SOME ASPECTS

OF THE

THEORY OF QUEUES

C.R. Heathcote

A thesis submitted to the
Australian National University for
the degree of Doctor of Philosophy
in the Department of Statistics.

Canberra
1960

Some Aspects of the Theory of Queues

ERRATA

Page 83. Correction to the sentence commencing on second last line of page 83 and ending at top of page 84.

Replace sentence:

"We can now apply the extension of Abel's theorem quoted previously (Chapter I, equation 3.5) to obtain

$$\lim_{t\to\infty} Q_0(t,\infty) = \lim_{s\to 0+} s\,Q_0^*(s,\infty) = \begin{cases} 1 & \text{if } \nu\bar{h} \le 1, \\ e^{-\omega_0(0)} & \text{if } \nu\bar{h} > 1 \end{cases}"$$

by the following:-

"Since $Q_0(t,\infty)$ is a non-decreasing function of t so is $\int_0^t u\,d\,Q_0(u,\infty)$. Then by Theorem 4.5 of Widder (1946), $\lim_{t\to\infty} Q_0(t,\infty) = e^{-\omega_0(0)}$."

Page 90. Add Footnote to last word on the page, "...normalized to unity*."

* "I am indebted to Dr. W.L. Smith for pointing out to me that the existence of these limits can be established in a relatively simple way by using his results on regenerative stochastic processes (Smith, 1958). In what follows we assume that $\lim_{t\to\infty} P_n(t,k) = p_n(k)$ exists, but we do not assume that $\sum_{n=0}^{\infty} p_n(k) = 1$."

Page 108. Add the following sentence to end of the first paragraph:- "As stated on page 105 we assume that the 'preemptive resume' rule holds."

Page 158. Add the following paragraph:-

"An arrival process with probabilities as in (2.5) has been considered by Pollaczek (1957), and called by him the Bernoulli distribution of arrival times. In the monograph referred to, Pollaczek obtained the distribution of waiting times for the single server queue with general renewal input and with input defined by (2.5). Pollaczek's methods

are different from those used here and his results are expressed in terms of contour integrals. I am indebted to Dr. D.G. Kendall for drawing my attention to this reference."

Page 170. Add the following references:

Pollaczek, F. (1957).
Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente a un guichet et par des phénomenès apparentes. Paris: Gauthier-Villars.

Smith, W.L. (1958).
Renewal theory and its ramifications. J.R. Statist. Soc. (B) 20, 243-302.

CONTENTS

PREFACE

This thesis was written whilst I was a research student at the Australian National University from March 1958 to December 1960. Work done during the first eight months of this period is not reported on here, partly because it is incomplete, but mainly because it is not relevant to the theory of queues.

Many of the problems with which the thesis is concerned have been discussed by other writers on the subject; however, most of the results presented here are new. Previous work is acknowledged in the appropriate part of the text. The contents of Chapters II, III, and IV have been published or submitted for publication, and work is continuing on the unsolved problems posed in Chapters I and V. Chapter II is based on an article which appeared in Biometrika, 46, (1959) written jointly by J.E. Moyal and myself and it is impossible to indicate which results are specifically my own. As a broad statement it seems fair to say that the basic ideas of that chapter were Mr Moyal's and that my responsibility lay more with the calculations and the extension to the many server queue. The work on which the other chapters are based is my own although many of the more difficult problems were only disposed of satisfactorily after discussion with Mr. Moyal.

In fact my deepest thanks are due to Mr. Moyal, not only for introducing me to queueing theory and proposing it as a field of study, but also for instructing me in technical matters and for his criticism and suggestions during the preparation of the thesis. I would also like to thank Professor P.A.P. Moran for many helpful discussions and encouragement; the Australian National University and General Motors Holden Ltd. for funds which made this work possible; to Mr. H. Guenther for drawing Figure 1; and last but not least, Miss J. Duffield for the excellent way in which she typed the manuscript.

C.R. Heathcote.

Department of Statistics
Australian National University
Canberra.

C.R. Heathcote

SUMMARY

Chapter I describes briefly the main problems of queueing theory and the methods used to treat them. The two most important quantities associated with a queue are are its length at time t and the waiting time of a customer. Of these the first is the more fundamental since the distribution of waiting time can always be found if queue length is known so that in this thesis we are primarily interested in the distribution of queue length. Queueing systems are divided into two classes depending on whether or not the input process is of renewal type. If the input (and the service process) are of this type we argue that of the standard Markovisation procedures available for the analysis of such systems the most useful is the inclusion of supplementary variables. Procedures of this sort are difficult to apply to more general systems and we point out that the problem of finding the distribution of queue length is essentially a combinatorial one which is still unsolved. The arguments of the thesis are presented briefly in §4. It appeared advisable to summarise these arguments after stating the problem rather than to give such a summary here.

Chapter II is concerned with queueing processes that are Markovian without modification, that is, M/M/1 and M/M/N . A unified theory of these processes is given which includes cases in which side conditions are imposed on queue length. Fluctuation in queue length is represented by an imaginary particle describing a random walk on the non-negative integers. In §3 we give the solution of the random walk problem with absorbing barriers at positions 0,N and in §4 we consider the case when the barriers at these two positions are reflecting ones. These cases are of importance in the study of queueing systems in which the size of the waiting room is limited. The process considered in §5 incorporates side conditions that change the nature of the random walk at the interior point N . The solution of this problem enables us to give the Laplace transform of the probability generating function of the many server system M/M/N. Throughout this chapter we endeavour to give explicit expressions for the temporal probabilities of queue length. In many cases these formulae are too complicated for immediate application but we do not consider approximation procedures. In §6 we show how the forward Kolmogorov equations can be used to analyse the N server queue, but the main purpose of this section is to serve as an introduction to the methods used in the next chapter.

Chapter III. Amongst queueing systems whose input and service processes are independent and of renewal type, we assert that the only two of real interest are those in which the input process is (i) Poisson, (ii) deterministic. The purpose of this chapter is to study the temporal development of single server queues with these inputs and general service time distribution. The method of supplementary variables is used to obtain a generalisation of the Pollaczek-Khinchine formulae for the temporal process $E_K/G/1$ in which the renewal input is defined by an interarrival distribution of χ^2 type. Results for the two processes of interest, $M/G/1$ and $D/G/1$, are obtained by specialising these formulae. The existence of the asymptotic equilibrium distribution is established by applying an Abelian argument to the Laplace transform solution of the temporal problem. We show that the necessary and sufficient condition for a true equilibrium distribution to exist is that the probability that a busy period end in finite time be unity and that its expected duration be finite. The approach used provides an alternative to the usual method of analysing such systems by means of the Markov chain imbedded in the queueing process.

Chapter IV is devoted to the study of a special type of queueing system in which interruptions are allowed to the servicing of customers. Situations of this sort arise when allowance is made for breakdowns in the service mechanism or when the queue discipline is such that certain customers have a preemptive priority right to service. The effect of such interruptions on the queue $M/E_k/1$ is considered in §§2-4, and tables of expected queue length in the equilibrium state are given when the service distribution is (i) negative exponential, (ii) constant. In §§5 and 6 the system is generalised so that the population of customers is divided into a hierarchy of R priority classes, although in this case we specialise all service distributions to negative exponential.

Chapter V. In §1 of this chapter we discuss the input process as a model for the arrival behaviour of customers. We advance arguments in favour of the assertion made earlier in the thesis that the only forms of renewal input of real interest are those which are Poisson or deterministic. A new input model, known as general independent arrivals (GIA), is proposed in §2. This seems to describe in a more realistic way than the description afforded by the renewal model the behaviour of customers when

the arrival times are not scheduled or controlled in any way. To date we have not been able to analyse the proposed model very thoroughly and as a result only partial results are presented in §§3 and 4. The main function of this chapter is to draw attention to the need for new models of the input process and to point out that this is one of the most important current problems of queueing theory.

CHAPTER I

GENERAL REMARKS

1. Preliminary Statement

The theory of queues is concerned with the stochastic processes that arise in the study of physical systems of a special sort. These systems have as their distinctive feature a sequential input of discrete units which suffer a delay in the system before being discharged and lost. We refer to the incoming units technically as customers, although this term may not be strictly appropriate in any particular application. By the service mechanism or service facility we mean an agency which operates on customers to discharge them from the system. The term queue denotes the ensemble of customers who have entered the system but have not been discharged.

The theory has developed out of attempts to formulate mathematical models of situations in which service is provided to meet randomly arising demands. It may happen that at certain times the service facility can only satisfy the demands made on it by arriving customers if the latter are prepared to wait. If customers are not prepared to wait we have what is called a loss system, and a queue does not form. On the other hand if customers do wait (at least for a certain time) when they cannot be attended to immediately, the system is called a queueing or waiting system and a queue of waiting customers develops.

The study of the attributes of this queue constitutes the subject matter of the theory of queues. We will be concerned with only certain aspects of the subject. No comprehensive and brief account seems possible at the present time and in particular we will not discuss related questions in the theory of storage, inventory control, and the like, although some of the results we obtain have applications in these fields. This thesis is written from an applied point of view. That is, for systems that are well-defined we seek explicit representations of the quantities of interest, and we do not investigate the general theoretical questions related to the stochastic processes arising in queueing theory.

2. The Main Problems of Queueing Theory

A mathematical description of a queueing system requires knowledge of the following:

(i) The input. Successive customers arrive at times $t_1, t_2, t_3, \ldots, t_n, \ldots$, $t_n \leq t_{n+1}$, $n=1,2,3,\ldots$, which are the events of a stochastic process called the input process. It is often assumed that the input constitutes a renewal process so that the interarrival times $\tau_n = t_{n+1} - t_n$ are independently and identically distributed. The input is discussed in more detail in § 1 of Chapter V.

(ii) The service facility, consists of one or more servers (or channels or counters) which operate on customers to discharge them from the system. The duration of this operation, the service or holding time, is in general a random variable the distribution of which may not be the same for all customers or all servers. If the service time of all customers are independently and identically distributed and are independent of the input, we will say that the service process is an independent renewal process.

(iii) The queue discipline is the rule under which customers wait for service to commence when waiting is necessary. The most common queue discipline is known as first come, first served under which new arrivals await their turn for service in order of arrival.

When (i) - (iii) and possibly additional side conditions are specified the main problem of queueing theory is to find the distribution of the following quantities associated with the queue:

(iv) The queue length at time t. This is denoted by $n(t)$ or n and is defined as the number of customers waiting or being served at t. Waiting line is the term used to denote the number waiting.

(v) The waiting time of a customer, which is the time between arrival and the instant service commences. The virtual waiting time at time t is the duration a customer would have to wait if he arrived at t. We use the symbol $\eta(t)$ to denote the virtual waiting time. The waiting time of the nth arriving customer is then $\eta(t_n - 0)$.

These two are the most important of the random functions associated with a queue. Of the other quantities we mention only the busy period which is defined as a time interval in which the service facility is continuously occupied. Figure 1 at the end of this chapter illustrates $n(t)$ and $\eta(t)$ in the case of a service facility consisting of one server. The third function graphed, $\xi(t)$, is the work load submitted to the server in the interval $[0,t)$. The ordinate $\xi(t)$ is the sum of the service times of customers arriving before t.

We use the shorthand notation introduced by Kendall (1953) to denote a particular queueing system. Thus GI/G/N indicates the system in which

(a) the input constitutes a renewal process with arbitrary distribution (general independent interarrival times),

(b) the service process is of independent renewal type with general distribution,

(c) the service facility consists of N servers.

The symbol GI/G/N alone is an incomplete specification and it is also necessary to state the queue discipline and the side conditions, if any. If no side conditions, such as placing an upper limit on possible queue length or waiting time, are imposed, we speak of the unrestricted queue. If the notation GI/G/N is used without qualification then it is understood that we are referring to the unrestricted queue with first come, first served queue discipline. Queues of this type will sometimes be called renewal queueing systems or more simply, renewal queues.

A queueing process is the stochastic process $\{\underline{n}(t), t \in T\}$ whose realisations are the random functions $n(t)$, the queue length at t, and in which the index set T is the positive real axis. It is a discrete process with continuous parameter and ranges over the set of non-negative integers. On the other hand the waiting time

process $\{\eta(t), t \in T\}$ is a continuous stochastic process whose range is the non-negative real axis. When we refer to the process GI/G/N we mean the process $\{\underline{n}(t)\}$ describing fluctuations in the length of the queue associated with the system GI/G/N .

It is convenient to describe the arrival and departure of customers in the same way. We assume that the input process is a well defined stochastic process $\{\underline{a}(t), t \in T\}$ in which $\underline{a}(t)$ represents the number of arrivals up to and including time t . A realisation $a(t)$ is a left-continuous, non-decreasing random step function with jumps of unit magnitude at the arrival epochs $t_1, t_2, \ldots$. We may similarly define the departure process $\{\underline{g}(t), t \in T\}$ whose realisations $g(t)$ are left-continuous, non-decreasing step functions, the unit jumps of which occur at the epochs $\theta_1, \theta_2, \ldots$ when customers leave the system. The departure process is not independent of the input since each departure time is the sum of an arrival time and the period spent in the system.. From a knowledge of the arrival and service times one can in principle construct the departure process. An alternative formulation is to consider the point process composed of two input processes as defined above, but operating in opposite directions. If $\{\underline{a}'(t)\}$ and $\{\underline{g}'(t)\}$ are two

such processes with epochs of events t_1', t_2', ... ; $\theta_1', \theta_2', \cdots$, respectively we say that this compound process defines a queueing system provided we always have

$$t_n' \leq \theta_n' \quad , n = 1,2,3,\cdots . \tag{2.1}$$

Any two stochastic processes of the type described above can give rise to a queueing problem provided the essential condition (2.1) is fulfilled.

Let us consider a system in which the service facility caters for a population of N customers whose arrival and departure times are respectively $t_1, t_2, \cdots, t_N$, and $\theta_1, \theta_2, \cdots, \theta_N$. Let $\{\underline{a}(t)\}$ and $\{\underline{g}(t)\}$ be such that for finite N all these epochs are almost certainly finite. Then a knowledge of the input and departure processes enables us to write down the joint distribution function

$$F_N(t_1, t_2, \cdots, t_N; \theta_1, \theta_2, \cdots, \theta_N) . \tag{2.2}$$

The sample space Ω is the subset of 2N dimensional Euclidean space defined by the inequalities

$$\left.\begin{aligned} 0 \leq t_1 \leq t_2 \leq \cdots\cdots \leq t_N \\ \theta_1 \leq \theta_2 \leq \cdots\cdots \leq \theta_N \\ t_i \leq \theta_i \quad , i = 1,2,\cdots,N . \end{aligned}\right\} \tag{2.3}$$

If the number of customers is infinite then Ω is a subset of Euclidean space with a countable number of dimensions. Queue length at time t, $n(t)$, is a functional on Ω since it is a mapping from this space into the set of non-negative integers $\{n, n=0,1,2,..,N\}$. Similarly $\eta(t)$ is a mapping from Ω into the non-negative real axis. The central problem of the theory of queues is to find the distributions of $n(t)$ and $\eta(t)$ from (2.2). Strictly speaking there are two problems; firstly, a precise study and definition of the functionals $n(t), \eta(t)$, and secondly, the devising of methods which will yield their distributions.

Of $n(t)$ and $\eta(t)$ the former is the more fundamental one. The reason for this is that, if queue length is known, then the waiting time of a customer is given by summing the service times of those already waiting and the balance of the service time of one of the customers currently in service. On the other hand queue length cannot be obtained from a knowledge of the waiting time, unless of course the service time is a constant. Referring to Figure 1 at the end of this chapter we see that $n(t)$ has discontinuities at both arrival and departure epochs, whereas the jumps of $\eta(t)$ occur only at the instants customers arrive. Because $\eta(t)$ is of simpler structure than $n(t)$ it is often possible to find its distribution

by methods which fail when applied to the study of queue length. In fact for a single server queue Benes (1960a) has shown that the function $\xi(t)$ contains sufficient information to determine $\eta(t)$ uniquely. $n(t)$ cannot be obtained in this way since $\xi(t)$ contains no information relating to departure epochs. As a result more is known about the distribution of waiting times than of queue length, and even for fairly special systems little is known about the distribution of the latter at arbitrary instants of time.

The above description is perfectly general since it involves no special assumptions about the queueing system. Essentially the problem is a combinatorial one since, for fixed t, $n(t)$ is the excess of arrivals over departures in $[0,t)$. Hence to find $n(t)$ we have to count the number of these events that occurred in $[0,t)$ subject to (2.3). We also note that it is a problem in finite time. If N is infinite and the input and service processes are independent and of renewal type then under certain conditions the equilibrium distribution of $n(t)$ exists when $t\to\infty$. On the other hand it may not be sensible to talk of equilibrium distributions for other types of input and service processes.

3. Methods of Queueing Theory

Except for systems in which both interarrival and service times have the negative exponential distribution queueing processes are not Markovian, and the most widely used methods are those which restate the original problem in terms of a Markov process. Two standard ways of doing this are available when the input and service processes are independent and of renewal type. They are

(i) the method of the imbedded Markov chain,

(ii) the method of supplementary variables.

The first of these was introduced by Kendall (1951) and has since been applied to a wide variety of queueing problems (for example, Kendall, 1953; Wishart, 1956; Gaver, 1959; Winsten, 1959; and Miller, 1960). The queue is considered only at those epochs at which the Markov property holds, and the process is analysed in terms of the stochastic matrix governing transitions at these epochs. An equivalent approach due to Lindley (1952) and Smith (1953) considers the queue at arrival epochs only and leads to an integral equation of Wiener-Hopf type for the distribution of the waiting time of an arriving customer. The method of the imbedded Markov chain and the Lindley-Smith approach are particularly suited to studying distributions associated with the nth

arriving (or departing) customer as $n \to \infty$. Results for finite time are difficult to obtain by these techniques.

The method of supplementary variables consists in characterising the states of the system by vectors so that the Markov property is restored in a phase space of higher dimension. It has long been known as a means of analysing non-Markovian processes (Bartlett, 1956) but seems to have been first applied to a specifically queueing problem by Cox (1955). Strictly speaking Erlang's differential-difference equation approach is a case of this method although the term is usually applied to situations in which the supplementary variables considered are continuous. The application of this method in queueing theory is similar in many respects to techniques used in the study of stochastically developing populations with age-dependent birth and death rates. In the case of a service time with distribution function $H(x)$, $H(0+) = 0$, one defines the first order conditional probability $\mu(x)\delta x$ that a customer completes service in the interval $(x, x+\delta x)$, given that service has not been completed earlier. The formal relationship between $\mu(x)$ and $H(x)$ is

$$\mu(x)\,\delta x = \frac{H(x+\delta x) - H(x)}{1 - H(x)} + o(\delta x),$$

which on taking the limit as $\delta x \to 0$ reduces to

$$\frac{d}{dx} H(x) = [1 - H(x)] \mu(x) .$$

Hence

$$H(x) = 1 - \exp\left\{-\int_0^x \mu(u)\,du\right\} , \qquad (3.1)$$

and for the density function we have

$$dH(x) = h(x)\,dx = \mu(x) \exp\left\{-\int_0^x \mu(u)\,du\right\} dx . \qquad (3.2)$$

If the derivative of $H(x)$ does not exist at a point x we formally define $\mu(x)$ by (3.1). In this case $\mu(x)$ and $h(x)$ contain Dirac delta functions. We can also define a probability $\lambda(y)\delta y$ related to the interarrival distribution in the same way, where y is the elapsed time since the last arrival. Then for example, the queueing process GI/G/1 is Markovian if the state of the system at time t is defined by the vector (n,x,y), where x,y are as above and n denotes queue length, $n=0,1,2,..$; $0 \leq x$; $0 \leq y$. Consideration of all possible events in the interval $(t,t+\delta t)$ yields the forward Kolmogorov differential-difference equations satisfied by the transition probabilities of this Markov process. The boundary conditions when x or y are zero describe the process at arrival or departure epochs.

Erlang's method (Brockmeyer, Halstrom, and Jensen, 1948) consists of approximating to the interarrival and service distributions by members of the χ^2 family with density

$$dE_k(t) = e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} \lambda dt \,, \quad k=1,2,3,\ldots . \qquad (3.3)$$

Distributions of this type can be considered as the convolution of k negative exponential distributions each with parameter λ, so that it is possible to use the device of dividing an interarrival or service period into k fictitious phases, the time spent in each having the negative exponential distribution. The joint specification with queue length of the phase of the customers arriving and being served defines the state of a Markov process with discrete phase space. Distributions defined by (3.3) are called Erlang distributions and are denoted by the symbol E_k. If $k=1$ we have the negative exponential distribution and we write $E_1=M$, the M standing for Markov. If all interarrival and service times have the negative exponential distribution the queueing process is Markovian without modification, and in this case we will speak of the Markov queue.

Since we are interested in the state of a queue at an arbitrary instant of time it is clear that the inclusion of supplementary variables is a preferable Markovisation

procedure than the method of the imbedded Markov chain. The differential-difference equations obtained by the former govern queue behaviour for all t, and we therefore use this method for the analysis of systems in which the input and service processes are of independent renewal type. As an aid to writing down these differential-difference equations it is often convenient to state the queueing problem as a hypothetical random walk. This is particularly useful in multidimensional problems (see Chapter IV), but even for the simple Markov queues discussed in the next chapter it leads to a unified treatment of various queueing systems. Queue length at time t is represented by the position of a hypothetical particle describing a random walk on the non-negative integers. A new arrival increases $n(t)$ by unity and a departure decreases $n(t)$ by unity. The virtual waiting time $\eta(t)$ can also be described in these terms. In this case we actually have a diffusion process of a special sort, since the particle drifts towards the origin with unit velocity and jumps away from the origin at arrival epochs (see Figure 1). The magnitude of each jump is the service time of an arriving customer. Restricting the random walk by barriers placed at the origin and perhaps also at other points, describes a queueing process under various side conditions.

An important advantage in using differential-difference equations to specify the queue length probabilities $P_n(t)$ is that, provided Laplace transforms are used extensively, solutions for finite time are essentially no more difficult to obtain than those in the equilibrium state. Let the asymptotic equilibrium probabilities be

$$\lim_{t\to\infty} P_n(t) = p_n \,, \tag{3.4}$$

and let the Laplace transforms of the temporal probabilities be

$$P_n^*(s) = \mathcal{L}[P_n(t)] = \int_0^\infty e^{-st} P_n(t)\, dt \,, \quad \mathrm{Re}\, s > 0.$$

Then if the system of interest contains only Erlang distributions, the difference equations satisfied by the p_n are almost identical to those satisfied by the transforms $P_n^*(s)$, $n=0,1,2,\ldots$, and it is a simple matter to pass from the solutions of one set to the solutions of the other. With some modifications this is also true of more general systems in which the transition probabilities depend on elapsed service or interarrival times.

For systems in which the input and service processes are of independent renewal type the limit in (3.4) exists, but the asymptotic equilibrium distribution $\{p_n\}$ may not be normalised to unity. The conditions under which

$\{p_n\}$ is a true probability distribution have been established by many authors, the most general result being due to Kiefer and Wolfowitz (1955) for GI/G/N . If the transform solutions $P_n^*(s)$ are known an alternative way of proving the existence of a true equilibrium distribution is by the use of Abelian or Tauberian arguments. In this case passage to the equilibrium state is most easily effected by using an extension of Abel's Theorem (Widder, 1946, Chapter V) which yields

$$\lim_{s\to 0+} sP_n^*(s) = \lim_{t\to\infty} P_n(t) = p_n \qquad (3.5)$$

provided the limit on the right hand side exists. It is in fact sufficient for (3.5) to be true that the right hand side exists as a Cesaro limit.

It is apparent that the Markovisation precedures discussed above may be difficult to apply when the input and service are not independent renewal processes. The only work in this direction is due to Winsten (1959) who showed that the method of the imbedded Markov chain can be extended to analyse some non-renewal queueing processes. Winsten's model is discussed in Chapter V when we consider processes of this type. In §2 of this chapter it was pointed out that the problem of finding the distribution of n(t) from the input and departure processes is essentially a combinatorial one. For the unrestricted Markov queue M/M/1

Champernowne (1956) obtained the temporal distribution of queue length by a direct probability argument, but his method does not seem to generalise to more complicated systems. The main reason for this is that the argument depends on finding the distribution of the supremum of a stochastic process, a task which seems impossible for processes other than the Poisson. An alternative approach is to treat the problem as an occupancy one of a restricted nature. Intervals between successive arrivals are designated as 'boxes', the lengths of which are the interarrival times. The number of 'balls' placed in a box denotes the number of services successfully completed in an interarrival period when occupancy is restricted by the inequality (2.1). Gani (1958) has successfully used this method to study first emptiness problems in the theory of dams with Poisson input, but again the method does not seem strong enough to cope with more general systems. It is possible that the combinatorial results of Sparre Andersen and Feller (see Feller, 1959, for references) can be applied in queueing theory, but at the moment it is not clear how this can be done.

Benes (1960a, 1960b) has effectively solved the combinatorial problem associated with the waiting time process. We have seen that specification of both arrival and depart-

ure times is necessary to find $n(t)$ but that sufficient information is contained in the arrival and service times only to determine $\eta(t)$. This enables Benes to replace the joint distribution (2.2) of arrival and departure times by the distribution of the single random function $\xi(t)$. The latter can be written down immediately for any given single server queueing process and Benes shows how to obtain the distribution of $\eta(t)$ from this. The major difficulty in applying this method to find the distribution of queue length is that in this case there appears to be no simple function analogous to $\xi(t)$ that is related to $n(t)$. This is parallel to the position for renewal queueing processes, where essentially only the non-Markovian nature of the input has to be removed to treat waiting time problems (see for example Takács, 1955).

4. Summary

We have not been able to investigate successfully the general problem stated in §2 and have accordingly classified queueing systems into two groups depending on the nature of the input. In the first group are those systems in which the input constitutes a renewal process, and in the second are those for which this is not true. Most of the systems considered here belong to the first class and only in Chapter V do we consider more general systems. We assume throughout that the service times are independently and identically distributed and are independent of the input. Further, we will always assume that the service facility consists of only a single server unless specifically stated to the contrary. The argument of the thesis is briefly as follows :

(i) Amongst those systems in which the input is a renewal process only two models are of real interest in the theory of queues, namely when the input is

(a) Poisson

(b) deterministic.

The argument in favour of this conclusion has been deferred until §1 of Chapter V since it seemed preferable to present the material concerning input processes in the one chapter.

(ii) Of the standard methods available for the analysis of queueing systems with recurrent input the most suitable is the inclusion of supplementary variables. Judging by the literature this method has not been fully appreciated in the past and we hope to demonstrate that it yields many important results in a straightforward way. To this end the method is applied in Chapter III to the unrestricted process $E_k/G/1$ to obtain a generalisation of the Pollaczek-Khinchine formula. A more detailed analysis is given of the two systems of major interest, $M/G/1$ and $D/G/1$.

(iii) An important problem in queueing theory is the examination of the effects on queue behaviour of side conditions imposed on a process. No comprehensive treatment of this problem appears possible at this stage and we consider two cases. In Chapter II a unified account is given of the Markov queue with side conditions on queue length. In Chapter IV we examine the effect of a queue discipline that permits interruptions to the servicing of customers.

