ON THE REDUCTION OF CONGESTION

by

Nasser Hadidi
University of Oslo, Oslo, Norway

and

Brian W. Conolly*

*) Present address; Via Narbostro 4, La Sera di Lerici, La Spezia, Italy
ON THE REDUCTION OF CONGESTION

Contents

I. Introduction
   1. Preliminary Remarks
   2. Suggested Means of Improvement
II. Conventional Methods
   3. The Appointment System, (D/G/s)
   4. Constant Service Times, (M/D/s)
III. State Dependent Queues
   5. The Model of Cox and Smith
   6. The Queue with Uniform Arrival Rate and State Dependence
   7. The Model with Linearly State Dependent Service Rate
      7.1 Transient Behaviour of Queue Length
      7.2 The Effective Service Time
      7.3 The Waiting Time Process
      7.4 The Busy Period and Output Processes
   8. Systems with Controlled Number of Channels
   9. Jackson's 'Jobshop Like' Queuing System
IV. Further Models
   10. Service-Inter-arrival Dependence
      10.1 The Waiting Time Process
      10.2 State Probabilities
      10.3 The Output Process
      10.4 The Initial Busy Period
   11. A 'Dual' Model
V. Numerical Comparisons
VI. Conclusions
    References
I. Introduction

The subject matter of Queueing Theory is sufficiently widely known to warrant little explanation. One is typically concerned with situations which can be described in terms of customers who demand service from a service facility. When this facility is limited to a finite number of servers the customers may have to wait or be turned away, to their inconvenience, annoyance and possible financial loss. On the other hand it may happen that one or more servers are idle, a situation possibly detrimental, both financially and from the organizational point of view, to the management of the facility.

It is intuitively obvious that the former situation becomes more aggravated as traffic intensity becomes high and the latter when traffic intensity is low. In this context traffic intensity is usually measured by the ratio of the mean service interval to the mean arrival interval. A very fundamental operational problem of queueing systems is accordingly that of finding ways of minimizing simultaneously both waiting time and idle time under a given level of traffic intensity.

1. Preliminary Remarks

The first probabilistic models of Queueing Theory appear to have been formulated by A.K. Erlang a mathematician and engineer employed by the Copenhagen Telephone Company in the first years of the present century. An account of his work is conveniently, if somewhat in-accessibly collected in V. Erlang recognised the need to describe demand in terms of a probability distribution of the intervals between successive arrivals, and of service in terms of a probability distribution of the length of time required to serve a customer. This leads
to the notation $X/Y/N$ introduced by D. G. Kendall (1953) in which $X$ describes probabilistically the nature of the inter-arrival intervals and $Y$ plays a similar role for service times. $N$ denotes the number of servers available. Particular cases of $X$ and $Y$ are denoted traditionally by $M$ and $D$; the former meaning a negative exponential distribution while the latter signifies that the interval in question has a fixed length.

The notation is usually taken to imply that both inter-arrival intervals and and service times are independently and identically distributed. There are many other facets of particular queueing situations which the notation does not describe and of these we might mention as a selection "queue discipline", or order of service; maximum queue size; the allocation of priorities to certain customers; and so forth.

Erlang found from observation that the most appropriate model for telephone traffic problems were $M/M/N$ and $M/D/N$. He was mainly interested in the "probability of loss", that is to say that the incoming call finds all lines (servers) busy. This can, of course, be reduced by increasing $N$ but then questions arise as to balancing the investment required to increase $N$ against the gain achieved from improved customer satisfaction.

The supermarket did not exist in Erlang time, but its formal similarity to a telephone exchange, with queues allowed, permits similar analysis and considerations. Speeding flow of customers can be achieved by increasing the number of channels but at the risk of creating server idleness when traffic intensity is low.

2. Suggested Means of Improvement

It is clear that one way to redress this practical
imbalance is to create conditions such that the rate of service is geared to the rate of arrival. Another possibility, perhaps less practical, could be to regulate demand according to service potential. Strangely enough, however, there is very little analysis on these lines to be found in the literally immense literature devoted to queueing and related topics. Attention seems in the first place to have been concentrated on the development of techniques for the solution of the basic problems in systems with independent service and demand in as great generality as possible; then on the variations of the main themes which derive from particular systems. Current research shows a tendency to explore and explain the underlying mechanisms and to discern common general features.

Circumstances led the present authors to investigate models which embody the spirit of gearing service to demand. This research has been directed not only at the analysis, but also at measuring practical consequences of the various systems. In the belief, based on extensive literature surveys, that work of this nature is rather rare it occurred to us that a review of our own work and that of other investigators who have considered models of a similar character, would fill what seems to be a rather astonishing gap in queueing literature, and point to areas of research in the practical aspects of queuing models which remain to be explored.

The review which follows will first in Section II cover briefly conventional models D/M/N (appointment systems) and M/D/N (partial regularity consequent on fixed service). D/M/N and M/D/N may be regarded as the simplest models attempting to achieve some measure of regularity without actually imposing
direct dependence between service and arrivals. This material
is well known but is included for convenience and reference
in Section V where numerical comparison is made between
various models.

Sections III and IV are largely condensed from the work
of the present authors which has or will appear in different
sources. Cognate models will be described.

The review is thus partly of theory and partly an
attempt to provide an operational evaluation of potential
gains. Few proofs will be given. In general systematic cover-
age is desirable and will be given at least of the stochastic
processes associated with waiting and idle time already dis-
cussed and also with the 'state' of the system (total number
of customers present, including those in service), busy period
(time interval during which the server is uninterruptedly
busy) and in some cases a description of output in some form
(number completing service in a given time interval, or possib-
by the interval between the completions of successive service).

In Section VI we shall attempt to summarize our findings
and suggest areas where further work seems to be needed.

II: Conventional Methods

A practical method commonly used in queueing situations
to reduce congestion is to control the arrival or service
patterns. The most usual control procedures which have recieve
extensive analysis are:
i) The establishment of an appointment system for customers;
ii) Allocating fixed service intervals.

We shall here summarize some aspects of these models.
3. The Appointment System, (D/G/s)

The queueing model D/G/s seems to have received its first attention in the literature because of its applicability to medical and related services. Apparently the first analytical treatment of the waiting time process for this model, with \( s=1 \), was given by Lindley (1952) as an example of his method of solving the problem for the general GI/G/1 system. The Wiener–Hopf type integral equation derived for the steady state waiting time distribution function \( (F(x)) \) takes the form

\[
F(x) = \int_{y \leq x} F(x-y) \, dG_1(y+1), \quad (3.1)
\]

in this case, where \( G_1(y) \) is the distribution function of the service times and the constant inter-arrival interval is taken to be unity. For the case where services have an Erlangian distribution \( E_k \),

\[
dG(y) = \frac{\mu^{n+1}}{n!} y^n e^{-\mu y}
\]

(i.e. \( 2\mu y \) enjoys a \( \chi^2 \) distribution with \( 2(n+1) \) degrees of freedom) he postulates and proves a solution of the form

\[
F(x) = 1 + \sum_{i=1}^{n} c_i e^{z_i x}, \quad (3.2)
\]

where the \( z_i \) are the roots (with real part <0) of

\[
e^z = (1+z/\mu)^{n+1} \quad (3.3)
\]

and the \( c_i \) are complex constants related to the roots by

\[
\frac{1}{\mu^{r+1}} + \sum_{i=1}^{n+1} \frac{c_i}{(\mu+z_i)^{r+1}} = 0, \quad (r=0,1,\ldots,n). \quad (3.4)
\]

Lindley then tabulates the probability of zero wait, and mean and variance of waiting time for \( n=0 \) and \( n=1 \) contrasting them with "random" (negative exponential) input.
Lindley's waiting time result was subsequently checked by Wishart (1956) who extended Luchak's (1956) work obtaining some results for GI/E_\text{K}/1.

