ON NETWORKS OF QUEUES

by

Frank Fish

August 1962

This research was supported by the Office of Naval Research under contract No. Nonr-855(09) for research in probability and statistics at the University of North Carolina, Chapel Hill, N. C. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Statistics
Mimeo Series No. 336
ACKNOWLEDGEMENTS

I wish to express my heartfelt thanks to Professor W. L. Smith, who suggested this topic, encouraged me throughout, meticulously read the manuscript and suggested improvements. His aid was invaluable.

I must also record my thanks to the Office of Naval Research for financial assistance and to the "Fulbright" authorities who made this year in the United States possible.

Finally, I am most grateful to Mrs. Doris Gardner for her careful typing of the manuscript and to Miss Martha Jordan for her aid in all things regulatory.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>11</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>5</td>
</tr>
<tr>
<td>I GENERAL QUEUEING THEORY</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 History</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Specification of a queue</td>
<td>2</td>
</tr>
<tr>
<td>1.4 A single server queue</td>
<td>3</td>
</tr>
<tr>
<td>1.5 Proof of the existence of a limiting</td>
<td>4</td>
</tr>
<tr>
<td>distribution of waiting-times</td>
<td></td>
</tr>
<tr>
<td>1.6 Uniqueness of the limiting distribution</td>
<td>9</td>
</tr>
<tr>
<td>function of the waiting times</td>
<td></td>
</tr>
<tr>
<td>1.7 The waiting time</td>
<td>11</td>
</tr>
<tr>
<td>II GENERAL QUEUEING SYSTEMS</td>
<td>16</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>16</td>
</tr>
<tr>
<td>2.2 The output of a single multi-server queue</td>
<td>16</td>
</tr>
<tr>
<td>2.3 The independence of the interdeparture</td>
<td></td>
</tr>
<tr>
<td>interval and the state of the system at</td>
<td></td>
</tr>
<tr>
<td>the departure</td>
<td>20</td>
</tr>
<tr>
<td>2.4 The independence of inter-departure</td>
<td></td>
</tr>
<tr>
<td>intervals</td>
<td>21</td>
</tr>
</tbody>
</table>
CHAPTER

2.5 A single server queue with random arrivals and
general service time------------------------- 22

2.6 The waiting time at second multi-server queue - 23

III MULTI-STAGE QUEUEING SYSTEMS -------------- 32

3.1 Introduction ----------------------------- 32

3.2 A series of queues ------------------------ 32

3.3 The ergodicity of queues in series ------- 32

3.4 The steady state solution ---------------- 42

3.5 A network of queues ---------------------- 50

IV THE EFFECT OF LIMITING THE SIZE OF THE WAITING

ROOM ---------------------------------------- 54

4.1 Introduction ----------------------------- 54

4.2 A waiting room of size N ----------------- 54

4.3 Several special cases --------------------- 61

V FEEDBACK------- -------------------------------- 67

5.1 Introduction ----------------------------- 67

5.2 Terminal feedback ------------------------ 68

5.3 Single stage feedback -------------------- 69

VI APPLICATIONS------- -------------------------------- 71

6.1 Introduction ----------------------------- 71

6.2 The spare and standby machine problem ----- 71

6.3 An inventory problem -------------------- 73

6.4 A generalisation of the machine interference
    problem ------------------------------------- 76

6.5 A consideration of some practical problems -- 83

BIBLIOGRAPHY ------------------------------- 94
INTRODUCTION

This M. S. thesis is devoted to an exposition of some of the principle results in queueing theory which serve to throw light on the underlying theme of a discussion of a network of queues. Although a lot of work has been done on this subject, it is felt that many authors seem unaware of other contributions to this field, owing, no doubt, to the somewhat scattered nature of the papers. Thus, it appears convenient, at this time, to collect some of this work together.