(iv) Probably the main current problem of queueing theory is the analysis of systems with correlated input and the specification of models which describe practical situations in a more realistic way than in the past. An input process depending on broader independence assump-

tions than those incorporated in the renewal model is put forward in Chapter V, and it is suggested that this model is applicable to a wide class of queues. To date we have not been able to analyse it very thoroughly, but the partial results obtained indicate that more information can probably be obtained in the future.

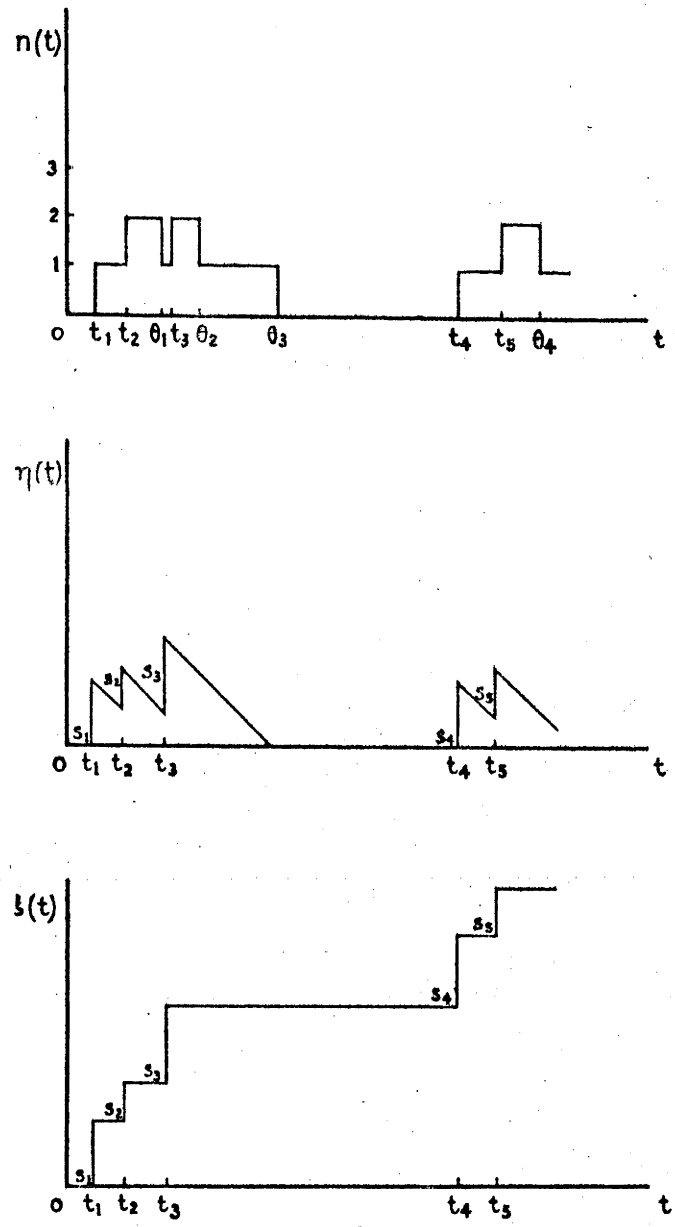

Figure 1. Illustration of $n(t)$, $\eta(t)$, and $\xi(t)$.

$t_1, t_2, \ldots$, are arrival times of successive customers.
$\theta_1, \theta_2, \ldots$, " departure " " " "
$s_1, s_2, \ldots$, " service " " " "

CHAPTER II

THE RANDOM WALK IN CONTINUOUS TIME AND ITS APPLICATION TO MARKOV QUEUES

1. Introduction

A particle will be said to describe a random walk in continuous time if it has constant chances λ, μ per unit time of taking one unit step respectively to the right or left on the real axis. For prescribed boundary conditions, such as absorbing or reflecting barriers at given positions, we ask for the distribution of the particle's position at time t after the start of the process. Specifically, we choose the origin on the real axis (as can be done without loss of generality) so that the initial position of the particle at $t=0$ is k, an integer, and for each of the processes studied we seek $P_{k,n}(t)$, the probability of a transition from k to n in time t. We use the 'backward' equation (Bartlett, 1956; Feller, 1957)

$$\frac{d}{dt} P_{kn}(t) = -(\lambda+\mu) P_{kn}(t) + \lambda P_{k+1\,n}(t) + \mu P_{k-1\,n}(t) . \qquad (1.1)$$

The probability generating function $G_k(z,t) = \sum_n z^n P_{kn}(t)$ then satisfies

$$\frac{\partial}{\partial t} G_k(z,t) = -\alpha G_k(z,t) + \lambda G_{k+1}(z,t) + \mu G_{k-1}(z,t) , \qquad (1.2)$$

where $\alpha = \lambda + \mu$. Clearly $G_k(z,0) = z^k$. Introducing the Laplace transform

$$G_k^*(z,s) = \mathcal{L}\left[G_k(z,t)\right] = \int_0^\infty e^{-st} G_k(z,t)\,dt\,, \quad \mathrm{Re}\ s > 0\,,$$

it follows from (1.2) that

$$(s+\alpha) G_k^* = \lambda G_{k+1}^* + \mu G_{k-1}^* + z^k. \tag{1.3}$$

The range of k depends on the problem under consideration. Thus for the unrestricted random walk $k = 0, \pm 1, \pm 2, \ldots$; whereas for a barrier at the origin k is restricted to the non-negative integers.

The system consisting of the backward equation, together with the side conditions, has a unique 'honest' solution ('honest' in the sense that $\sum_n P_{kn}(t) = 1$, or equivalently $G_k(1,t) \equiv 1, G_k^*(1,s) \equiv s^{-1}$); in fact 'honesty' implies uniqueness and conversely. This is well known to be the case for (1.1) without side conditions (cf. Bartlett, Feller, loc. cit.). With side conditions it can be proved to follow from the general theory of discontinuous Markov processes given in Moyal (1957) by transforming the system into an equivalent integral equation of the type considered in that paper; the explicit solution for the single absorbing boundary case is given as an example of the general theory in Moyal (1960). The method of obtaining explicit

solutions developed here is however much less laborious, especially for the more complicated side conditions, but we still rely on the general theory for the fact that the 'honesty' of each solution thus obtained implies its uniqueness. The last remark applies also to the renewal type queueing processes considered in subsequent chapters.

The Markov queueing process M/M/1, in which the density functions of the interarrival and service distributions are respectively

$$dA(t) = \lambda e^{-\lambda t} dt,$$

$$dH(t) = \mu e^{-\mu t} dt,$$

can be described in terms of such a random walk. To do this it is necessary to impose a barrier at the origin so that the random walk takes place only on the non-negative integers. Then the position of the particle at time t is the same as the number of customers in the queue at t and $P_{kn}(t)$, $k, n=0,1,2,\ldots$, is the probability that queue length changes from k to n in an interval of duration t. The nature of the barrier imposed at the origin depends on the particular quantity of interest. An absorbing barrier is appropriate if one sought the distribution of first passage times through the origin, such as the duration of a busy period. Alternately this barrier must be reflecting if it is required to find the distribu-

tion of queue length at an arbitrary instant of time. A situation of interest in queueing theory is that in which the number of customers waiting at any instant is limited to some finite value due, for example, to a waiting room of fixed capacity. Then another barrier must be imposed at N, say, in addition to the one at the origin. Whether this second barrier is absorbing or reflecting depends again on the question at issue, the former being appropriate if one seeks the distribution of the time taken for the waiting room to become full and the latter if the distribution of queue length is of interest. If the waiting room is full new arrivals do not join the queue and the number of potential customers lost for this reason is clearly of importance to the economics of a queueing system and in the theory of loss systems.

Up to the present time no unified theory has been given for the simple queue with side conditions on queue length additional to the essential barrier at the origin. By considering the general solution of the difference equation (1.3) we give such a unified theory which includes previous work on the subject as special cases and readily yields the solution for any other consistent form of boundary conditions. The solution of the random walk problem when there are absorbing barriers at 0 and N is given in §3, and in §4 we derive the solution when the barriers

at these two states are reflecting ones. In § 5 a more general process is considered in which the transition intensities λ and μ depend linearly on the particle's position in a bounded interval, and are constant outside that interval. The solution of this random walk problem gives the Laplace transform of the probability generating function to the N server queueing process M/M/N . This process has also been considered by Karlin and McGregor (1958) using a different method.

It is well-known that asymptotic equilibrium distributions exist for this type of process. We obtain them directly from the transform solution of (1.2) by using the Abelian result (3.5) of Chapter I. We mention in passing that there is another type of asymptotic behaviour of the solutions which is of interest, namely, the passage to a 'diffusion' process. This passage is effected as follows: Let

$$\alpha = \lambda+\mu = \sigma^2/h^2 \ , \ \lambda-\mu = a/h \ , \ x = nh \ , \ y = kh .$$

We assume that each step is of magnitude h , that σ and a are constants, and we look for the distribution of the displacement x conditional on y as $h \to 0$. Let

$$f(x|y,t) = \lim_{h\to 0} P_{y/h,\,x/h}^{(t)} \ .$$

Carrying out this limiting procedure in (1.1) one arrives formally at the backward diffusion equation with drift term

$$\frac{\partial}{\partial t} f(x|y,t) = \left[\frac{\sigma^2}{2} \frac{\partial^2}{\partial y^2} + \alpha \frac{\partial}{\partial y} \right] f(x|y,t). \qquad (1.4)$$

It has been shown by Heathcote and Moyal (1959) that passage to the limit in the solutions of the random walk problems yields the solutions of (1.4) with appropriate boundary conditions. This approach provides an alternative to the well-known method of images (Bartlett, 1956).

2. General Solution of the Random Walk Problem

To obtain the general solution of the difference equation (1.3) we need a particular solution and the solution of the homogeneous equation obtained by omitting the term z^k. Since the transition probabilities clearly remain invariant under translations of the origin it follows that $G_k(z,t) = z^k G_0(z,t)$ and hence

$$G_k^*(z,s) = z^k G_0^*(z,s) \quad , \quad k = 0, \pm 1, \pm 2, \ldots .$$

Substituting in (1.3) we have

$$G_0^*(z,s) = (s+\alpha-\lambda z-\mu z^{-1})^{-1} = \phi(z,s)$$

so that a particular solution of (1.3) is

$$G_k^*(z,s) = z^k \phi(z,s) . \tag{2.1}$$

Substituting $G_k^* = u^k$ in the homogeneous equation we obtain

$$u^k (s+\alpha-\lambda u-\mu u^{-1}) = 0 , \tag{2.2}$$

the two solutions of which are the zeros of $\phi(z,s)$, namely

$$\left.\begin{aligned} u_1(s) &= (2\lambda)^{-1}\left[s+\alpha-\sqrt{(s+\alpha)^2-4\lambda\mu}\,\right] \\ \text{and} \qquad & \\ u_2(s) &= (2\lambda)^{-1}\left[s+\alpha+\sqrt{(s+\alpha)^2-4\lambda\mu}\,\right] . \end{aligned}\right\} \tag{2.3}$$

The general solution of (1.2) is therefore

$$G_k^*(z,s) = z^k \phi(z,s) + A(z,s)\, u_1^k(s) + B(z,s)\, u_2^k(s). \qquad (2.4)$$

We require that $G_k^* \to 0$ as $s \to \infty$, so, since $u_1 \to 0$ and $u_2 \to \infty$ as $s \to \infty$, A must be bounded, while B must tend to 0 with $1/s$ faster than u_2^{-k}. For problems involving only one boundary we set $B \equiv 0$ so that (2.4) reduces to

$$G_k^*(z,s) = z^k \phi(z,s) + A(z,s)\, u_1^k(s), \qquad (2.5)$$

where A is determined by the boundary condition. Boundary conditions at two points determine both A and B.

If no boundary conditions are imposed we have the unrestricted random walk, the solution of which is (2.1). The Laplace transforms of the probabilities $P_{k,r}(t)$ may be readily obtained by expanding (2.1) in a Laurent series for values of z in an annulus centred on the origin, which contains the circle $|z| = 1$ for all s whose real part is positive. Clearly it is sufficient to do this for $k=0$. We have then

$$G_0^*(z,s) = \lambda^{-1}(u_2-u_1)^{-1}\left\{ \sum_{r=0}^{\infty} (u_1/z)^r + \sum_{r=1}^{\infty} (z/u_2)^r \right\}. \qquad (2.6)$$

We note that $G_k^*(1,s) = 1/s$, so that the solution is unique. The inverse Laplace transforms may be found directly from tables (Erdelyi, 1954) to give for the probabilities

$$P_{or}(t) = \beta^r e^{-\alpha t} I_{|r|}(2t\sqrt{\lambda\mu}) \quad , r = 0, \pm 1, \pm 2, \ldots , \qquad (2.7)$$

where $\beta = (\lambda/\mu)^{1/2}$ and $I_r(x)$ is the modified Bessel function of the first kind. Throughout the rest of this chapter the argument of Bessel functions will be suppressed, it being understood that it is $2t\sqrt{\lambda\mu}$. The mean and variance of the distribution (2.7) are easily found to be $(\lambda-\mu)t$ and $\alpha t = (\lambda+\mu)t$ respectively. Inverting (2.1) we have the generating function

$$G_k(z,t) = z^k \exp\left[-t(\alpha - \lambda z - \mu z^{-1})\right] \qquad (2.8)$$

which may be expanded to verify the results obtained above.

3. The Random Walk with Absorbing Barriers

If the states 0 and N , say, are absorbing barriers, the process stops whenever the particle reaches either 0 or N so the boundary conditions are

$$G_0^* = 1/s \quad , \quad G_N^* = z^N/s \quad . \tag{3.1}$$

The values of the constants A and B appropriate to this process are found by substituting the general solution (2.4) in (3.1); with these values of A and B (2.4) becomes

$$G_k^* = (z^k - u_1^k)\phi + s^{-1}u_1^k + (s^{-1}-\phi)(z^N-u_1^N)(u_2^k-u_1^k)(u_2^N-u_1^N)^{-1} \tag{3.2}$$

$$= (u_2^N-u_1^N)^{-1}\Bigg\{ s^{-1}(u_1u_2)^k(u_2^{N-k}-u_1^{N-k}) + \sum_{r=1}^{k} z^r[\lambda(u_2-u_1)]^{-1}(u_2^r-u_1^r) \times (u_2^{N-k}-u_1^{N-k})(u_1u_2)^{k-r} + \sum_{r=k+1}^{N-1} z^r[\lambda(u_2-u_1)]^{-1}(u_2^k-u_1^k)(u_2^{N-r}-u_1^{N-r}) + z^N s^{-1}(u_2^k-u_1^k) \Bigg\} . \tag{3.3}$$

Inverting the Laplace transforms gives for the probabilities

$$P_{k0}(t) = \beta^{-k}\sum_{j=0}^{\infty}\int_0^t \tau^{-1}e^{-\alpha\tau}\Big\{(2jN+k)I_{2jN+k} - [2(j+1)N-k]I_{2(j+1)N-k}\Big\}\,d\tau , \tag{3.4}$$

$$P_{kr}(t) = \beta^{r-k}\sum_{j=0}^{\infty} e^{-\alpha t}\Big\{I_{2(j+1)N-k+r} + I_{2jN+k-r} - I_{2(j+1)N-k-r} - I_{2jN+k+r}\Big\} , \tag{3.5}$$

$$r = 1,2,\cdots,k,$$

$$P_{kr}(t) = \beta^{r-k} \sum_{j=0}^{\infty} e^{-\alpha t} \left\{ I_{2(j+1)N+k-r} + I_{2jN-k+r} - I_{2(j+1)N-k-r} - I_{2jN+k+r} \right\}, \qquad (3.6)$$

$$r = k+1, k+2, \ldots, N-1,$$

$$P_{kN}(t) = \beta^{N-k} \sum_{j=0}^{\infty} \int_0^t \tau^{-1} e^{-\alpha\tau} \left\{ [(2j+1)N-k] I_{(2j+1)N-k} - [(2j+1)N+k] I_{(2j+1)N+k} \right\} d\tau. \qquad (3.7)$$

The absorption probabilities $P_{k,0}(t)$ and $P_{k,N}(t)$ are symmetrical in λ , μ and k, N-k as expected. The expected value of the position r at time t conditional on initial position k is

$$E(r|k,t) = k + (\lambda-\mu)t + (\lambda-\mu) \sum_{j=0}^{\infty} \int_0^t \tau^{-1}(t-\tau)\, e^{-\alpha\tau}$$

$$\times \left\{ \beta^{-k} \left\{ [2(j+1)N-k] I_{2(j+1)N-k} - (2jN+k) I_{2jN+k} \right\} \right.$$

$$\left. + \beta^{N-k} \left\{ [(2j+1)N+k] I_{(2j+1)N+k} - [(2j+1)N-k] I_{(2j+1)N-k} \right\} \right\} d\tau. \qquad (3.8)$$

The generating function of the stationary distribution obtained using (3.5) of Chapter I is

$$\lim_{t\to\infty} G_k(z,t) = \lim_{s\to 0+} sG_k^*(z,s) = \pi_k + z^N \gamma_k$$

$$= \begin{cases} \dfrac{(\mu/\lambda)^k - (\mu/\lambda)^N}{1-(\mu/\lambda)^N} + \dfrac{z^N[1-(\mu/\lambda)^k]}{1-(\mu/\lambda)^N} & , \lambda \neq \mu, \\ 1 - k/N + z^N k/N & , \lambda = \mu, \end{cases} \tag{3.9}$$

which yields for the expectation $E(r|k,\infty)$

$$E(r|k,\infty) = \begin{cases} N[1-(\mu/\lambda)^k][1-(\mu/\lambda)^N]^{-1} & , \lambda \neq \mu, \\ k & , \lambda = \mu. \end{cases} \tag{3.10}$$

Apart from applications to queueing theory, the boundary value problem discussed above is the continuous time analogue of the classical 'gambler's ruin' problem, in which two gamblers with initial capital k and N-k respectively compete. As may be deduced from general arguments the asymptotic results we derive are identical with those of Feller (1957, chapter XIV). The probabilities of ultimate ruin are π_k and γ_k respectively given by (3.9). Since $\pi_k + \gamma_k = 1$ ultimate absorption at either one of the boundaries is certain, and the expected lapse of time before this occurs is

$$E(t|k) = \int_0^\infty t\left[P'_{k0}(t) + P'_{kN}(t)\right] dt$$

$$= \begin{cases} \left[k - \dfrac{N\left[1-(\mu/\lambda)^k\right]}{1-(\mu/\lambda)^N}\right](\mu-\lambda)^{-1} & , \lambda \neq \mu, \\ k(N-k)(2\lambda)^{-1} & , \lambda = \mu. \end{cases} \qquad (3.11)$$

The solution of the problem with a single absorbing barrier, say at the origin, may either be found from the above by letting $N \to \infty$ or directly using (2.5). Taking the limit as $N \to \infty$ in (3.2) we have for the generating function

$$G_k^* = (z^k - u_1^k)\,\phi + s^{-1}u_1^k. \qquad (3.12)$$

The last term in the right hand side of (3.2) tends to 0 as $N \to \infty$ since $u_1 < |z| < u_2$ and $u_1 < 1$, $u_2 > 1$ for Re $s > 0$. The probabilities in this case are

$$P_{k0}(t) = k\beta^{-k}\int_0^t \tau^{-1} e^{-\alpha\tau} I_k \, d\tau, \qquad (3.13)$$

$$P_{kr}(t) = \beta^{r-k} e^{-\alpha t}\left[I_{r-k} - I_{r+k}\right], \quad r = 1,2,3,\ldots, \qquad (3.14)$$

with expectation

$$E(r|k,t) = k+(\lambda-\mu)t-(\lambda-\mu)k\beta^{-k}\int_0^t \tau^{-1}(t-\tau)\,e^{-\alpha\tau}I_k\,d\tau\,. \qquad (3.15)$$

These three equations yield incidentally the relations

$$\sum_{n=1}^{\infty}\beta^n\left[I_{n-k}-I_{n+k}\right] = \beta^k e^{\alpha t} - k\int_0^t \tau^{-1}e^{\alpha(t-\tau)}I_k\,d\tau\,, \qquad (3.16)$$

and, putting $\lambda=\mu$,

$$\sum_{n=1}^{\infty} n\left[I_{n-k}(\alpha t)-I_{n+k}(\alpha t)\right] = k e^{\alpha t}, \quad k=1,2,3,\ldots. \qquad (3.17)$$

Asymptotic results are well known (see Feller, loc. cit.) and are

$$P_{k0}(\infty) = \begin{cases} (\mu/\lambda)^k & , \lambda \geq \mu, \\ 1 & , \lambda \leq \mu, \end{cases} \qquad (3.18)$$

$$P_{kr}(\infty) = 0 \quad , \; r=1,2,3,\ldots, \qquad (3.19)$$

$$E(r|k,\infty) = \begin{cases} +\infty & , \lambda > \mu, \\ k & , \lambda = \mu, \\ 0 & , \lambda < \mu. \end{cases} \qquad (3.20)$$

Probability densities of first passage times through a barrier are given by differentiating the appropriate absorption probability. For example, in the single barrier case, the probability density of the first passage time through the origin is, from (3.13),

$$\rho_k(t)\,dt = k\beta^{-k}\,t^{-1}\,e^{-\alpha t}\,I_k\quad dt\,. \tag{3.21}$$

4. The Random Walk with Reflecting Barriers

We now take the barriers at 0 and N to be reflecting ones, in which case the boundary conditions satisfied by (2.4) are

$$(s+\lambda)G_0^* = \lambda G_1^* + 1 \quad , \quad (s+\mu)G_N^* = \mu G_{N-1}^* + z^N. \qquad (4.1)$$

Solving for the constants A and B as in § 3 we obtain the solution

$$G_k^* = \phi\left[(z^N - z^{N+1})X_k + z^k + (1-z^{-1})Y_k\right], \qquad (4.2)$$

where

$$X_k = \lambda s^{-1}\left[(u_2^{k+1} - u_1^{k+1}) - (u_1u_2)(u_2^k - u_1^k)\right](u_2^{N+1} - u_1^{N+1})^{-1},$$

$$Y_k = \lambda s^{-1}(u_1u_2)^{k+1}\left[(u_2^{N+1-k} - u_1^{N+1-k}) - (u_2^{N-k} - u_1^{N-k})\right](u_2^{N+1} - u_1^{N+1})^{-1}.$$

(4.2) may be written as

$$G_k^* = \phi\left[z^k + (1-z^{-1})u_1^{k+1}(1-u_1)^{-1}\right] + \phi X_k\left[z^N - z^{N+1} + (1-z^{-1})u_1^{N+1}\right]$$

$$\longrightarrow \phi\left[z^k + (1-z^{-1})u_1^{k+1}(1-u_1)^{-1}\right] \quad \text{as } N \to \infty. \qquad (4.3)$$

(4.3) is thus the solution for a single reflecting barrier at the origin, a result which may be checked by direct methods.

Expanding (4.2) and inverting the coefficient of z^r

yields for the probabilities

$$P_{kr}(t) = \beta^{r-k} e^{-\alpha t} I_{k-r} + \beta^{r-k} e^{-\alpha t} \sum_{j=0}^{\infty} \left[I_{2(j+1)(N+1)+k-r} + I_{2(j+1)(N+1)-k-r} \right]$$

$$+ \lambda \beta^{r-k} \sum_{j=0}^{\infty} \int_0^t e^{-\alpha\tau} \left\{ I_{2(j+1)(N+1)-k-r-2} - 2\beta^{-1} I_{2(j+1)(N+1)-k-r-1} + \beta^{-2} I_{2(j+1)(N+1)-k-r} \right.$$

$$\left. + I_{2j(N+1)+k+r+2} - 2\beta^{-1} I_{2j(N+1)+k+r+1} + \beta^{-2} I_{2j(N+1)+k+r} \right\} d\tau .$$

$$r = 0, 1, 2, \ldots, N. \qquad (4.4)$$

In order to lighten the calculations, we give the expected value for the special case k=0 only;

$$E(r|0,t) = (\lambda-\mu)t + \lambda \sum_{j=0}^{\infty} \int_0^t \tau^{-1}(t-\tau) e^{-\alpha\tau} \Phi_{0j}(\tau)\, d\tau , \qquad (4.5)$$

where

$$\Phi_{0j}(\tau) = \beta^{-3} \left\{ (2jN-1) I_{2jN-1} - [2(j+1)N+1] I_{2(j+1)N+1} \right\}$$

$$+ \beta^{-2} \left\{ 2(j+1)N I_{2(j+1)N} - 2jN I_{2jN} \right\}$$

$$+ \beta^{-(N-1)} \left\{ [(2j+1)N+1] I_{(2j+1)N+1} - [(2j+1)N-1] I_{(2j+1)N-1} \right\} .$$

Asymptotic results are obtained by applying (3.5) of Ch.I to (4.2). Thus the generating function of the stationary distribution is

$$\lim_{t\to\infty} G_{1k}(z,t) = \begin{cases} \dfrac{(\lambda-\mu)\left[z^{N+1} - (\mu/\lambda)^{N+1}\right]}{(\lambda z-\mu)\left[1 - (\mu/\lambda)^{N+1}\right]} , & \lambda \neq \mu , \\ \dfrac{1 - z^{N+1}}{(1-z)(N+1)} , & \lambda = \mu . \end{cases} \qquad (4.6)$$

and hence

$$\lim_{t\to\infty} P_{kr}(t) = \begin{cases} \dfrac{(\lambda-\mu)(\mu/\lambda)^{N-r}}{\lambda[1-(\mu/\lambda)^{N+1}]} & , \lambda \neq \mu, \\ \\ (N+1)^{-1} & , \lambda = \mu . \end{cases} \tag{4.7}$$

The expectation of r in the stationary distribution is

$$E(r,\infty) = \begin{cases} \left\{ N - \dfrac{(\mu/\lambda)[1-(\mu/\lambda)^{N}]}{1-\mu/\lambda} \right\} [1-(\mu/\lambda)^{N+1}]^{-1} & , \lambda \neq \mu, \\ \\ N/2 & , \lambda = \mu. \end{cases} \tag{4.8}$$

In the context of queueing theory the random walk considered above is the single server queue problem with the restriction that the size of the queue is N. In this case new arrivals do not join the queue if there are already $N-1$ customers waiting for service. We list below the Laplace transforms of some formulae of interest. Inversion of these results involves simple but tedious calculations which we do not carry out.