The multiple server case was considered by Kendall (1953). He too discussed this system as a special example of his analysis of the GI/M/s model. For s=1,2,3 Kendall tabulates the probability of not having to wait and the ratio of mean waiting time to mean service time contrasting them with their random arrival counterparts. These results are naturally based on the roots of equation (3.3), (with n=0 in Kendall's case) which he calls the ($\mu,z$) equation.

From the point of view of application of the regular input system Bailey (1952) and (1954), and Welch and Bailey (1952) demonstrated the remarkable improvements that could be achieved in the operation of a hospital outpatient department. The slightly more general deterministic model in which the scheduled customer may arrive at any epoch in the two inter-arrival intervals prior and subsequent to his scheduled arrival time has been investigated by Winsten (1959) and Mercer (1960).

In a later Section we shall compare numerically the various control procedures. It is therefore appropriate here to also summarize other parts of the theory required to deal with D/M/1

Let Q(t) be the state of the system (i.e. total number of customers present including the one in service, if any). We usually specify Q(0) = 1, though this is not important for the steady state theory. If arrivals occur at epochs $t_n$, where $t_1=0$ is the epoch of the first arrival, then we write
\[ Q_n = Q(t_n - 0) \] and
\[ p_r^{(n)} = \Pr(Q_n = r | Q_0 = 0), \quad (0 \leq r < n) \]
and
\[ p_r(t) = \Pr(Q(t) = r | Q(0) = 1). \]

In addition we recall that the mean interarrival interval is taken to be unity and the mean service interval is \( \mu^{-1} \).

The probabilities \( p_r^{(n)} \) and \( p_r(t) \) are dealt with adequately in the literature (c.f. Prabhu (1965)). It is known that when \( \rho = 1/\mu < 1 \) both \( \lim_{n \to \infty} p_r^{(n)} = \overline{p}_r \) and \( \lim_{t \to \infty} p_r(t) = \overline{q}_r \) are non-zero. However
\[ \overline{p}_r = (1-x_0) x_0^r, \quad r \geq 0 \quad (3.5) \]
and
\[ \overline{q}_r = \rho(1-x_0) x_0^{r-1}, \quad r \geq 1 \quad (3.6) \]
\[ \overline{q}_0 = 1-\rho, \]
where \( x_0 \) is the root with smallest modulus of the functional equation
\[ x = e^{-(1-x)/\rho} \]
(i.e. equation (3.4) with \( n=0 \)). These results can be obtained easily by the difference equation technique of Conolly (1958). Formulae suitable for numerical calculation can be obtained by application of Lagrange's theorem (c.f. Whittaker and Watson (1946)).

The theory of busy period for \( D/M/1 \) can be deduced from Conolly (1959). Here we limit ourselves to the mean, \( E(t) \), and variance, \( V(t) \), of the length in time of a busy period. These are given by
\[ E(t) = \{\mu(1-x_0)\}^{-1} \]
and
\[ V(t) = (\rho+x_0)/\{\mu^2(1-x_0)^2(\rho-x_0)\}^{-1}. \]
4. Constant Service Times; M/D/s

We shall now consider the system with constant service times, Poisson input and s servers. For s=1,2,3 such a model was first considered by Erlang in connection with telephone calls of a fixed duration. He basically treats the problem of waiting time. Subsequently this model was examined again by Fry (1928), Molina (1927), Crommelin (1932) and Everett (1953) among others. Fry gave an elaborate argument regarding the state of the system which he subsequently used to investigate the probability of delay of a certain magnitude. An account of his works on this system is given in his book (1928) which also contains valuable information on other models. Molina (1927) by implementing obvious methods, found steady state probabilities for the state of the system. Then by arguing that upon the arrival of a customer who has to wait there could be a multiple of s plus m (< s) customers he derives the p.d.f. of the waiting time of that customer.

Perhaps the most elegant treatment of the model was given by Crommelin (1932). Let the constant service interval be taken to be 1/\mu and let the mean arrival rate be \lambda. Further let \overline{q}_r have the same definition as in Section 3 and

\[ a_r = \sum_{i=0}^{r} \overline{q}_i. \tag{4.1} \]

Then obviously

\[ \overline{q}_r = a_s \rho^r e^{-\rho} \frac{e^{\rho}}{r!} + e^{-\rho} \sum_{i=1}^{r} \overline{q}_{s+i} \frac{\rho^{r-i}}{(r-i)!}, \tag{4.2} \]

where \rho = \lambda/\mu. Now introduce the generating functions

\[ f(z) = \sum_{r \geq 0} z^r \overline{q}_r, \]

- 10 -
\[ F(z) = \sum_{r \geq 0} z^r a_r \]

and

\[ Q_n(z) = \sum_{r=0}^{n} z^r q_r. \]

After some manipulations we obtain from (4.2)

\[ f(z) = \frac{Q_s(z) - z^s a_s}{1 - z^s e^{\rho(1-s)}}. \tag{4.3} \]

Crommelin proceeds to show that

\[ z^s = e^{-\rho(1-z)} \tag{4.4} \]

(an equation formally identical to (3.3)) has exactly \( s \) roots in and on the unit circle. Denoting these roots by \( \xi_l \) and having the natural side condition \( f(1) = 1 \) he then obtains

\[ f(z) = -\frac{s-\rho}{(1-\xi_1)(1-\xi_2)\ldots(1-\xi_s)} \times \frac{(z-1)(z-\xi_1)\ldots(z-\xi_{s-1})}{1 - z^s e^{\rho(1-s)}}, \tag{4.5} \]

which upon expanding as powers of \( z \) would yield \( q_r \). The probability of not having to wait is easily seen to be \( a_{s-1} \).

We are next concerned with \( F(t) \) which is the probability that a customer arriving in the steady state has to wait less than \( t \). To find \( F(t) \) Crommelin first let \( t = T + \tau \) where \( T \) is an integral multiple of \( 1/\mu \) and \( \tau \) is a proper fraction of it.

He then introduced \( b_r(\tau) \) which is defined as follows:

Suppose that upon the arrival of a customer who has to wait there is a given number (naturally more than \( s \)) of customers in the system; \( b_r(\tau) \) is then the probability that after the ellapse of a time \( \tau \), \( r \) or less of them are still in the system. Then

\[ F(t) = F(T+\tau) = b_{TS+s-1}(\tau), \tag{4.6} \]
because we require that after \( \tau \) the state of the system be less than \( T_s + s-1 \); that is after \( T + \tau \) it is not more than \( s-1 \).

Now \( a_r \) is related to \( b_r \) as follows:

\[
a_r = e^{-\rho \tau} \sum_{i=0}^{r} b_i \left( \frac{(\rho \tau)^{r-i}}{(r-i)!} \right),
\]

(4.7)

which with

\[
G(z) = \sum_{r \geq 0} z^r b_r(\tau)
\]

gives

\[
F(z) = e^{-\rho \tau} \sum_{r \geq 0} \sum_{i=0}^{r} z^r b_i(\tau) \left( \frac{(\rho \tau)^{r-i}}{(r-i)!} \right)
\]

\[
= e^{-\rho \tau} \sum_{r \geq 0} z^r b_r(\tau) e^{\rho \tau z}
\]

\[
= G(z) e^{\rho \tau (z-1)}.
\]

To obtain \( F(t) \) from \( G(z) \), having \( F(z) \) is now a matter of routine. We obtain

\[
F(T+\tau) = \sum_{i=0}^{T} \sum_{j=0}^{s-1} a_j \frac{(-\rho (i+\tau))^{T-is+s-1-j} e^{\rho (i+j)}}{(T-is+s-1-j)!}.
\]

(4.8)

Naturally if \( \tau = 0 \) (the customer whose waiting time we are considering arrives at the epoch the service commences)

\( F(z) = G(z) \).

