SOME ASPECTS OF QUEUEING AND STORAGE PROCESSES

A thesis in partial fulfilment of the

requirements for the degree of Master of Science

in statistics at

Massey University

Trevor Milton CRAIG

March 1986

ABSTRACT

In this study the nature of systems consisting of a single queue are first considered. Attention is then drawn to an analogy between such systems and storage systems. A development of the single queue viz queues with feedback is considered after first considering feedback processes in general. The behaviour of queues, some with feedback loops, combined into networks is then considered. Finally, the application of such networks to the analysis of interconnected reservoir systems is considered and the conclusion drawn that such analytic methods complement the more recently developed mathematical programming methods by providing analytic solutions for sub systems behaviour and thus guiding the development of a system model.

CONTENTS

ERRATA

Page 53 line 2 should read as follows:
'the reservoir running dry during the given period [o,T]. Extending the interval to [0,T'] indicates that a dam'

In the appended lists of references for chapters 2 and 4, the name PHATARFOD, R.M. is misspelt as PHATARFOD, E.M. and as PHATAFOD, R.M.

CHAPTER 1 - QUEUEING SYSTEMS

1.1 INTRODUCTION

When customers arrive at a station where a particular service is offered, a queue of customers can form if the demand for service exceeds the ability of the server to supply the service immediately. A queueing system is formed when one or more such demand/supply structures operate in conjunction with one another. Although queueing systems often appear in which people are the customers, e.g. the queue formed by people waiting for service from a teller in a bank, the same demand/supply structure can be recognised in more diverse systems. One such example is a telephone system in which the placing of a telephone call corresponds to the arrival of a 'customer' and the 'service' provided is the provision of a telephone circuit for the duration of the call. When the call is completed the caller vacates the system by freeing the circuit for use by other callers. If the latter had rung during the initial call, they would have formed a queue waiting for the circuit to become free, or would have balked on finding the line was 'busy'. If they had found the line was busy for longer than they cared to wait, they could have cancelled their call, thus reneging from the queue. The terms customer and server are thus often applied figuratively. The rapid development of telephone systems at the turn of the century led to a need by telephone engineers for rules for determining the number of connecting lines, operators, etc, in order to handle the demand adequately but in an economical way. Until this time, rules of thumb based on experience had been applied, but the increasing demand for telephone services and complexity in telephone systems necessitated a more systematic approach. This was initiated in 1917 by A.K. Erlang who used probability distributions to describe the variation in the number of calls arriving/unit time and the variation in the length of these calls. He was thus able to determine a distribution function for the number of calls waiting and the distribution of waiting times. The probability distributions used by Erlang were either negative exponential or constant. Later work extended his results to other distributions for which the mathematical treatment proved to be much less tractable. Saaty (1961) gives details of this earlier development and an extensive bibliography up to 1961. Kleinrock (1976) also covers this material as an introduction to more recent work on the behaviour of more elaborate queueing systems viz computer networks.

The large variety of queueing networks as well as, within each queue, the infinity of combinations of interarrival time distributions, service time distributions, queue disciplines and number of servers has led to a large queueing theory literature (see 1.6). The development of digital computers has enabled systems, which because of their complexity were difficult to treat analytically, to be studied by simulation methods. Again, because of the variety of models and methods to be considered, a large literature on computer simulation has developed to which reference will be made later. In this study use is made of analytic and simulation methods to investigate the behaviour of a hydro-storage system. This is modeled as a queue-network incorporating a feedback mechanism between separate queues.

Although arrival and service processes in real queueing systems are generally found to be stochastic in nature, it is useful to consider first the limiting case in which the timing of future events is known.

1.2 DETERMINISTIC QUEUEING MODELS

Queueing models in which the interarrival times and service times are constant (i.e. deterministic) are useful as approximations to real queues in which, for a period at least, the variation in the interarrival and service times is limited. Such models, by being conceptually simpler than stochastic models, also enable a clearer view of the interaction between the arrival and service streams to be obtained. Figure 1 below illustrates a queueing system which is empty at time t=o and which has first-in-first-out (FIFO) queueing discipline. Also the arrival rate is less than the service rate (in order that the queue length should not keep increasing) and as each service is completed a new one is begun. Then clearly the number in the system n(t) = (no. of arrivals in (o,t]) - (no. of services completed in (o,t])

$$= [\lambda t] - [\mu t - \frac{\mu}{\lambda}] \quad \text{where } [x] = \text{integer part of } x,\ x \geq o$$

If the system size is limited to k-1 say, then this equation is valid only until time t_i where $n(t_i) = k$. Any customers arriving until the end of the current service will balk and the system size will remain at k-1. At the time of the next service completion, n(t) will drop to k-2, unless an arrival occurs at the same instant in which case n(t) remains at k-1. Arrival and service completion events coincide if and only if $1/\lambda$ is a multiple of $1/\mu$

i.e.
$$n(t) = \begin{cases} 0 & (t < 1/\lambda) \\ [\lambda t] - [\mu t - \mu/\lambda] & (1/\lambda \le t < t_i) \\ k-1 & (t \ge t_i) \end{cases} \qquad (1.1)$$

Since in the example shown on Figure 1, $\frac{1}{\lambda} = 4$, $\frac{1}{\mu} = 8$ and k=5 Equation (1.1) shows that in this case $t_i = 32$ and so for $t \geq 32$ the system is in a steady state.