The expected length of the waiting line L at time t, conditional on initial length $k-1$ is

$$E(L|k-1,t) = E(r|k,t) + P_{k0}(t) - 1$$

and has the Laplace transform

$$\mathcal{L}[E(L|k-1,t)] = (\lambda-\mu)s^{-2} + s^{-1}[k-1+Y_k - X_k] + \frac{[u_1^k + u_1^N(1-u_1)X_k + u_2^{-1}(u_2-1)Y_k]}{\lambda(u_2-u_1)}, \tag{4.9}$$

and by (3.5), of Chapter I

$$E(L|k-1,\infty) = \left\{ N - [1-(\mu/\lambda)^N][1-\mu/\lambda]^{-1} \right\} [1-(\mu/\lambda)^{N+1}]^{-1}.$$

The distribution of the duration of a busy period, $\rho_k(t)$ with $k=1$, is obtained by considering the random walk with an absorbing barrier at the origin; i.e. with the boundary conditions

$$G_0^* = s^{-1} \quad , \quad (s+\mu)G_N^* = \mu G_{N-1}^* + z^N. \tag{4.10}$$

Proceeding as before the solution of (2.4) with these boundary conditions is

$$G_k^* = s^{-1}V_k + \phi \left[(z^N - z^{N+1})W_k + z^k - V_k \right], \tag{4.11}$$

where

$$V_k = \left[u_1^k u_2^N (u_2 - 1) + u_1^N u_2^k (1-u_1) \right] \left[u_2^N (u_2-1) + u_1^N (1-u_1) \right]^{-1}$$

and

$$W_k = (u_2^k - u_1^k) \left[u_2^N (u_2-1) + u_1^N (1-u_1) \right]^{-1}.$$

The probability that the particle will be absorbed at the origin before time t, conditional on initial state

k , is $\mathcal{L}^{-1}[V_k/s]$. Thus the probability density, $\rho_k(t)$ of the duration of a busy period has the Laplace transform

$$\mathcal{L}[\rho_k(t)] = V_k(s). \tag{4.12}$$

The formulae are simplified when $N \to \infty$ and we have the familiar single server queue problem in which no limit is placed on queue length. The appropriate generating function is given by (4.3), and the transition probabilities by

$$P_{kr}(t) = \beta^{r-k}\left\{e^{-\alpha t}I_{k-r} + \lambda\int_0^t e^{-\alpha\tau}\left[\beta^{-2}I_{k+r} - 2\beta^{-1}I_{k+r+1} + I_{k+r+2}\right]d\tau\right\}.$$
$$r = 0,1,2,\ldots. \tag{4.13}$$

(4.13) is identical with the result given by Ledermann and Reuter (1954, p.366), obtained by the use of spectral theory. This result has also been obtained by the following authors: Bailey (1954), by solving the forward equations using a generating function technique; Clarke (1956), by solving an integral equation of Volterra type; Champernowne (1956), by a direct probability argument; and Conolly (1958), by an argument similar to the one used here. The expected val-

ue of r is

$$E(r|k,t) = k + (\lambda-\mu)t + \sum_{j=0}^{\infty} \beta^{-(k+1+j)} (k+1+j) \int_0^t \tau^{-1} e^{-\alpha\tau} I_{k+1+j} \, d\tau . \qquad (4.14)$$

The probability density of the duration of a busy period, $f_k(t)$, is given by differentiating (3.13), thus

$$f_k(t)dt = k \beta^{-k} t^{-1} e^{-\alpha t} I_k \, dt . \qquad (4.15)$$

An alternative way of writing the generating function given by (4.3) is

$$G_{1k}^*(z,s) = \frac{\mu(1-z) P_{k0}^*(s) - z^{k+1}}{\lambda (z-u_1)(z-u_0)} , \qquad (4.16)$$

the transform of the null probability being

$$P_{k0}^*(s) = u_1^{k+1} [\mu(1-u_1)]^{-1} . \qquad (4.17)$$

Applying the Abelian result (3.5) of Chapter I to these expressions, we find for the equilibrium distribution

$$\lim_{s\to 0+} s P_{k0}^*(s) = \lim_{t\to\infty} P_{k0}(t) = \begin{cases} 1-\lambda/\mu & , \lambda < \mu , \\ 0 & , \lambda \geqslant \mu , \end{cases} \qquad (4.18)$$

$$\lim_{s\to 0+} s G_{1k}^*(z,s) = \lim_{t\to\infty} G_k(z,t) = \begin{cases} (1-\lambda/\mu)(1-\lambda z/\mu)^{-1} & , \lambda < \mu, \\ 0 & , \lambda \geqslant \mu . \end{cases} \qquad (4.19)$$

Hence all probabilities of the equilibrium distribution are zero if $\lambda \geqslant \mu$, and if $\lambda < \mu$ we obtain the classical result of a geometric distribution independent of the initial conditions.

5. Queues with N Servers

The Markov queue with N servers (Feller, 1957, pg. 415) is defined in our notation by the equations

$$(s+\lambda+k\mu)\,G_k^* = \lambda G_{k+1}^* + k\mu\, G_{k-1}^* + z^k \quad , k=0,1,\ldots,N, \qquad (5.1)$$

$$(s+\lambda+N\mu)\,G_k^* = \lambda G_{k+1}^* + N\mu\, G_{k-1}^* + z^k \quad , k=N,N+1,\ldots. \qquad (5.2)$$

These equations also describe a random walk on the non-negative real axis in which the side conditions specify a change in the rules governing the process at the position N . This situation is not uncommon in queueing theory. Another instance is the two server system in which arriving customers join the shorter of the two queues in front of the servers. In this case the random walk takes place on the positive quarter plane (n_1,n_2) and the side conditions are such that the nature of the walk depends on whether $n_1 > n_2$ or $n_1 < n_2$.

Before giving the solution $G_k^*(z,s)$ of the system (5.1), (5.2) we also mention briefly the generalisation of (1.3) to the case where the coefficients of the difference equation are linear functions of k ;

$$[k(a_1+b_1)+s+a_0+b_0]\,G_k^* = (ka_1+a_0)G_{k+1}^* + (kb_1+b_0)G_{k-1}^* + z^k. \qquad (5.3)$$

The homogeneous equation obtained from (5.3) is the well-known hypergeometric difference equation, and its solutions have been studied in detail by several authors (e.g. Batchelder, 1927). In fact there are twenty-four such solutions corresponding to the twenty-four solutions of the hypergeometric differential equation (Batchelder, 1927, pg. 101), and we may choose two of these appropriate to our problem, say $f_k(s)$ and $h_k(s)$. A particular solution of the non-homogeneous equation may be found by taking the Laplace transform of the generating function $G_k(z,t)$ of the original birth and death process. $G_k(z,t)$ can be found readily by standard methods and has the convoluted binomial form

$$G_k(z,t) = z^{b_0/b_1}\left\{\frac{b_1[e^{(a_1-b_1)t}-1] + z[a_1-b_1e^{(a_1-b_1)t}]}{a_1-b_1}\right\}^{k-b_0/b_1} \times \left\{\frac{[a_1e^{(a_1-b_1)t}-b_1] + za_1[1-e^{(a_1-b_1)t}]}{a_1-b_1}\right\}^{a_0/a_1-k}. \tag{5.4}$$

The general solution of (5.3) is then

$$G^*_k(z,s) = \psi_k(z,s) + A(z,s)f_k(s) + B(z,s)h_k(s), \tag{5.5}$$

where $\psi_k(z,s) = \mathcal{L}[G_k(z,t)]$. As before A and B may be chosen to satisfy given boundary conditions.

There are several interesting variants of (5.3). For example, a linear birth and death process with $a_0 = 0 = b_0$ and reflecting barriers at $k=N_1, N_2$; $N_2 > N_1$, could be used as a linearized model to investigate the logistic process (Kendall, 1949). The essential feature of Kendall's logistic model is that the coefficients are quadratic in k, so that the states N_1 and N_2 are natural reflecting barriers. Explicit solutions can be obtained for the linear boundary value problem and could prove useful as an approximation to the quadratic case.

Since (5.1) and (5.2) are special cases of (5.3) their solutions are, respectively,

$$G_k^*(z,s) = \psi_k(z,s) + A(z,s)\, v_k(s), \quad k = 0,1,\ldots,N, \tag{5.6}$$

$$G_k^*(z,s) = z^k \phi(z,s) + B(z,s)\, u_1^k(s), \quad k = N, N+1, \ldots. \tag{5.7}$$

$u_1(s)$ and $\phi(z,s)$ are the same as in previous paragraphs with $N\mu$ substituted for μ;

$$\psi_k(z,s) = \mathcal{L}\left\{\left[1-(1-z)e^{-\mu t}\right]^k \exp\left[-(\lambda/\mu)(1-z)(1-e^{-\mu t})\right]\right\}$$

$$= e^{-\lambda(1-z)/\mu} \sum_{j=0}^{k} \binom{k}{j} \frac{(z-1)^j}{\mu}\, {}_1F_1\left(s/\mu + j;\, s/\mu + j + 1;\, \lambda(1-z)/\mu\right). \tag{5.8}$$

$$\upsilon_k(s) = \int_0^1 e^{-\lambda x/\mu}\, x^k (1-x)^{s/\mu - 1}\, dx$$

$$= B(k+1, s/\mu)\, {}_1F_1(k+1; k+1+s/\mu; -\lambda/\mu), \tag{5.9}$$

where $B(m,n)$ and ${}_1F_1(m;n;x)$ are the Beta and confluent hypergeometric functions, respectively. The boundary at $k=0$ being a natural one, A and B are determined by the equation for $G_N^*(z,s)$, yielding the solution

$$G_k^*(z,s) = \begin{cases} \psi_k + \dfrac{\upsilon_k[z^{N-1}(z-u_1)\phi + u_1\psi_{N-1} - \psi_N]}{\upsilon_N - u_1\upsilon_{N-1}}, & k = 0,1,\ldots,N, \\ z^k\phi + \dfrac{u_1^{k+1-N}[z^{N-1}(z\upsilon_{N-1} - \upsilon_N)\phi + \upsilon_N\psi_{N-1} - \upsilon_{N-1}\psi_N]}{\upsilon_N - u_1\upsilon_{N-1}}, & k = N+1, N+2, \ldots. \end{cases} \tag{5.10}$$

The Laplace transform of the expected number of customers in the system at time t, given k initially, is

$$\mathcal{L}[E(r|k,t)] =$$

$$\begin{cases} \dfrac{\lambda}{\mu s} + \dfrac{\mu k - \lambda}{\mu(s+\mu)} + R_k\left\{\dfrac{\mu N - \lambda}{\lambda\mu(u_2^{-1})} + \dfrac{1}{\lambda(u_2^{-1})^2} - \dfrac{(1-u_1)(N-\lambda/\mu)+u_1}{s+\mu}\right\}, & k = 0,1,\ldots,N, \\ \dfrac{k+\lambda-\mu}{s} + S_k\left\{\dfrac{(\lambda-\mu)(\upsilon_{N-1}-\upsilon_N)}{s^2} - \dfrac{\mu\{(N-1-\lambda/\mu)\upsilon_N - (N-\lambda/\mu)\upsilon_{N-1}\}}{s(s+\mu)}\right\}, & k = N+1, N+2, \ldots. \end{cases} \tag{5.11}$$

where

$$R_k = v_k [v_N - u_1 v_{N-1}]^{-1},$$

$$S_k = u_1^{k+1-N} [v_N - u_1 v_{N-1}]^{-1}.$$

The equilibrium distribution, which exists for $N\mu > \lambda$ may be found directly from (5.10) using (3.5) of Chapter I. The calculations involved in taking the limit are considerably simplified by using the following recurrence relation for confluent hypergeometric functions

$$(x/\gamma)\,{}_1F_1(\alpha+1;\gamma+1;x) = {}_1F_1(\alpha+1;\gamma;x) - {}_1F_1(\alpha;\gamma;x). \qquad (5.12)$$

Using (5.12) repeatedly in the terms $u_1\psi_{N-1} - \psi_N$ and $v_N - u_1 v_{N-1}$, and expanding as a power series, we have

$$\lim_{t\to\infty} G_k(z,t) = p_0 \left\{ \sum_{r=0}^{N-1} \frac{(\lambda/\mu)^r z^r}{r!} + \sum_{r=N}^{\infty} \frac{(\lambda/\mu)^r z^r}{N!\, N^{r-N}} \right\}, \quad \mu N > \lambda, \qquad (5.13)$$

where

$$p_0 = \left\{ \sum_{r=0}^{N-1} \frac{(\lambda/\mu)^r}{r!} + \sum_{r=N}^{\infty} \frac{(\lambda/\mu)^r}{N!\, N^{r-N}} \right\}^{-1}.$$

(5.13) agrees with the result given by Feller (1957, p. 415).

The barrier at the origin of the system (5.1) and (5.2) is a natural reflecting one. By imposing an absorb-

ing barrier at the origin

$$G_0^* = 1/s \tag{5.14}$$

and solving (5.1), (5.2), for $k > 0$, and (5.14) for $k=0$, we can find the probability density $\rho_k(t)$, of the first passage time through the origin, i.e. the density of the duration of a busy period. Say this solution is $G_k^* = G_k^*(z,s)$, $k \geq 1$. Then

$$\rho_k(t) = \frac{d}{dt} P_{k0}(t) = \mathcal{L}^{-1}\left[sG_{1k}^*(0,s)\right]. \tag{5.15}$$

[Note that $\int_0^\infty \rho_k(t)dt = 1$ if and only if $N\mu \geq \lambda$.]

The moments of the distribution $\rho_k(t)$ may be found directly from (5.15) using the following elementary properties of the Laplace transform: If $\mathcal{L}[f(t)] = g(s)$ then

$$\left.\begin{aligned} \mathcal{L}\left[t^n f(t)\right] &= (-1)^n \frac{d^n}{ds^n} g(s), \\ \text{and} \qquad \mathcal{L}\left[\int_0^t f(\tau)d\tau\right] &= s^{-1} g(s). \end{aligned}\right\} \tag{5.16}$$

Using (5.16), the expected value of t^n conditional on k, $E(t^n|k)$, is

$$E(t^n|k) = \left[(-1)^n \frac{d^n}{ds^n}\left(sG_{1k}^*(0,s)\right)\right]_{s=0}. \tag{5.17}$$

The solution of this absorbing barrier problem follows the same lines as before, with the modification that we require the general solution of (5.1), $k > 0$, with two arbitrary constants. It is easily verified that a second solution of the homogeneous equation obtained from (5.1) is

$$\omega_k(s) = \sum_{j=0}^{k} \binom{k}{j} (\mu/\lambda)^j \Gamma(j+s/\mu) . \qquad (5.18)$$

The general solution of (5.1), $k=1,2,\ldots,N$, is then

$$G_k^*(z,s) = \psi_k(z,s) + A(z,s)\upsilon_k(s) + C(z,s)\,\omega_k(s), \qquad (5.19)$$

where ψ_k and υ_k are given by (5.8) and (5.9) respectively. C can be eliminated by substituting (5.19) in (5.14) to give

$$G_k^* = \left[\omega_k + s(\omega_0\psi_k - \omega_k\psi_0)\right](s\omega_0)^{-1} + (\omega_0\upsilon_k - \omega_k\upsilon_0)\omega_0^{-1} A . \qquad (5.20)$$

Elimination of the constants A and B from (5.20) and (5.6) gives the solution $G_k^*(z,s)$. We give the expression for the required quantity $G_k^*(0,s)$ for $k \geq N$ only:

$$G_k^*(0,s) = \frac{u_1^{k+1-N}\left\{(\omega_0\upsilon_N - \omega_N\upsilon_0)(\omega_{N-1} + s\omega_0\psi_{N-1} - s\omega_{N-1}\psi_0) - (\omega_0\upsilon_{N-1} - \omega_{N-1}\upsilon_0)(\omega_N + s\omega_0\psi_N - s\omega_N\psi_0)\right\}}{s\omega_0\left[\omega_0\upsilon_N - \omega_N\upsilon_0 - u_1(\omega_0\upsilon_{N-1} - \omega_{N-1}\upsilon_0)\right]}, \qquad (5.21)$$

where the function $\psi_j(z,s)$ appearing in (5.21) has been evaluated at $z=0$.

Application of (5.17) to (5.21) yields the moments of the duration of a busy period for $k \geqslant N$. Results for $1 \leq k < N$ may of course be obtained by the same method.

6. The Forward Equations

So far we have used the backward Kolmogorov equations satisfied by the transition probabilities $P_{kn}(t)$. It is sometimes more convenient to use the forward equations, particularly when considering more general processes in which the transition intensities λ, μ are age-dependent, that is when they depend on elapsed interarrival or service times. To illustrate the use of the forward equations we consider again the N server Markov queue with the initial conditions

$$P_{on}(0) = \delta_{on} . \tag{6.1}$$

Writing $P_n(t)$ for the transition probabilities of this process we see that the transforms $P_n^*(s)$ satisfy the equations

$$(s+\lambda) P_o^*(s) = \mu P_1^*(s) + 1 , \tag{6.2}$$

$$(s+\lambda+n\mu) P_n^*(s) = \lambda P_{n-1}^*(s) + (n+1)\mu P_{n+1}^*(s) \quad , n=1,2,\ldots,N-1, \tag{6.3}$$

$$(s+\lambda+N\mu) P_n^*(s) = \lambda P_{n-1}^*(s) + N\mu P_{n+1}^*(s) \quad , n=N, N+1, \ldots . \tag{6.4}$$

We separate the probability generating function into two parts, one generating the first N probabilities and the

other generating the probabilities when all servers are occupied;

$$\overline{J}^{*}(z,s) = \sum_{n=0}^{N-1} z^n P_n^{*}(s) ,$$

$$L^{*}(z,s) = \sum_{n=N}^{\infty} z^n P_n^{*}(s) .$$

From (6.2)-(6.4) the equations for these partial generating functions are respectively

$$\mu(1-z)\frac{\partial}{\partial z}\overline{J}^{*}(z,s) - (s+\lambda-\lambda z)\overline{J}^{*}(z,s) = z^{N-1}\left[\lambda z P_{N-1}^{*}(s) - N\mu P_N^{*}(s)\right] - 1, \quad (6.5)$$

$$L^{*}(z,s) = \frac{z^N\left[\lambda z P_{N-1}^{*}(s) - N\mu P_N^{*}(s)\right]}{\lambda(z-u_1)(u_2-z)} . \quad (6.6)$$

$u_1(s)$, $u_2(s)$ are as in (2.3) with μ replaced by μN. One of the unknowns $P_{N-1}^{*}(s)$, $P_N^{*}(s)$ can be eliminated from (6.5) and (6.6) by the following argument. Since $L^{*}(z,s)$ is a generating function it converges for at least $|z| \leq 1$, so that zeros in z within the unit circle of numerator and denominator of the right hand side of (6.6) coincide. The only zero of the denominator within the unit circle is $z = u_1(s)$. Hence equating the numerator to zero when z has this value we

find

$$N\mu P_N^*(s) = \lambda u_1(s) P_{N-1}^*(s).$$

Rewriting (6.5), (6.6) using this value of $P_N^*(s)$ we have

$$\mu(1-z)\frac{\partial}{\partial z}\bar{J}^*(z,s) - (s+\lambda-\lambda z)\bar{J}^*(z,s) = \lambda z^{N-1}(z-u_1)P_{N-1}^*(s) - 1, \qquad (6.7)$$

$$L^*(z,s) = z^N P_{N-1}^*(s)\,[u_2(s)-z]^{-1}. \qquad (6.8)$$

The Laplace transform of the probability generating function of the process is

$$G^*(z,s) = \bar{J}^*(z,s) + L^*(z,s).$$

Solution of the differential equation (6.7) then yields $G^*(z,s)$ in terms of the single unknown quantity $P_{N-1}^*(s)$.

CHAPTER III

SINGLE SERVER QUEUES WITH RECURRENT INPUT

1. Introduction

In Chapter I we distinguished between those queueing systems in which the input constitutes a renewal or recurrent process and those in which the interarrival times are not necessarily independent. We asserted that of the first class only two are of real interest, namely when the input process is

(i) Poisson

(ii) deterministic.

Single server queues with inputs (i) and (ii) and general service distribution are denoted respectively by M/G/1 and D/G/1 . They are special cases of the system $E_k/G/1$ in which the recurrent input is defined by the Erlang-k interarrival distribution

$$dE_k(t) = e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} \lambda \, dt . \qquad (1.1)$$

A unified account of M/G/1 and D/G/1 is possible if we can first find the quantities of interest associated with $E_k/G/1$. In fact the latter probably includes most single server recurrent input queues of practical importance.

The aim of this chapter is to find the distributions in continuous time of queue length and virtual waiting time of $E_k/G/1$. These results will then be specialised to yield the required distributions for M/G/1

and D/G/1 . The main result we obtain is a generalisation of the classical Pollaczek-Khinchine formula for the generating function of the queue length probabilities (equation 3.11) and from this we derive an expression for the distribution of the virtual waiting time. It is shown in §5 for M/G/1 and D/G/1 that the Laplace transform of the distribution of the duration of a busy period satisfies an equation similar to one of importance in the study of branching processes. In §6 we show by the simple Abelian argument outlined in Chapter I that the equilibrium distribution exists if and only if the probability that a busy period ends in finite time is unity and that its expected duration is finite. The conditions for the existence of the equilibrium distribution have been found by many authors but it is suggested that for the processes studied the approach used here has several advantages. The argument is a straightforward one and yields explicit formulae (which include much previous work as special cases) without difficulty. Furthermore the connection between queueing theory and the general theory of recurrent events seems more natural when use is made of first passage times and the roots of branching process type equations.

As stated in Chapter I we use the method of supplementary variables. Cox (1955) used this method to study M/G/1 in the equilibrium state and many of the techniques used here are based on his work. In this paper Cox also showed how the inclusion of supplementary variables can be used to analyse many server queueing processes. Keilson and Kooharian (1960) used this method to obtain almost complete results for the temporal queue length probabilities of M/G/1 . Other work on the temporal development of queueing systems with general service distribution has been confined to M/G/1 . For this system Takács (1955), followed by several other writers, studied the distribution of virtual waiting time, and Gaver (1959) has considered problems in finite time using the method of the imbedded Markov chain. Equilibrium results for one or both of the systems of interest have been found by several authors, sometimes by specialising the service distribution to the form (1.1). We refer specifically to Kendall (1951, 1953), Lindley (1952), Smith (1953), and Wishart (1956).

2. Equations for the Queue Length Probabilities

Let the distribution function of the service time be $H(x)$ with $H(0+)=0$, and assume that the expected service time is finite,

$$\bar{h} = \int_0^\infty t \, dH(t) < \infty.$$

The letter x will be used to denote the elapsed service time of the customer currently in service, and we define the first order conditional probability $\mu(x)\,\delta x$ of a service completion occurring in $(x, x+\delta x)$ by (3.1), (3.2) of Chapter I.

Let the recurrent input be defined by the inter-arrival distribution (1.1). Then the queueing process $E_k/G/1$ is Markovian if the states of the system are defined by the vector (m,n,x). The meaning of this is as follows: (1.1) may be considered the convolution of k negative exponential distributions each with parameter λ. Then we use the device introduced by Erlang of assuming that customers pass through k fictitious arrival phases before being admitted to the queue, the transition from phase m to phase $m+1$ having the Markov property. The letter m, $m=1,2,\ldots,k$, will be used to denote the phase of an arriving customer, and customers pass through phases $1,2,\ldots,k$ in that order. As a new arrival completes the kth phase and joins the queue the

next customer automatically enters phase 1. If n denotes the number in the queue then m and n must be specified jointly to remove the non-Markovian nature of the input. To account for the general service distribution we have to supplement knowledge of m and n by also specifying x, the elapsed service time of the current customer. Then the process whose possible states at time t are (m,n,x), $m=1,2,\ldots,k$; $n=0,1,2,\ldots$; $0 \leq x$, is a Markov process. Let the transition probabilities of this process be $P_{mn}(t,x)$. If $P_n(t)$ denotes the probability of n customers in the queue at time t, conditional on the initial number, then

$$P_n(t) = \sum_{m=1}^{k} \int_0^\infty P_{mn}(t,x)\, dx. \qquad (2.1)$$

Writing down the differential-difference equations satisfied by $P_{mn}(t,x)$ is facilitated by considering the following hypothetical random walk. A two-dimensional random walk takes place on the strip defined by the points (m,n), $m=1,2,\ldots,k$; $n=0,1,2,\ldots$. Corresponding to the queueing process $E_k/G/1$ only certain transitions are possible. Transitions $(m,n) \longrightarrow (m+1,n)$ occur with intensity λ, whereas the intensity for a unit decrease in n is $\mu(x)$. The only transition leading to an increase in n is $(k,n) \longrightarrow (1,n+1)$, since all k arrival phases have to be completed before a new arrival joins

the queue. The probability of more than one event in the interval $(t, t+\delta t)$ is $o(\delta t)$. The forward Kolmogorov equations are obtained by considering all possible transitions in $(t, t+\delta t)$. For example, if $n \geq 1$, $m=2,\ldots,k$, then

$$P_{mn}(t+\delta t, x+\delta t) = [1-(\lambda+\mu(x))\delta t]P_{mn}(t,x) + \lambda\delta t P_{m-1\,n}(t,x) + o(\delta t).$$

When $n \geq 1$ and $m=1$ we find

$$P_{1n}(t+\delta t, x+\delta t) = [1-(\lambda+\mu(x))\delta t]P_{1n}(t,x) + \lambda\delta t P_{k\,n-1}(t,x) + o(\delta t).$$

When $n=0$ the transition probabilities no longer depend on x since there is no current customer. The same argument yields for $n=0$, $m=2,..,k$,

$$P_{m0}(t+\delta t) = (1-\lambda\delta t)P_{m0}(t) + \lambda\delta t P_{m-1\,0}(t) + \delta t\int_0^\infty \mu(x)P_{m1}(t,x)dx + o(\delta t).$$

The last term on the right hand side is the first order probability that a service is completed in $(t, t+\delta t)$. These considerations lead to the equations (valid for $x > 0$)

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)P_{mn}(t,x) = -[\lambda+\mu(x)]P_{mn}(t,x) + \lambda P_{m-1\,n}(t,x),\ n\geq 1,\ m=2,\ldots,k, \quad (2.4)$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)P_{1n}(t,x) = -[\lambda+\mu(x)]P_{1n}(t,x) + \lambda P_{k\,n-1}(t,x),\ n\geq 2, \quad (2.5)$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)P_{11}(t,x) = -[\lambda+\mu(x)]P_{11}(t,x), \quad (2.6)$$

$$\frac{d}{dt}P_{m_0}(t) = -\lambda P_{m_0}(t) + \lambda P_{m-1_0}(t) + \int_0^\infty \mu(x) P_{m_1}(t,x)\,dx \,, \quad m=2,\ldots,k, \tag{2.7}$$

$$\frac{d}{dt}P_{1_0}(t) = -\lambda P_{1_0}(t) + \int_0^\infty \mu(x) P_{1_1}(t,x)\,dx. \tag{2.8}$$

If the number of waiting customers is limited to some finite number N say, it is necessary to place an additional reflecting barrier at $n=N$.

The boundary conditions at $x=0$ are found by considering the process at the instant a departure occurs. We have

$$P_{m_n}(t,0) = \int_0^\infty \mu(x) P_{m\,n+1}(t,x)\,dx \,, \quad n\geq 2 \,, \quad m=1,2,\ldots,k, \tag{2.9}$$

$$P_{1_1}(t,0) = \int_0^\infty \mu(x) P_{1_2}(t,x)\,dx + \lambda P_{k_0}(t) \,, \tag{2.10}$$

$$P_{m_1}(t,0) = \int_0^\infty \mu(x) P_{m_2}(t,x)\,dx \,, \quad m=2,\ldots,k. \tag{2.11}$$

Finally the initial conditions when $t=0$ must be specified. In general the initial state will be (m_0,n_0,x_0) so that

$$P_{m_n}(0,x) = \delta_{m m_0}\,\delta_{n n_0}\,\delta(x-x_0) \,,$$

where δ_{ij} is the Kronecker delta and $\delta(x)$ is the Dirac delta function. For simplicity we assume that the system is empty at $t=0$,

$$P_n(0) = \delta_{no} . \tag{2.12}$$

More specifically let

$$P_{m_o}(0) = \delta_{m1} . \tag{2.13}$$

An important consequence of (2.13) is that we always have $x \leqslant t$ so that the range of the integrals appearing in (2.7) - (2.12) is $(0,t)$.