Erlang had obtained (4.8) for \( s=1,2,3 \). To obtain the mean waiting time is also a matter of routine. Crommelin showed that

\[
E(w) = \frac{1}{\rho} \sum_{i=1}^{s-1} (1-\xi_i)^{-1} + \frac{\rho^2 - s(s-1)}{2\rho (s-\rho)}.
\]

(4.9)

This result too was obtained by Erlang earlier for \( s=1,2,3 \).

Everett (1953) apparently not satisfied with the
"complicated" form (4.5) as the generating function of state probabilities gave a very crude method for iterative solution of the system (4.2). The approximation is based on the fact that \( \bar{q}_{l+1}/\bar{q}_l \) is proportional to a constant \( k \) which depends only on \( s \) and \( \lambda \). This yields a rapidly convergent approach to \( \bar{q}_l \) for large values of \( l \).

Naturally for \( s=1 \), (4.5) reduces to the special case of the so called Póllaczek–Khinchine generating formula

\[
f(z) = \frac{(1-\rho)(1-z)}{1-ze^{\rho(1-z)}}.
\]

Further it is not too difficult to show (c.f. Takács (1962)) that for this model the probabilities \( \bar{p}_r \) defined in Section 3 are identical with \( \bar{q}_r \).

The mean and variance, \( E(t) \) and \( V(t) \) of the duration of a busy period in time are

\[
E(t) = \left\{ \mu(1-\rho) \right\}^{-1}
\]

and

\[
V(t) = \rho \left\{ \mu^2(1-\rho)^3 \right\}^{-1}
\]

as can be easily verified from the numerous treatments of the system M/G/1, (c.f. Prabhu (1960)).
### III. State Dependent Queues

In this section we shall review and discuss congestion models which for convenience will be referred to as 'state dependent' queues. This terminology is used to imply that either or both the mechanisms of arrival or service depend instantaneously in some manner on the number in, or state of, the system. Such models seem to have received a rather modest, restricted and on the whole unsystematic treatment arising in the main from their genesis in the context of particular applications. Moreover it is by no means clear that existing analysis has always recognized the potentiality of such models for the regularization and reduction of chaos which should be the aim of a well ordered operational system. The pattern of our description will primarily be factual but will keep in mind this desideratum of regularity.

#### 5. The Model of Cox and Smith

The most general formulation of a model with state dependent arrival and service patterns appears to be that treated rather briefly in the book by Cox and Smith (1961). Subsequently, special cases or particular aspects of the general model were considered by Conway and Maxwell (1961), Jackson (1963), Yadin and Naor (1967), Gebhard (1967), Hadidi and Conolly (1969) and Hadidi (1969).

The general model of Cox and Smith specifies probability differentials $\lambda_n dt + O(dt^2)$, $(n \geq 0)$ and $\mu_n dt + O(dt^2)$, $(n \geq 1)$ for arrival and service completions respectively in a small time interval $(t, t + dt)$. This implies negative exponential distribution both for inter-arrival interval and service times but
with state \( n \) dependent parameters, \( \lambda_n \) and \( \mu_n \) are supposed to be time \( t \) independent. An equivalent expression of this mechanism is to state that the instantaneous arrival and service rates are \( \lambda_n \) and \( \mu_n \) respectively. Further assumptions of the model are a single server and infinite waiting room size. It is a straightforward matter to formulate the set of equations describing state probabilities at a given time \( t \). The reader is likely to have seen this formulation more than once, but it is repeated here for self-consistency and convenience.

Let \( Q(t) \) be the number of customers present in the system (including the one in service) at time \( t \). The development of the process in the interval \( (t, t+s) \) is independent of the events which take place in \( (0, t) \). Thus \( Q(t) \) is a denumerable Markov Process with states \( 0, 1, 2, \ldots \). Let its transition probabilities be denoted by

\[
P_{ij}(t) = \Pr[Q(t) = j | Q(0) = i], \quad i \geq 0, \ j \geq 0.
\]

According to the assumptions of the model we have

\[
\begin{align*}
P_{i,i+1}(dt) &= \lambda_i dt + O(dt^2), \quad (i \geq 0), \\
P_{i,i}(dt) &= 1 - (\lambda_i + \mu_i) dt + O(dt^2), \quad (i \geq 1), \\
P_{00}(dt) &= 1 - \lambda_0 dt + O(dt^2), \\
P_{i,i-1}(dt) &= \mu_i dt + O(dt^2), \quad (i \geq 1).
\end{align*}
\]

There are no other possible infinitesimal transitions. The Chapman-Kolmogorov equations of the process are typified by

\[
P_{ij}(t+dt) = \sum_{k \geq 0} P_{ik}(t) P_{kj}(dt), \quad (5.2)
\]

which give, by virtue of (5.1),
\[ P_{l,j}(t+dt) = P_{l,j-1}(t) \lambda_{j-1} dt + P_{l,j}(t)\left[1-(\lambda_j+\mu_j)dt\right] + \]
\[ +P_{l,j+1}(t)\mu_{j+1} dt + O(dt^2). \quad (5.3) \]

We shall now assume that the process starts with a single customer in the system; i.e., that \( Q(0)=1 \) with probability one. Thus with
\[ p_j(t) = \sum_{i \geq 0} \Pr[Q(t)=j \mid Q(0)=i] \times \Pr[Q(0)=i], \]
we have from (5.3)
\[ p_j(t+dt) = p_{j-1}(t) \lambda_{j-1} dt + p_j(t)[1-(\lambda_j+\mu_j)dt] + p_{j+1}(t)\mu_{j+1} dt \]
which, upon reorganizing, gives the forward Kolmogorov equations of the process
\[ \dot{p}_j(t) + (\lambda_j+\mu_j) p_j(t) = p_{j-1}(t) \lambda_{j-1} + p_{j+1}(t)\mu_{j+1}, \quad j \geq 1. \quad (5.4) \]
(The dot means differentiation with respect to time.)

(5.4) has to be complemented with an equation for \( j=0 \).

This is obviously
\[ \dot{p}_0(t) + \lambda_0 p_0(t) = \mu_1 p_1(t). \quad (5.5) \]

The 'birth and death' equations (5.4) and (5.5) completely determine the state probabilities at any given time \( t \). The question of existence of a solution of this set and its uniqueness is not by any means easy and the reader is referred to Feller (1940) for a detailed discussion. For appropriate choice of \( \lambda_n \) and \( \mu_n \) a unique solution exists which satisfies the regularity condition
\[ \sum_{n \geq 0} p_n(t) = 1. \]

To obtain this solution in general, however, for finite time has
so far proved unsuccessful and in consequence most authors have confined themselves with the provision of "steady state" result

$$\bar{p}_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \bar{p}_0 , \quad j \geq 1$$

derived from (5.4) and (5.5) by setting $\dot{p}_j(t)=0$ and supposing that $\lim_{t \to \infty} p_j(t)=\bar{p}_j$ for $j \geq 0$. This result is valid provided that the series

$$S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots$$

converges, and in this case

$$\bar{p}_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} S^{-1} . \quad (5.7)$$

Conway and Maxwell (1961) considered a special case in which $\mu_n = c \mu$ and $\lambda_n = \lambda$ for all $n$. $c$ is a constant indicating the degree to which the extent the service rate is affected by the system state. They term $1/\mu$ "normal" service rate and $c$ a "pressure coefficient". The case $c=0$ then corresponds to $M/M/1$. With these they evaluate (5.6) and then tabulate numerically the mean number of customers in the system, and idleness probabilities for different values of $\rho=\lambda/\mu$. They next mention the arrival rate state dependent case in which $\lambda_n = (n+1)^{-b} \lambda$ while $\mu_n = \mu$ for all $n$.