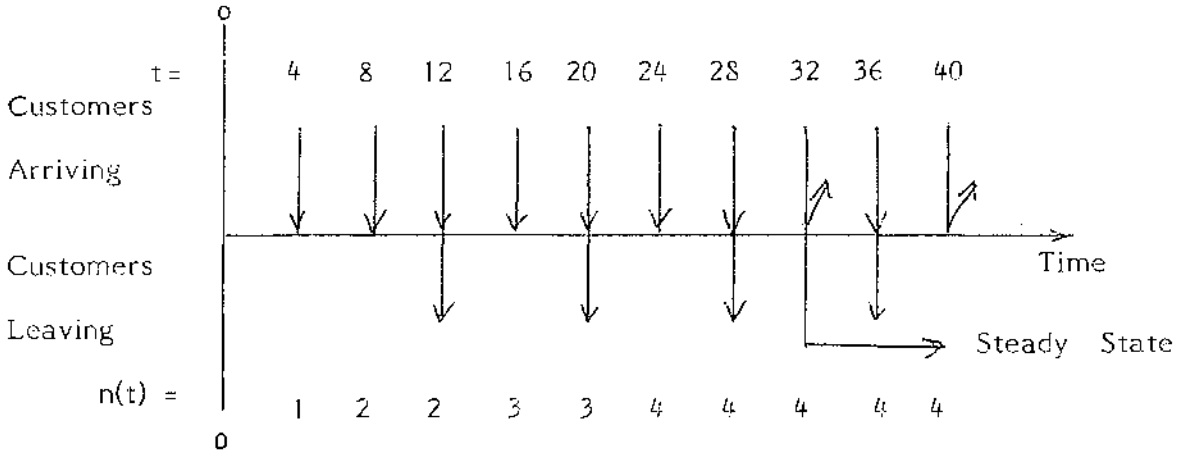

Figure 1

In addition to n(t), the number in the system at time t, another important measure of the queue performance is the waiting time experienced by customers forced to join the queue. Writing $W_q^{(n)}$ for the time spent in the queue by the n^{th} customer, $S^{(n)}$ for the same customer's service time and $T^{(n)}$ for the time between the arrivals of the n^{th} and $(n+1)^{th}$ customers the following recurrence relation exists between these values:

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)} & (W_q^{(n)} + S^{(n)} - T^{(n)} > 0) \\ 0 & (W_q^{(n)} + S^{(n)} - T^{(n)} \le 0) \end{cases} \qquad (1.2)$$

This is illustrated in Figure 2 and holds whether the times $S^{(n)}$ and $T^{(n)}$ are deterministic or stochastic:

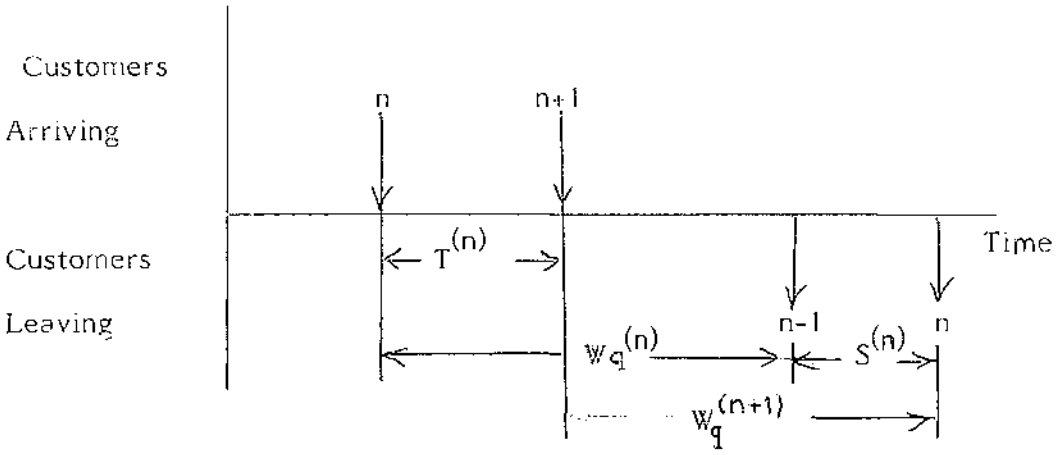

Figure 2

In the present example for n=8 equation (1.2) yields $W_q^{(n)}$ as follows:

$S^{(n)} = 8$ and $T^{(n)} = 4$ so that $W_q^{(n+1)} = W_q^{(n)} + 4$

i.e. $\Delta W_q^{(n)} = W_q^{(n+1)} - W_q^{(n)}$

$= 4$

so $W_q^{(n)} = 4n-4$

since $W_q^{(1)} = 0$

If n 8 each arrival (which does not balk) finds k-2 customers already in the system and each requiring a service time $S^{(n)} = 8$ thus if k=5,

$$W_q^{(n)} = \begin{cases} 4(n-1) & n<8 \\ 24 & n \geq 8 \end{cases}$$

In the case in which $1/\mu$ is not a multiple of $1/\lambda$ a diagramatic method as used in Figure 1 reveals that the system size undergoes a cyclic pattern of change, the length of the cycle being equal to the least common multiple of $1/\mu$ and $1/\lambda$. Further complications can be introduced by having the system start off in a non-empty state or by changing the queue discipline, etc. Exact solutions are always obtainable by graphing however since all factors are deterministic. As the number of complications increases or the queues are combined into networks, graphical methods become more cumbersome and methods of approximation, to be considered later in connection with stochastic queueing models, become appropriate.

1.3 GRAPHICAL REPRESENTATION OF CUMULATIVE FUNCTIONS

Although the diagrams in section 1.2 clarify the relationships between individual arrival and service events, the nature of cumulative processes over a period of time is less clear. Three cumulative processes are of particular interest:

$A(t)$ = Cumulative quantity or number to arrive by time t,

$D(t)$ = Cumulative quantity or number to enter service by time t,

$D^*(t)$ = Cumulative quantity or number to have left service by time t;

all of these functions are monotonic non-decreasing.

Clearly $Q(t)$ = quantity in queue or queue length at time t

$= A(t) - D(t)$

and $S(t)$ = quantity or number in service

$= D(t) - D^*(t)$

so that $A(t) \geq D(t) \geq D^*(t)$ since $Q(t) \geq 0$, $S(t) \geq 0$

These relationships hold for any queue discipline or number of servers or number of servers and are illustrated in Figure 3 for the example in Section 1.2.

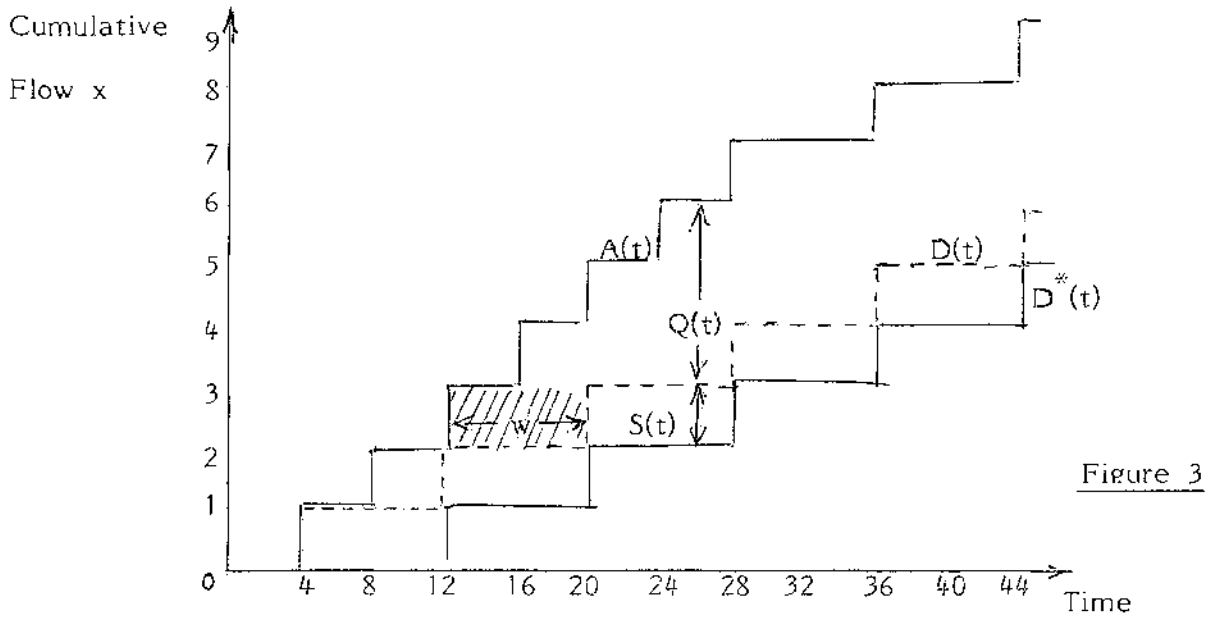

Figure 3

For FIFO queue discipline the horizontal distance marked w in Figure 3 represents A(12)-D(20) i.e. w = time spent in the queue by the 3rd customer. The height of the shaded rectangle is 1 unit, so that waiting time accumulated by all customers up to time t is equal numerically to the area between the A(t) and D(t) curves. For non FIFO queue disciplines or queues with more than one server, the simple interpretation of the length marked w in Figure 3 may not be valid and an alternative approach is needed. This is achieved by considering x as the independent variable and constructing as in Figure 4 a graph of $\Delta(x)$ where $\Delta(x)$ is the time of departure from the queue of x^{th} cumulative arrival, i.e. for $j-1 < x < j$, $\Delta(x)$ = departure time from the queue of the j^{th} arrival whereas $D^{-1}(x)$ was the time of the x^{th} cumulative departure. Thus x is now the label given to the customer on arrival.