We seek the generating function

$$\begin{aligned} F(z,t) &= \sum_{n=0}^{\infty} z^{kn} P_n(t) \\ &= \sum_{m=1}^{k} P_{m_o}(t) + \sum_{m=1}^{k} \sum_{n=1}^{\infty} z^{kn} \int_0^{\infty} P_{mn}(t,x)\, dx . \end{aligned} \tag{2.14}$$

It is convenient to take the dummy variable as z^k instead of the usual z .

3. Temporal Distribution of Queue Length and Waiting Time

Define the partial generating functions

$$G_m(z,t,x) = \sum_{n=1}^{\infty} z^{kn} P_{mn}(t,x) \exp\left[\int_0^x \mu(u)\,du\right] \quad , \quad m = 1,2,\ldots,k.$$

From (2.4) - (2.6) the $G_m(z,t,x)$ satisfy

$$\frac{\partial G_1}{\partial t} + \frac{\partial G_1}{\partial x} = -\lambda G_1 + \lambda z^k G_k \,, \tag{3.1}$$

$$\frac{\partial G_m}{\partial t} + \frac{\partial G_m}{\partial x} = -\lambda G_m + \lambda G_{m-1} \,, \quad m = 2,\ldots,k. \tag{3.2}$$

The cyclic nature of these equations suggests the substitution

$$G_m(z,t,x) = z^{1-m} G_1(z,t,x) \quad , \quad m = 2,\ldots,k,$$

in which case (3.1), (3.2) reduce to the single equation

$$\frac{\partial G_1}{\partial t} + \frac{\partial G_1}{\partial x} = -\lambda(1-z) G_1 .$$

This is a partial differential equation of Lagrange type whose general solution is

$$G_1(z,t,x) = e^{-\lambda(1-z)x} \Phi(z,t-x).$$

Therefore

$$\sum_{n=1}^{\infty} z^{kn} P_{mn}(t,x) = z^{1-m} \Phi(z,t-x) \exp\left\{-\lambda(1-z)x - \int_0^x \mu(u)du\right\}. \qquad (3.3)$$

The unknown function $\Phi(z,t)$ depends on the side conditions. We now derive a connection between $\Phi(z,t)$ and the null probability $P_0(t) = \sum_{m=1}^{k} P_{m0}(t)$. From (2.9) - (2.11) the conditions on $G_m(z,t,x)$ at $x=0$ are

$$G_1(z,t,0) = \Phi(z,t) = z^{-k}\int_0^\infty G_1(z,t,x)\,dH(x) + \lambda z^k P_{k0}(t) - \int_0^\infty \mu(x) P_{11}(t,x)\,dx, \qquad (3.4)$$

$$G_m(z,t,0) = z^{1-m}\Phi(z,t) = z^{-k}\int_0^\infty G_m(z,t,x)\,dH(x) - \int_0^\infty \mu(x) P_{m1}(t,x)\,dx, \; m=2,\dots,k. \qquad (3.5)$$

The equations for the null probabilities (2.8), (2.9) can now be rewritten using these last three equations;

$$\frac{d}{dt}P_{10}(t) = -\lambda P_{10}(t) + \lambda z^k P_{k0}(t) - \Phi(z,t) + z^{-k}\int_0^\infty e^{-\lambda(1-z)x}\,\Phi(z,t-x)\,dH(x), \qquad (3.6)$$

$$\frac{d}{dt}P_{m0}(t) = -\lambda P_{m0}(t) + \lambda P_{m-1,0}(t) - z^{1-m}\Phi(z,t) + z^{1-m-k}\int_0^\infty e^{-\lambda(1-z)x}\,\Phi(z,t-x)\,dH(x), \; m=2,\dots,k. \qquad (3.7)$$

Denote Laplace transforms by asterisks; thus for example

$$P^*_{m0}(s) = \mathcal{L}[P_{m0}(t)] = \int_0^\infty e^{-st} P_{m0}(t)\,dt \quad , \quad \mathrm{Re}\ s > 0.$$

An exception will be made in the case of the Laplace-Stieltjes transform of the service time distribution, for which the symbol $\psi(s)$ will be used:

$$\psi(s) = \int_0^\infty e^{-st}\, dH(t).$$

Because of the initial conditions (2.12) assumed, the integrals on the right hand sides of (3.6) and (3.7) are actually convolution integrals since $\Phi(z, t-x) = 0$ for $x > t$. For more general initial conditions $\Phi(z,-x)$ is non-zero and account would have to be taken of the age of the current customer when observation of the system commenced at $t=0$. It follows from (3.6) and (3.7) that for initial conditions (2.13), the Laplace transforms $P^*_{m0}(s)$ satisfy the difference equations

$$(s+\lambda)P^*_{10}(s) - \lambda z^k P^*_{k0}(s) = 1 - A(z,s), \qquad (3.8)$$

$$(s+\lambda)P^*_{m0}(s) - \lambda P^*_{m-1,0}(s) = -z^{1-m} A(z,s) \ , \ m = 2, \ldots, k, \qquad (3.9)$$

where

$$A(z,s) = \left[1 - z^{-k}\psi(s+\lambda-\lambda z)\right] \underline{\Phi}^*(z,s).$$

The general solution of (3.9) is

$$P^*_{mo}(s) = C\left[\lambda/(s+\lambda)\right]^{m-1} - z^{1-m} A(z,s)\,(s+\lambda-\lambda z)^{-1}.$$

The constant C is determined by substituting in (3.8)

$$C = (s+\lambda)^{k-1}\left[(s+\lambda)^k - (\lambda z)^k\right]^{-1}.$$

The transform of the null probability is therefore

$$P^*_o(s) = \sum_{m=1}^{k} P^*_{mo}(s)$$

$$= \frac{(s+\lambda)^k - \lambda^k}{s\left[(s+\lambda)^k - (\lambda z)^k\right]} - \frac{(1-z^{-k})\left[1-z^{-k}\psi(s+\lambda-\lambda z)\right]\underline{\Phi}^*(z,s)}{(1-z^{-1})(s+\lambda-\lambda z)}. \tag{3.10}$$

From (3.10) and (3.3) we obtain finally the Laplace transform of the generating function F(z,t) defined in (2.14):

$$F^*(z,s) = \frac{(1-z^k)P^*_o(s) - z^k\gamma(z,s)}{1 - z^k\left[\psi(s+\lambda-\lambda z)\right]^{-1}}, \tag{3.11}$$

where

$$\gamma(z,s) = \frac{\left\{\left[\psi(s+\lambda-\lambda z)\right]^{-1} - 1\right\}\left[(s+\lambda)^k - \lambda^k\right]}{s\left[(s+\lambda)^k - (\lambda z)^k\right]}.$$

Equation (3.11) is the basic formula we have been seeking. It is a generalisation, for initial conditions (2.12), of the classical Pollaczek-Khinchine formula, found originally for the equilibrium distribution of the process M/G/1 (see Kendall, 1951).

The unknown $P_0^*(s)$ can be found by the following argument. $F^*(z,s)$ is a generating function so that at least for $|z| \leq 1$, $\mathrm{Re}\, s > 0$, it is an analytic function of z. Then by Cauchy's Theorem

$$\int_C F^*(z,s)\, dz = 0,$$

the contour C being the unit circle. Therefore, from (3.11),

$$P_0^*(s) = s^{-1}\left[(s+\lambda)^k - \lambda^k\right]\left(I_1/I_2\right) \quad , k = 1,2,3,\ldots, \qquad (3.12)$$

where

$$I_1 = \int_C \frac{dz}{\left[(s+\lambda)^k - (\lambda z)^k\right]\left\{1 - z^k\left[\psi(s+\lambda-\lambda z)\right]^{-1}\right\}} ,$$

$$I_2 = \int_C \frac{dz}{1 - z^k\left[\psi(s+\lambda-\lambda z)\right]^{-1}} .$$

The only singularities of these two integrands within the unit circle are the zeros in z of the equation

$$\psi(s+\lambda-\lambda z) - z^k = 0 . \qquad (3.13)$$

By Rouchés Theorem (3.13) has exactly k zeros within the unit circle if $\mathrm{Re}\, s > 0$. The evaluation of $P_0^*(s)$ thus requires a knowledge of the residues of the two integrands at the k zeros of (3.13).

Let $\eta(t)$ denote the <u>virtual waiting time</u> of a customer arriving at time t, that is the time a customer has to wait if his arrival occurs at t. Queue discipline has so far been irrelevant but for what follows we assume that the 'first come, first served' rule holds. Then, if $W(t,x) = P_r\{\eta(t) \leq x\}$,

$$W(t,x) = \epsilon(x)P_0(t) + \sum_{n=1}^{\infty} \int_0^t du\, P_n(t,u) \int_0^x \frac{H^{*(n-1)}(x-v)\, h(u+v)}{1-H(u)}\, dv, \quad (3.14)$$

where $H^{*(n-1)}(x)$ is the (n-1) fold convolution of the service distribution H(x), and $\epsilon(x)$ is the unit step function

$$\epsilon(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

(3.14) is obtained in the following way: if the system is empty at t an arriving customer does not have to wait so that there is a probability mass $P_0(t)$ at x=0. On the other hand suppose that at time t there are already $n > 0$ customers in the system, the elapsed service time of the current customer being u. The probability density

of the current customer finishing service at age $u+v$ is $\left[1-H(u)\right]^{-1}h(u+v)$, and the waiting time is the sum of this partial service time and the $n-1$ service times of those customers waiting at t.

Introducing the Laplace-Stieltjes transform

$$W^{*}(t,\alpha) = \int_0^\infty e^{-\alpha x}\, dW(t,x),$$

we obtain from (3.14)

$$W^{*}(t,\alpha) = P_0(t) + \sum_{n=1}^{\infty} \psi_{(\alpha)}^{n-1} \int_0^t \frac{du\, P_n(t,u)\, e^{\alpha u}}{1-H(u)} \int_u^\infty e^{-\alpha v}\, dH(v).$$

Writing $\xi^k = \psi(\alpha)$ and using (3.3), we have

$$W^{*}(t,\alpha) = P_0(t) + \frac{(1-\xi^{-k})}{(1+\xi)} \int_0^t du\, \Phi(\xi, t-u)\, e^{-u(\lambda-\lambda\xi-\alpha)} \int_u^\infty e^{-\alpha v}\, dH(v).$$

Since the integral on the right hand side is in the form of a convolution, it is convenient to introduce the Laplace transform with respect to t,

$$W^{**}(s,\alpha) = \int_0^\infty e^{-st}\, W^{*}(t,\alpha)\, dt.$$

Hence, using (3.10) and elementary properties of the Laplace transform,

$$W^{**}(s,\alpha) = \left\{ \frac{(s+\lambda-\lambda\xi)[(s+\lambda)^k - \lambda^k]}{s[(s+\lambda)^k - (\lambda\xi)^k]} - \alpha P_0^*(s) \right\} (s+\lambda-\lambda\xi-\alpha)^{-1}. \qquad (3.15)$$

(3.15) is the basic result concerning waiting times and holds for initial conditions (2.13).

4. Results for M/G/1 and D/G/1

Results for the two processes M/G/1 and D/G/1 are obtained from the preceding formulae by substituting $k=1$ in the first case and making $k \to \infty$ in the second respectively. It is convenient at this stage to alter the notation slightly in order to exhibit explicitly the value of k we are concerned with. In the case of a process $E_k/G/1$ we write

$$P_n(t) = P_n(t,k),$$

$$F_k(z,t) = \sum_{n=0}^{\infty} z^{kn} P_n(t,k),$$

and, similarly, $W_k(t,x)$ for the waiting time distribution. For M/G/1 we have

$$F_1^*(z,s) = \frac{(1-z)P_0^*(s,1) - z(s+\lambda-\lambda z)^{-1}\{[\psi(s+\lambda-\lambda z)]^{-1} - 1\}}{1 - z[\psi(s+\lambda-\lambda z)]^{-1}}, \qquad (4.1)$$

$$W_1^{**}(s,\alpha) = [1 - \alpha P_0^*(s,1)][s+\lambda-\lambda\psi(\alpha)-\alpha]^{-1}.$$

Inverting the latter yields

$$W_1^*(t,\alpha) = e^{-t[\lambda-\lambda\psi(\alpha)-\alpha]}\left\{1 - \alpha\int_0^t e^{\tau[\lambda-\lambda\psi(\alpha)-\alpha]} P_0(\tau,1)\,d\tau\right\}. \qquad (4.2)$$

This expression for the Laplace-Stieltjes transform of the waiting time distribution was first obtained by Takács (1955) using a different method. Takác's result is more general than (4.2) since it holds when λ varies with time. When $k=1$ the expression for the transform of the null probability (3.12) simplifies to

$$P_0^*(s,1) = [s+\lambda-\lambda z_0(s)]^{-1} = \sum_{j=0}^{\infty} \lambda^j (s+\lambda)^{-(j+1)} z_0^j(s) , \qquad (4.3)$$

where $z=z_0(s)$ is the unique zero within the unit circle of (3.13) with $k=1$. If the service distribution is known, $P_0^*(s,1)$ can be evaluated explicitly by the Lagrange Inversion Formula (Whittaker and Watson, 1940, page 132).

Consider the process D/G/1 . Write $z=1-w/k$ so that

$$F_\infty^*(w,s) = \lim_{k\to\infty} \sum_{n=0}^{\infty} (1-w/k)^{nk} P_n^*(s,k)$$

$$= \sum_{n=0}^{\infty} e^{-wn} P_n^*(s,\infty) .$$

If λ is replaced by νk in the interarrival density (1.1) then as $k \to \infty$ we have arrivals occurring regularly at the instants $\nu^{-1}, 2\nu^{-1}, 3\nu^{-1}, \ldots$. Making these substitutions in (3.11) and proceeding to the limit we find the moment generating function

$$\bar{F}^{*}_{\infty}(\omega,s) = \frac{(1-e^{-\omega})P^{*}_{0}(s,\infty) - e^{-\omega}\gamma_{\infty}(\omega,s)}{1 - e^{-\omega}[\psi(s+\nu\omega)]^{-1}}, \qquad (4.4)$$

where

$$\gamma_{\infty}(\omega,s) = \frac{(1-e^{-s/\nu})\{[\psi(s+\nu\omega)]^{-1} - 1\}}{s\,[1 - e^{-(\omega+s/\nu)}]} .$$

Writing $e^{-w}=z$ yields the probability generating function. The result for the waiting time distribution is, from (3.15),

$$W^{**}_{\infty}(s,\alpha) = \left\{\frac{(1-e^{-s/\nu})[s-\nu\log\psi(\alpha)]}{s[1-\psi(\alpha)e^{-s/\nu}]} - \alpha P^{*}_{0}(s,\infty)\right\}[s-\alpha-\nu\log\psi(\alpha)]^{-1}. \qquad (4.5)$$

The unknown $P^{*}_{0}(s,\infty)$ can be found by essentially the same argument as before. $F^{*}_{\infty}(w,s)$ converges for at least $\mathrm{Re}\ w \geq 0$, $\mathrm{Re}\ s > 0$, so that zeros in w of numerator and denominator coincide in this region. The zeros of the denominator are those of

$$e^{-\omega} - \psi(s+\nu\omega) = 0, \qquad (4.6)$$

and it is in fact sufficient to consider this equation for real w , $0 \leq w < \infty$. In the neighbourhood of the origin $e^{-\omega} > \psi(s+\nu\omega)$, and furthermore $\psi(s+\nu\omega)$ is a convex function and $\lim_{\omega\to\infty} \psi(s+\nu\omega) = 0$. This implies that (4.6) has at most one root on the finite part of the positive real axis. If $\psi(s)$ is a meromorphic function then there is exactly one root since in that case there exists a $w=w_0(s)$ such that

$$e^{-\omega} > \psi(s+\nu\omega) \qquad \text{for} \quad \omega < \omega_0$$

and

$$e^{-\omega} < \psi(s+\nu\omega) \qquad \text{for} \quad \omega > \omega_0 .$$

Then $w_0(s)$ is the required unique root. On the other hand if no restriction is placed on $\psi(s)$ other than that it be the Laplace-Stieltjes transform of a probability distribution, then it is easy to construct examples where (4.6) has no positive root. An obvious (if trivial) instance is when the service time is also deterministic of duration μ^{-1} say. Then

$$\psi(s+\nu\omega) = \exp\left[-(s+\nu\omega)\mu^{-1}\right].$$

and the left hand side of (4.6) does not vanish unless $\mu > \nu$. It will be shown later that if the service distribution is such that (4.6) has no positive root for $\text{Re } s > 0$ then queue length increases indefinitely and the

equilibrium distribution does not exist. For this reason we confine our attention to those service distributions for which equation (4.6) has exactly one positive root $\omega = \omega_0(s)$. Included in this class are distributions whose Laplace-Stieltjes transform is a meromorphic function. Equating the numerator of the right hand side of (4.4) to zero when $\omega = \omega_0(s)$ we find

$$P_0^*(s,\infty) = \frac{(1-e^{-s/\nu})\, e^{-\omega_0}}{s\left[1-e^{-(\omega_0+s/\nu)}\right]} . \tag{4.7}$$

Moments are obtained by differentiating the appropriate generating function. For M/G/1, D/G/1 denote expected queue length at time t, conditional on the null state initially, by $E_1[n(t)|0]$, $E_\infty[n(t)|0]$ respectively, and similarly $E_1[\eta(t)|0]$, $E_\infty[\eta(t)|0]$ for the expected virtual waiting time. We find

$$\mathcal{L}\{E_1[n(t)|0]\} = \lambda s^{-2} - \psi(s)[1-\psi(s)]^{-1}[s^{-1} - P_0^*(s,1)], \tag{4.8}$$

$$\mathcal{L}\{E_\infty[n(t)|0]\} = e^{-s/\nu}[s(1-e^{-s/\nu})]^{-1} - \psi(s)[1-\psi(s)]^{-1}[s^{-1} - P_0^*(s,\infty)]. \tag{4.9}$$

The transforms of the expected waiting times can be inverted to yield

$$E_1[\eta(t)|0] = \lambda \bar{h} t - t + \int_0^t P_0(\tau,1)\, d\tau, \tag{4.10}$$

$$E_\infty[\eta(t)|0] = u(\nu^{-1},t)\bar{h} - t\int_0^t P_0(\tau,\infty)\,d\tau, \qquad (4.11)$$

where $u(\nu^{-1},t)$ is the step function

$$u(\nu^{-1},t) = n \quad , n\nu^{-1} \le t < (n+1)\nu^{-1};\ n=0,1,2,\ldots.$$

The last two formulae permit the following interpretation (c.f. Benes, 1960b). The first term on the right hand sides is a temporal analogue of the traffic intensity $\rho = \lambda k^{-1}\bar{h}$ and is a measure of the rate at which arriving customers are serviced. The virtual waiting time is reduced by the server at a uniform rate and the term $t - \int_0^t P_0(\tau,\cdot)\,d\tau$ is just that part of the interval $[0,t)$ for which the server is occupied. Informally speaking we may then write the right hand sides of (4.10), (4.11) as

$$\left\{\text{Expected service time of customers arriving in } [0,t)\right\} - \left\{\text{Expected busy time of server in } [0,t)\right\}.$$

5. Busy Period Distributions

A busy period is a period of time in which the server is continuously occupied. The distribution of this period can be found by solving the system (2.4) - (2.13) modified by assuming (i) the initial condition that the first customer has just commenced service, and (ii) an absorbing barrier at $n=0$. The required cumulative distribution function of a busy period is the null probability of this modified system. To avoid confusion re-label the transition probabilities of this process as $Q_{mn}(t,x)$, so that the null probability we seek is $Q_0(t)$. The modifications are the following: (2.7), (2.8) are replaced by the single equation

$$\frac{d}{dt}Q_{m0}(t) = \int_0^\infty \mu(x)\, Q_{m1}(t,x)\,dx \quad , m=1,2,\cdots,k. \tag{5.1}$$

Boundary conditions (2.9), (2.11) still hold but (2.10) is replaced by

$$Q_{11}(t,0) = \int_0^\infty \mu(x)\, Q_{12}(t,x)\,dx \qquad \text{if } t>0, \tag{5.2}$$

$$Q_{11}(0,0) = 1 .$$

By the same procedure as before we find, instead of (3.10),

$$z^{-k}(1-z^{-k})(1-z^{-1})^{-1}\Phi^*(z,s) = \left[sQ_0^*(s)-z^k\right]\left[\psi(s+\lambda-\lambda z)-z^k\right]^{-1}, \tag{5.3}$$

and, instead of (3.11),

$$F^*_{(z,s)} = \sum_{n=0}^{\infty} z^{kn} Q^*_n(s)$$

$$= \frac{\left\{\lambda z^k(1-z) - [s(1-z^k) + \lambda(1-z)]\psi(s+\lambda-\lambda z)\right\} Q^*_0(s) + z^{2k}[1-\psi(s+\lambda-\lambda z)]}{(s+\lambda-\lambda z)[z^k - \psi(s+\lambda-\lambda z)]} . \tag{5.4}$$

When $k \longrightarrow \infty$, (4.4) is replaced by

$$F^*_\infty(\omega,s) =$$

$$\frac{\left\{\nu\omega e^{-\omega} - [\nu\omega + s(1-e^{-\omega})]\psi(s+\nu\omega)\right\} Q^*_0(s,\infty) + e^{-2\omega}[1-\psi(s+\nu\omega)]}{(s+\nu\omega)[e^{-\omega} - \psi(s+\nu\omega)]} . \tag{5.5}$$

The required transforms $Q^*_0(s,1)$, $Q^*_0(s,\infty)$, obtained by the same argument as in §4 and with the same restriction when the input is deterministic, are

$$\begin{aligned} Q^*_0(s,1) &= s^{-1} z_0(s) , \\ Q^*_0(s,\infty) &= s^{-1} e^{-w_0(s)} , \end{aligned} \tag{5.6}$$

where $z_0(s)$, $w_0(s)$ are the unique zeros of (3.13) with $k=1$ and (4.6) respectively. The probability density functions of the duration of a busy period are then respectively

$$\begin{aligned} \frac{d}{dt} Q_0(t,1) &= \mathcal{L}^{-1}[z_0(s)] , \\ \frac{d}{dt} Q_0(t,\infty) &= \mathcal{L}^{-1}[e^{-w_0(s)}] . \end{aligned} \tag{5.7}$$

The connection between the null probability of the queueing process and the duration of a busy period through the zeros $z_0(s)$, $w_0(s)$, is clearly shown by substituting the appropriate member of (5.6) in (4.3), (4.7) . An interpretation of the null probability $P_0(t,\cdot)$ is that it is the distribution function of the time taken for the process to be in the null state after completing j busy periods, irrespective of j . (4.3) and (4.7) are then instances of the formulae for the passage times of renewal processes (Bartlett, 1956, § 3.3).

The probability that a busy period will end in finite time is similar to the probability of ultimate extinction of a stochastically developing population. The equations determining $z_0(s)$, $w_0(s)$ are essentially the same as the branching process equation, the smallest root of which is the probability of ultimate extinction (see Feller, 1957, Ch. XII; Harris, 1948). The parallel situation here is that we are concerned with the behaviour as $t \to \infty$ of $Q_0(t,\cdot)$, or equivalently as $s \to 0+$ of $z_0(s)$ and $w_0(s)$. Consider first D/G/1, and again assume that the service distribution is such that (4.6) has exactly one positive root $w=w_0(s)$. As

s decreases along the real axis $\psi(s+\nu w)$ increases so that $w_0(s)$ decreases. Thus the limit

$$\lim_{s \to 0+} w_0(s) = w_0(0)$$

exists and, by continuity, is the largest non-negative root of the equation

$$e^{-w} - \psi(\nu w) = 0. \tag{5.8}$$

Since $\psi(s)$ is the Laplace-Stieltjes transform of a probability distribution, (5.8) has at least one non-negative root, since the left hand side always vanishes when $w=0$. If $w_0'(0) > 0$ the convexity of $\psi(\nu w)$ implies

$$\left[-\frac{d}{dw} \psi(\nu w) \right]_{w=0} = \nu \bar{h} > \left[-\frac{d}{dw} e^{-w} \right]_{w=0} = 1,$$

and conversely. On the other hand, if $\nu \bar{h} \leq 1$ then

$$e^{-w} \leq \psi(\nu w) \tag{5.9}$$

for all $w > 0$, with equality only holding when $\psi(s) = e^{-s/\nu}$. This corresponds to a deterministic service of the same duration as the interarrival interval. If this trivial case is excluded, then the inequality in (5.9) is strict, and for $\nu \bar{h} \leq 1$ we have $w_0(0)=0$. We can now apply the extension of Abel's theorem quoted previously (Chapter I,

equation 3.5) to obtain

$$\lim_{t\to\infty} Q_0(t,\infty) = \lim_{s\to 0+} s Q_0^*(s,\infty) = \begin{cases} 1 & \text{if } \nu\bar{h} \le 1, \\ e^{-w_0(0)} & \text{if } \nu\bar{h} > 1. \end{cases}$$

We have proved the following:

Theorem 1: If the service distribution is such that the equation (4.6) has a non-negative root, the probability that a busy period of the process D/G/1 ends in finite time is

$$\lim_{t\to\infty} Q_0(t,\infty) = q_0(\infty) = e^{-w_0(0)}, \qquad (5.10)$$

where $w_0(0)$ is the largest non-negative root of (5.8). If $\rho = \nu\bar{h} \le 1$ then $q_0(\infty) = 1$, and if $\rho > 1$ then $q_0(\infty) < 1$.

The quantity $\rho = \nu\bar{h}$ is known as the traffic intensity. When $\rho = 1$ it is easy to show that the expected duration of a busy period is infinite, although normalisation of its distribution is to unity. To see this observe that if $\rho = 1$ (5.8) has a multiple root at the origin so that the expected duration

$$\left[-\frac{d}{ds} s Q_0^*(s,\infty) \right]_{s=0}$$

does not exist. When $\rho<1$ this expectation is

$$E(t) = \overline{h}(1-\rho)^{-1}. \tag{5.11}$$

This result may be stated as a corollary to Theorem 1:

Corollary: If $\rho<1$ the expected duration of a busy period is given by (5.11). If $\rho=1$ the expected duration is infinite.