The important concept of a steady state distribution of customers and waiting times is introduced, the conditions under which it exists are found, for a variety of systems, and the solution of the steady state equations is displayed in each case. That these solutions are unique is shown. The account goes on to discuss series of queues in the same way and leads naturally to our main topic of a network of queues. We discuss this, give an account of some of the applications of the theory and close the thesis with a deduction of the solutions to two common practical problems. These problems are

(i) to find an optimum arrangement for a fixed amount of service power, and

(ii) to find, for a fixed input rate, the best arrangement of service rate and size of waiting room in the case of two machines in series, separated by a waiting room, where the first machine is blocked if the waiting room overflows.
CHAPTER ONE

GENERAL QUEUEING THEORY

1.1. Introduction.

The theory of queues has long interested mathematicians and statisticians alike and, consequently, there exists a large literature on the subject. Because of the astonishing range of applications this is dissipated through many journals on a wide variety of topics. In this chapter we will display some of the basic properties insofar as they affect the topic under consideration.

1.2. History.

The earliest work of importance was that of A. K. Erlang in 1903 under the auspices of the Copenhagen Telephone Company and his results in Telephone Engineering include many of our basic formulae. Great strides were then made by Pollaczek and Khintchine, working in the early 1930's. Much of this early work was the outcome of studies of Telephone and Telegraphic problems, and an excellent survey of some of the early congestion problems is to be found in many engineering textbooks. Later workers, whose work we mainly reproduce here, include both D. G. Kendall \cite{137}, and D. V. Lindley \cite{167}, both writing in 1951.

\footnote{This research was supported by the Office of Naval Research under contract No. N000-855(09) for research in probability and statistics at the University of North Carolina, Chapel Hill, N. C. Reproduction in whole or in part is permitted for any purpose of the United States Government.}
1.3. **Specification of a queue.**

Following Kendall [13], we will specify a queue by its three basic properties.

(a) The input process.

(b) The queue-discipline.

(c) The service mechanism.

**The input process.**

Throughout this account we will, in the main, restrict ourselves to a consideration of a random (Poisson) input to the queues. The advantages of the Poisson input lie in the facts that it is the most common, and theoretically the most simple type of input one can consider.

**The queue-discipline.**

With one or two exceptions we shall always assume the queue-discipline to be the usual first come-first served type.

**The service mechanism.**

There are two special cases of importance.

(1) Negative exponential service times.

(2) Constant service times.

Most of our interest will be centered on the former case, but the latter will also receive some attention. Both (1) and (2) were considered by Erlang, who also employed an intermediate hypothesis in which the service time was of the $\chi^2$ form:

$$
\frac{k^k}{k!} \cdot \frac{1}{\Gamma(k)} \cdot e^{-uk/b} \cdot u^{k-1} du,
$$

$$(0 < u < \infty)$$
(which has an interesting interpretation in terms of \( k \) successive 'stages' of service). It will be noted that (1) and (2) are the limiting forms of this distribution when \( k = 1 \) and \( k = \infty \) respectively.

1.4. A single server queue.

Let us consider a queue which has a single server. We make two assumptions.

(1) The time intervals between the arrivals of successive customers are independent random variables with identical probability distributions, and finite means. We denote by \( t_r \) the interval between customers numbered \( r \) and \( r+1 \). This specifies the input.

If a customer arrives whilst the server is occupied, he takes his place in the queue immediately behind anyone else who may be waiting, waits until they have been served, and then immediately begins his own service time. Let \( s_r \) be the service time of the \( r \)-th customer.

(2) The \( \{ s_r \} \) are identically and independently distributed with finite expectation, and the two sets of random variables \( \{ s_r \} \), \( \{ t_r \} \) (\( r = 1, 2, \ldots \)) are independent. This completely specifies the system.

We wish to investigate the waiting time of the customers, i.e. the time that the customer waits in line before he starts his service. In particular, we are interested in whether this distribution tends to a limiting form as the number of customers increases. There are several proofs in the literature that, under certain conditions this is the case. Kendall \( \text{[13]} \), utilizes the work of Feller \( \text{[5]} \), and
his general theory of recurrent events, to investigate the ergodic properties of the queue after reducing the problem to one of a Markov chain in 'discrete time'. He notes that the epochs of departure are points of regeneration, (i.e. what Feller \cite{5} calls 'recurrent events'). Yosida and Kakutani \cite{20} give a very elegant discussion of the limiting behaviour in the case of a Markov Process with an enumerable, infinite number of possible states (which we have here if we allow an infinite queue). However, they remark that their method is, in general, inadequate if the number of possible states is finite.