It is evident that (5.6) yields the same values for $\bar{p}_j$ in the following three cases:

i) $\lambda_n = \lambda$ (all $n$) ; $\mu_n = c \mu$

ii) $\lambda_n = (n+1)^{-a} \lambda$ ; $\mu_n = \mu$ (all $n$)

iii) $\lambda_n = (n+1)^{-b} \lambda$ ; $\mu_n = n^c \mu$ and $b+c=\alpha$.

Thus it is sufficient, as they remark, when studying the state • • •
probabilities to incorporate all three variations in a single theory.

It does not seem to be feasible to pursue the theory further in any direction for general $\mu_n$ and $\lambda_n$. Accordingly we now fix the arrival rate at a constant value $\lambda$ which is both time and state independent, and in the next Section investigate the consequences of models involving only variable service rates.

6. The Queue with Uniform Arrival Rate and State Dependent Service Rate

As mentioned above $\lambda_n$ is now assigned a constant time and state independent value $\lambda$. A model of this character has been considered by Gebhard (1967). He evaluates the consequences of techniques which he terms "Single Control Level" and "Bilevel Hysteretic Control". "Single Control Level" is defined by assigning $\mu_n$ a value $\mu_1$ for all $n$ less than or equal to a given $N$ and a value $k_1 \mu_1$ for all other $n$. $k_1$ is supposed to be a given constant. "Bilevel Hysteretic Control" is such that the rate $k_1 \mu_1$ is maintained so long as $n$ is greater or equal to an upper control level $N_2$ and then decreased to $\mu_1$ when the queue length drops to $N_1$. For such a case he evaluates (5.6) and the mean and variance of the number of customers in the queue in the steady state. He then judges the efficiency of the two control models on the basis of the number of changes which the server has to make from rate $\mu_1$ to $k_1 \mu_1$ and the proportion of time the server has to operate at the greater of the two rates.

Harris (1967) considers a model which is similar in nature. In his case the service rate is fixed by the number in
the system at the beginning of service of a particular customer and does not change as a consequence of arrivals during the service time of that customer. In effect he considers a type of $M/G/1$ process in which the service time parameter becomes a stochastic process indexed on the number of customers in the queue at the moment service is begun. This system partially adapts the rate at which service is provided to demand, but of course does not provide for the further adaptation which would be called for by increasing pressure of demand created by an unusually heavy stream of arrival during a particular service interval.

Few, if any, transient condition result have been obtained analytically for a general $\mu_n$. Let $\mu_n = \sigma_n - 1$ so that the effect of the state of the system is represented only by $\sigma_n$ and $\mu$ is time and state independent. Further suppose $t_n$ is the arrival epoch of the $n$th customer and $Q_n = Q(t_n - 0)$. Hadidi and Conolly (1968) give formulae for

$$ p_r^{(n)} = \Pr(Q_n = r) \quad (6.1) $$

This permits probabilistic statements to be made about the state of the system at all stages of its evolution. The analysis depends on evaluating the matrix of transition probabilities over the time interval separating the arrival epoch of the $n$th and $(n+1)$th customers and subsequently solving a difference equation of the form

$$ U_{n+1} = a_n U_n + b_n $$

with suitable initial conditions.

Let $\rho = \lambda / \mu$,

$$ \alpha_i = \frac{\rho}{\rho + \rho_i}, \quad i = 0, 1, ... $$

- 19 -
\[ \beta_i = \frac{\alpha_i}{\sigma_0} \alpha_{i-1} \alpha_i, \quad i=1,2,\ldots \]

and \( \beta_0 = \alpha_0 \), then

\[ p_r(n+1) = \left( \frac{\alpha_0}{\rho} \right)^k \alpha_0 \alpha_1 \ldots \alpha_{r-1} \sum_{i_1=0}^{r} \sum_{i_2=0}^{i_1+1} \ldots \sum_{i_k=0}^{i_{k-1}+1} \beta_{i_1} \beta_{i_2} \ldots \beta_{i_k}, \quad (6.2) \]

for \( r=1,2,\ldots,n-1 \), where \( k=n-r \);

\[ p_n(n+1) = \alpha_0 \alpha_1 \ldots \alpha_{n-1} \]

and

\[ p_0(n+1) = \frac{\alpha_0}{\rho} p_1(n+1). \]

Some remarks are also made on waiting time. However, the general nature of the model permits only a closed form result to be developed for the mean waiting time of the nth customer to join the system. This is

\[ E_n(w) = \frac{1}{\lambda} \sum_{r=0}^{n} p_r(n) \left( \frac{1}{\rho} \right)^{r+1} \sum_{i_1 \geq r} \sum_{i_2 \geq i_1-1} \ldots \sum_{i_{r+1} \geq i_r-1} A_{r,i_1} A_{i_1-1,i_2} \ldots A_{i_{r-1},i_{r+1}} x \]

where

\[ A_{ij} = \alpha_i + \alpha_{i+1} + \ldots + \alpha_j \]

and

\[ B_{ij} = \sigma_j \alpha_i \alpha_{i+1} \ldots \alpha_j. \]

The complexity of these results discourage further work on general \( \mu_n \). However, if a further and realistic particularisation of the dependence of \( \mu_n \) on \( n \) is made, considerable simplifications become feasible. This is dealt with in the next Section.

7. The Model with Linearly State Dependent Service Rate

We recall that the momentary arrival rate is still the time and state independent constant \( \lambda \). We now assume momentary
service rate to be a linear multiple of the number of customers in the system, i.e. that

\[ \mu_n = cn\mu' \]

\( \mu' \) would represent the 'normal' or 'standard' rate and \( c \) is a real number chosen in such a way that an increase of \( 1/c \) customers increases the service rate by \( \mu' \) units. It is convenient however to let \( c\mu' = \mu \), so that \( \mu_n = n\mu \).

7.1 Transient Behaviour of Queue Length

As the momentary service rate is now \( n\mu \) it is evident that equations describing the state of the system for this process are formally identical with those describing the state of the system for the infinite service facility \( M/M/\infty \), whose solution is known (c.f. Khintchine (1960)), and which are thus available for our purposes. In point of fact the set (5.4) reduces to

\[ \dot{p}_j(t) + (\lambda + j\mu)p_j(t) = p_{j-1}(t)\lambda + p_{j+1}(t)(j+1)\mu, \quad j \geq 1, \]

and

\[ \dot{p}_0(t) + \lambda p_0(t) = \mu p_1(t), \]

from which we have

\[ p_j(t) = e^{-\rho \Lambda(t)} \frac{[\rho \Lambda(t)]^j}{(j-1)!} \left( e^{-\mu t} + \frac{\rho \Lambda(t)}{j} \right), \quad (7.1) \]

for \( j \geq 1 \), and

\[ p_0(t) = \Lambda(t) e^{-\rho \Lambda(t)}, \quad (7.2) \]

where \( \rho \) has the same definition as in Section 6 \( (\lambda/\mu) \), and

\[ \Lambda(t) = 1 - e^{-\mu t}, \quad (7.3) \]

bearing in mind that \( Q(0) = 1 \).