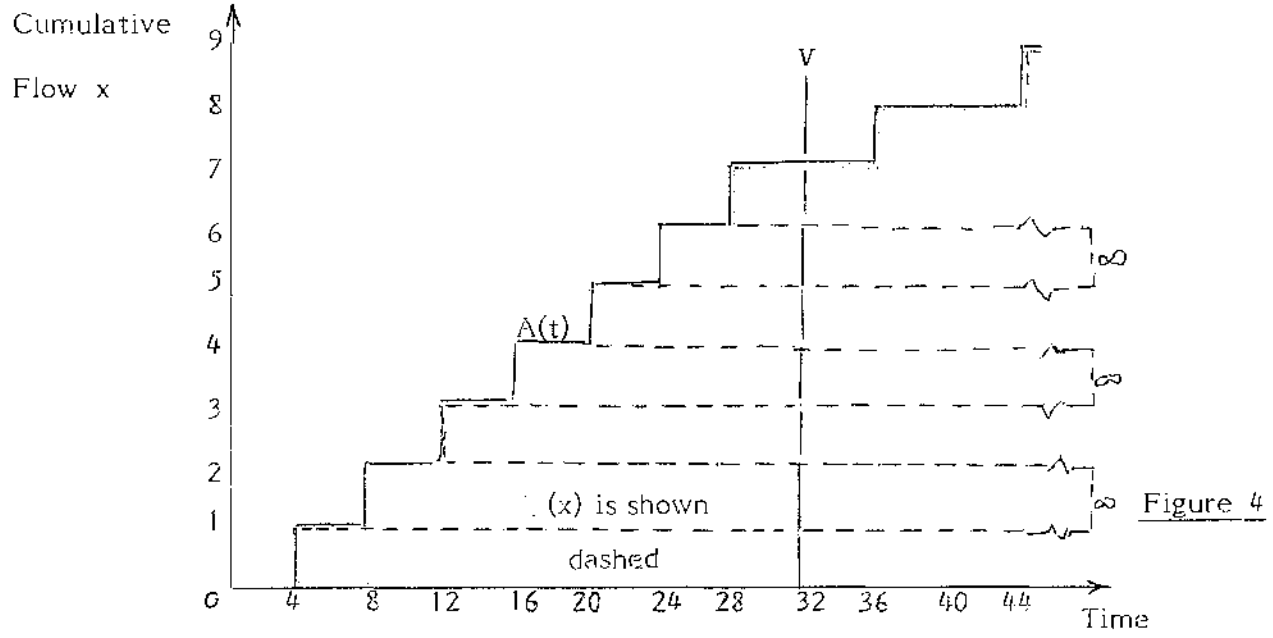

Figure 4

The shape of $\Delta(x)$ depends on the queue discipline and if this is FIFO the graphs $\Delta(x)$ and $D^{-1}(x)$ are identical. In Figure 4 the graph of $\Delta(x)$ corresponds to a last-in-first-out (LIFO) queue discipline applied to the queue graphed in Figure 1. In this case the 2nd, 4th and 6th customers wait in the queue forever for service, all other new arrivals arriving just at the instant a service is completed and so having priority to enter service.

Thus $$\Delta(x) - A^{-1}(x) = \begin{cases} 0 \text{ if } x \neq 2, 4 \text{ or } 6 \\ \infty \text{ if } x = 2, 4 \text{ or } 6 \end{cases}$$

corresponding to the horizontal measurement w in Figure 3.

A vertical line V drawn at t=32 in Figure 4 indicates that the cumulative total remaining in the queue at time t is now the sum of possibly several segments rather than just one (i.e. Q(t)) as in Figure 3. The same method applies to service completions if there is more than one server. Clearly, graphs such as Figures 3 and 4 can be drawn for stochastic as well as deterministic queues if the arrival and departure times of each customer are known. However, even if only the cumulative functions A(t) and D(t) or $\Delta(x)$ are known, useful averages representing the queue behaviour can be found from such graphs. Evaluating the area between the graphs of A(t) and D(t) in Figure 3 using vertical strips and then horizontal strips illustrates Little's formula $L = \lambda w$ (J.D.C. Little, "A Proof for the Queueing Formula : $L = \lambda w$" Operations Research 9 383-387, 1961.) i.e. average queue length = arrival rate x average queue time/customer since in the example $3 = \frac{1}{8} \times 24$. Before considering these graphical methods in the fluid approximation of queues, it is necessary to consider the influence of the introduction of random variation into the arrival and service streams.