The proof we will follow is that of D. V. Lindley \cite{16}.

1.5. Proof of the existence of a limiting distribution of waiting-times.

Let \( u_r = s_r - t_r \) , (the difference between service and inter-arrival times.)

and let \( w_r \) be the waiting time of the \( r \)-th customer, i.e. the length of time he spends in the queue waiting to begin his service time.

We have, easily

\[
\begin{align*}
w_{r+1} = w_r + u_r & \quad \text{if } w_r + u_r > 0 \\
= 0 & \quad \text{if } w_r + u_r \leq 0
\end{align*}
\]

(1.5.1)

In virtue of our two assumptions the \( \{ u_r \} \) are identically and independently distributed and \( \mathbb{E}[|u_r|] \) is finite. Also, \( u_r \) is independent of \( w_r \). But, \( w_r \) is necessarily positive and hence \( F_r(x) = 0 \) for \( x < 0 \), where \( F_r(x) = \mathbb{P}(w_r \leq x) \) is the distribution
function of the random variable $w_r$. But, $F_r(o)$ is the probability that the $r$-th customer will not have to wait. Thus, in general, $F(x)$ has a discontinuity at the origin.

We assume that the first customer does not have to wait, that is, the process starts with his entry into the service system.

Hence,

$$v_1 = o \quad \text{and} \quad F_1(x) = 1 \quad \text{for} \quad x \geq o.$$  

Now, from (1.5.1), we have, for any $x > o$,

$$F_{r+1}(x) = P(w_{r+1} \leq x)$$

$$= P(w_{r+1} = o) + P(o < w_{r+1} \leq x)$$

$$= P(w_r + u_r \leq o) + P(o < w_r + u_r \leq x)$$

$$= P(w_r + u_r \leq x)$$

(1.5.2)

$$= \int_{w_r + u_r \leq x} dF_r(w_r) dG(u_r)$$

$$= \int_{u_r \leq x} F_r(x - u_r) dG(u_r).$$

where $G(x)$ is the distribution function of any $u_r$.

Thus, knowing the distribution of $G(x)$, we can calculate $F_r(x)$ successively for increasing $r$. We have then the result that the waiting time distribution depends not on the individual distributions of the service-times and inter-arrival-times, but only on the distribution of the difference between them.
Nowhere have we used the fact that \( s_r \) and \( t_r \) are independent, and hence we could use weaker assumptions, but this would not serve any useful purpose. Below, we utilise the factorisation of the distribution of the \( u \)'s in a special case.

Lindley shows that our queueing problem can be related to a random walk with an absorbing barrier. In fact, he shows that

\[
(1.5.3) \quad F_{r+1}(x) = P(U_s \leq x \text{ for all } s \leq r)
\]

where, after replacing our original \( u_s \) by \( u_{r+1-s} \) for \( 1 \leq s \leq r \), i.e. inverting the series about its median, we define \( U_s \) to equal

\[
\sum_{s=1}^{r} u_s.
\]

The queueing problem thus corresponds to an infinity of random walks with different barriers, one for each value of \( x \) considered.

If \( E_r \) is the event

\[
\left\{ U_s \leq x \text{ for all } s \geq r \right\},
\]

the sequence of events \( \{E_r\} (r = 1, 2, \ldots) \) is decreasing, tending to the limit event \( E \), as \( r \) increases where \( E \) is the event,

\[
\left\{ U_s \leq x \text{ for all } s \geq 1 \right\}
\]

Hence, by a well-known property of a probability measure, there exists

\[
\lim_{r \to \infty} F_{r+1}(x) = \lim_{r \to \infty} P(E_r) = P(E)
\]

If we denote this limit by \( F(x) \), it clearly satisfies, for \( x \geq 0 \).
(1.5.4) \[ F(x) = \int_{u \leq x} F(x-u) \, d \, G(u) \]

or,

(1.5.5) \[ F(x) = \int_{y \geq 0} F(y) \, d \, G(x-y) \]

It follows that, since \( F(x) = P(U_s \leq x \text{ for all } s \geq 1) \), \( F(x) \) is a non-negative, non-decreasing function with \( F(x) = 0 \) for \( x < 0 \).