The steady state result is also known to be

\[ \bar{p}_j = \frac{e^{-\rho \Lambda}}{j!}, \quad j \geq 0, \quad (7.4) \]
and was obtained by Erlang. It is also interesting to note that the probabilities \( p_j^{(n)} \) introduced in Section 6 reduce to a form in this case which conveniently produce

\[
\lim_{n \to \infty} p_j^{(n)} = e^{-\rho_j p_j}. \tag{7.5}
\]

That is in this model as well we have

\[
\lim_{t \to \infty} p_j(t) = \lim_{n \to \infty} p_j^{(n)},
\]

a relation which holds for \( M/M/1 \) and \( M/G/1 \), but not for \( G I / M / 1 \). Further comments on this model's analogy with \( M / M / \infty \) is made in Hadidi and Conolly (1968).

It is worth recalling that this model corresponds to a special case of that of Conway and Maxwell in which the "pressure coefficient" is equal to unity. As mentioned earlier, however, these authors merely concern themselves with the evaluation of \( p_j \).

7.2 The Effective Service Time

At this point we propose to interpolate some remarks on what we shall call the effective service time of this process. As has been seen, the length of a service interval depends on how many customers are present at the beginning of that interval as well as on how many arrivals takes place during the interval. The question thus arises as to how to make a probabilistic description of a "typical" service time. It is shown in Hadidi (1969) that if we consider the duration of service of a customer in the steady state condition, then its p.d.f. \( b(t) \) is given by

\[
b(t) = e^{-t} + \left( \mu e^{-\mu t} - (1 - e^{-\mu t})^2 \right) \exp(-t - \rho e^{-\mu t}), \tag{7.6}
\]

where without loss of generality \( \lambda \) is taken to be unity and accordingly \( \rho = 1/\mu \). The first two moments of the effective service time are given by
\[ E(t) = 1 - e^{-\rho} \]  
\[ E(t^2) = 2[1 - \rho^2 - \gamma(\rho, \rho)] \]  
where \( \gamma(\cdot, \cdot) \) denotes the incomplete Gamma function 
\[ \gamma(a, x) = \int_0^x e^{-t} t^{a-1} dt. \]

We shall in a later Section stress the significance of (7.7) in connection with comparison with other models.

7.3 The Waiting Time Process

The waiting time process, at least in some cases and from the point of view of customers, constitutes a fair measure of effectiveness of any queueing system. As such it ought to receive prime attention in systems designed to achieve regularization.

For the model under discussion the waiting time could be analysed as follows:

Let \( \gamma_{kr}(w) \) be the p.d.f. of exactly \( r \) services having combined length \( w (>0) \) given that the first service interval starts with \( k \) customers in the system. \( \gamma_{kr}(w) \) then satisfies the integral difference equation

\[ \gamma_{kr}(w) = \sum_{j \geq 0} \int_0^w \gamma_{k1}^{(j)}(s) \gamma_{k+j-1, r-1}(w-s) ds, \]  

where \( \gamma_{k1}^{(j)}(s) \) is associated with a service interval which starts with \( k \) customers in the system and during which \( j \) arrivals take place. In this way \( f(w) \), the p.d.f. of waiting time including service in the steady state is given by

\[ f(w) = \sum_{r \geq 0} P_r \gamma_{r+1, r+1}(w), \]  

where \( P_r \) is given by (7.5). \( \gamma_{k1}^{(j)}(s) \) is easily obtainable and thus we merely have to solve (7.9). It is shown in Hadidi (1969).
that
\[ f(w) = \mu \text{Exp}(\rho - 2\rho e^{-\lambda w} (1 + \mu) w) \left( \{1 + 2\rho \Lambda(w)\} I_0(2\sqrt{\phi}) + \{\sqrt{\phi} + \rho \Lambda(w)\}^2 \frac{1}{\sqrt{\phi}} I_1(2\sqrt{\phi}) \right), \] (7.11)
where \( \Lambda(w) \) is given by (7.3),
\[ \phi(w) = \rho e^{-\lambda w} (w - \rho \Lambda(w)), \]
and \( I_v(.) \) is the modified Bessel function of order \( v \). The expected result
\[ E(w) = \rho \] (7.12)
can be obtained from (7.11). A less compact expression can also be obtained for the variance. Numerical values will be given in a later Section.

7.4 The Busy Period and Output Processes

Let \( E \) be the event "a transition from 0-1 occurs". In this model \( E \) is recurrent. Thus we immediately have from a theorem of Feller (1961) that, in words, the limiting probability of occurrence of \( E \) is equivalent to the ratio of probability of its ever occurring to the mean recurrence time. When conditions are met for \( E \) to be persistent
\[ \frac{1}{\lambda P_0} = \frac{1}{E(BP) + E(I)} \] (7.13)
where \( E(BP) \) and \( E(I) \) stand for expected values of busy period and idle period and \( P_0 \) is as introduced in Section 7. As inter-arrival intervals are negative exponentially distributed with unit mean, we have \( E(I) = 1 \) and hence
\[ E(BP) = e^\rho - 1 \] (7.14)
Now let \( u(t)dt \) and \( f(t)dt \) be respectively the probabilities of occurrence of \( E \) and occurrence of \( E \) for the first time in the interval \((t, t+dt)\). From
\[ u(t) = f(t) + \int_0^t f(s) u(t-s)ds \] (7.15)
we have
\[ f^u(z) = \frac{u^u(z)}{1 + u^u(z)}, \quad (7.15) \]

where \( * \) notation refers to Laplace transformation with respect to time \( t \). On the other hand

\[ u^u(z) = \lambda p^u_0(z) \]

and
\[ t'(t) = \int_0^t b(s) \lambda e^{-\lambda(t-s)} \, ds, \]

where \( b(t) \) is the p.d.f. of duration in time of the busy period and \( p^u_0(t) \) has the obvious meaning. A formal expession for \( b^u(z) \) is thus given by
\[ b^u(z) = \frac{(\lambda + z)p^u_0(z)}{1 + \lambda p^u_0(z)}, \quad (7.16) \]

where
\[ p^u_0(z) = \mu e^{-\rho} \sum_{r \geq 0} \frac{\rho ^r}{r!} \frac{1}{(z + \rho r + \mu)(z + \mu r)}. \]

The second moment of the duration of a busy period is obtained from (7.16) to be
\[ E(t^2) = \frac{2}{\lambda \mu} e^\rho \sum_{r \geq 1} \frac{\rho ^r}{r! r}, \quad (7.17) \]

Finally a word about the output process. This model shares the well known steady state property of \( M/M/1 \) that the duration of time between successive service completions (output) enjoys the inter-arrival distribution with the same parameter. The argument is based on \( q_r \), the limiting distribution that the \( n \)th customer leaves behind on his service completion, and the fact that \( T(u) \), the steady state output p.d.f. is given by
\[ T(u) = q_0 \int_0^u \lambda e^{-\lambda x} \gamma_1(u-x) \, dx + \sum_{r \geq 1} q_r \gamma_r(u), \quad (7.18) \]

where \( \gamma_r(u) \) is defined as in Section 7.3.