We now consider the other process of interest, M/G/1 . In this case additional information on busy periods can be obtained by exploiting further the analogy with population extinction problems. Kendall (1951, page 168) has previously noted that not only does the duration of a busy period correspond to the time to extinction of a population with a single ancestor, but also that the number of customers served in this duration corresponds to the cumulated population total. By the same sort of argument used in branching process theory (see Harris, 1948; Otter, 1949) we now derive an expression for the joint probability $B_j(x)$ that j customers are served in a busy period and that its duration is $\leq x$, $j=1,2,3,\ldots$; $0 \leq x$. Denote the two marginal probabilities by

$$B_j = \int_0^\infty dB_j(x) = Pr\{j \text{ customers served in busy period}\}$$

and

$$B(x) = \sum_{j=1}^{\infty} B_j(x) = Pr\{\text{duration of busy period} \leq x\}.$$

The queue discipline is irrelevant to the development of a busy period so that the number of new arrivals occurring in a service period is analogous to the number of offspring produced by a single individual of a stochastically developing population. If a service period is of duration y, the probability that n customers arrive whilst this service is in progress is, for M/G/1,

$$a_n(y) = e^{-\lambda y} \frac{(\lambda y)^n}{n!}.$$

Then

$$B_1(x) = \int_0^x a_0(y)\, dH(y),$$

since the joint probability $B_1(x)$ is the probability that no new arrivals occur during the service period of the initial customer which is of duration $\leq x$. By direct enumeration we find the further terms

$$B_2(x) = \int_0^x B_1(x-y)\, a_1(y)\, dH(y),$$

$$B_3(x) = \int_0^x B_2(x-y) a_1(y)\, dH(y) + \int_0^x B_1^{*(2)}(x-y)\, a_2(y)\, dH(y),$$

$$B_4(x) = \int_0^x B_3(x-y)\, a_1(y)\, dH(y) + \int_0^x \{B_1 * B_2 + B_2 * B_1\}_{(x-y)}\, a_2(y)\, dH(y)$$
$$+ \int_0^x B_1^{*(3)}(x-y)\, a_3(y)\, dH(y),$$

where $B_j^{*(n)}(x)$ is the n fold convolution of $B_j(x)$ and $\{B_i * B_j\}_{(x)}$ is the convolution of $B_i(x)$ and $B_j(x)$. The expressions for $B_j(x)$ are simplified on using the Laplace-Stieltjes transforms

$$\alpha_n(s) = \int_0^\infty e^{-sx}\, a_n(x)\, dH(x),$$

$$B_j^*(s) = \int_0^\infty e^{-sx}\, dB_j(x).$$

We obtain

$$B_1^*(s) = \alpha_0(s),$$

$$B_j^*(s) = \sum_{n=1}^{j-1} \alpha_n(s) \sum_{\substack{r_1+r_2+\cdots+r_n \\ = j-1}} B_{r_1}^*(s)\, B_{r_2}^*(s) \cdots B_{r_n}^*(s), \quad j = 2, 3, \ldots.$$

Hence the generating function

$$G^*(z,s) = \sum_{j=1}^{\infty} z^j B_j^*(s) \tag{5.12}$$

satisfied the functional equation

$$G^*(z,s) = z \sum_{n=0}^{\infty} \alpha_n(s) \left[G^*(z,s)\right]^n$$
$$= z\, \psi\left[s + \lambda - \lambda G^*(z,s)\right].$$

(5.13)

In the notation used previously

$$G^*(z=1,s) = s\,Q_0^*(s,1) = z_0(s).$$

The probability that a busy period ends in finite time is

$$\lim_{s\to 0+} G^*(1,s) = \lim_{s\to 0+} z_0(s) = q_0(1). \tag{5.14}$$

Corresponding to Theorem 1 we now have

Theorem 2: If, for the process M/G/1, $G^*(z,s)$ is the generating function defined by (5.12) then $u=G^*(z,s)$, $u_0=G^*(z,0)$, $u_1=G^*(1,s)$ are respectively the unique zeros within the unit circle of the equations

$$z\psi(s+\lambda-\lambda u) - u = 0 \quad , \mathrm{Re}\, s>0,\ |z|<1, \tag{5.15}$$

$$z\psi(\lambda-\lambda u_0) - u_0 = 0 \quad , \ |z|<1, \tag{5.16}$$

$$\psi(s+\lambda-\lambda u_1) - u_1 = 0 \quad , \quad \mathrm{Re}\, s>0. \tag{5.17}$$

Furthermore, the probability $q_0(1)$ defined by (5.14) is the smallest positive root of the equation

$$\psi(\lambda-\lambda q_0) - q_0 = 0. \tag{5.18}$$

If $\rho=\lambda\overline{h}>1$ then $q_0(1)<1$, and if $\rho\le 1$ then $q_0(1)=1$.

Proof of the theorem follows directly by the use of Rouches Theorem and well-known results from the theory of branching processes, and indeed Takács (1960) has stated and proved a very similar theorem in his study of the dual process GI/M/1 . (5.15) has also been found by Gaver (1959) using the method of the imbedded Markov chain, and some of the other results embodied in the theorem have been obtained by several authors, of whom we mention only Takacs (1955). When $\rho=1$ the situation is the same as for D/G/1 and the root at $q_0=1$ of (5.18) is a multiple one. The corollary to Theorem 1 applies also to Theorem 2, and in fact the expected number of customers served in a busy period is

$$E(j) = (1-\rho)^{-1}.$$

In the terminology of the theory of recurrent events the null state is certain but has infinite mean recurrence time when $\rho=1$.

6. Equilibrium Results

Denote the queue length probabilities of the equilibrium distribution by

$$\lim_{t\to\infty} P_n(t,k) = p_n(k),$$

with generating function

$$F_k(z) = \sum_{n=0}^{\infty} z^{kn}\, p_n(k).$$

In §5 we proved that if $\rho<1$, return to the origin in finite expected time is an event with probability one. We now show that this condition is necessary and sufficient for the existence of the equilibrium distribution of queue length. Applying the Abelian result used before to (4.1) and (4.4) we have

$$\lim_{s\to 0+} s F_1^*(z,s) = \lim_{t\to\infty} F_1(z,t) = \frac{(1-z)\,p_0(1)}{1 - z/\psi(\lambda-\lambda z)}, \qquad (6.1)$$

$$\lim_{s\to 0+} s F_\infty^*(\omega,s) = \lim_{t\to\infty} F_\infty(\omega,t) = \frac{(1-e^{-\omega})\,p_0(\infty)}{1 - e^{-\omega}/\psi(\nu\omega)}. \qquad (6.2)$$

The limits on the right hand side exist but the distributions defined by them may not be normalised to unity.

Theorem 3: The equilibrium distribution defined by (6.1), (6.2) are normalised to unity and are independent of the initial conditions if and only if $\rho<1$. If $\rho \geqslant 1$ the probabilities $p_n(\cdot)$ are identically zero. If the equilibrium distribution exists the necessary and sufficient condition for the existence of its rth moment, $r=1,2,\ldots$, is that the $(r+1)$th moment of the service time distribution exist.

Proof: Consider first D/G/1 . Application of Theorem 1 and its corollary to (4.7) yields

$$\lim_{s\to 0+} s P_0^*(s,\infty) = \lim_{s\to 0+} \frac{(1-e^{-s/\nu})\, s\, Q_0^*(s,\infty)}{1 - e^{-s/\nu} s\, Q_0^*(s,\infty)}$$

$$= \begin{cases} 0 & \text{if} \quad \rho \geqslant 1, \\ 1-\rho & \text{if} \quad \rho < 1. \end{cases}$$

Therefore

$$\lim_{s\to 0+} s F_\infty^*(\omega,s) = F_\infty(\omega) = \begin{cases} 0 & \text{if} \quad \rho \geqslant 1, \\ \dfrac{(1-e^{-\omega})(1-\rho)}{1-e^{-\omega}/\psi(\nu\omega)} & \text{if} \quad \rho < 1. \end{cases} \tag{6.3}$$

The non-zero member of the right hand side is easily found to be normalised to unity, and conversely if $F_\infty(0) = 1$ then $\rho < 1$. The initial conditions are contained in the term $e^{-\omega}\gamma_\infty(\omega,s)$ of (4.4) and

$$\lim_{s\to 0+} s e^{-\omega}\gamma_\infty(\omega,s) = 0$$

independently of the value of ρ . To prove the second half of the theorem we obtain a formula connecting the first r moments of the queue length distribution with the first (r+1) moments of the service distribution. Let the moments of the two distributions be respectively

$$m_r(\infty) = \left[(-1)^r \frac{d^r}{d\omega^r} F_\infty(\omega)\right]_{\omega=0}$$

$$h_r = \left[(-1)^r \frac{d^r}{ds^r} \psi(s)\right]_{s=0} .$$

Differentiating the generating function r+1 times we find

$$m_r(\infty) = \frac{(1-\rho)\sum_{j=0}^{r}\binom{r+1}{j}\nu^j h_j - \sum_{j=2}^{r+1}\binom{r+1}{j}(1-\nu^j h_j)\, m_{r+1-j}^{(\infty)}}{(r+1)(1-\rho)} \quad , r=1,2,\ldots . \quad (6.4)$$

The proof follows by induction on r .

The same argument holds for M/G/1 by the application of Theorem 2 to (4.1). Corresponding to (6.4), the formula for the rth <u>factorial moment</u> $m_{[r]}(1)$ is

$$m_{[r]}(1) = \lambda^r h_r + \frac{\sum_{j=2}^{r+1} \binom{r+1}{j} \lambda^j h_j \, m_{[r+1-j]}^{(1)}}{(r+1)(1-\rho)} \quad , r=1,2,\ldots . (6.5)$$

The above theorem is our equivalent to Theorem 4 of Kendall (1951, pg. 161) in which the equilibrium behaviour of the queue is classified in terms of the ergodic properties of the imbedded Markov chain.

It was stated in §4 that when the input is deterministic we consider only the class of service distributions for which equation (4.6) has exactly one positive root. The reason for this restriction is now apparent. $\Psi(s+\nu w)$ is defined only for $\mathrm{Re}\ w \geqslant -\nu^{-1}\,\mathrm{Re}\ s$. Then as $s \to 0+$ the non-existence of a positive root $w_0(s)$ implies

$$\left[-\frac{d}{dw}\Psi(\nu w)\right]_{w=0} = \nu\bar{h} > \left[-\frac{d}{dw}e^{-w}\right]_{w=0} = 1 .$$

Thus return to the origin is not certain and the equilibrium distribution does not exist.

The distributions of virtual waiting time in the equilibrium state clearly exist under the conditions of Theorem 3. From (4.2), (4.5) the Laplace-Stieltjes transform of the two distributions are respectively

$$W_1^*(\alpha) = (1-\rho)\left\{1-(\lambda/\alpha)\left[1-\psi(\alpha)\right]\right\}^{-1}, \tag{6.6}$$

$$W_\infty^*(\alpha) = (1-\rho)\left\{1+(\nu/\alpha)\log\psi(\alpha)\right\}^{-1}. \tag{6.7}$$

All the formulae derived here apply to the state of the process observed at an arbitrary instant of time. For M/G/1 these results are asymptotically the same as those obtained by considering the queue only at arrival epochs (Takács, 1955), and in fact (6.1), (6.6) are the original Pollaczek-Khinchine formulae. This is not the case for the process D/G/1 and we refer to the papers previously cited of Kendall, Lindley, Smith, and Wishart, for the results relating to the nth customer as $n \to \infty$.

Smith (1953) has shown that the service distribution plays an important role in determining the analytical character of the waiting time distribution. This point is borne out for the process $E_k/G/1$ by the formulae derived

in this chapter. It is apparent that the influence of the service distribution is exerted through the zeros of the branching process type equation (3.13) since these zeros essentially determine the stochastic properties of the queue. In the case of finite time, both the generating function of the queue length probabilities and the distribution of virtual waiting time are expressed in terms of the null probability which is itself given by (3.12) and so depends on the zeros of (3.13). For M/G/1 and D/G/1 we have just seen that it is the asymptotic behaviour of the smallest zero of (3.13) with $k=1$ and (4.6) respectively that determines the existence of the equilibrium distribution.

7. Example

The simplest example is when the service distribution is negative exponential,

$$dH(t) = \mu e^{-\mu t} dt, \tag{7.1}$$

so that

$$\psi(s) = \mu(s+\mu)^{-1}.$$

With this value of $\psi(s)$ formulae (4.1), (4.3) reduce to (4.16), (4.17) of Chapter II respectively for the process M/M/1 with initial state k=0 ;

$$F_1^*(z,s) = \frac{\mu(1-z)P_0^*(s,1) - z}{\lambda z^2 - z(s+\lambda+\mu) + \mu}, \tag{7.2}$$

$$P_0^*(s,1) = [s+\lambda-\lambda z_0(s)]^{-1} = u_1(s)[\mu(1-u_1(s))]^{-1}. \tag{7.3}$$

The second expression for the right hand side follows by noting that $z_0(s)$ is identical to the root $u_1(s)$ defined in (2.3) of Chapter II, and then by using the relationships

$$s+\lambda+\mu = \lambda(u_1+u_2),$$
$$\mu = \lambda u_1 u_2 .$$

The Laplace transform of the distribution of virtual waiting time is from (4.2)

$$W_1^*(t,\alpha) = e^{-t\alpha(\lambda-\mu-\alpha)(\alpha+\mu)^{-1}}\left[1-\alpha\int_0^t e^{\tau\alpha(\lambda-\mu-\alpha)(\alpha+\mu)^{-1}} P_0(\tau,1)\,d\tau\right]. \quad (7.4)$$

For the busy period of M/M/1 we find from Theorem 2,

$$\sum_{j=1}^{\infty} z^j\, dB_j(x) = \mu e^{-x(\lambda+\mu)} \sum_{j=1}^{\infty} \frac{z^j (\lambda\mu x^2)^{j-1}}{j!\,(j-1)!}\, dx$$

$$= \left(\sqrt{\frac{\mu z}{\lambda}}\right) \frac{e^{-x(\lambda+\mu)}}{x} I_1(2x\sqrt{\lambda\mu z})\, dx, \quad (7.5)$$

where $I_1(x)$ is the modified Bessel function of the first kind. The marginal distributions of the duration of a busy period and the number of customers served are respectively

$$dB(x) = \left(\sqrt{\frac{\mu}{\lambda}}\right) \frac{e^{-x(\lambda+\mu)}}{x} I_1(2x\sqrt{\lambda\mu})\, dx, \quad (7.6)$$

$$B_j = \frac{(2j-2)!}{j!\,(j-1)!} \frac{\lambda^{j-1}\mu^j}{(\lambda+\mu)^{2j-1}}, \; j=1,2,3,\ldots. \quad (7.7)$$

(7.6) is identical to (4.15) of Chapter II with k=1 .

From (4.4) the Laplace transform of the moment generating function of D/M/1 is

$$F^*_\infty(w,s) = \frac{\mu(1-e^{-w})P^*_0(s,\infty) - (s+\nu w)e^{-w}(1-e^{-s/\nu})}{\mu - (s+\mu+\nu w)e^{-w}} . \qquad (7.8)$$

The difficulty in this case lies in specifying $P^*_0(s,\infty)$. We require the largest non-negative root $w_0(s)$ of the transcendental equation (c.f. 4.7)

$$e^w - \rho w = 1 + s/\mu . \qquad (7.9)$$

Although $w_0(s)$ cannot be found in closed form some of the general points made earlier in the chapter may be illustrated in the following way. Let the Taylor expansion of $w_0(s)$ about $s=0$ be

$$w_0(s) = a_0 + a_1 s + a_2 s^2 + a_3 s^3 + \cdots .$$

Differentiating (7.9) we find for the first few terms

$$w_0(s) = a_0 + \frac{s}{\mu(e^{a_0}-\rho)} - \frac{s^2 e^{a_0}}{\mu^2(e^{a_0}-\rho)^3} + \frac{s^3 e^{a_0}(2e^{a_0}+\rho)}{\mu^3(e^{a_0}-\rho)^5} - \cdots , \qquad (7.10)$$

where a_0 satisfies

$$e^{a_0} = 1 + \rho a_0 . \qquad (7.11)$$

Figure 2 illustrates the nature of the solution a_o of (7.11) . If $\rho > 1$ then $a_o > 0$ and the probability that a busy period ends in finite time is less than unity, namely e^{-a_o} . If $\rho \leq 1$ the only non-negative root of (7.11) is $a_o = 0$ so that $e^{-a_o} = 1$. However, if $\rho \simeq 1$ all other coefficients in the right hand side of (7.10) are infinite, implying that no moments of the busy period distribution exist. If $\rho < 1$, (7.10) reduces to

$$W_o(s) = \frac{s}{\mu(1-\rho)} \left\{ 1 - \frac{s}{\mu(1-\rho)^2} + \frac{s^2(2+\rho)}{\mu^2(1-\rho)^4} - \cdots \right\} . \qquad (7.12)$$

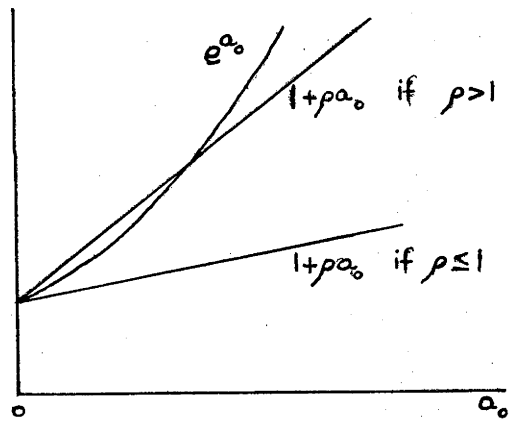

Figure 2. See text for explanation.

CHAPTER IV

PREEMPTIVE PRIORITY QUEUEING

1. Statement of the Problem

The point has been made before that most of the literature of queueing theory is concerned with models of only limited applicability. In the preceding chapter we discussed a system which is one of the most general for which an explicit solution can be constructed by elementary methods, at least at the present time. We were concerned with the standard problem of finding the distributions of queue length and waiting time for the unrestricted queue without side conditions. Queue behaviour is often difficult to study if the model of interest includes special features, such as a particular queue discipline, and in this case it may be necessary to limit the scope of the model by specialising the input and service processes. In this chapter we are interested in discovering to what extent interruptions to the servicing of customers affect the distribution of queue length. We consider a system in which the servicing of a customer is subject to interference due to (i) breakdowns in the service mechanism, or (ii) a queue discipline which assigns priority rights of a preemptive nature to a certain group of customers. To date we have not been able to find the distribution of queue length for such systems if the input and service processes are arbitrary and

we impose the following two restrictions;

(i) all input processes are Poisson

(ii) service times follow the Erlang-k distribution (sometimes with $k = 1$)

$$dE_k(t) = e^{-\mu t} \frac{(\mu t)^{k-1}}{(k-1)!} \mu dt . \qquad (1.1)$$

Throughout we assume that the service facility consists of a single server.

The method used here applies whenever the interarrival or service distributions can be expressed in terms of phases although the calculations are very lengthy for distributions more general than (1.1). We treat the problem as a multidimensional random walk of a restricted nature and as pointed out in Chapter 1, Erlang's method of differential-difference equations is well suited to problems stated in this way. It is worth pointing out that results for a wide class of interarrival and service distributions can be found without much effort if we are content to accept formulae expressed in terms of phases. For example, Luchak (1958) has obtained many interesting results for queueing systems (without service interruptions) in which the service distribution can be written in the form

$$dH(t) = \sum_{k=1}^{\infty} c_k e^{-\mu t} \frac{(\mu t)^{k-1}}{(k-1)!} \mu dt \quad , \quad \sum_{j=1}^{\infty} c_j = 1 .$$

However, the results of Luchak are in terms of service phases and not customer numbers, and can therefore be misinterpreted to give a very misleading picture of queue behaviour. In general, there appears to be no simple transformation from the generating function of the distribution of phases to that of the distribution of the queue length of customers. Since it is the latter that is of interest we restrict the generality of the models considered here to those for which meaningful results can be obtained without too involved a calculation.

Queueing systems in which service interruptions are a distinguishing feature are called Preemptive Priority systems in distinction to the priority discipline in which a priority customer proceeds to the head of the waiting line on arrival, but waits until the service of the current customer has ended. The latter is known as 'head of the line' priority queueing. The effect of interruptions to the servicing of customers is of interest in the practical application of queueing theory. An example of a preemptive priority system is a communications agency in which messages classified as urgent are dispatched immediately on receipt by the service facility irrespective of the state of the queue of routine messages. The question could arise as to whether or not it is preferable or more

economical to introduce additional servers to cope with urgent work rather than retain a queue discipline that increases the usual delay of non-priority items. From the point of view of the customer being served a breakdown in the service mechanism is equivalent to the arrival of a priority customer so that, with some modifications, the same mathematical model can be used to describe both breakdowns and preemptive queueing.

The formal description of a preemptive priority model is as follows. A service facility caters for a population of customers divided into R priority classes, R = 2,3,... , (R = 1 corresponds to no priorities), which are labelled serially in order of precedence 1,2,..,R. On arrival, a customer of class i commences service immediately provided no members of classes 1,2,..,i are present. If customers belonging to any of these classes are present the new arrival joins the queue of members of the same priority class. The servicing of the class i queue does not commence until the system is empty of customers of classes 1,2,...,i-1. Further, a rule must be given governing the manner in which the service of an interrupted customer is resumed. If service starts from the beginning every time a customer re-enters the service mechanism we say that the 'preemptive repeat' rule holds. An alternative procedure is that whereby service is recommenced at the point reached when the interruption occurred, and

this is known as the 'preemptive resume' rule. We make the additional assumption that the queue discipline within each class is 'first come, first served'. If R=2 we speak of the priority class and the non-priority class.

If all service times follow the negative exponential distribution the distinction between the 'preemptive repeat' and 'preemptive resume' rules is not relevant, since the service process is Markovian. When the service distributions do not have this simple form we assume for simplicity that the 'preemptive resume' rule holds. Breakdowns in the service mechanism are included in this model if such an event is formally interpreted as the arrival of a priority customer with precedence over all others. This implies that times between the occurrence of breakdowns and their repair times are equivalent to interarrival and service times respectively.

This queueing process is conveniently described in terms of a hypothetical particle describing a random walk of a restricted nature in R dimensional Euclidean space. Let $n_i(t) = n_i$ be the number of customers of the ith class in the system at time t, $n_i = 0,1,2,\ldots$; $i = 1, 2,\ldots, R$. The position of the particle at t, $(n_1,n_2,\ldots,n_R)$, describes the number of customers of the various classes in the system. The arrival or departure of a class i customer increases or decreases n_i by

unity respectively. Since all $n_i \geqslant 0$, reflecting barriers must be imposed along each axis, although these must be modified to include an absorbing barrier if first passage times are of interest. The random walk is further restricted by the fact that if $n_j > 0$ queues of classes $j + 1$, $j + 2$, $j + 3$,..., R can only increase, and n_j can only decrease if $n_1 = n_2 = \ldots = n_{j-1} = 0$. It is this special feature of the problem that permits a solution by the standard use of generating functions. The only component unaffected by these restrictions is n_1 since class 1 enjoys priority over all others. The breakdown model discussed in §3 is a case where, in addition to the above, other restrictions are also imposed.

Preemptive priority queueing was apparently first considered by Barry (1956) and in more detail by White and Christie (1958) and Stephan (1958). These authors discussed the equilibrium behaviour of the negative exponential queue when $R = 2$. The present writer subsequently extended their results (Heathcote, 1959, 1960a, 1960b) and this Chapter is essentially an account of the work contained in these papers. In passing we mention that a thorough study of the alternative 'head of the line' priority system has been carried out by Kesten and Runnenberg (1957) and Miller (1960). These authors have obtained more general results than we have been able to find for preemptive queueing, but the application of their methods

to our problem does not appear to be feasible. In the paper cited Miller points out that there seems to be no natural way of applying the method of the imbedded Markov chain to systems involving interruptions to service, although he does obtain some results for preemptive queues. We refer to this paper of Millers and to Stephan (1958) for a discussion of the distribution of waiting times.

2. Two Priority Classes

We consider first preemptive priority queueing with two priority classes in which the service distribution of the non-priority customers is defined by (1.1) with parameters μ_2 and k. Let λ_1, λ_2 be the rates of the Poisson input of the priority and non-priority customers respectively, and assume further that the service distribution associated with the priority class is negative exponential with parameter μ_1. Using Erlang's device of considering a service time with distribution given by (1.1) as composed of k negative exponential phases, we see that fluctuations in queue length constitute a Markov process provided the service phase of the current non-priority customer is specified.