Finally, we must discover when \( F(x) \) is a distribution function, so that an equilibrium distribution of waiting times exists. There are three cases.

(i) \( E \sqrt{\frac{u}{n}} > 0 \).

By the strong law of large numbers \( \lim_{n \to \infty} \frac{U_n}{n} = E(u) \), with probability one and thus, except for an event with zero probability, we can find, for any sequence \( \{ u_1, u_2, \ldots \} \), an \( n \) such that \( U_n > \frac{1}{2} n \, E \sqrt{\frac{u}{n}} \) for all \( n > n_0 \). Thus, for any \( x \), by choosing \( n_0 \) sufficiently large, \( U_n > x \) for all \( n > n_0 \) with probability one. Thus, \( F(x) = 0 \) for all \( x \). In other words, the distribution function of waiting times does not tend to a limiting distribution; the waiting time increases beyond all bounds.

(ii) \( E \sqrt{\frac{u}{n}} < 0 \).

As in the first case, the strong law of large numbers applies and shows that for any positive \( \delta \) there exists an \( n_0 \) such that,

(1.5.6) \[ P(U_n \leq 0, \text{ for all } n \geq 0) > 1 - \frac{\delta}{2} \]

By considering the distribution of \( U_1, \ldots, U_n \); a positive \( x \)
can be found so that

\[(1.5.7) \quad P(U_n \leq x, \text{ for all } n \leq n_0) > 1 - \delta/2 .\]

(1.5.6) and (1.5.7) together show that

\[P(U_n < x, \text{ for all } n \geq 1) > 1 - \delta ,\]

and since \( \lim_{x \to \infty} F(x) \leq 1 \) we must have \( \lim_{x \to \infty} F(x) = 1 \); consequently, \( F(x) \) is a limiting distribution function.

(iii) \( E\sqrt{u} = 0 .\)

This case cannot be treated by the strong law which only shows that the \( U_n/n \to 0 \), and provides no information about the probability of the event \( E \). More delicate methods are needed, and have been supplied in a paper by Chung and Fuchs \( \sqrt{u} \), who show, as a special case of some general results, that if \( E\sqrt{u} = o \), and not all values assumed by \( u \) are integral multiples of a fixed number, then

\[(1.5.8) \quad P( |U_n - x| < \varepsilon, \text{ for an infinity of } n) = 1\]

for any \( x \) and any \( \varepsilon > 0 \). If \( u \) only assumes integral multiples of a fixed number then (1.5.8) only holds for \( x \) having these same values, and the case where \( u = 0 \) is excluded. This result has the immediate consequence that \( F(x) = o \) for all \( x \), since any value will be exceeded by \( U_n \), for some \( n \), with probability one. This leaves us with the case \( u = 0 \), which is trivial.

We have, therefore, the following result:-
A necessary and sufficient condition that the distribution function of the waiting time tends to a non-degenerate limit, as the number of customers increases, is that $\mathbb{E}^{-u} < 0$ or $u = 0$ certainly. If $\mathbb{E}^{-u} > 0$, and $u \neq 0$, the probability of waiting a time not more than $x$, tends to zero, for any $x$.