The steady state joint p-d-f of two successive output intervals which merely turn out to be the product of two single ones (as in \( M/M/1 \)) can also be found in a similar manner. For details see Hadidi (1969 a).
8. Systems with Controlled Number of Channels

The device of controlling the number of channels according to demand has also been suggested as a means of keeping the queueing situation under control.

Romani (1957) considered a potentially infinite server system in which the queue was never allowed to grow beyond a certain limit. When the queue reached a given size and a further arrival occurred automatically a new server was called on duty. Romani found the stationary state of the system probabilities for such a model.

Moder and Phillips (1962) and Yadin and Naor (1967) have extended Romani's works. The former authors consider queueing systems in which both inter-arrival internal and service times are negative exponentially distributed and an increase in the number of servers occurs from a fixed minimum when the queue reaches a specific length \( N \). A maximum of \( S \) servers is supposed to be available and when all are occupied no further increase is possible. By arguing that the rate of transition into a subset of possible states is equal to the rate of transition out of that same subset in the steady state they find the joint probability that the system is in state \( n \) and \( s \) servers are busy. They then consider some 'measures of effectiveness' such as the total time the server is idle and the mean number of customers in the system.

It is evident that as far as the state of the system is concerned this model is identical with the special case of that of Cox and Smith discussed in detail in Section 5 with \( \lambda_n = \lambda \) for all \( n \) and \( \mu_n = c\mu \) where \( c \) is the number of servers occupied. Yadin and Naor (1967) are concerned with a vector
\( \mu \) of possible service capacities (i.e., rates) and two other vectors \( R \) and \( S \) whose purpose respectively is to determine the size of the queue which change the capacities from \( \mu_k \) to \( \mu_{k+1} \) and from \( \mu_j \) to \( \mu_{j+1} \). When the system is operating at the rate \( \mu_k \) they say that the system is in phase \( k \) and they find the steady state joint probability that the system is in phase \( k \) and the queue length is \( j \); a result which is formally identical with that of Moder and Phillips.

In a further work (1963) the same authors consider what they call a "(0, R) doctrine" which amounts to closing of the station when the server becomes idle and subsequently reopening when the system contains \( R \) customers. Their main result here is the mean queueing (waiting) time and the mean number of customers in the system in the steady state. They also analyse a cost structure and outline some optimization procedures.

9. Jackson's "Jobshop-Like" Queueing System

The model Jackson (1963) considers is in essence identical with the ones discussed in previous Sections, except that he is concerned with several stations. At each station the mechanism of arrival and service is that explained in connection with the model of Cox and Smith in Section 5. Each customer might require service at one, some or all the stations. The main result is the steady state joint probability of queue size at the several stations. He considers the model Conway and Maxwell have studied as an example. A further generalization he makes in the same paper allows for the automatic injection of a customer in the system whenever the total number of customers falls below a specific limit.
Several special cases are also treated as examples.

It should be emphasized that all the results mentioned in Sections 8 and 9 could be obtained within the framework of the theory of Section 5. The terminology of individual authors was used in connection with their results partly to illustrate this very fact.

IV: Further Models

The material in this Section is condensed from Conolly (1968) and Conolly and Hadali (1969).

10. Service–Inter–arrival Dependence

In a sense all the models we review in this paper embody the idea of some sort of correlation between service and arrival patterns, but the expression "correlated queue" is used here, for convenience, in reference to a very particular single server system. The only independent stream of events is that of arrivals and is supposed Poisson with mean interval $\lambda^{-1}$. It is also supposed that the initial epoch $t = 0$ is an arrival epoch and that this initial arrival $A_1$ immediately begins to receive service of duration $\rho T_1$, where $T_1$ is a random variable with p.d.f. $\exp(-\lambda T)/\lambda$ and $\rho$ is a time independent constant selected in advance by the server. The second arrival $A_2$ may be supposed to arrive after a time $T_2$, again determined by the p.d.f. $\exp(-\lambda T)/\lambda$, and his service is of duration $\rho T_2$, with the same choice of $\rho$. The process evolves in a like manner; the service time allocated to arrival $A_n$ being the same multiple $\rho$ of the inter–arrival interval $T_n$ between $A_n$ and $A_{n-1}$. The traffic intensity thus has the numerical value $\rho$, and is in effect determined by the server, who, it is important to notice from the practical point of view is able to observe
in total the length of the inter-arrival interval which
determines the duration of a particular customers service
before that service commences. If, as would be generally
impractical, \( A_n \) received service in proportion to \( T_{n+1} \), at
least the waiting time part of the analysis could have been
effected by traditional methods. The genesis of this model was
in a non-queueing context (Conolly (1967)) and its interpret-
ation as a queueing model is due to the ingenuity of R.Cruon.
The practical implications in terms of reduction in waiting
time and server idle time were not appreciated until later.
However it is now obvious that the nature of the model is such
that an unusually closely space cluster of arrivals will
receive more rapid service than a more widely spaced cluster,
though traffic intensity remains the same. Since the server's
idle time is distributed like an inter-arrival interval it is
perhaps more correct to say that in this system he tends to be
idle less often.

We do not, in this paper, elaborate on the applicability
of such a model and how and where it could actually be imple-
mented. Our main purpose is its analysis as well as its prac-
tical consequences. Numerical results will be given in a
later Section. We now describe briefly the theory.

10.1 The Waiting Time Process

Waiting time includes service and a first come first
discipline servedVis assumed. Let the waiting time of the nth arrival \( A_n \)
be denoted by \( w_n \), with p.d.f. \( f_n(y) \). Then it is shown in
Conolly (1968) that

\[
\frac{\rho}{\lambda} f_{n+1}(y) = e^{-\lambda y/\rho} \left[ \int_0^{y/\rho} f_n(x) \, dx + \int_0^y e^{\lambda x/\rho} f_n(x + \frac{y-x}{\rho}) \, dx \right],
\]

(10.1)
with solution
\[ f_n(y) = \frac{\lambda}{p} \left( \sum_{r=0}^{n-1} g_{nr} \exp\left\{-y(s_r/p)\right\} \right), \quad (10.2) \]
where
\[ g_{nr} = \frac{(-1)^{r+1}(n-r/2-1)}{s_0s_1 \cdots s_{n-1}}, \]
and the functions \(s_n\) which occur repeatedly in the analysis are defined by
\[ s_n = \sum_{i=0}^{n} r^{-i}. \quad (10.3) \]

If \(p<1\) then a steady state value can be found, viz.
\[ f(y) = \lim_{n \to \infty} f_n(y) = \frac{\lambda}{p} \left( \sum_{r \geq 0} g_r \exp\left\{-y(s_r/p)\right\} \right), \quad (10.4) \]
with
\[ g_r = \frac{(-1)^{r+1}r^2}{(1-p^2)(1-p^3) \cdots (1-p^r)}. \quad (10.5) \]
The distribution function \(F_n(y)\) of \(w_n\) is related to the p.d.f. \(f_n(y)\) by
\[ F_n(y) = \frac{\rho e^{\lambda y}}{\lambda \rho s_n} f_{n+1}(\rho y). \quad (10.6) \]

The derivation of these results was carried out by a Laplace transform technique and it is of interest to observe that the Laplace transform \(f_n^*(z)\) has the form
\[ f_n^*(z) = \frac{s_0s_1 \cdots s_{n-1}}{(s_0+\rho z/\lambda)(s_1+\rho z/\lambda) \cdots (s_{n-1}+\rho z/\lambda)}. \]

