$ is identical to the limiting distribution of the virtual waiting time obtained by Miller [5].

The relationship of the waiting time and the queue size is such that $\sum_{j=0}^{\infty} P_k(j)z^j$ and $\Omega_k(s)$ should satisfy the equation

$$\sum_{j=0}^{\infty} P_k(j)z^j = \phi(\lambda_k(1 - z))\Omega_k(\lambda_k(1 - z))$$

where $\phi(s) = \phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$. Comparison of Theorems 2 and 4 show that our results do satisfy this check.

5. The virtual waiting times. The transient and asymptotic behavior of the virtual waiting times $\{\eta_k(t)\}$ $k = 1, \cdots, r$ was characterized by Miller [5]. We will, in this section, show how these processes can also be reduced to the comparable process in a simple generalization of an M/G/1 system.

Because of the pre-emptive nature of the service discipline $\{\eta_1(t): t \geqq 0\}$ behaves exactly as the virtual waiting time of an M/G/1 queueing system with input parameter λ_1 and service time distribution function $G_1(x)$. The transient and asymptotic behavior of this process has been extensively studied; see, e.g., Takács [6].

Hence, we will consider $\{\eta_k(t): t \geqq 0\}$ for $k \geqq 2$. Let us first suppose that $\eta_k(t) > 0$. In this case $\eta_k(t)$ equals the remaining service time of all those waiting and being served in the first k queues plus the service times of all those 1-through $(k - 1)$-customers arriving while they are being served plus the service times of all those 1-through $(k - 1)$-customers arriving while these arrivals are being served, etc. In other words, $\eta_k(t)$ has the distribution function of a busy period of the first $k - 1$ queues with an initial waiting time equal to the remaining service time of all the customers in the first k queues. Hence, if $\eta_k(t) > 0$, the arrival of a 1-through $(k - 1)$-customer will not affect it since we have already accounted for the time required to service all such customers. However, the arrival of a k-customer will increase $\eta_k(t)$ by the service-plus-interruption time of a k-customer. We saw in Section 3 that this service-plus-interruption time has a distribution function with Laplace-Stieltjes transform $\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$.

Next, suppose that $\eta_k(t) = 0$. In this case, the arrival of any 1-through k-customer will increase $\eta_k(t)$ by a busy period of the first $k - 1$ queues with the service time of the arriving customer as the initial waiting time of the busy period. Now, if we consider the arrival process of 1-through k-customers, we know from the remarks of Section 2 that it is a Poisson process with parameter Λ_k and that the service times, taken in the order of arrival, constitute a sequence of independent, identically distributed, random variables with a distribution function $H_k(x)$. Hence, by an argument parallel to those given in Section 3, if $\eta_k(t) = 0$, it is affected by a Poisson arrival process with parameter Λ_k and an arrival of this process will increase it by a random time whose distribution function has Laplace-Stieltjes transform $\Phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$.

Now, from the remarks of Section 2, k-customers arrive in the process of all 1-through k-arrivals independently and with probability λ_k/Λ_k. Hence, the process $\{\eta_k(t): t \geqq 0\}$ behaves exactly as the virtual waiting time in the following single queue, single server system. The input is Poisson with parameter Λ_k. If the queue is empty, an arriving customer always joins it and has a service time whose distribution function has Laplace-Stieltjes transform $\Phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$. If the queue is not empty, an arriving customer joins it with probability λ_k/Λ_k and departs without service with probability $1 - \lambda_k/\Lambda_k$. If he joins the queue, his service time distribution function has Laplace-Stieltjes transform $\phi_k\{s + \Lambda_{k-1}(1 - \gamma_{k-1}(s))\}$.

The above single queue, single server system is a special case of the following generalized M/G/1 system with the addition of balking. There are two service time distribution functions, $G_e(x)$ and $G_b(x)$. If a customer arrives and finds the server idle, his service time has distribution function $G_e(x)$ while if he arrives and finds the server busy, his service time has distribution function $G_b(x)$. The virtual waiting time of this process with the addition of balking is characterized in Appendix B of Welch [7].

6. Acknowledgment. The author would like to acknowledge his indebtedness to Professor Lajos Takács for suggesting the topic of priority queues and for many helpful discussions during the course of this work. He would also like to thank the referee for suggestions which led to clarification of the original presentation.

APPENDIX

In the body of the paper we reduce the queue size processes and the waiting time processes of pre-emptive resume priority queues to comparable processes in the following generalized M/G/1 system. If a customer arrives and finds the server busy, he is served immediately after the departure of the customer ahead of him; however, if he arrives and finds the server idle, he must wait a random delay time before his service begins. These additional delay times constitute a sequence of independent, identically distributed random variables, independent of the service times. This process was previously studied by Finch [1].

We will now give results which characterize the transient and asymptotic behavior of the queue size and waiting times for this process. Proofs of the results on the asymptotic behavior can be found in Finch [1]. Welch [7] (Appendix B) contains proofs of both the transient and asymptotic results. We let the service time have expectation α and a distribution function $G(x)$ with Laplace-Stieltjes transform $\phi(s)$. We let the additional delay time have expectation β and a distribution function $H(x)$ with Laplace-Stieltjes transform $\mu(s)$. We let λ be the parameter of the Poisson input process.