1.4 PROBABILISTIC DESCRIPION OF ARRIVAL AND SERVICE PROCESSES

The state of a queueing system in which arrival and/or service times vary in a random way, is not precisely predictable at future moments in time. The probability of a queue having a particular length at a particular time t in the future can be related to the probability of it changing from one length to another during a given interval as follows:

Let $P_t(\ell_1)$ = the prob. that the queue has length ℓ_1 at time t

Let $T_t(\ell_1,\ell_2,a)$ = the prob. of changing from length ℓ_1 to ℓ_2 during the interval $[t,t+a)$,

then by the theorem of total probability,

$$P_{t+a}(\ell_2) = \sum P_t(\ell_1)T_t(\ell_1,\ell_2,a) \quad \text{[summing over all lengths } \ell_1\text{]} \qquad (1.3)$$

The probability $T_t(\ell_1,\ell_2,a)$ will depend on the number of arrivals and departures during $[t,t+a)$. In turn the number of arrivals will depend on the length of the interval and the time at which the last customer arrived before t, the start of the interval. Similarly, the number of departures will depend on the length of the interval and the length of time any services in progress at the start of the interval had been going on. This dependancy makes solving (1.3) difficult except for particular distributions of arrivals and service times. It is useful therefore to consider the case in which this dependancy is absent.

Let f(t) be the p.d.f. of the interarrival intervals

and $F(t) = \int_0^t f(u)du$ i.e. the d.f. and let $p(t, \delta t)$

be the probability of an arrival during $(t, t+(\delta t)$

Then $p(t, \delta t)$ = the proportion of the dist. in the interval $(t\ t+\delta t)$ / the proportion of the dist. in excess of t

then $p(t, \delta t) = \dfrac{f(t)\,\delta t}{1-F(t)}$

If now $P(t, \delta t)$ is independent of t, then $P(t, \delta t) = Kt$ say, where K is a constant.

$\therefore K\delta t = f(t)\,\delta t/(1-F(t))$

and hence $f(t) = Ke^{-Kt}$

As it can be shown (Parzen (1962)) that the negative exponential distribution is the only continuous p.d.f. with this Markov property, the central importance of this p.d.f. in queueing theory is apparent. Implied in the above derivation are the following three assumptions which define a Poisson process $\{N(t), t\geq 0\}$

(i) prob. of an arrival during the interval $[t,t+\delta t) = \lambda\delta t + o(\delta t)$

i.e. $p(t, \delta t) = \lambda\delta t + o(\delta t)$ where λ is a constant independant of $N(t)$ and $\lim_{\delta t\to 0} \frac{o(\delta t)}{\delta t} = 0$

(ii) prob. of more than one arrival during $[t,t+\delta t) = o(\delta t)$

(iii) the number of arrivals in non-overlapping intervals are statistically independant.

i.e. the process has independant increments. With these assumptions, the probability $P_n(t)$ of $n(n \geq 0)$ arrivals during a time interval of length t can be found by the following method which appears often in queueing theory.

$$\begin{aligned} P(t+\delta t) = {} & \Pr\{n \text{ arrivals in } t \text{ and zero in } \delta t\} \\ & + \Pr\{n-1 \text{ arrivals in } t \text{ and } 1 \text{ in } \delta t\} \\ & + \Pr\{n-2 \text{ arrivals in } t \text{ and } 2 \text{ in } \delta t\} \\ & + \ldots + \Pr\{0 \text{ arrivals in } t \text{ and } n \text{ in } \delta t\}, \quad n \geq 1 \end{aligned} \tag{1.4}$$

by (i), (ii) and (iii) (1.4) becomes

$$P_n(t+\delta t) = P_n(t)[1-\lambda\delta t + o(\delta t)] + P_{n-1}(t)[\lambda\delta t + o(\delta t)] + o(\delta t) \tag{1.5}$$

where the last term represents $P_n\{n-j \text{ arrivals in } t \text{ and } j \text{ in } \delta t;\ 2 \leq j \leq n\}$

If n=o then $P_o(t+\delta t) = P_o(t)[1-\lambda\delta t - o(\delta t)]$ (1.6)

Rewriting (1.5) and (1.6) and taking limits as $\delta t \to 0$ gives

$$\frac{dP_o(t)}{dt} = -\lambda P_o(t)$$

and $$\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t), \quad n \geq 1$$

The general solution of these linear first order differential equations using the boundary conditions $P_o(0) = 1$ and $P_n(0) = 0$, $n \geq 1$ is

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad n \geq 0 \tag{1.7}$$

From (1.7) it follows that $E[N(t)] = \lambda t$ i.e. N(t) has a mean arrival rate λ. That the Poisson process has stationary increments i.e. for $t > s$, N(t) - N(s) and N(t+h) - N(s+h) are identically distributed, is seen by noting that assumption (iii) i.e. independent increments implies that there is no loss of generality if N(s) and N(s+h) are assumed to be zero. If the above derivation is carried out under assumptions (i), (ii) and (iii) the same formula results for N(t) as for N(t+h).