10.2 State Probabilities

Let again \(Q(t)\) be the number of customers in the system (including service) at time \(t\) and if \(t_n\) is the arrival epoch of \(A_n\) let \(Q_n = Q(t_n-0).\) No analysis exists for the quantities \(p_r(t) = \Pr(Q(t)=r),\)
but if
\[ p_r(n) = \Pr(Q_n=r), \quad (0 \leq r \leq n-1) \]
the following results have been found:

\[ p_0^{(n)} = (\rho^{n-1} s_{n-1})^{-1}; \quad (10.7) \]

\[ p_r^{(n)} = \frac{s_r}{s_{n-r-1}} - \frac{s_{r-1}}{s_{n-r}}, \quad (10.8) \]

for \( 1 \leq r \leq n-2 \); and

\[ p_{n-1}^{(n)} = 1 - s_{n-2}/s_1; \quad (10.9) \]

where \( s_n \) is defined by (10.3),

\[ S_n = \sum_{i=1}^{n} s_i^{-1}, \quad (10.10) \]

and \( S_n^{(m)} \) is the homogeneous form of degree \( m \) of the sequence \( \{s_i^{-1}\}, (0 \leq i \leq n) \). Thus for example,

\[ S_1 = 1/s_0 + 1/s_1, \]
\[ S_2 = 1/s_0^2 + 1/s_0 s_1 + 1/s_1, \]
\[ S_3 = 1/s_0^3 + 1/s_0^2 s_1 + 1/s_0 s_1^2 + 1/s_1^3, \]

and so on.

If \( \rho < 1 \), then

\[ \lim_{n \to \infty} p_r^{(n)} = \bar{p}_r = (1-\rho)(S^{-1})^{(r)}, \quad r \geq 1; \quad (10.11) \]

and

\[ \bar{p}_0 = 1-\rho. \]

Here

\[ S = \lim_{n \to \infty} S_n. \]

The method of proof consists essentially of finding

\[ p_k^{(n)} = \operatorname{Pr}(Q_n \geq k) \]

which can be obtained by purely probabilistic arguments in terms of the waiting time distribution and then

\[ p_k^{(n)} = p_k^{(n)} - p_k^{(n+1)} . \]

10.3 The Output Process

The waiting time result can also be used to obtain the
p.d.f. \( K_n(T) \) of the interval of time \( T \) separating the departure epochs of \( A_n \) and \( A_{n+1} \), and the joint p.d.f. \( h_n(T, U) \) of the intervals \( T \) and \( U \) separating the departure epochs of \( A_n, A_{n+1} \) and \( A_{n+2} \). The argument is probabilistic. We obtain

\[
\rho/\lambda K_n(T) = \sum_{r=0}^{n-1} \{ g_{nr} s_{r+1} \exp(-\lambda T s_{r+1}/\rho) \} / s_r(s_r s_{r+1}) + \\
+ \exp(-\lambda T/\rho s_1) \sum_{r=0}^{n-1} g_{nr}(s_1 s_{r+1}). \tag{10.12}
\]

for \( h_n(T, U) \)

The corresponding expression is lengthy and to save space the reader is referred to Conolly and Hadidi (1969).

10.4 The Initial Busy Period

By definition the 'correlated' queueing process starts with a busy period. This has been analysed and the main result is given below. No similar analysis exists for subsequent busy periods which are complicated by the fact that the interarrival interval determining the initial service belongs to a truncated negative exponential distribution whose characteristics are in turn determined by the preceding idle and busy periods. Some unpublished numerical experiments however suggest that as time proceeds busy periods tend to have the same statistical characteristics both for a single process and taken over a number of processes (i.e. they are stationary and ergodic) with mean \( \rho/\lambda (1-\lambda) \).

Let \( b_n(t) \) be the joint probability and p.d.f. that the initial busy period has length \( t \) and consists of \( n \) services, and let \( b_n^\circ(z) \) be its Laplace transform. We then have

\[
\rho^n b_n^\circ(z) = (s_0 s_1 \ldots s_{n-2}) / [(s_1 s_0 z/\lambda)(s_2 z s_1 z/\lambda) \ldots (s_n z s_{n-1} z/\lambda)]
\]

with 1 in the numerator when \( n=1 \), and

- 32 -
\[ b_n(t) = \lambda \sum_{j=0}^{n-1} (-1)^j \rho^{n-3}/2 -j(j-1)/2 \ s_{n-j-1} s_{n-j} s_{n-j} \ldots \]
\[ \ldots s_{n-2} \exp\{\lambda s_{n-j} t(s_{n-j-1})^{-1}\}/\{s_0 s_1 \ldots s_{j-1}\} , \]

with \( \lambda \) in the denominator when \( j=0 \). We observe that the Laplace transform is more elegant and its form is quite easy to conjecture from first principles but so far a simple inductive proof, which might have been expected to exist, has not been found. The direct proof is carried out in the time domain using the extension of a technique of Prabhu (1960).

11. A "Dual" Model

The "correlated" queue model envisages a single independent stream of arrivals with service deterministically dependent on inter-arrival intervals. Models on the other hand can be conceived which consist of an independent stream of epochs which it is convenient to describe as "service opportunities". The notion of service opportunities is natural in such physical situations as obtain at bus stops, railway stations, elevators, where the service is in a sense instantaneous. Moreover it could happen that no service takes place at all if there is no customer present. Such a situation may be described as a lost service opportunity. Management wants to avoid lost service opportunities (empty buses) and customers to avoid waiting. Actual situations are thus under discussion and the dual correlated model assumes an aspect of realism.

A further example could be a production line of units which have to be serviced by a machine where it is costly for the machine to be idle.

To implement the above discussion we suppose that \( \sigma_n \)
(n ≥ 0) to be a sequence of epochs each of which provide a service opportunity with the convention that σ₀ corresponds to time t = 0 and is a fictitious service opportunity which together with σ₁ provides the initial interval u₁ on which to base the first arrival interval. This first interval will in the model be \( u₁/ρ \) where ρ is a time independent constant and can easily be seen to be a measure of traffic intensity. ρ is supposed to be under the control of the arrivals. For a stable system the requirement is as usual that ρ<1, and will be supposed in what follows.

Consider the situation at the second service opportunity σ₂. It will be agreed that the first arrival interval will be measured from σ₁ and hence the first customer arrives at \( t₁ = u₁s₁ \) where s₁ is as given by (10.10).

\[
\begin{array}{cccccc}
\sigma₀ & u₁ & \sigma₁ & u₂ & \sigma₂ \\
\hline
 & \downarrow & & \downarrow & \downarrow \\
\sigma₀ & u₁ & \sigma₁ & u₂ & \sigma₂ \\
& \downarrow & & \downarrow & \downarrow \\
& u₁/ρ & \rightarrow & t₁ & \rightarrow \\
& u₁/ρ & \rightarrow & t₁ & \rightarrow \\
\end{array}
\]

Either of the situations depicted above are possible. In the first there is one customer waiting at σ₂; in the second no one. In the first it is clear that when σ₂ occurs the arrival organizer knows \( u₂ = σ₂ - σ₁ \) and hence the next inter-arrival interval \( u₂/ρ \). This he proceeds to measure from \( t₁ \) to obtain \( t₂ \) and since \( ρ<1 \) it is not possible for the position to arise that \( t₂ < σ₂ \), which would be absurd.