Let $P_{rmn}(t)$ be the probability of r priority and n non-priority customers in the system at time t, the current non-priority customer being in phase m; $r=0,1,2,\ldots$; $n=0,1,2,\ldots$; $m=k,k-1,..,2,1$. The phases are numbered in reverse order, a customer passing first through phase k, then phase $k-1$, and finally phase 1 before leaving the system. The suffix m appears only when $n>0$ and $P_{ro}(t)$ denotes the probability of r priority and zero non-priority customers. The forward differential-difference equations satisfied by $P_{rmn}(t)$

are obtained by considering all possible transitions from the point (r,m,n) in the interval $(t, t+\delta t)$. In fact Morse (1958, page 72) has given the difference equations satisfied by the equilibrium probabilities of $M/E_k/1$, and these can easily be extended to incorporate the effect of the priority discipline. We again use Laplace transforms;

$$P^*_{rmn}(s) = \mathcal{L}\left[P_{rmn}(t)\right] = \int_0^\infty e^{-st} P_{rmn}(t)\, dt.$$

Since

$$\mathcal{L}\left[\frac{d}{dt} P_{rmn}(t)\right] = s P^*_{rmn}(s) - P_{rmn}(0),$$

it is apparent that the pure difference equations satisfied by $P^*_{rmn}(s)$ are almost identical with those satisfied by the equilibrium probabilities. The only equation substantially changed is that for the initial state. Hence it is no more difficult to solve the equations for the temporal process than it is to find the equilibrium probabilities. We assume that initially the system is empty;

$$P_{rmn}(0) = 0 \quad , \quad P_{r0}(0) = \delta_{r0}. \tag{2.1}$$

The method used carries over directly for more general initial conditions. Writing $\alpha = s + \lambda_1 + \lambda_2$ the equations satisfied by $P^*_{rmn}(s)$ under initial conditions (2.1) are

$$-\alpha P_{00}^{*} + \mu_1 P_{10}^{*} + \mu_2 P_{011}^{*} = -1,$$

$$-(\alpha+\mu_2) P_{0m1}^{*} + \mu_1 P_{1m1}^{*} + \mu_2 P_{0m+11}^{*} = 0, \quad m=1,2,\ldots,k-1,$$

$$-(\alpha+\mu_2) P_{0k1}^{*} + \lambda_2 P_{00}^{*} + \mu_1 P_{1k1}^{*} + \mu_2 P_{012}^{*} = 0,$$

$$-(\alpha+\mu_2) P_{0mn}^{*} + \lambda_2 P_{0mn-1}^{*} + \mu_1 P_{1mn}^{*} + \mu_2 P_{0m+1n}^{*} = 0, \quad n=2,3,\ldots, \quad m=1,2,\ldots,k-1,$$

$$-(\alpha+\mu_2) P_{0kn}^{*} + \lambda_2 P_{0kn-1}^{*} + \mu_1 P_{1kn}^{*} + \mu_2 P_{01n+1}^{*} = 0, \quad n=2,3,\ldots,$$

$$-(\alpha+\mu_1) P_{r0}^{*} + \lambda_1 P_{r-10}^{*} + \mu_1 P_{r+10}^{*} = 0, \quad r=1,2,\ldots,$$

$$-(\alpha+\mu_1) P_{rm1}^{*} + \lambda_1 P_{r-1m1}^{*} + \mu_1 P_{r+1m1}^{*} = 0, \quad r=1,2,\ldots, \quad m=1,2,\ldots,k-1,$$

$$-(\alpha+\mu_1) P_{rk1}^{*} + \lambda_1 P_{r-1k1}^{*} + \lambda_2 P_{r0}^{*} + \mu_1 P_{r+1k1}^{*} = 0, \quad r=1,2,\ldots,$$

$$-(\alpha+\mu_1) P_{rmn}^{*} + \lambda_1 P_{r-1mn}^{*} + \lambda_2 P_{rmn-1}^{*} + \mu_1 P_{r+1mn}^{*} = 0, \quad r=1,2,\ldots, \quad n=2,3,\ldots, \quad m=1,2,\ldots,k.$$

We seek the joint generating function

$$F^*(s;x,y,z) = \sum_{r=0}^{\infty} x^r P^*_{r0}(s) + \sum_{r=0}^{\infty} x^r J_r^*(s;y,z), \qquad (2.11)$$

where

$$J_r^*(s;y,z) = \sum_{m=1}^{k} \sum_{n=1}^{\infty} y^m z^n P^*_{rmn}(s).$$

The first sum on the right hand side of (2.11) can be evaluated immediately. From (2.7)

$$P^*_{r0}(s) = P^*_{00}(s)\, u_1^r(s), \qquad (2.12)$$

where

$$u_1(s) = (2\mu_1)^{-1}\left[\alpha + \mu_1 - \sqrt{(\alpha+\mu_1)^2 - 4\lambda_1\mu_1}\,\right].$$

Then

$$\sum_{r=0}^{\infty} x^r P^*_{r0}(s) = P^*_{00}(s)\,[1 - x u_1(s)]^{-1}. \qquad (2.13)$$

Multiplying (2.3)-(2.6), (2.8)-(2.10) by the appropriate power of $y^m z^n$, summing, and using (2.12), we find that the $J_r(s;y,z)$ satisfy the difference equations

$$\mu_1 J^*_{r+1} - (\alpha+\mu_1-\lambda_2 z) J^*_r + \lambda_1 J^*_{r-1} = -\lambda_2 z y^k u_1^r P^*_{00}, \quad r = 1,2,\ldots, \qquad (2.14)$$

$$\mu_1 J_1^* - (\alpha+\mu_2-\lambda_2 z-\mu_2 y^{-1}) J_0^* = \mu_2 (1 - y^k z^{-1}) \sum_{n=1}^{\infty} z^n P^*_{0n} + y^k\left[(\alpha-\lambda_2 z-\mu_1 u_1) P^*_{00} - 1\right]. \qquad (2.15)$$

The general solution of (2.14) is

$$J_r^* = A\,\omega_1^r + B\,\omega_2^r - \gamma^k u_1^r P_{00}^* ,$$

where A, B are constants and w_1, w_2 are the roots of the characteristic equation, namely,

$$\begin{aligned} \omega_1(s,z) &= (2\mu_1)^{-1}\left[\alpha+\mu_1-\lambda_2 z - \sqrt{(\alpha+\mu_1-\lambda_2 z)^2 - 4\lambda_1\mu_1}\,\right], \\ \omega_2(s,z) &= (2\mu_1)^{-1}\left[\alpha+\mu_1-\lambda_2 z + \sqrt{(\alpha+\mu_1-\lambda_2 z)^2 - 4\lambda_1\mu_1}\,\right]. \end{aligned} \tag{2.16}$$

Since there is only one boundary condition, (2.15), the constant B may be set identically zero. Then the value of A appropriate to (2.15) is

$$A(s;\gamma,z) = \frac{\mu_2(1-\gamma^k z^{-1})\sum_{n=1}^{\infty} z^n P_{01n}^*(s) \;-\; \gamma^k\left[(1-\gamma^{-1})\mu_2 P_{00}^*(s) + 1\right]}{\mu_1(1-\omega_2) - \mu_2(1-\gamma^{-1})} .$$

The unknown sum $\sum_{n=1}^{\infty} z^n P_{01n}^*(s)$, a function of s and z only, is found by choosing a value of γ which makes the denominator, and hence the numerator, vanish, namely

$$\gamma = \left[1 + (\omega_2 - 1)\mu_1\mu_2\right]^{-1} .$$

Hence

$$\sum_{n=1}^{\infty} z^n P_{01n}^*(s) = z\mu_2^{-1}\left[\mu_1(\omega_2-1)P_{00}^*(s) - 1\right]\left\{1 - z\left[1 + (\omega_2-1)\mu_1\mu_2^{-1}\right]^k\right\}^{-1} .$$

Writing $\beta(s,z) = [\omega_2(s,z) - 1]\mu_1\mu_2^{-1}$,

the required generating function is found to be

$$F^*(s;x,\gamma,z) = \frac{z\gamma[\gamma^k(1+\beta)^k - 1] + P_{00}^*(s)\{\mu_1\gamma(z-\gamma^k)(\omega_2-1) - \mu_2\gamma^k(1-\gamma)[z(1+\beta)^k - 1]\}}{(1-x\omega_1)[z(1+\beta)^k - 1][\mu_1\gamma(\omega_2-1) - \mu_2(1-\gamma)]}$$

$$+ \frac{(1-\gamma^k)}{1-x\omega_1} P_{00}^*(s) . \qquad (2.17)$$

The joint generating function of the numbers of priority and non-priority customers, irrespective of the phase of the current customer, is

$$F^*(s;x,1,z) = \frac{\mu_1(1-z)(\omega_2-1)P_{00}^*(s) + z[1-(1+\beta)^k]}{\mu_1(1-x\omega_1)(\omega_2-1)[1-z(1+\beta)^k]} . \qquad (2.18)$$

As a check, we note that the marginal distribution of the numbers of priority customers is the classical result

$$F^*(s;x,1,1) = \frac{1-\omega_1(s,1)}{s[1-x\omega_1(s,1)]} . \qquad (2.19)$$

The generating function of particular interest, that of the non-priority customers, is

$$F^*(s;1,1,z) = \frac{\mu_1(1-z)(\omega_2-1)P_{00}^*(s) + z[1-(1+\beta)^k]}{(s+\lambda_2-\lambda_2 z)[1-z(1+\beta)^k]} . \qquad (2.20)$$

The solution is now complete except for the unknown $P_{00}^*(s)$. We find $P_{00}^*(s)$ by applying the regu-

larity condition to the generating function in the same way as in Chapter III. $F^*(s,1,1,z)$ converges for at least $|z| \leq 1$, and by Rouchés Theorem the only zero within the unit circle of the denominator is the unique one of the equation

$$1 - z\left[1 + \beta(s,z)\right]^k = 0 .$$

If this zero is $z = z_0(s)$, then

$$P_{00}^*(s) = \mu_1^{-1}\left[\omega_2(s, z_0(s)) - 1\right]^{-1}. \qquad (2.21)$$

The Lagrange Inversion Formula (Whittaker and Watson, 1940, page 132) can be used to evaluate $P_{00}^*(s)$.

When the service time is constant we write $\mu_2 = \delta_2 k$ and take the limit as $k \to \infty$. Then

$$\lim_{k\to\infty}\left[1 + \beta(s,z)\right]^k = \exp\left\{\mu_1 \delta_2^{-1}\left[\omega_2(s,z) - 1\right]\right\} . \qquad (2.22)$$

The preceding formulae hold also for constant service time on using (2.22).

The Laplace transform of the expected number of non-priority customers at time t, conditional on 0 initially, is

$$\mathcal{L}\left[E\left(n(t)|0\right)\right] = \lambda_2 s^{-2} + \frac{1 - \mu_2 \beta(s,1) P_{00}^*(s)}{s\left\{1 - \left[1 + \beta(s,1)\right]^k\right\}} . \qquad (2.23)$$

3. Model for Breakdowns

The preceding model is now modified to describe in a more realistic manner the way in which breakdowns occur. The essential difference between a breakdown and a priority arrival is that breakdowns can occur only when a customer is in service, and also a queue of breakdowns cannot form. If r denotes the number of breakdowns, then r can take only the values 0 and 1 . The random walk in the r,n plane takes place only on the lines $r=0$ and $r=1$, excluding the point $(1,0)$. If $r=0$, the three permissible transitions from the point $(0,n)$, $n=1, 2, \ldots$, are to the points $(0,n-1)$, $(0,n+1)$, and $(1,n)$. If $r=1$ only two transitions are allowable, to $(1, n+1)$ and $(0,n)$. From the point $(0,0)$ the only transition possible is to $(0,1)$.

If $Q^{*}_{rmn}(s)$ denotes the Laplace transform of the transition probability $Q_{rmn}(t)$ of this process, and $\alpha = s+\lambda_1+\lambda_2$, then for an empty system initially, the difference equations satisfied by $Q^{*}_{rmn}(s)$ are

$$-(s+\lambda_2)Q^*_{00} + \mu_2 Q^*_{011} = -1, \tag{3.1}$$

$$-(\alpha+\mu_2)Q^*_{0m1} + \mu_1 Q^*_{1m1} + \mu_2 Q^*_{0m+11} = 0, \quad m=1,2,\dots,k-1, \tag{3.2}$$

$$-(\alpha+\mu_2)Q^*_{0k1} + \lambda_2 Q^*_{00} + \mu_1 Q^*_{1k1} + \mu_2 Q^*_{012} = 0, \tag{3.3}$$

$$-(\alpha+\mu_2)Q^*_{0mn} + \lambda_2 Q^*_{0mn-1} + \mu_1 Q^*_{1mn} + \mu_2 Q^*_{0m+1n} = 0, \quad n=2,3,\dots, \; m=1,2,\dots,k-1, \tag{3.4}$$

$$-(\alpha+\mu_2)Q^*_{0kn} + \lambda_2 Q^*_{0kn-1} + \mu_1 Q^*_{1kn} + \mu_2 Q^*_{01n+1} = 0, \quad n=2,3,\dots, \tag{3.5}$$

$$-(s+\lambda_2+\mu_1)Q^*_{1m1} + \lambda_1 Q^*_{0m1} = 0, \quad m=1,2,\dots,k-1, \tag{3.6}$$

$$-(s+\lambda_2+\mu_1)Q^*_{1mn} + \lambda_1 Q^*_{0mn} + \lambda_2 Q^*_{1mn-1} = 0, \quad n=2,3,\dots, \; m=1,2,\dots,k. \tag{3.7}$$

Proceeding as before we find the joint generating function

$$G^*(s;x,z) = Q^*_{00}(s) + \sum_{r=0}^{1}\sum_{m=1}^{k}\sum_{n=1}^{\infty} x^r z^n Q^*_{rmn}(s)$$

$$= Q^*_{00}(s) - \frac{z(s+\mu_1+\lambda_2-\lambda_2 z+\lambda_1 x)\left[1-(1+\theta)^k\right]\left[(s+\lambda_2-\lambda_2 z)Q^*_{00}(s)-1\right]}{\mu_2(s+\mu_1+\lambda_2-\lambda_2 z)\left[1-z(1+\theta)^k\right]\theta}, \quad (3.8)$$

where

$$\theta = \theta(s,z) = \frac{(s+\mu_1+\lambda_1+\lambda_2-\lambda_2 z)(s+\lambda_2-\lambda_2 z)}{\mu_2(s+\mu_1+\lambda_2-\lambda_2 z)}.$$

Then the generating function defining the distribution of customer numbers is

$$G^*(s;1,z) = \frac{(1-z)Q^*_{00}(s) + z\left[1-(1+\theta)^k\right](s+\lambda_2-\lambda_2 z)^{-1}}{1-z(1+\theta)^k}. \quad (3.9)$$

$Q^*_{00}(s)$ can be found by the same argument as before. Expected queue length at time t in this case has the Laplace transform

$$\mathcal{L}\left[E(n(t)|0)\right] = \lambda_2 s^{-2} + \frac{1 - s\,Q^*_{00}(s)}{s\left\{1-\left[1+\theta(s,1)\right]^k\right\}}. \quad (3.10)$$

4. Comparison with the Unrestricted Queue

Let $R_n(t)$, $n=0,1,2,\ldots$, denote the transition probabilities of the unrestricted process $M/E_k/1$, with generating function

$$J(z,t) = \sum_{n=0}^{\infty} z^n R_n(t) .$$

If the interarrival and service parameters of this process are respectively λ_2, μ_2, and the system is initially empty, then the Laplace transform of $J(z,t)$ is given by equation (4.1) of Chapter III with

$$\psi(s) = \mu_2^k (s+\mu_2)^{-k} .$$

Thus

$$J^*(z,s) = \frac{(1-z)R_0^*(s) - z(\gamma^k - 1)(s+\lambda_2-\lambda_2 z)^{-1}}{1-z\gamma^k} , \qquad (4.1)$$

where

$$\gamma = \gamma(z,s) = (s+\mu_2+\lambda_2-\lambda_2 z)\mu_2^{-1} .$$

The Laplace transform of expected queue length at time t is

$$\mathcal{L}[E(n(t)|0)] = \lambda s^{-2} - \frac{\mu_2^k [s^{-1} - P_0^*(s)]}{(s+\mu_2)^k - \mu_2^k} . \qquad (4.2)$$

The comparable results for the preemptive priority model are (2.20), (2.23), and for the breakdown model (3.9), (3.10) respectively.

The equilibrium distributions of queue length for the preemptive and breakdown models are found by the same arguments as before, and we denote their generating functions by $F(z)$ and $G(z)$ respectively. The conditions for the existence of these distributions are found to be

$$\text{(i) preemptive model,} \quad 1 > \rho_1 + \rho_2 \,, \tag{4.3}$$

$$\text{(ii) breakdown model,} \quad 1 > \rho_2(1+\rho_1) \,,$$

where $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_2 k/\mu_2$. We recall that the condition for the existence of the equilibrium distribution of $M/E_k/1$ without interruption is

$$1 > \rho_2 \,. \tag{4.4}$$

Let the generating function of this equilibrium distribution be

$$\lim_{t\to\infty} J(z,t) = \overline{J}(z) \,.$$

Assuming that (4.3), (4.4) hold, the three equilibrium generating functions are then respectively

$$\bar{J}(z) = \frac{(1-z)(1-\rho_2)}{1 - z\left[1+(1-z)\lambda_2\mu_2^{-1}\right]^k} , \qquad (4.5)$$

$$F(z) = \frac{\mu_1(1-z)(1-\rho_1-\rho_2)}{\left[1-\omega_1(0,z)\right]\left\{1 - z\left[1+\beta(0,z)\right]^k\right\}} , \qquad (4.6)$$

$$G(z) = \frac{(1-z)(1-\rho_2-\rho_1\rho_2)}{1 - z\left[1+\theta(0,z)\right]^k} . \qquad (4.7)$$

(4.5) - (4.7) hold for $k=1,2,\ldots$. If the service time is of constant duration δ_2^{-1} we have, by replacing μ_2 by $\delta_2 k$ and taking the limit as $k \to \infty$,

$$\bar{J}(z,\infty) = \frac{(1-z)(1-\rho_2)}{1 - z\, e^{(1-z)\rho_2}} , \qquad (4.8)$$

$$F(z,\infty) = \frac{\mu_1(1-z)(1-\rho_1-\rho_2)}{\left[1-\omega_1(0,z)\right]\left\{1 - z\exp\left[\mu_1\delta_2^{-1}\left(\omega_2(0,z)-1\right)\right]\right\}} , \qquad (4.9)$$

$$G(z,\infty) = \frac{(1-z)(1-\rho_2-\rho_1\rho_2)}{1 - z\exp\left\{\dfrac{\rho_2(1-z)(\mu_1+\lambda_1+\lambda_2-\lambda_2 z)}{\mu_1+\lambda_2-\lambda_2 z}\right\}} . \qquad (4.10)$$

Let the expected values of the three equilibrium distributions for parameter k be respectively $\bar{a}_k, \bar{n}_k, \bar{b}_k$. Then

$$\bar{a}_k = \left[\frac{d}{dz}\bar{J}(z)\right]_{z=1} = \frac{\rho_2\left[2-(1-k^{-1})\rho_2\right]}{2(1-\rho_2)}, \qquad (4.11)$$

$$\bar{n}_k = \left[\frac{d}{dz}F(z)\right]_{z=1} = \frac{\rho_2\left[2(1-\rho_1+\rho_1\mu_2/\mu_1 k)-(1-k^{-1})\rho_2\right]}{2(1-\rho_1)(1-\rho_1-\rho_2)}, \qquad (4.12)$$

$$\bar{b}_k = \left[\frac{d}{dz}G(z)\right]_{z=1} = \frac{\rho_2\left[2(1+\rho_1+\rho_1\rho_2\mu_2/\mu_1 k)-(1-k^{-1})\rho_2(1+\rho_1)^2\right]}{2(1-\rho_2-\rho_1\rho_2)}. \qquad (4.13)$$

The expected length of the priority queue is of course

$$\bar{r} = \rho_1(1-\rho_1)^{-1}. \qquad (4.14)$$

Two other expected values of interest are those of the 'head of the line' priority model. These are available only when all service times are negative exponential. Denoting the expected queue lengths of the priority and non-priority customers for this system by E(r), E(n) respectively, we quote from Miller (1960, equations (2.7), (2.8));

$$E(r) = \rho_1(1+\rho_2\mu_1/\mu_2)(1-\rho_1)^{-1},$$

$$E(n) = \frac{\rho_2\left[1-\rho_1+\rho_1\mu_2/\mu_1-\rho_1(1-\rho_1-\rho_2)\right]}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

$$= \bar{n}_1 - \rho_2\bar{r}.$$

The expectations $\bar{n}_1, \bar{b}_1$ and $\bar{n}_\infty, \bar{b}_\infty$ are tabulated for different values of ρ_1 and ρ_2 in Tables 1 and 2 respectively. Expected waiting times are obtained from the tabulated quantities on division by the input parameter λ_2. The column under $\rho_1 = 0{\cdot}0$ is the expected queue length when there are no service interruptions. Comparing this column with the figures when $\rho_1 > 0$ we see that if the traffic intensity ρ_2 is moderate or large, expected queue length is considerably increased even for small values of the traffic intensity ρ_1. Also, if ρ_1 is small, there is no marked difference between the figures for the preemptive and breakdown models, although this difference becomes more apparent as ρ_1 increases.

ρ_2 \ ρ_1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.111	0.139	0.179	0.238	0.333	0.500	0.833	1.667	5.000	∞
	0.111	0.125	0.139	0.153	0.167	0.182	0.198	0.213	0.229	0.246
0.2	0.250	0.318	0.417	0.571	0.833	1.333	2.500	6.667	∞	
	0.250	0.287	0.326	0.368	0.411	0.457	0.506	0.558	0.613	
0.3	0.429	0.556	0.750	1.071	1.667	3.000	7.500	∞		
	0.429	0.506	0.591	0.684	0.786	0.900	1.027	1.169		
0.4	0.667	0.889	1.250	1.905	3.333	8.000	∞			
	0.667	0.814	0.985	1.183	1.413	1.700	2.044			
0.5	1.000	1.389	2.083	3.571	8.333	∞				
	1.000	1.278	1.625	2.071	2.667	3.500				
0.6	1.500	2.222	3.750	8.571	∞					
	1.500	2.047	2.829	4.036	6.150					
0.7	2.333	3.889	8.750	∞						
	2.333	3.561	5.863	11.744						
0.8	4.000	8.889	∞							
	4.000	7.867	27.200							
0.9	9.000	∞								
	9.000	107.100								

Table 1. Mean queue length of M/M/1 with service interruptions. Upper entry in each cell is $\bar{n}_1$, lower entry $\bar{b}_1$.

ρ_2 \ ρ_1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.106	0.132	0.170	0.226	0.317	0.475	0.792	1.583	4.750	∞
	0.106	0.118	0.131	0.143	0.156	0.169	0.182	0.196	0.210	0.223
0.2	0.225	0.286	0.375	0.514	0.750	1.120	2.250	6.000	∞	
	0.225	0.256	0.288	0.322	0.357	0.393	0.431	0.470	0.511	
0.3	0.364	0.472	0.638	0.911	1.417	2.550	6.375	∞		
	0.364	0.425	0.489	0.559	0.634	0.716	0.805	0.904		
0.4	0.533	0.711	1.000	1.524	2.667	6.400	∞			
	0.533	0.641	0.763	0.902	1.062	1.250	1.476			
0.5	0.750	1.042	1.563	2.679	6.250	∞				
	0.750	0.942	1.175	1.468	1.850	2.375				
0.6	1.050	1.556	2.625	6.000	∞					
	1.050	1.407	1.903	2.654	3.945					
0.7	1.517	2.528	5.688	∞						
	1.517	2.272	3.658	7.144						
0.8	2.400	5.333	∞							
	2.400	4.640	15.680							
0.9	4.950	∞								
	4.950	58.095								

Table 2. Mean queue length of M/D/1 with service interruptions. Upper entry in each cell is $\bar{n}_\infty$, lower entry $\bar{b}_\infty$.

5. R Priority Classes

Let the number of priority classes be R, an arbitrary positive integer. Difficulties not present in the two dimensional cases just considered are incurred when R is arbitrary, and in this section we impose the further restriction that all service times follow the negative exponential distribution. Our method holds for more general distributions but we consider only this special case to keep the calculations brief. Let the parameters of the interarrival and service distributions of the ith priority class be λ_i, μ_i respectively, $i=1,2,..,R$. We seek the joint probability $P(t; n_1,n_2,...,n_R)$ that at time t there are n_i customers of class i in the system, conditional on given initial conditions, $n_i=0,1,2,...$; $i=1,2,...,R$. In what follows the argument t of this probability will be suppressed. The equivalent probability in the equilibrium state is denoted by $p(n_1,n_2,...,n_R)$. It has been shown by White and Christie (1958) that the condition for the existence of the equilibrium distribution is

$$1 > \sum_{i=1}^{R} \rho_i ,$$

where $\rho_i = \lambda_i/\mu_i$.

The forward differential-difference equations are obtained in the usual way. For all $n_k=0$, $k=1,2,..,j$; $n_{j+1} > 0$; $n_{j+i} \geq 0$, $i=2,3,...,R-j$; where j is fixed,

$1 \leq j \leq R-1$, we have

$$-(\mu_{j+1} + \sum_{i=1}^{R}\lambda_i)P(0,\dots,0,n_{j+1},n_{j+2},\dots,n_R) + \mu_1 P(1,0,\dots,0,n_{j+1},\dots,n_R)$$

$$+\mu_2 P(0,1,0,\dots,0,n_{j+1},\dots,n_R)+\dots+\mu_j P(0,\dots,0,1,n_{j+1},\dots,n_R)$$

$$+\mu_{j+1}P(0,\dots,0,n_{j+1}+1,n_{j+2},\dots,n_R)+\lambda_{j+1}P(0,\dots,0,n_{j+1}-1,n_{j+2},\dots,n_R)$$

$$+\lambda_{j+2}P(0,\dots,0,n_{j+1},n_{j+2}-1,n_{j+3},\dots,n_R)+\dots+\lambda_R P(0,\dots,0,n_{j+1},\dots,n_{R-1},n_R-1)$$

$$= \frac{d}{dt}P(0,\dots,0,n_{j+1},\dots,n_R).$$

To rationalise the notation, write for the coefficient of μ_i , $i=1,2,\dots,j$,

$$P(0,\dots,0,1,0,\dots,0,n_{j+1},n_{j+2},\dots,n_R) = P(\delta_i, n_{j+1}).$$

This indicates that the first j components are all zeros except the ith which is unity, and that components from the (j+1)th on are not necessarily zero. Similarly for the coefficient of λ_{j+i} , $i=1,2,\dots,R-j$,

$$P(0,\dots,0,n_{j+1},n_{j+2},\dots,n_{j+i-1},n_{j+i}-1,n_{j+i+1},\dots,n_R) = P(0_j, n_{j+i}-1).$$

We also write $S_R = \sum_{k=1}^{R}\lambda_k$.

The equations of the system are then

$$-(\mu_1+S_R)P(n_1,\dots,n_R)+\mu_1 P(n_1+1,n_2,\dots,n_R)+\sum_{i=1}^{R}\lambda_i P(n_1,\dots,n_i-1,n_{i+1},\dots,n_R)$$

$$= \frac{d}{dt}P(n_1,\dots,n_R) \ , \ n_1>0, n_k\geq 0, \ k=2,3,\dots,R, \qquad (5.1)$$

$$-(\mu_2+S_R)P(0,n_2)+\mu_1 P(\delta_1,n_2)+\mu_2 P(0,n_2+1,n_3,\dots,n_R)+\sum_{i=2}^{R}\lambda_i P(0,n_i-1)$$

$$= \frac{d}{dt}P(0,n_2) \ , \ n_2>0, n_k\geq 0, \ k=3,4,\dots,R, \qquad (5.2)$$

$$-(\mu_{j+1}+S_R)P(0_j,n_{j+1})+\sum_{i=1}^{j}\mu_i P(\delta_i,n_{j+1})+\mu_{j+1}P(0_j,n_{j+1}+1,n_{j+2},\ldots,n_R)$$
$$+\sum_{i=j+1}^{R}\lambda_i P(0_j,n_i-1) \;=\; \frac{d}{dt}P(0_j,n_{j+1}),$$
$$n_{j+1}>0,\ n_{j+k}\geq 0,\ k=2,3,\ldots,R-j,\quad j=2,3,\ldots,R-2, \qquad (5.3)$$

$$-(\mu_R+S_R)P(0_{R-1},n_R)+\sum_{i=1}^{R-1}\mu_i P(\delta_i,n_R)+\mu_R P(0_{R-1},n_R+1)$$
$$+\lambda_R P(0_{R-1},n_R-1)=\frac{d}{dt}P(0_{R-1},n_R),\qquad n_R>0, \qquad (5.4)$$

$$-S_R P(0_R)+\sum_{i=1}^{R}\mu_i P(\delta_i)\;=\;\frac{d}{dt}P(0_R). \qquad (5.5)$$

The difference equations satisfied by the equilibrium probabilities $p(n_1,n_2,..,n_R)$, independent of t, are obtained from (5.1)-(5.5) by writing all time derivatives as zero. We have noted before the similarity between the equations in the Laplace transforms $P(s;n_1,n_2,..,n_R)$ and those satisfied by $p(n_1,n_2,..,n_R)$. Hence if the solution of the latter set can be found it is a simple matter to derive from this the transforms $P^*(s;n_1,n_2,..,n_R)$. We consider then the equations satisfied by the equilibrium probabilities, and assume that the condition for the existence of the equilibrium distribution holds. In fact White

and Christie (1958, page 81) have shown that the null probability, denoted by p_0, is

$$p_0 = 1 - \sum_{i=1}^{R} \rho_i \,. \tag{5.6}$$

We seek the joint generating function

$$F(z_1,z_2,\ldots,z_R) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\cdots\sum_{n_R=0}^{\infty} z_1^{n_1} z_2^{n_2}\cdots z_R^{n_R}\, p(n_1,n_2,\ldots,n_R) \,. \tag{5.7}$$