The close association between the Poisson process and the exponential distribution can be seen from (1.7) as follows. If T is the random variable 'time between successive arrivals' then

$$\Pr\left\{T \geq t\right\} = \Pr\left\{\text{zero arrivals in time } t\right\} = P_o(t) = e^{-\lambda t}$$

Letting A(t) be the d.f. of T, it follows that

$$A(t) = \Pr\left\{T \leq t\right\} = 1-e^{-\lambda t}$$

thus T has the exponential distribution with mean $1/\lambda$, which is intuitive since the mean arrival rate is λ. Conversely it can be shown that if the interarrival times are independent and have the same exponential distribution, then the arrival rate follows the Poisson distribution. This Poisson/exponential arrival process is sometimes referred to as completely random arrivals. This is because of the following property of a Poisson process. Given that k arrivals have occured during an interval [0,T] the k times $T_1 < T_2 < \ldots < T_k$ at which the arrivals occured are distributed as the order statistics of k uniform random variables on [0,T]

This is shown as follows:

Writing $f_{T_1,T_2,\ldots,T_k}(t_1,t_2,\ldots,t_k|k \text{ arrivals in } [0,T])\, dt_1\, dt_2 \ldots dt_k \equiv f_T(t|k)dt$

$$= \Pr\left\{t_1 \leq T_1 \leq t_1 + dt_1, \ \ldots\ , t_k \leq T_k \leq t_k + dt_k | k \text{ arrivals in } [0,T]\right\}$$

$$= \frac{\lambda dt_1 e^{-\lambda dt_1} \ldots \lambda dt_k e^{-\lambda dt_k} \quad e^{-\lambda(T-dt_1-dt_2 - \ldots -dt_k)}}{(\lambda T)^k e^{-\lambda T}/k!}$$

which reduces to $f_T(t/k) = k!/T^k$ which is the joint density function of the order statistics of k random variable on [0,T]

The arrival process considered above can also be used to describe the service pattern if in assumptions (i) - (iii), the term arrival is replaced by service and if the probabilities are conditioned on the system being non-tempty. In addition, the basic Poisson/exponential process can be generalised in several ways which include (a) truncating the infinite range (b) allowing to depend on t, i.e. the process becomes non-homogeneous (c) allowing that more than one occurence in t has probability greater than o(t) i.e. a batch process.

More general renewal processes than the Poisson process can be used to describe arrival/service patterns as can non-renewal processes. These will be considered later where appropriate but the Poisson process often appears in the description of a queueing system either directly or in an imbedded process or as a first approximation. The reason for this is not just the tractable properties of this process but also because many real-life processes obey at least approximately the requirements (i) - (iii) listed previously. Information theory provides an additional argument. This is that the information content for the distribution f(x), defined as $\int_0^\infty f(x)\log f(x)dx$, is lease for the exponential function and as such provides a conservative description of arrival and service patterns.

WAITING TIME AND BUSY TIME DISTRIBUTIONS

These features of queue behaviour are, unlike the measures of effectiveness, dependant on the queue discipline. It is also noted that the time a ficitious customer would have to wait were he to arrive at an arbitrary point in time, i.e. the virtual waiting time, has a steady state distribution equal to that of the waiting time of an actual customer iff theinput is Poisson. For the present model the waiting time distribution $W_q(t)$ is part discrete and part continuous.

$$W_q(t) = \begin{cases} 1-\rho & t=0 \\ 1-\rho e^{-\mu(1-\rho)t} & t>0 \end{cases}$$

The mean waiting time (via a Riemann-Stieltjes integration) is $W_q = \frac{\lambda}{\mu(\mu-\lambda)}$

Similarly for the total time spent in the system,

$$W(t) = (\mu-\lambda)e^{-(\mu-\lambda)t} \quad t>0$$

and $W = E[T] = \frac{1}{\mu-\lambda}$

These results exemplify again Little's formula

$$L_q = \lambda W_q \qquad \text{or} \qquad L = \lambda W$$

which is valid under much less stringent restrictions than those of the present model.

Since a busy period continues as long as there is at least one item in the system $P_0(t)$ is seen to be the d.f. of the busy period and $P'_0(t)$ the p.d.f..

giving $P'_0(t) = \frac{2\sqrt{\mu\lambda}\, e^{-(\lambda+\mu)} I_1(2\sqrt{\mu\lambda}\, t)}{t}$ and also

$$E[T_{Busy}] = \frac{1}{\mu-\lambda}$$

1.5 PROBALISTIC DESCRIPTION OF QUEUES

The simplest probabilistic queuing model is the single server model with exponential interarrival and service times and having a first-in-first-out (FIFO) queue discipline.