We first give two lemmas characterizing the behavior of the queue size. We define

$$\xi_n = \text{queue size immediately after the } n\text{th departure, } n = 0, 1, \cdots .$$

The value $n = 0$ corresponds to the initial departure point.

LEMMA 2. *The process* $\{\xi_n : n = 0, 1, 2, \cdots\}$ *is a homogeneous Markov chain. The generating functions*

$$L_n(z) = \sum_{k=0}^{\infty} P\{\xi_n = k\} z^k \qquad |z| \leqq 1$$

are given by

$$\sum_{n=0}^{\infty} L_n(z) w^n = \frac{zL_0(z)}{z - w\phi(\lambda(1 - z))} + \frac{wL_0\{g(w)\}\phi(\lambda(1 - z))[z\mu(\lambda(1 - z)) - 1]}{[z - w\phi(\lambda(1 - z))][1 - g(w)\mu\{\lambda(1 - g(w))\}]} \qquad |w| < 1$$

where $g(w)$ *is the unique root with minimum absolute value of the equation* $z = w\phi(\lambda(1 - z))$.

LEMMA 3. *If* $\lambda\alpha < 1$, *then the Markov chain is ergodic and independent of the initial distribution, we have* $\lim_{n\to\infty} P\{\xi_n = k\} = P_k$, $k = 0, 1, 2, \cdots$ *where* $\{P_k\}$ *is a probability distribution with generating function*

$$\sum_{k=0}^{\infty} P_k z^k = \frac{1 - \lambda\alpha}{1 + \lambda\beta} \frac{\phi(\lambda(1 - z))[z\mu(\lambda(1 - z)) - 1]}{z - w\phi(\lambda(1 - z))} \qquad |z| \leqq 1.$$

If $\lambda\alpha \geqq 1$, *then* $\lim_{n\to\infty} P\{\xi_n = k\} = 0$, *for all* k.

We now give two lemmas characterizing the waiting time. We define

$$\eta_n = \text{waiting time of the } n\text{th departing customer}$$

$$\tilde{W}_n(x) = P\{\eta_n \leqq x\}, \quad \tilde{\Omega}_n(s) = \int_0^{\infty} e^{-sx}\, d\tilde{W}_n(x) \qquad \text{Re } (s) \geqq 0.$$

LEMMA 4. *The* $\tilde{\Omega}_n(s)$ *are given by the following generating function:*

$$\sum_{n=1}^{\infty} \tilde{\Omega}_n(s) w^n = \frac{w(\lambda - s)\tilde{\Omega}_1(s)}{\lambda - s - w\lambda\phi(s)} + \frac{wg(w)\tilde{\Omega}_1\{\lambda(1 - g(w))\}[(\lambda - s)\mu(s) - \lambda]}{[\lambda - s - w\lambda\phi(s)][1 - g(w)\mu\{\lambda(1 - g(w))\}]}$$

$$\text{Re } (s) \geqq 0, \qquad |w| < 1,$$

where $g(w)$ is the unique root with minimum absolute value of the equation $z = w\phi(\lambda(1 - z))$.

LEMMA 5. *If $\lambda\alpha < 1$, then, independent of $\tilde{W}_1(x)$, $\tilde{W}_n(x)$ converges to a distribution function $\tilde{W}(x)$ as $n \to \infty$, where $\tilde{W}(x)$ has Laplace-Stieltjes transform*

$$\tilde{\Omega}(s) = \frac{1 - \lambda\alpha}{1 + \lambda\beta} \frac{(\lambda - s)\mu(s) - \lambda}{\lambda - s - \lambda\phi(s)}.$$

If $\lambda\alpha \geqq 1$, then independent of $\tilde{W}_1(x)$, we have for all x

$$\lim_{n\to\infty} \tilde{W}_n(x) = 0.$$

REFERENCES

[1] FINCH, P. D. (1959). A probability limit theorem with application to a generalization of queueing theory. *Acta Math. Acad. Sci. Hungar.* **10** 317–325.

[2] GAVER, D. P., JR. (1962). A waiting line with interrupted service, including priorities. *J. Roy. Statist. Soc. Ser. B* **24** 73–90.

[3] JAISWAL, N. K. (1961). Pre-emptive resume priority queue. *Operations Res.* **9** 732–742.

[4] KEILSON, J. (1962). Queues subject to service interruption. *Ann. Math. Statist.* **33** 1314–1322.

[5] MILLER, R. G., JR. (1960). Priority queues. *Ann. Math. Statist.* **31** 86–103.

[6] TAKÁCS, L. (1962). *Introduction to the Theory of Queues*. Oxford Univ. Press.

[7] WELCH, P. D. (1963). Some contributions to the theory of priority queues. Ph.D. Thesis, Columbia University, IBM Research Report RC-922, IBM Research Center, Yorktown Heights, New York.