It is convenient to introduce the partial generating functions

$$F(n_1,n_2,\ldots,n_j;\, z_{j+1},z_{j+2},\ldots,z_R) = \sum_{n_{j+1}=0}^{\infty}\sum_{n_{j+2}=0}^{\infty}\cdots\sum_{n_R=0}^{\infty} z_{j+1}^{n_{j+1}} z_{j+2}^{n_{j+2}}\cdots z_R^{n_R}\, p(n_1,n_2,\ldots,n_R)$$

$$= \sum_{n_{j+1}=0}^{\infty} z_{j+1}^{n_{j+1}} F(n_1,n_2,\ldots,n_{j+1};\, z_{j+2},z_{j+3},\ldots,z_R) \,.$$

$$j = 1,2,\ldots,R-1 \,. \tag{5.8}$$

These generating functions are obtained from (5.1)-(5.5) by multiplying each equation by the appropriate powers of $z_1,z_2,\ldots,z_R$ and summing. We write down only the equations satisfied by $F(n_1;z_2,\ldots,z_R) \equiv F(n_1)$ which will be required later;

$$-(\mu_1+S_R-\sum_{i=2}^{R}\lambda_i z_i)F(n_1) + \mu_1 F(n_1+1) + \lambda_1 F(n_1-1) = 0 \,, \qquad n_1 = 1,2,3,\ldots \,, \tag{5.9}$$

$$-\left[S_R - \sum_{i=2}^{R}\lambda_i z_i - \mu_2(z_2^{-1}-1)\right]F(0) + \mu_1 F(1)$$

$$= \mu_R(z_R^{-1}-1)p_0 + \sum_{i=2}^{R-1}\left[\mu_i(z_i^{-1}-1) - \mu_{i+1}(z_{i+1}^{-1}-1)\right]F(0_i,z_{i+1}) \,. \tag{5.10}$$

where, in accordance with the notation previously defined,

$$F(0_i, z_{i+1}) = F(0,0,\ldots,0,z_{i+1},z_{i+2},\ldots,z_R).$$

We find for the joint generating function

$$F(z_1,z_2,\ldots,z_R) = \sum_{n_1=0}^{\infty} z_1^{n_1} F(n_1; z_2,z_3,\ldots,z_R)$$

$$= \frac{\sum_{i=1}^{R-1} [\mu_i(z_i^{-1}-1) - \mu_{i+1}(z_{i+1}^{-1}-1)] F(0_i, z_{i+1}) + \mu_R(z_R^{-1}-1)p_0}{\mu_1(z_1^{-1}-1) - \sum_{i=1}^{R} \lambda_i(1-z_i)} . \qquad (5.11)$$

(5.11) expresses the Rth dimensional generating function $F(z_1,z_2,\ldots,z_R)$ in terms of the (R-i)th dimensional functions $F(0_i, z_{i+1})$. If a recurrence relation between $F(0_i, z_{i+1})$ and $F(0_{i+1}, z_{i+2})$, i=1,2,..,R-2 can be found which will give these functions in terms of p_0 , then the problem is solved. The first step towards obtaining such a recurrence relation is in fact given by solving the difference equations (5.9), (5.10) in $F(n_1)$. The characteristic equation of (5.9) is

$$\mu_1 v^2 - v\left(\mu_1 + S_R - \sum_{i=2}^{R} \lambda_i z_i\right) + \lambda_1 = 0,$$

with roots

$$v_1(z_2,z_3,\ldots,z_R) = (2\mu_1)^{-1}\left[\mu_1 + S_R - \sum_{i=2}^{R} \lambda_i z_i - \sqrt{\left(\mu_1 + S_R - \sum_{i=2}^{R} \lambda_i z_i\right)^2 - 4\mu_1\lambda_1}\right] , \qquad (5.12)$$

$$\upsilon_2(z_2,z_3,\ldots,z_R)=(2\mu_1)^{-1}\left[\mu_1+S_R-\sum_{i=2}^{R}\lambda_i z_i+\sqrt{\left(\mu_1+S_R-\sum_{i=2}^{R}\lambda_i z_i\right)^2-4\mu_1\lambda_1}\right]. \quad (5.13)$$

If $|z_i|\leq 1$, $i=2,3,\ldots,R$, and $\mu_1>\lambda_1$, then $|\upsilon_1|<1\leq|\upsilon_2|$. (5.12), (5.13) are a generalised form of $w_1(s,z), w_2(s,z)$ given by (2.16). Since n_1 is not bounded above and the only boundary condition, (5.10), is at $n_1=0$, we may discard the root greater than unity, υ_2, and take as the general solution

$$F(n_1) = F(0)\,\upsilon_1^{n_1}.$$

$F(0)$ is found by substituting this solution in (5.10). Using the relation

$$\mu_1+S_R-\sum_{i=2}^{R}\lambda_i z_i = \mu_1(\upsilon_1+\upsilon_2)$$

we have

$$F(0) = F(0,z_2,z_3,\ldots,z_R) = F(0_1,z_2)$$
$$= \frac{\sum_{i=2}^{R-1}\left[\mu_i(z_i^{-1}-1)-\mu_{i+1}(z_{i+1}^{-1}-1)\right]F(0_i,z_{i+1})+\mu_R(z_R^{-1}-1)P_0}{\mu_2(z_2^{-1}-1)+\mu_1(1-\upsilon_2)}. \quad (5.14)$$

This is the first of the recurrence relations we seek and gives $F(0_1,z_2)$ in terms of $F(0_i,z_{i+1})$, $i=2,3,\ldots,R-1$.

Since $F(0_1,z_2)$ is a generating function it must converge everywhere within the unit sphere $|z_i|\leq 1$ $i=2,3,\ldots,R$. This implies that zeros in z_i of numerator and denominator in this region coincide. In

particular, this is true of z_2 irrespective of the values of $z_3, z_4, \ldots, z_R$. By Rouché's Theorem the denominator has only one zero in z_2 within the unit circle. If this zero is

$$z_2 = \alpha_2(z_3, z_4, \ldots, z_R)$$

then the numerator also vanishes when $z_2 = \alpha_2$. Therefore

$$\left[\mu_2(\alpha_2^{-1}-1) - \mu_3(z_3^{-1}-1)\right] F(0_2, z_3) + \mu_R(z_R^{-1}-1) p_0$$
$$+ \sum_{i=3}^{R-1} \left[\mu_i(z_i^{-1}-1) - \mu_{i+1}(z_{i+1}^{-1}-1)\right] F(0_i, z_{i+1}) = 0,$$

and hence

$$F(0_2, z_3) = \frac{\sum_{i=3}^{R-1} \left[\mu_i(z_i^{-1}-1) - \mu_{i+1}(z_{i+1}^{-1}-1)\right] F(0_i, z_{i+1}) + \mu_R(z_R^{-1}-1) p_0}{\mu_3(z_3^{-1}-1) - \mu_2(\alpha_2^{-1}-1)}. \tag{5.15}$$

Repeating the same argument we find for $i=3,4,\ldots,R-2$,

$$F(0_i, z_{i+1}) = \frac{\sum_{k=i+1}^{R-1} \left[\mu_k(z_k^{-1}-1) - \mu_{k+1}(z_{k+1}^{-1}-1)\right] F(0_k, z_{k+1}) + \mu_R(z_R^{-1}-1) p_0}{\mu_{i+1}(z_{i+1}^{-1}-1) - \mu_i(\alpha_i^{-1}-1)}. \tag{5.16}$$

$\alpha_i = \alpha_i(z_{i+1}, z_{i+2}, \ldots, z_R)$ is the zero in z_i within the unit circle of the denominator of $F(0_{i-1}, z_i)$. The result

for $F(0_{R-1}, z_R)$ is

$$F(0_{R-1}, z_R) = \frac{\mu_R (z_R^{-1} - 1) p_0}{\mu_R (z_R^{-1} - 1) - \mu_{R-1} (\alpha_{R-1}^{-1} - 1)} \cdot \qquad (5.17)$$

Starting from (5.17) it is now simply a matter of substituting back to find $F(0_i, z_{i+1})$ in terms of p_0 ;

$$F(0_i, z_{i+1}) = p_0 \alpha_i \prod_{j=i+1}^{R} \frac{\mu_j (\alpha_j - z_j)}{[\mu_j \alpha_{j-1} (1 - z_j) - \mu_{j-1} z_j (1 - \alpha_{j-1})]} , \qquad (5.18)$$

$i = 2, 3, .. R-1$, and where $\alpha_R = 1$.

We also have

$$F(0_1, z_2) = \frac{\mu_2 (\alpha_2 - z_2) F(0_2, z_3)}{\alpha_2 [\mu_2 (1 - z_2) + \mu_1 z_2 (1 - v_2)]} , \qquad (5.19)$$

and

$$F(z_1, z_2, \ldots, z_R) = \sum_{n_1=0}^{\infty} z_1^{n_1} F(n_1; z_2, z_3, \ldots, z_R) = \sum_{n_1=0}^{\infty} (z_1 v_1)^{n_1} F(0_1, z_2)$$

$$= F(0_1, z_2) (1 - z_1 v_1)^{-1} . \qquad (5.20)$$

Combining these results we have the final solution

$$F(z_1, z_2, \ldots, z_R) = \frac{p_0 \alpha_1}{1 - z_1 v_1} \prod_{i=2}^{R} \frac{\mu_i (\alpha_i - z_i)}{[\mu_i \alpha_{i-1} (1 - z_i) - \mu_{i-1} z_i (1 - \alpha_{i-1})]} \cdot \qquad (5.21)$$

For conciseness we have written $\alpha_1 = \sigma_2^{-1}$ and $\alpha_R = 1$ in (5.21). The study of this process is continued in the next section.

6. Explicit Solutions when Service Rates are equal.

To complete the solution of the R dimensional pre-emptive priority problem of §5 it is necessary to find explicit expressions for the quantities $\alpha_i(z_{i+1}, z_{i+2}, \ldots, z_R)$, $i = 2, 3, \ldots, R-1$. We recall from (5.14) that α_2 is the smallest zero in z_2 of

$$\mu_2(1-z_2) + \mu_1 z_2(1-v_2) = 0, \tag{6.1}$$

which on expansion is

$$z_2^3\lambda_2(\mu_1-\mu_2) + z_2^2\left[\lambda_1\mu_1+\lambda_2\mu_2-(\mu_1-\mu_2)(\mu_2+S_R-\sum_{i=3}^{R}\lambda_i z_i)\right]$$
$$+z_2\mu_2\left[2(\mu_1-\mu_2)-\mu_1-S_R-\sum_{i=3}^{R}\lambda_i z_i\right] + \mu_2^2 = 0. \tag{6.2}$$

It is apparent that to find $\alpha_2, \alpha_3, \ldots, \alpha_{R-1}$ we require the solution of a series of equations similar to (6.2). These can be found directly or by the Lagrange Inversion Formula, but the calculations are lengthy. Since the method is clear we simplify the problem by assuming that all service rates μ_i are equal to the same value μ. The effect of this assumption is to reduce (6.2) and all subsequent equations for the α_i to quadratics. In particular, if $\mu_1 = \mu_2 = \mu$, (6.2) becomes

$$z_2^2(\lambda_1+\lambda_2) - z_2\left(\mu + S_R - \sum_{i=3}^{R}\lambda_i z_i\right) + \mu = 0.$$

Then

$$\alpha_2(z_3,\dots,z_R) = \frac{1}{2(\lambda_1+\lambda_2)}\left[\mu+S_R-\sum_{i=3}^{R}\lambda_i z_i-\sqrt{\left(\mu+S_R-\sum_{i=3}^{R}\lambda_i z_i\right)^2-4\mu(\lambda_1+\lambda_2)}\right].$$

(6.3)

We find for $j=2,3,\dots,R-1$

$$\alpha_j(z_{j+1},\dots,z_R) = \frac{1}{2S_j}\left[\mu+S_R-\sum_{i=j+1}^{R}\lambda_i z_i-\sqrt{\left(\mu+S_R-\sum_{i=j+1}^{R}\lambda_i z_i\right)^2-4\mu S_j}\right].$$

(6.4)

The formula for the joint generating function, (5.21), in this case has the simple form

$$F(z_1,z_2,\dots,z_R) = \frac{p_0\alpha_1}{1-z_1\mu_1}\prod_{i=2}^{R}\left(\frac{\alpha_i-z_i}{\alpha_{i-1}-z_i}\right), \qquad (6.5)$$

(6.5)

where the α_i are now given by (6.4).

We list below some formulae of interest:

The joint distribution of the lengths of queues of classes 1 and R is given by

$$F(z_1,1,1,\dots,1,z_R) = \left(\frac{1-u_1}{1-z_1\mu_1}\right)\left(\frac{1-\alpha_{R-1}}{\alpha_{R-1}-z_R}\right)\frac{p_0}{\rho_R}, \qquad (6.6)$$

where

$$u_1(z_R) = U_1(1,1,\dots,1,z_R) = (2\mu)^{-1}\left[\mu+\lambda_1+\lambda_R-\lambda_R z_R-\sqrt{(\mu+\lambda_1+\lambda_R-\lambda_R z_R)^2-4\mu\lambda_1}\right].$$

The covariance is

$$C(n_1,n_R)=\frac{\rho_1\rho_R}{(1-\bar{J}_R)}\left\{\frac{1-\bar{J}_R-(1-\rho_1)^2(1-\bar{J}_{R-1})^{-1}}{(1-\rho_1)^3}+\frac{(1+\bar{J}_{R-1})^2(1-\bar{J}_{R-1})^2}{2\bar{J}_{R-1}(1-\bar{J}_{R-1}^3)}\right\}, \quad (6.7)$$

where $\bar{J}_i=\sum_{K=1}^{i}\rho_K$. The marginal distribution of n_1 is of course the classical result

$$F(z_1,1,1,\ldots,1)=(1-\rho_1)(1-\rho_1 z_1)^{-1}.$$

The marginal distribution of n_R has generating function

$$F(1,1,\ldots,1,z_R)=(1-\alpha_{R-1})(\alpha_{R-1}-z_R)^{-1}(\rho_0/\rho_R), \quad (6.8)$$

and its expected value is

$$\bar{n}_R=\rho_R\left[1-\sum_{i=1}^{R}\rho_i\right]^{-1}\left[1-\sum_{i=1}^{R-1}\rho_i\right]^{-1}. \quad (6.9)$$

The distribution of n_R conditional on $n_1=n_2=\ldots=n_{R-1}=0$ is given by

$$F(z_R|0,0,\ldots,0)=\sum_{n_R=0}^{\infty}z_R^{n_R}\,p(n_R|0,0,\ldots,0)$$

$$=\frac{(1-z_R)\alpha_{R-1}}{\alpha_{R-1}-z_R}\left\{\frac{1-\sum_{i=1}^{R}\rho_i}{1-\sum_{i=1}^{R-1}\rho_i}\right\}. \quad (6.10)$$

(6.10a)

The distribution of the total number of customers in the system irrespective of class is found by substituting $z_1=z_2=\ldots=z_R=z$ in (6.5) to yield

$$F(z)=\left[1-\sum_{i=1}^{R}\rho_i\right]\left[1-z\sum_{i=1}^{R}\rho_i\right]^{-1}. \quad (6.11)$$

Thus if $n_1+n_2+..+n_R=n$, then

$$p(n) = \left[1-\sum_{i=1}^{R}\rho_i\right]\left[\sum_{i=1}^{R}\rho_i\right]^n .$$

The Laplace transform of the generating function of the queue length distribution in finite time is easily derived from these results. Let $F^*(s;z_1,z_2,..,z_R)$ be this generating function, and denote the transform of the null probability by

$$P_0^*(s) = \int_0^\infty e^{-st} P(t;0,0,\cdots,0)\,dt.$$

We assume that the initial state is $(m,0,0,..,0)$, $m=0,1,2,...$, so that

$$P(0;n_1,n_2,\cdots,n_R) = \delta_{n_1 m}\,\delta_{n_2 0}\cdots\delta_{n_R 0} . \qquad (6.10)$$

Then $F^*(s;z_1,z_2,..,z_R)$ is given by the formulae for $F(z_1,z_2,..,z_R)$ provided the following two alterations are made

(i) replace S_R by $s+S_R$

(ii) replace $(1-z_R)p_0$ by $(1-z_R)P_0^*(s) - z_R^{m+1}\mu_R^{-1}$.

We consider in detail only the case when all $\mu_i=\mu$. To avoid confusion relabel the roots α_i, ω_j by β_i, ω_j respectively;

$$\beta_i(s;z_{i+1},\cdots,z_R) = (2S_i)^{-1}\left[s+\mu+S_R-\sum_{j=i+1}^{R}\lambda_j z_j - \sqrt{\left(s+\mu+S_R-\sum_{j=i+1}^{R}\lambda_j z_j\right)^2 - 4\mu S_i}\,\right],$$

$$w_1(s;z_2,\ldots,z_R) = (2\mu)^{-1}\left[s+\mu+S_R-\sum_{i=2}^{R}\lambda_i z_i-\sqrt{\left(s+\mu+S_R-\sum_{i=2}^{R}\lambda_i z_i\right)^2-4\mu\lambda_1}\right].$$

$w_2(s;z_2,\ldots,z_R)$ is the conjugate of $w_1(s;z_2,\ldots,z_R)$. Then,

$$F^*(s;z_1,z_2,\ldots,z_R) = \frac{\left[(1-z_R)P_0^*(s)-z_R^{m+1}\mu^{-1}\right]\beta_1}{(1-z_1w_1)(\beta_{R-1}-z_R)}\prod_{i=2}^{R-1}\left(\frac{\beta_i-z_i}{\beta_{i-1}-z_i}\right), \qquad (6.11)$$

where $\beta_1 = w_2^{-1}$. Corresponding to (6.6) we find for the Rth priority class

$$F^*(s;1,1,\ldots,1,z_R) = \frac{(1-\beta_{R-1})\left[\mu(1-z_R)P_0^*(s)-z_R^{m+1}\right]}{(s+\lambda_R-\lambda_R z_R)(\beta_{R-1}-z_R)}. \qquad (6.12)$$

The transform of the null probability $P_0^*(s)$ is obtained by application of the regularity condition used before to the right hand side of (6.12). For $Re\ s>0$ the only zero of the denominator within the unit circle is that of

$$\beta_{R-1}(s;z_R)-z_R = 0,$$

namely,

$$\beta_R(s) = (2S_R)^{-1}\left[s+\mu+S_R-\sqrt{(s+\mu+S_R)^2-4\mu S_R}\right]. \qquad (6.13)$$

Hence

$$P_0^*(s) = \beta_R^{m+1}\left[\mu(1-\beta_R)\right]^{-1} = \mu^{-1}\sum_{j=m+1}^{\infty}\beta_R^j.$$

Then by direct inversion from tables (Erdelyi et al, 1954),

$$P_0(t) = \mu^{-1} \sum_{j=m+1}^{\infty} (\mu/S_R)^{j/2} \frac{j e^{-t(\mu+S_R)}}{t} I_j(2t\sqrt{\mu S_R}) , \qquad (6.14)$$

where $I_j(x)$ is the modified Bessel function of the first kind. The expected queue of class R customers, for initial conditions (6.10), is

$$E(n_R(t)|m) = m + (\lambda_R - \mu)t + \mu \int_0^t P_0(\tau)\, d\tau + \sqrt{\mu S_R} \int_0^t \frac{(t-\tau)}{\tau} e^{-\tau(\mu+S_{R-1})} I_1(2\tau\sqrt{\mu S_{R-1}})\, d\tau . \qquad (6.15)$$

The two comparable formulae for the process M/M/1 without service interruptions are, from Chapter II, respectively

$$P_{m0}(t) = \mu^{-1} \sum_{j=m+1}^{\infty} (\mu/\lambda)^{j/2} \frac{j e^{-t(\mu+\lambda)}}{t} I_j(2t\sqrt{\mu\lambda}),$$

$$E(r|m,t) = m + (\lambda - \mu)t + \mu \int_0^t P_{m0}(\tau)\, d\tau .$$

7. Duration of Busy Periods

To end this chapter we investigate the duration of a busy period for the simplified preemptive model when $R=2$ and both service times follow the negative exponential distribution. We are particularly interested in the case when the service rates are not equal. Let the service distributions of the priority and non-priority customers be respectively

$$dH_1(t) = \mu_1 e^{-\mu_1 t} dt,$$

$$dH_2(t) = \mu_2 e^{-\mu_2 t} dt,$$

and denote the number of customers of each class present in the system by r,n respectively. We seek the distribution of the first passage time from the point $(0,m)$ to $(0,0)$, so that there is an absorbing barrier at the origin. The space in which the hypothetical random walk takes place is then the quarter plane bounded by reflecting barriers along the lines $r=0$ and $n=1$, and the point $(0,0)$.

Let $B_{rn}(t)$ be the transition probabilities of this process with initial conditions

$$B_{rn}(0) = \delta_{r0}\delta_{nm}. \qquad (7.1)$$

The forward equations satisfied by the Laplace transforms $B^*_{rn}(s)$ are then

$$-(s+\mu_1+\lambda_1+\lambda_2)B^*_{rn}+\mu_1 B^*_{r+1\,n}+\lambda_1 B^*_{r-1\,n}+\lambda_2 B^*_{r\,n-1}=0, \quad \begin{matrix} r=1,2,\dots, \\ n=2,3,\dots, \end{matrix} \tag{7.2}$$

$$-(s+\mu_1+\lambda_1+\lambda_2)B^*_{r1}+\mu_1 B^*_{r+1\,1}+\lambda_1 B^*_{r-1\,1}=0, \quad r=1,2,3,\dots, \tag{7.3}$$

$$-(s+\mu_2+\lambda_1+\lambda_2)B^*_{0n}+\mu_1 B^*_{1n}+\lambda_2 B^*_{0\,n-1}+\mu_2 B^*_{0\,n+1}=-\delta_{nm}, \quad n=2,3,\dots, \tag{7.4}$$

$$-(s+\mu_2+\lambda_1+\lambda_2)B^*_{01}+\mu_1 B^*_{11}+\mu_2 B^*_{02}=-\delta_{1m}, \tag{7.5}$$

$$-sB^*_{00}+\mu_2 B^*_{01}=0. \tag{7.6}$$

By the same method as before we find for the joint generating function

$$G^*(s;x,z)=B^*_{00}(s)+\sum_{r=0}^{\infty}\sum_{n=1}^{\infty}x^r z^n B^*_{rn}(s)$$

the expression

$$G^*(s;x,z)=\frac{[\mu_2-z(\mu_2-\mu_1+\mu_1\omega_2)+sz]B^*_{00}(s)-z^{m+1}(1-x\omega_1)^{-1}}{\mu_2-z(\mu_2-\mu_1+\mu_1\omega_2)}, \tag{7.7}$$

where $w_1(s,z)$ is defined by (2.16), namely

$$w_1(s,z) = (2\mu_1)^{-1}\left[s+\mu_1+\lambda_1+\lambda_2-\lambda_2 z - \sqrt{(s+\mu_1+\lambda_1+\lambda_2-\lambda_2 z)^2 - 4\mu_1\lambda_1}\right].$$

Applying the same regularity condition to the right hand side of (7.7), we find that the denominator has exactly one zero inside the unit circle if $\mathrm{Re}\, s > 0$. If this zero is $z=z_0(s)$, then equating the numerator to zero when z has this value we have

$$B^*_{00}(s) = s^{-1} z_0^m(s).$$

The probability density function of the first passage time we require is

$$\frac{d}{dt}B_{00}(t) = \mathcal{L}^{-1}\left[z_0^m(s)\right]. \qquad (7.8)$$

$z_0(s)$ is the root of smallest absolute value of an equation similar to (6.1) which can be expanded as

$$z^3\lambda_2(\mu_1-\mu_2) + z^2\left[\lambda_1\mu_1+\lambda_2\mu_2-(\mu_1-\mu_2)(s+\mu_2+\lambda_1+\lambda_2)\right]$$
$$-z\mu_2\left[s+\mu_1+\lambda_1+\lambda_2-2(\mu_1-\mu_2)\right] + \mu_2^2 = 0. \qquad (7.9)$$

We do not solve (7.9) for $z_0(s)$ but instead find the moments of the distribution function $B_{00}(t)$ by means of the following property of the Laplace transform:

$$E(t^n|m) = \lim_{t\to\infty}\int_0^t \tau^n\, dB_{00}(\tau) = \left[(-1)^n\frac{d^n}{ds^n}z_0^m(s)\right]_{s=0}.$$

Assuming a series expansion for $z_0(s)$ of the form

$$z_0(s) = a_0 + a_1 s + a_2 s^2 + \ldots \quad , \tag{7.10}$$

the coefficients a_i can be found by equating like powers of s to zero after substituting (7.10) in (7.9). The first three a_i are given by the following equations:

$$a_0^3\lambda_2(\mu_1-\mu_2) + a_0^2\left[\mu_2(\lambda_1+\lambda_2)-(\mu_1-\mu_2)(\lambda_2+\mu_2)\right] + a_0\left[\mu_1-2\mu_2-\lambda_1-\lambda_2\right] + \mu_2^2$$

$$= (a_0-1)\left[a_0^2\lambda_2(\mu_1-\mu_2) + a_0\mu_2(\mu_2-\mu_1+\lambda_1) - \mu_2^2\right] = 0 \quad , \tag{7.11}$$

$$a_1\left\{3a_0^2\lambda_2(\mu_1-\mu_2) + 2a_0\left[\mu_2(\lambda_1+\lambda_2)-(\mu_1-\mu_2)(\lambda_2+\mu_2)\right] + \mu_2(\mu_1-2\mu_2-\lambda_1-\lambda_2)\right\}$$

$$- a_0^2(\mu_1-\mu_2) - a_0\mu_2 = 0 \quad , \tag{7.12}$$

$$a_2\left\{3a_0^2\lambda_2(\mu_1-\mu_2) + 2a_0\left[\mu_2(\lambda_1+\lambda_2)-(\mu_1-\mu_2)(\lambda_2+\mu_2)\right] + \mu_2(\mu_1-2\mu_2-\lambda_1-\lambda_2)\right\}$$

$$+ 3a_0a_1^2\lambda_2(\mu_1-\mu_2) + a_1^2\left[\mu_2(\lambda_1+\lambda_2) - (\mu_1-\mu_2)(\lambda_2+\mu_2)\right]$$

$$- 2a_0a_1(\mu_1-\mu_2) - a_1\mu_2 = 0. \tag{7.13}$$

Examination of these equations shows that the form of $z_0(s)$ depends on the range of $\rho_1+\rho_2 = \lambda_1/\mu_1 + \lambda_2/\mu_2$ in a manner analogous to that of the unrestricted single server problem. Thus normalisation is to unity, i.e. $z_0(0) = a_0 = 1$ if and only if $\rho_1+\rho_2 \leq 1$. When equality holds we have a confluent case and (7.11) factorises

further into

$$(a_0-1)^2\left[a_0\lambda_2(\mu_1-\mu_2)+\mu_2^2\right] = 0. \tag{7.14}$$

The other a_i, $i=1,2,\ldots$, become infinite so no moments exist. When $\rho_1+\rho_2<1$, moments of all orders exist; in particular, since $a_0=1$

$$E(t|m) = -ma_1, \tag{7.15}$$

$$E(t^2|m) = m(m-1)a_1^2 + 2ma_2, \tag{7.16}$$

where m is the initial number of non-priority customers. The mean and variance of the distribution $B_{00}(t)$ are then respectively

$$E(t|m) = m\left[\mu_2(1-\rho_1-\rho_2)\right]^{-1}, \tag{7.17}$$

$$V(t|m) = \frac{m\left[1+\rho_2-\rho_1(1-2\mu_2/\mu_1)\right]}{\mu_2^2(1-\rho_1-\rho_2)^3}. \tag{7.18}$$

Writing $m=1$ yields results for the distribution of the duration of a busy period.

If $\mu_1=\mu_2=\mu$, the explicit expression for $z_0(s)$ is

$$z_0(s) = \frac{1}{2(\lambda_1+\lambda_2)}\left[s+\mu+\lambda_1+\lambda_2-\sqrt{(s+\mu+\lambda_1+\lambda_2)^2-4\mu(\lambda_1+\lambda_2)}\right]. \tag{7.19}$$

In this case the probability density of the first passage time is, from (7.8),

$$dB_{cc}(t) = \left(\frac{\mu}{\lambda_1+\lambda_2}\right)^{m/2} \frac{me^{-t(\mu+\lambda_1+\lambda_2)}}{t} I_m(2t\sqrt{\mu(\lambda_1+\lambda_2)})\, dt. \qquad (7.20)$$

(7.20) is a simple generalisation of equation (4.15) of Chapter II for the unrestricted process M/M/1.