To find an equation relating the state probabilities

$P_n(t) = \Pr\{n \text{ in the system at time } t\}$ a similar method to that used to obtain (1.4) is employed.

Applying the assumptions set out in § 1.4 gives

$$P_n(t+\delta t) = P_n(t)(1 - \lambda\delta t - \mu\delta t) \quad P_{n+1}(t)(\mu\delta t) + P_{n-1}(t)(\lambda\delta t) + o(\delta t),\ n \geq 1$$

Similarily $P_0(t+\delta t) = P_0(t)(1-\lambda\delta t) + P_1(t)(\mu\delta t) + o(\delta t)$

Rewriting and taking limits as $\delta t \to 0$ gives

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \lambda P_{n-1}(t),\ n \geq 1$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \qquad (1.8)$$

Bailey (1956) gave the following solution to (1.8)

$$P_{in}(t) = e^{-(\lambda+\mu)t}\left[\left(\frac{\mu}{\lambda}\right)^{(i-n)/2} I_{n-i}(2\sqrt{\lambda\mu}\,t) + \left(\frac{\mu}{\lambda}\right)^{(i-n+1)/2} I_{n+i+1}(2\sqrt{\lambda\mu}\,t) + \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \sum_{l=n+i+2}^{\infty}\left(\frac{\mu}{\lambda}\right)^{l/2} I_l(2\sqrt{\lambda\mu}\,t)\right],\ n \geq 1 \qquad (1.9)$$

where i is the number in the system at time t=o, and

$I_n(y) = k^{-n}J_n(ky)$ where $J_n(y)$ is the regular Bessel function. Using the assymptiotic approximation $I_n(y) \sim \frac{e^y}{\sqrt{2\pi y}}$ it can be shown that as $t \to \infty$, $P_n(t) \to (1-\frac{\lambda}{\mu})(\frac{\lambda}{\mu})^n$ (1.10)

The symbol ρ is often used to represent the rate which although dimensionless is often given in 'erlangs' in honour of A.K. Erlang. Thus (1.9) becomes $P_n = (1-\rho)\rho^n, \rho<1$]

Considering the complexity of (1.9) and its derivation, in this the simplest of probabilistic queueing models, it is fortunate that it is often the limiting distribution which is of most interest. By taking the limit as $t \to \infty$ in (1.8), the resulting equations can be used to determine (1.10) directly. There are a number of ergodic theorems which consider the existence of steady-state solutions but, as in this case, the conditions under which a queueing process is ergodic often becomes apparent from other considerations.

Having determined the steady-state probability distribution of the system size, representative characteristics of the system called measures of effectiveness can be calculated. These include the expected number in the system (L) and the expected number in the queue, (L_q) (i.e. customers actually waiting - excluding the customer being served).

$L = E[N] = \sum_{n=0}^{\infty} n P_n = \frac{\rho}{1-\rho}$ Similarly, $L_q = E[N_q] = \frac{\rho^2}{1-\rho}$

Also of interest is the expected size of non-empty queues i.e. $L'_q = E[N_q | N_q \neq 0]$

Let $P'_n = Pr\{ \text{n in the system} \mid n \geq 2 \}$

then $L'_q = \sum_{n=2}^{\infty} (n-1) P'_n = \frac{1}{1-\rho}$

The classification of more general queue models was facilitated by a notation due to Kendall.

1.6 KENDALLs NOTATION

D.G. Kendall (1953) introduced a notation later modified to the following form A|B|m|K|L to clarify different queue models. In the notation A and B represent the distribution function of the inter arrival time and the service time respectively. m represents the number of servers, K the system capacity (queue plus service) and L the size of the customer population. The arrival and service streams are considered to be sequences of random variables having independent and identical distributions. Morse (1957) discusses sampling of a queueing system to obtain the A and B distributions and describes the Erlang and hyper-exponential distributions which provide a reasonable representation of sampling distributions found in practice. The letters M = Markov (i.e. Poisson process), G = general distribution, D = deterministic are used in positions A and B in the notation.

The moments and often the distributions of many of the random variables which are involved in the description of queue behaviour of many different queue models have been determined. The complexity of the derviation of these results tends to increase as either A or B or both differ from a Markov process. Brief references to this work follows.

Page (1972) considers various queueing models which have Erlang, Poisson or deterministic inter arrival or service time distributions. Graphs and tables of system variables are provided. Some comment is made about priority queues with reference to Jaiswal (1968) for a more complete discussion. Takacs (1961) in addition to batch arrival processes discusses the application of queueing models to particle counting and considers queues with infintiely many servers. Riordan (1962) includes a discussion on virtual waiting time and queue disciplines other than FIFO.