Continuing in this way we see that at σ₂ the system may contain 0 or 1 customers; at σ₃, 0, 1 or 2; and in general at σₙ, 0, 1, ..., n-1, since it is supposed that only one customer at
a time is disposed of at a service opportunity. If \( u \) has a negative exponential distribution with mean \( \rho / \lambda \), little calculation shows that, for example

\[
p_{n-1}^{(n)} = \left[ s_1 (\rho s_2)^{n-2} \right]^{-1},
\]

where \( p_r^{(n)} \) has the same meaning as in Section 10.2. However, aside from events of special interest, one of which is given below, it does not seem to be feasible to conjecture and prove other general formulae.

The probability \( q_n \) of the important practical event that immediately before the 2nd, 3rd, ..., nth service opportunity there is always a single customer waiting, an event which ensures no lost opportunities and minimal waiting time for customers is given by

\[
q_n = \rho^{n(n-1)/2} / s_n
\]

This model seems worth investigating.
V. Numerical comparisons

We are now in a position to compile a few tables illustrating numerically the operational behavior of various models. It is in general difficult to pick a 'measure of effectiveness' which would be an indication of gains in all possible situations. To have a clear indication of potential gains all the characteristics of the particular system in mind are to be taken into account.

In the following we concern ourselves with single server systems only and set the mean inter-arrival interval to unity. We further assume that two models are truely comparable if they are taken at the same traffic intensity. The evaluation of traffic intensity does not create any problems in the conventional models D/M/1, M/D/1 and what was termed 'correlated' queue in Section 10 for by definition it would be the ratio of mean service interval to mean inter-arrival interval. However for the state dependent queue with constant arrival rate as defined and analysed in Section 7 the instantaneous service rate in \((\mu)^{-1}\) which is naturally subject to change at all the epochs that a change occurs in the state of the system. What we propose to do, however, is to substitute for the mean service time the mean service interval as obtained from the 'equivalent' service time distribution (7.6), i.e. (7.7). Thus for the state dependent system (abbreviated as SD, here) by which we shall mean, through out this Section, the model discussed in Section 7, traffic intensity \(\rho^*\) is given by

\[
\rho^* = 1 - e^{-\rho}
\]

For all the other systems \(\rho^* = \rho\). This fact should be taken into account and the conclusions interpreted accordingly.

A thorough numerical comparison between various models both in transient and steady state conditions is given in Conolly and Hadidi (1969) and Hadidi and Conolly (1969). The following is mostly taken out of these sources and the values are given only at two values of traffic intensity, \(\rho^*\), namely 0.5 and 0.95. We, here, confine ourselves only with steady state results.

Table 1 gives the state probabilities \(q_r\) which we
recall were \( \lim_{n \to \infty} p_r^{(n)} \) and \( p_r^{(n)} \) was the probability that the nth customer finds \( r \) in the system upon his arrival.

**Table 1**

<table>
<thead>
<tr>
<th>State of the system</th>
<th>Probabilities (Steady state)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho^* = 0.5 )</td>
</tr>
<tr>
<td></td>
<td>D/M/1</td>
</tr>
<tr>
<td>( \bar{p}_0 )</td>
<td>0.797</td>
</tr>
<tr>
<td>( \bar{p}_1 )</td>
<td>0.162</td>
</tr>
<tr>
<td>( \bar{p}_2 )</td>
<td>0.033</td>
</tr>
<tr>
<td>( \bar{p}_3 )</td>
<td>0.007</td>
</tr>
<tr>
<td>( \bar{p}_4 )</td>
<td>0.001</td>
</tr>
<tr>
<td>( \sum_{i=0}^{q} \bar{p}_i )</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We observe that the idleness probabilities are all the same except in D/M/1 where it is considerably higher. With this is naturally coupled the fact that D/M/1 is less likely to contain fewer customers than the other system.

Table 2 gives the mean and variance of waiting time including service. The SD values are considerably lower than D/M/1, M/D/1 or correlated systems. Keeping in mind that the idleness probability is not increased for SD and in fact is much higher for D/M/1 clearly indicates the potential gains one might expect to achieve by arranging for the service patterns to be aligned with demand.

**Table 2**

<table>
<thead>
<tr>
<th>Mean and variance of steady state waiting time (including service)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho^* )</td>
</tr>
<tr>
<td>D/M/1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.95</td>
</tr>
</tbody>
</table>
We shall finally compile Table 3, the mean and variance of busy period for various systems.

Table 3

<table>
<thead>
<tr>
<th>$p^*$</th>
<th>$E(t^1)$</th>
<th>$Var(t^1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D/M/1</td>
<td>M/D/1</td>
<td>D/M/1</td>
</tr>
<tr>
<td>0.5</td>
<td>0.628</td>
<td>1.000</td>
</tr>
<tr>
<td>0.95</td>
<td>9.510</td>
<td>19.000</td>
</tr>
</tbody>
</table>

VI. Conclusions

It has been the subject matter of this paper to summarize and briefly analyze queueing models which in some sense embody the spirit of matching service rate to demand. The potential gains that could be achieved by such mechanisms are evident from numerical values of the previous section.

These could be measured in terms of substantial reduction in mean and variance of waiting time without actually increasing the probability of idleness. This dual benefit which improves the conditions for the lot of customers without extra cost to the management, in terms of higher probability of idleness, is the main feature of dropping the assumption of independence from arrival and service patterns. Coupled with this is the fact that the two models with such mechanisms have a much higher probability of containing fewer customers and hence a smaller probability of containing more. Thus for example, the model abbreviated with 'SD' and 'corr.' have probability 0.184 and 0.267 respectively of containing more than four customers when traffic intensity is 0.95, while the corresponding probabilities for D/M/1 and M/D/1 are respectively 0.590 and 0.644. This is a remarkable improvement in terms of the size of the waiting room and other operational features.

No attempt has been made here to explore ways of implementing such schemes in actual queueing situations and naturally further work remains to be done in studying cost-effectiveness which ought to be done for particular systems in
mind. A limited amount of such studies is reported in Gebhard (1967).

The material reported in this paper ought to be taken as indicative of the type of improvements on conventional systems which could be obtained. Much remains to be done in connection with more realistic specific models which take into account the physical limitations of the systems involved. It seems clear, nevertheless, that the results and findings which have been achieved and briefed here form a solid basis for such further investigations.
References

1. Bailey, N.T.J. (1952) "A Study of Queues and Appointment Systems in Hospital Out patient Depart-


<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Crommelin, C.D. (1927)</td>
<td>&quot;Delay Probability Formulae when the Holding Times are Constant&quot;</td>
</tr>
<tr>
<td>13</td>
<td>Everett, J.L. (1953)</td>
<td>&quot;State Probabilities in Congestion Problems Characterized by Constant Holding Times&quot;</td>
</tr>
<tr>
<td>16</td>
<td>Fry, T.C. (1928)</td>
<td>&quot;Probability and its Engineering Uses&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. van Nostrand, New York.</td>
</tr>
<tr>
<td>17</td>
<td>Gebhard, R.F. (1967)</td>
<td>&quot;A Queuing Process with Bilevel Hysteretic Service Rate Control&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Appl. Statist. 18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>J. Appl. Prob. 6</td>
</tr>
<tr>
<td>20</td>
<td>Hadidi, N. (1969 a)</td>
<td>&quot;The Output Process of a Certain Queueing Model&quot; (Submitted for Publication)</td>
</tr>
<tr>
<td>21</td>
<td>Harris, C.M. (1967)</td>
<td>&quot;Queues with State Dependent Stochastic Service Rates&quot;</td>
</tr>
</tbody>
</table>