Table 3 exhibits the mean and variance defined by (7.17) and (7.18) when $m=1$ and $\mu_1=\mu_2=1$. The tabulated values are for a fixed $\rho_2 = 0\cdot1$ and varying interruption rate ρ_1. For values of $\rho_2 = 0\cdot1+\delta$ where $\delta = \cdot1, \cdot2, \ldots, \cdot8$, the appropriate values of the table commence 10δ entries down from the top. As may be expected the effect on the variance of service interruptions is far more marked than their effect on mean duration.

ρ_1	V(t/1)	E(t/1)
0.0	1.509	1.111
0.1	2.344	1.250
0.2	3.790	1.429
0.3	6.481	1.666
0.4	12.000	2.000
0.5	25.000	2.500
0.6	62.963	3.333
0.7	225.000	5.000
0.8	1900.000	10.000

Table 3: Variance and mean of duration of a busy period for fixed $\rho_2 = 0\cdot1$.

CHAPTER V

QUEUES WITH CORRELATED INTERARRIVAL TIMES

1. The Input Process

The queueing systems considered so far in this thesis have been distinguished by the fact that their input processes were of the renewal or recurrent type. In Chapter I we asserted that the only systems of this sort that are of real interest are those in which the input is Poisson or deterministic. It is now necessary to justify this assertion and to discuss input processes of other types. We retain the assumption that the service times of customers are independently and identically distributed and are independent of the input.

If in fact the input is a renewal process then the interarrival times $\tau_n = t_{n+1} - t_n$, $n = 1, 2, 3, \ldots$, are independently and identically distributed. Another way of stating this is that each arrival time t_n is the sum of $n-1$ independent random variables, each with the same distribution, in addition to the interval between $t=0$ and the first arrival time (which may also follow this distribution). This implies that customers arrive only in a way which ensures the independence of interarrival times. An arrival pattern of this sort may occur in certain circumstances but clearly does not hold in general, even when the interarrival distribution is arbitrary. The major objection to a renewal input is that the independence assumptions it incorporates do not make for a

reasonable model of many queueing situations. For a mathematical model to be valid it must reflect the essential features of the reality it purports to represent. Since the input process is the idealisation of the arrival behaviour of customers it is first necessary to classify under broad headings what appear to be the most important ways in which customers do in fact arrive. Three such headings are

(i) completely random arrivals,

(ii) scheduled arrivals (including the possibility of a customer being late or early),

(iii) general independent arrivals.

The third of these, denoted by the symbol GIA , will be explained later. To avoid confusion we recall that some authors have used the term 'general independent arrivals' to describe the GI input in which the interarrival times are independently and arbitrarily distributed. A more appropriate term for the latter is renewal input, denoted by the letter R . Arrival patterns classified under (i) and (ii) lead to input processes which are Poisson and deterministic, respectively, or variants of these.

This classification is by no means exhaustive, but is useful not only because it includes many cases of interest, but also because models based on (i)-(iii) are

amenable to mathematical treatment. We do not include patterns accurately described by a general renewal process amongst the more important types of arrival behaviour. Situations in which interarrival times are independently and identically distributed appear to be of a very special kind, for example when two service facilities are in series so that the arrival times at the second facility are the departure times from the first. Care is needed in the definition of interarrival times in models of this sort since Finch (1959) has shown that, except in very special cases, successive interdeparture intervals of a single server renewal queue are not independently distributed.

The limited value of queueing models with recurrent input has long been recognised (see Winsten, 1959, and the discussion to the paper). It seems fair to say that the attention paid to models of this sort in which the input is neither Poisson nor deterministic has been justified in the past by the hope that they approximate in some way to more general behaviour. The usefulness of a renewal input is that the mathematical difficulties due to the presence of dependent random variables are avoided, but whe-

ther or not the approximation is a good one depends very much on the particular case. The mathematical problem is to devise new methods, probably of a combinatorial nature, which do not rely on the special independence properties of the input. It was stated in § 3 of Chapter I that Winsten (1959) and Benes (1960a, 1960b) have obtained results in this direction and it is possible that more attention to the basic combinatorial problem will yield the required new methods. In these circumstances it seems hardly worthwhile retaining the GI input as an approximation procedure.

The importance of input processes which are Poisson or deterministic is that they do provide reasonable models of the arrival behaviour of customers in given situations. The fact that these two processes are of renewal type follows from their special properties and not from independence assumptions about arrival patterns. The Poisson process (sometimes with time-dependent parameters) has long been used by telephone engineers and others and has apparently proved a satisfactory model. A deterministic input arises when customers are scheduled to arrive at constant intervals and has application in the study and design of appointment systems and the like in which the arrival of customers is organised in some way. Winsten (1959) analysed the single server queue with deterministic input essen-

tially by the method of the imbedded Markov chain and showed how this approach can be used to study the case when scheduled customers are allowed to be late. The time a customer can be late follows a lateness distribution $L(x)$ which is the same for all customers. Then instead of $t_n = n\nu^{-1}$, ν a positive constant, the actual arrival times are $t_n + e_n$, where $Pr\{e_n \leq x\} = L(x)$, and the interarrival times are thus correlated. By assuming different forms for $L(x)$ it is clear that a wide variety of arrival behaviour can be described by this model. The only difficulty is that calculations are very lengthy if the distributions involved are not simple (Mercer, 1960).

Winsten's model is pertinent to systems in which the arrival of customers is planned in advance, or at least subject to some control. In the next section we introduce a model which appears to describe more accurately the case in which the arrival pattern of customers is not organised in any way. Before doing so we make two points. Firstly, that it does not seem possible at the present time to analyse in detail general models not incorporating the assumption of independent service times. Relaxing this assumption is a more serious matter than doing the same with the input, since the service process is conditional in the sense that it only operates when at least one customer is pre-

sent. Furthermore assuming independent service times seems a reasonable first approximation to reality. Only partial results are available for simple models incorporating service times dependent on queue length (see Saaty, 1959, pg. 354). The second point refers to the work of Benes previously mentioned in Chapter I. We recall that Benes makes no assumptions about the input (or the service process) and for the single server queue he has obtained an integro-differential equation giving the distribution of virtual waiting time $\eta(t)$ in terms of the distribution of the 'work load' $\xi(t)$. Thus for a given input process the waiting time distribution is obtained by solving the integro-differential equation of Benes. However as previously noted it is not possible to find the distribution of queue length by this method.

2. The Input GIA

The queueing model we now describe is designed to apply to situations in which the arrival behaviour of customers is not subject to control. We consider a system in which a single server caters for a population of customers $\mathcal{C}$. If one takes as the independence assumption of the input that the arrival time t_C of customer $C \in \mathcal{C}$ is chosen without knowledge of the arrival times of all other members of $\mathcal{C}$, then observing the behaviour of customers $A, B, C, \ldots$ one has the arrival times $t_A, t_B, t_C, \ldots$. Provided there has been no collusion between customers these observed values are all independent. If these observations are ordered

$$t_1 \leq t_2 \leq t_3 \leq \cdots ,$$

then it is sensible to speak of the nth arriving customer. Knowledge of the input process is thus contained in the ordering of independent observations and in general the nth ordered arrival time will not be the sum of independently and identically distributed random variables. If it is further assumed that

$$Pr[t_C \leq t] = A(t) \quad , \quad \text{all } C \in \mathcal{C},$$

then the set of ordered arrival times $\{t_i\}$ constitute a sample of ordered independent observations from the parent distribution $A(t)$. In this case we call $A(t)$ the arrival distribution.

We will denote a single server queueing system with this input and service process of independent renewal type by the notation GIA/G/1 . The letters GIA stand for general independent arrivals, the word general denoting that the arrival distribution is arbitrary. The assumptions specific to this model can be stated concisely as follows;

(i) the arrival time of a customer follows an arrival distribution $A(t)$ which is the same for all customers,

(ii) this arrival time is selected independently to that of all other customers.

An example in which independence assumptions of this nature appear to be particularly appropriate is the passage of traffic through a toll station. Consider such a system over the finite interval of time $[0,T]$, say midnight to midnight of consecutive days. Then a plausible arrival distribution is one with a bimodal density function, the modes appearing at peak hours. It seems more reasonable to postulate the independent arrival of vehicles at the

toll station than to assume a renewal input or the lateness model of Winsten. The assumption that the arrival distribution is the same for all customers would seldom if at all hold in practice but seems a reasonable first approximation. The form of the arrival distribution can be estimated by observing the actual arrival times of customers. It is interesting to note in this connection that if a renewal model is adopted a priori then the quantities to be observed are the interarrival intervals. For Winsten's model it is necessary to measure the deviations of arrival times from preassigned values in order to estimate the lateness distribution. A matter of some importance (which is not discussed here) is to discover to what extent predictions of queue behaviour based on each of three models differ from each other.

A mathematical description of the GIA input is as follows: Let $A(t)$, $0 \leq t$, be an absolutely continuous distribution function with density

$$dA(t) = a(t)\,dt.$$

If the population $\mathcal{C}$ of customers consists of N members, the probability density of the arrival time t_r of the

rth arriving customer is

$$dF_r(t) = f_r(t)\,dt = \frac{N!}{(r-1)!(N-r)!}\,[A(t)]^{r-1}[1-A(t)]^{N-r}a(t)\,dt\,,\ r=1,2,\cdots,N. \quad (2.1)$$

The times between the arrival of customers are the distances between order statistics

$$y_r = t_r - t_{r-1}\,,\ r=1,2,\cdots,N,$$

with $t_0=0$. The probability density of y_r is

$$dG_r(y) = g_r(y)\,dy = \frac{N!\,dy}{(r-2)!(N-r)!}\int_0^\infty [A(x)]^{r-2}[1-A(x+y)]^{N-r}a(x+y)a(x)\,dx. \quad (2.2)$$

The probability that r arrivals occur in the interval $[0,t)$ is

$$C_r(t) = \frac{N!}{r!(N-r)!}\,[A(t)]^r[1-A(t)]^{N-r}. \quad (2.3)$$

In the case of a renewal input process the probability on the right hand side of (2.3) is the r fold convolution of the interarrival distribution.

In the model we are now considering completely random arrivals are described by a uniform arrival distribution

$$A(t) = \begin{cases} t/T\,, & 0 \le t \le T, \\ 1\,, & T \le t. \end{cases} \quad (2.4)$$

We use the symbol UIA to denote the input which has this arrival distribution. In this case

$$f_r(t)\,dt = \frac{N!}{(r-1)!(N-r)!}\left(t/\bar{T}\right)^{r-1}\left(1-t/\bar{T}\right)^{N-r} d\left(t/\bar{T}\right), \qquad (2.5)$$

$$g_r(y)\,dy = \left(N/\bar{T}\right)\left(1-y/\bar{T}\right)^{N-1} dy, \qquad (2.6)$$

and the distribution of y_r is independent of r. If T and N are allowed to tend to infinity together we find that (2.5), (2.6) approach the appropriate distributions of the Poisson input. Writing $\bar{T}=N\lambda^{-1}$ we have

$$\lim_{N\to\infty} f_r(t)\,dt = e^{-\lambda t}\,\frac{(\lambda t)^{r-1}}{(r-1)!}\,\lambda\,dt,$$

$$\lim_{N\to\infty} g_r(y)\,dy = \lambda e^{-\lambda y}\,dy.$$

Thus for large N and T it is possible to approximate to the UIA input by a Poisson process, the error involved being the same as the exponential approximation to the binomial. Simple approximations of this sort by renewal processes do not in general exist for other types of GIA input. It is also important to note that we are essentially concerned with a queue problem in finite time since

it may not always be sensible to talk about an equilibrium state for this model. If N is finite then with probability one all customers are served and discharged from the system in finite time.

3. The system GIA/M/1

We have indicated previously that methods are not yet available for the analysis of queueing processes such as GIA/G/1 . It seems impossible to apply either of the standard Markovisation procedures to such a general process and we consider only GIA/M/1 . This latter is given a quasi-Markovian character by specialising the service distribution to negative exponential. Let the negative exponential service distribution be

$$dB(t) = \mu e^{-\mu t} dt. \qquad (3.1)$$

A consequence of this is that if there are r customers in the queue at the beginning of a time interval of length t , the probability that m of them complete service in this interval is

$$c_m(t) = e^{-\mu t} \frac{(\mu t)^m}{m!} \quad , \; m = 0, 1, 2, \ldots, r-1.$$

The probability that all r complete service before time t is

$$b_r(t) = \int_0^t e^{-\mu\tau} \frac{(\mu\tau)^r}{r!} d\tau = 1 - \sum_{j=0}^{r-1} c_j(t) .$$

We assume the $\mathcal{C}$ consists of N members and that the arrival distribution is A(t) .

The main difference between the present system and renewal queues as far as the writing down of equations for the queue length probabilities is concerned is that we now have to distinguish between different interarrival periods. As in the method of the imbedded Markov chain we first consider the queue only at arrival epochs. Let $q_{nj}(t)\,\delta t$ be the probability density of the joint event that the jth customer arrives in $(t-\delta t, t)$ to yield a queue length of $n, n=1,2,..,j$; $j=1,2,...,N$. Then the $q_{nj}(t)$ satisfy the following recurrence relations

$$q_{11}(t) = N[1-A(t)]^{N-1}a(t), \tag{3.2}$$

$$q_{1j}(t) = \sum_{m=1}^{j-1}\int_0^t q_{m,j-1}(t-\gamma)\,g_j(\gamma)\Big[1-\sum_{r=0}^{m-1}c_r(\gamma)\Big]d\gamma \quad , j=2,3,\dots,N, \tag{3.3}$$

$$q_{nj}(t) = \sum_{m=n-1}^{j-1}\int_0^t q_{m,j-1}(t-\gamma)\,g_j(\gamma)\,c_{m-n+1}(\gamma)\,d\gamma \quad , \begin{matrix} n=2,3,\dots,j; \\ j=2,3,\dots,N. \end{matrix} \tag{3.4}$$

$g_j(y)$ is the density function of the jth distance defined in (2.2). (3.2) is the probability density of the arrival time of the first customer, and on this arrival occurring queue length must be unity. The other equations are obtained by considering the number of departures that can occur between the arrival of the (j-1)th and jth customers. If this interarrival time is y (with density $g_j(y)$) and there were m customers in the queue at time

t-y when the (j-1)th arrival occurred, then m-n+1 departures must take place in time y to have a queue length of n on the arrival of the jth customer at time t . The possible values of m are n-1,n,n+1,..,j-1 ; and m must be summed over this range. Starting from $q_{11}(t)$, the other $q_{nj}(t)$ can be found recursively from (3.3) and (3.4).

As a check on the derivation of these formulae we note that the probability that the jth arrival occurs in $(t-\delta t, t)$ is given from (3.3), (3.4) by summing over n . Thus

$$\sum_{n=1}^{j} q_{nj}(t) = q_{\cdot j}(t) = \int_0^t g_j(y) \sum_{m=1}^{j-1} q_{mj-1}(t-y)\, dy.$$

For j=2 we have

$$q_{\cdot 2}(t) = \int_0^t g_2(y)\, N\,[1-A(t-y)]^{N-1} a(t-y)\, dy$$

$$= \frac{N!}{1!\,(N-2)!}\,[A(t)]\,[1-A(t)]^{N-2}\, a(t).$$

Induction on j yields

$$q_{\cdot j}(t) = \frac{N!}{(j-1)!\,(N-j)!}\,[A(t)]^{j-1}\,[1-A(t)]^{N-j}\, a(t).$$

The probability that the jth arriving customer finds a queue of length n ahead of him on arrival is

$$P_{nj} = \int_0^\infty q_{n+1\,j}(t)\,dt \quad , \; n = 0, 1, 2, \ldots, j-1. \tag{3.5}$$

For completeness we derive expressions for the probabilities $P_n(t)$ of n customers in the queue at the arbitrary time t . Assume that the last arrival before time t is the jth and that it occurred at time t-u . If this arrival yielded queue length m then for n customers at t , m-n departures must occur in the interval u and the (j+1)th customer must arrive later than t . The probability density associated with this joint event is

$$q_{mj}(t-u)\left[1 - G_{j+1}(u)\right] c_{m-n}(u) \quad , \; m = n, n+1, \ldots, j,$$

where $G_{j+1}(u) = \int_0^u g_{j+1}(x)\,dx \quad , \; j = 1, 2, \ldots, N-1,$

and $G_{N+1}(u) = 0$. When n=0 it is also necessary to take into the account the possibility of no arrivals occurring in time t . Hence

$$P_0(t) = \left[1 - A(t)\right]^N + \sum_{j=1}^{N} \sum_{m=1}^{j} \int_0^t q_{mj}(t-u)\left[1 - G_{j+1}(u)\right]\left[1 - \sum_{r=0}^{m-1} c_r(u)\right] du, \tag{3.6}$$

$$P_n(t) = \sum_{j=n}^{N} \sum_{m=n}^{j} \int_0^t q_{mj}(t-u)\left[1 - G_{j+1}(u)\right] c_{m-n}(u)\,du, \; n = 1, 2, \ldots, N-1, \tag{3.7}$$

$$P_N(t) = \int_0^t q_{V_{NN}}(t-u)\, c_0(u)\, du \,. \qquad (3.8)$$

Unfortunately it is difficult to simplify these expressions.

4. The Distribution of Waiting Time

To conclude this thesis we indicate briefly how the general methods developed by Benes can be applied to the system GIA/G/1 to find the distribution of the virtual waiting time $\eta(t)$. We have stated previously that Benes obtained a representation of

$$W(t,x) = P_r[\eta(t) \leq x]$$

in terms of the distribution of the 'work load' $\xi(t)$ (see Figure 1 at the end of Chapter I). The Benes equations (Benes, 1960b, pg.140) for an empty system initially are

$$W(t,x) = P_r[\xi(t) - t \leq x] - \frac{\partial}{\partial x}\int_0^t P_r[\xi(t) - \xi(u) - t + u \leq x \,|\, \eta(u) = 0]\, W(u,0)\, du, \tag{4.1}$$

$$\int_0^t P_r[\xi(t) - \xi(u) \leq t - u \,|\, \eta(u) = 0]\, W(u,0)\, du = \int_0^t P_r[\xi(t) \leq u]\, du. \tag{4.2}$$

The first step is to solve (4.2) for $W(t,0)$, the probability that a customer arriving at t does not have to wait. Once this is known the continuous component of the distribution is given by substitution in (4.1). The term

$\Pr[\xi(t)-\xi(u)\le t-u \mid \eta(u)=0]$ is the probability that the sum of the service times of the customers arriving in the interval $[u,t)$ is less than or equal to $t-u$, conditional on an empty queue at time u. It is the compound probability

$$\Pr[\xi(t)-\xi(u)\le t-u \mid \eta(u)=0] = \sum_{r=0}^{\infty} \Pr[r \text{ arrivals in } [u,t)]\, H^{*(r)}_{(t-u)}, \qquad (4.3)$$

where $H^{*(r)}_{(x)}$ is the r fold convolution of the service distribution and $H^{*(0)}_{(x)} = 1$.

Let $A(t)$ be the arrival distribution of the system GIA/G/1. If the population of customers consists of N individuals we have from (4.3)

$$\Pr[\xi(t)-\xi(u)\le t-u \mid \eta(u)=0] = \sum_{r=0}^{N} \binom{N}{r} [A(t)-A(u)]^r [1-A(t)+A(u)]^{N-r} H^{*(r)}_{(t-u)},$$

and

$$\Pr[\xi(t)\le x] = \sum_{r=0}^{N} \binom{N}{r} [A(t)]^r [1-A(t)]^{N-r} H^{*(r)}_{(x)}.$$

Equation (4.2) for the null probability $W(t,0)$ of GIA/G/1 is then

$$\sum_{r=0}^{N} \binom{N}{r} [A(t)]^r [1-A(t)]^{N-r} \int_0^t \left\{ 1-\left(1-\frac{A(u)}{A(t)}\right)^r \left(1+\frac{A(u)}{1-A(t)}\right)^{N-r} W(u,0) \right\} H^{*(r)}_{(t-u)}\, du \qquad (4.4)$$

$$= 0.$$

The continuous part of the waiting time distribution is from (4.1)

$$W(t,x) = \sum_{r=0}^{N} \binom{N}{r} [A(t)]^r [1-A(t)]^{N-r} \left\{ H^{*(r)}_{(x+t)} - \frac{\partial}{\partial x} \int_0^t \left(1 - \frac{A(u)}{A(t)}\right)^r \left(1 + \frac{A(u)}{1-A(t)}\right)^{N-r} W(u,0)\, H^{*(r)}_{(x+t-u)}\, du \right\}. \tag{4.5}$$

(4.5) is similar to the results of Chapter III for renewal queues in which the waiting time distribution is expressed in terms of the null probability. Corresponding to the branching process type equation whose roots determine the null probability we now have (4.4), a Volterra equation of the first kind. The solution to this equation is not easy to obtain even for simple arrival and service distributions, and we do not investigate the problem further here.

It is apparent that explicit results are not easy to find for queueing systems with GIA input. In conclusion we point out again that the justification for introducing the model lies in the fact that it seems a reasonable approximation to a wide class of actual queueing situations. Furthermore the mathematics involved are not so formidable as to make the problem hopeless and it is possible that a more detailed analysis will result from further work.

REFERENCES

1. Bailey, N.T.J. (1954)
A continuous time treatment of a simple queue using generating functions.
J.R. Statist. Soc. (B), 16, 288-291.

2. Benes, V.E. (1960a)
Combinatory methods and stochastic Kolmogorov equations in the theory of queues with one server. Trans. Amer. Math. Soc. 94, 282-294.

3. Benes, V.E. (1960b)
General stochastic processes in traffic systems with one server. Bell System Tech. J. 39, 127-160.

4. Barry, J.Y. (1956)
A priority queueing problem. Opns. Res. 4, 385.

5. Bartlett, M.S. (1956)
An Introduction to Stochastic Processes.
Cambridge University Press.

6. Batchelder, P.M. (1927)
An Introduction to Linear Difference Equations.
Harvard University Press.

7. Brockmayer, E., Halstrom, H.L., and Jensen, A. (1948)
The Life and Works of A.K. Erlang. Copenhagen.
Copenhagen Telephone Company.

8. Champernowne, D.G. (1956)
An elementary method of solution of the queueing problem with a single server and constant parameters. J.R. Statist. Soc. (B), 18, 125-128.

9. Clarke, A.B. (1956)
A waiting line process of Markov type. Ann. Math. Statist. 27, 452-459.

10. Connolly, B.W. (1958)
A difference equation technique applied to the simple queue. J.R. Statist. Soc. (B), 20, 165-167.

11. Cox, D.R. (1955)
The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. Proc. Camb. Phil. Soc. 51, 433-441.

12. Erdelyi, A. (1954)
Tables of Integral Transforms, 1. New York: McGraw-Hill.

13. Feller, W. (1957)
An Introduction to Probability Theory and its Applications, 2nd ed. New York: John Wiley and Sons.

14. Feller, W. (1959)
Probability and Statistics (ed. U. Grenander)

15. Finch, P. (1959)
The output process of the queueing system M/G/1. J.R. Statist. Soc. (B). 21, 375-380.

16. Gani, J. (1958)
Elementary methods for an occupancy problem of storage. Math. Annalen, 136, 454-465.

17. Gaver, D.P. (1959)
Imbedded Markov chain analysis of a waiting line process in continuous time. Ann. Math. Statist. 30, 698-720.

18. Harris, T.E. (1948)
Branching processes. Ann. Math. Statist. 19, 474-494.

19. Heathcote, C.R. and Moyal, J.E. (1959)
The random walk in continuous time and its application to the theory of queues. Biometrika, 46, 400-411.

20. Heathcote, C.R. (1959)
The time-dependent problem for a queue with preemptive priorities. Opns. Res. 7, 670-680.

21. Heathcote, C.R. (1960a)
A simple queue with several preemptive priority classes. Opns. Res. 8,

22. Heathcote, C.R. (1960b)
Preemptive priority queueing. Submitted to Biometrika.

23. Karlin, S., and McGregor, J. (1958)
Many server queueing processes with Poisson input and exponential service times. Pacific J. Math. 8, 87-118.

24. Keilson and Kooharian (1960)
On time dependent queueing processes. Ann. Math. Statist. 31, 104-112.

25. Kendall, D.G. (1949)
Stochastic processes and population growth. J.R. Statist. Soc. (B), 11, 230-264.

26. Kendall, D.G. (1951)
Some problems in the theory of queues. J.R. Statist. Soc. (B), 13, 151-185.

27. Kendall, D.G. (1953)
Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann. Math. Statist. 24, 338-354.

28. Kesten, H. and Runnenberg, J.T. (1957)
Priority in waiting line problems. Proc. Kon. Ned. Akad. van Weten. (A). 60, 312-336.

29. Kiefer, J. and Wolfowitz, J. (1955)
On the theory of queues with many servers. Trans. Amer. Math. Soc. 78, 1-18.

30. Ledermann, W. and Reuter, G.E.H. (1954)
Spectral theory for the differential equations of simple birth and death processes. Phil. Trans. A, 246, 312-369.

31. Lindley, D.V. (1952)
The theory of queues with a single server. Proc. Camb. Phil. Soc. 48, 277-289.

32. Luchak, G. (1958)
The continuous time solution of the equations of the single channel queue with a general class of service time distributions by the method of generating functions. J.R. Statist. Soc. (B), 20, 176-181.

33. Mercer, A. (1960)
A queueing problem in which the arrival times of the customers are scheduled. J.R. Statist. Soc. (B). 22, 108-113.

34. Miller, R.G. (1960)
Priority queues. Ann. Math. Statist. 31, 86-103.

35. Morse, P.M. (1958)
Queues, Inventories and Maintenance. New York: John Wiley and Sons.

36. Moyal, J.E. (1957)
Discontinuous Markov processes.
Acta Mathematica, 98, 221-264.

37. Moyal, J.E. (1960)
Incomplete discontinuous Markov processes.
To appear in Quart. J. Math. Oxford.

38. Otter, R. (1949)
The multiplicative process. Ann. Math. Statist 20, 206-224.

39. Saaty, T.L. (1959)
Mathematical Methods of Operations Research.
New York: McGraw-Hill.

40. Smith, W.L. (1953)
On the distribution of queueing times.
Proc. Camb. Phil. Soc. 49, 449-461.

41. Stephan, F.F. (1958)
Two queues under preemptive priority.
Opns. Res. 6, 399-418.

42. Takacs, L. (1955)
Investigation of waiting time problems by reduction to Markov processes. Acta Math. Acad. Sci. Hung. 6, 101-128.

43. Takacs, L. (1960)
Transient behaviour of single-server queueing processes with recurrent input and exponentially distributed service times. Opns. Res. 8, 231-245.

44. White, H. and Christie, L.S. (1958)
Queueing with preemptive priorities or with breakdown. Opns. Res. 6, 79-96.

45. Whittaker, E.T. and Watson, G.N. (1940)
Modern Analysis. 4th ed. Cambridge University Press.

46. Widder, D.V. (1946)
The Laplace Transform. Princeton University Press.

47. Winsten, C.B. (1959)
Geometric distributions in the theory of queues. J.R. Statist. Soc. (B) 21, 1-35.

48. Wishart, D.M.G. (1956)
A queueing system with χ^2 service time distribution. Ann. Math. Statist. 27, 768-779.