Morse (1957) considers the derviation of sampling distributions from arrival and service processes. Erlang and hyper-exponential distributions are used as models and useful tables of these distributions are provided. Saaty (1961) considers the ergodic properties of queues and in addition to a study of queues with Poisson and non-Poisson input and service processes, provides an interesting discussion of less common queueing models. These include cooperating parallel channels and cyclic queues. A final chapter indicates the wide range of problems to which queueing models have been applied including semi-conductor noise, hospitals and the demand for medical care, as well as an introduction to dams and storage systems which is further considered in Chapter 2. Kosten (1973) considers the M|G|m model under several different queue disciplines and queue size restrictions.

An introduction to computer simulation methods includes a comparison of simulation and analytic methods in investigating the behaviour of queues. Gross and Harris (1974) provide a systematic coverage of Markovian queue models. They then procede to a discussion of semi-Markovian and ergodic processes in queues with general arrival and/or service processes as well as the use of approximation methods with such models. A chapter on computer simulation and simulation languages is followed by a final chapter detailing a case study involving queueing theory and simulation which provides and example of optimizing a queueing model using a cost critereon. Kleinrock (1975) Volume 1 provides a fuller discussion of the more general models G|M|m and G|G|1 and in Volume 2 considers queue networks and their application to time-shared computers

NEWELL (1972, 1982) has considered queueing models in an engineering context emphasising approximation and graphical methods. Borokov (1976) has presented general and unified mathematical treatments of aquite wide class of queueing models. Kingman (1966) discusses an algebriac approach to generalising treatment of the G|G|m queue model for m 1.Syski (1967) discusses the Pollaczek method in queueing theory and Prahbu (1974 discusses another technique of wide application in queueing theory, the Wiener-Hopf method. In addition to developing unifying treatments and general methods of analysis, more recent papers have treated topics in the optimising of queue operation, queueing network theory, simulation of queues as well as the analysis of queueing models arising in real systems. Reference to papers on these and related topics can be found in the followng and similar publications. Management Science, Operations Research, Information, ORSNZ, Journal of Advanced Probability, Journal of Applied Probability, Naval Research Logical Quarterly Journal. In the present context approximation methods and computer simulation are of particular relevance and are considered below.

1.7 APPROXIMATION METHODS, BOUNDS AND INEQUALITIES

In those cases in which an exact expression can be found for an expected value or the distribution of a variable of interest in a queueing model, it is often difficult to evaluate numerically and further, the assumptions about the conditions under which the result was derived may be difficult to verify. Consequently considerable effort has been spent in developing approximations, bounds and inequalities which are robust to changes from underlying assumptions and which are relatively quick to calculate. In practical situations the control of a queueing system becomes most critical when the traffic density is greatest i.e. as $\rho \to 1$. Kleinrock (1975) Volume 2 gives a discussion of results obtained for the G|G|1 queue model in the heavy traffic case. Central to to these results is that the waiting time distribution is approximately exponentially distributed with mean wait given by: $(\sigma_a^2 + \sigma_b^2)/2(1-\rho)\bar{t}$

where σ_a^2, σ_b^2 are the variances of the interarrival and service times respectively, ρ is the traffic density and $\bar{t}$ is the mean interarrival time. It is also shown that this mean wait forms an assymptotically sharp upper bound for the mean wait in any G|G|1 queue. The case for a lower bound on the mean waiting time is less clear cut and the results obtained depend on the nature of the input process. Bounds on the tail of the distribution of waiting time are given and bounds on the mean waiting time for the

G|G|m (m 1) model are derived. All of these results are approximations and bounds for the exact solution. An alternative procedure is to find exact solutions to an approximation of the original problem. As was noted in § 1.2 the basic recurrence relation for waiting times

$$W^{(n+1)} = \max\left[0, W^{(n)} + S^{(n)} - T^{(n)}\right]$$

holds for deterministic or stochastic arrival and service processes. Consequently an approximation to the system behaviour can be obtained by approximating the stochastic processes which control the operation of the system. A 'first-order' approximation is obtained by replacing the stochastic processes by their possibly time-dependent averages. This is called the fluid approximation method. By allowing each stochastic process to be represented by both its mean and its variance a 'second-order' approximation is obtained. This is termed a diffusion approximation. These methods enable the elementary deterministic methods outlined previously to facilitate the analysis of queues involving complex stochastic processes. To apply these approximation methods it is necessary first to estimate the appropriate parameters from the relevant stochastic processes. Gross and Harris (1974) discuss those aspects of statistical inference which relate to parameter estimation in queueing sytems. Computer simulation of queues provides another method of investigating queueing systems. As simulation amounts to statistical sampling, the approximations and bounds outlined above are helpful also in drawing inferences from the sample statistics produced by the computer simulation. (Fishman, 1974). An application of these methods is described in Chapter 5.