1.6. Uniqueness of the limiting distribution function of the waiting times.

Suppose the first customer waits a time $y$, and let $F_r(x/y)$ denote the conditional distribution function of customer number $r$. Then a calculation along the lines of that used to establish (1.5.3) shows that,

$$F_{r+1}(x/y) = P(U_s \leq x, \text{ for } 1 \leq s < r; U_r \leq x-y)$$

$$\geq P(U_s \leq x, \text{ for } 1 \leq s < r) - P(U_r > x-y).$$

Hence,

$$\lim_{r \to \infty} \inf F_r(x/y) \geq F(x),$$

since $P(U_r > x-y)$ tends to zero by the strong law of numbers whenever $E^{-u} < 0$. But also,

$$F_{r+1}(x/y) \leq P(U_s \leq x, \text{ for } 1 \leq s < r),$$

and so,

$$\lim_{r \to \infty} \sup F_r(x/y) \leq F(x).$$
(1.6.1) and (1.6.2) establish the independence of the waiting time, y, of the first customer, and that of any subsequent customer. If y has a distribution function H(y),

\[ F_r(x/H(y)) = \int F_r(x/y) dH(y), \]

and hence,

\[ \lim_{r \to \infty} F_r(x/H(y)) = \int \lim_{r \to \infty} F_r(x/y) dH(y) = F(x), \]

by Lebesgue's theorem on dominated convergence. These results show that there is one and only one distribution function satisfying (1.5.5) whenever \( \int y dG(y) < 0 \). That there exists at least one has been shown. To see that it is unique, suppose there were another solution, and that it was the waiting-time distribution of the first customer, then it would be the waiting-time distribution of every customer, and hence the limiting distribution, which is impossible.

Thus, we can draw the following conclusion.

The system has a unique stationary distribution if \( E[u^2] < 0 \), or \( u = 0 \). Therefore, recalling the equivalence of our process and a random walk, that the event A, that a customer does not have to wait, will occur and that it will have finite mean recurrence time. On the other hand, if \( E[u^2] > 0 \), then A is a transient event, and with probability one there is some point when every subsequent customer will have to queue. Finally, if \( E[u^2] = 0 \), then A is a certain event with an infinite mean recurrence time. This completes our study of the existence of the distribution of waiting times.
1.7 The waiting time

Let us now assume that,

\[ E[\bar{v}] < 0 , \]

i.e.

\[ E[s_{n-1}] < E[t_{n-1}] . \]

Thus, if we define the traffic intensity \( \rho \) by \( E[s_{n-1}]/E[t_{n-1}] \), then \( \rho < 1 \). Suppose also that the input is random with mean rate \( \lambda \), and that we have a general service time with mean rate \( \mu \). Then we are assuming that,

\[ \mu/\lambda < 1 . \]

The epochs at which customers leave are points of regeneration, as defined by Bartlett and Kendall \( \sum_{\lambda} \). Consider one such and let \( q \) be the size of the queue this customer leaves behind him. Let the service time of the next customer be \( v \), and suppose that \( r \) customers arrive during this time. Then, conditionally, \( r \) has a Poisson distribution with mean \( v/\lambda \), where \( v \) has the service time distribution.

Denoting by \( q' \) the size of the queue left behind by the second customer, we have in statistical equilibrium (if it can exist),

\[ q' = \max (q-1, 0) + r , \]

or,

\[ q' = q-1 + s + r , \]

where \( s = s(q) \) is zero for all non-zero \( q \), and \( s(0) = 1 \). As a consequence of the definition of \( s(q) \), we have
(1.7.3) \( \delta^2 = \delta \) and \( q(1-\delta) = q \).

Assuming now that an equilibrium solution exists and that the equilibrium values of \( E[q] \) and \( E[q^2] \) are finite, then, on forming the expectation of both sides of (1.7.2) we have,

\[
E[q'] = E[q] = E[q^2] - 1 + E[\delta^2] + E[r'] ,
\]
or,

(1.7.4) \( E[\delta^2] = 1 - E[r'] = 1 - \rho \)

This is the probability that a departing customer leaves an empty counter behind him.

Squaring both sides of (1.7.2) and using (1.7.3) we obtain

\[
q' = q^2 \quad 2q(1-r) + (r-1)^2 + 8(2r-1)
\]

and on forming expectations,

\[
E[q^2] = E[r'] + E[r(1-r)]/2 \quad \{1-E[r-1]\} .
\]

(1.7.5) \[
= \frac{\mu}{\lambda} + \frac{\text{var}(v) + \mu^2}{2\lambda (\lambda - \mu)}
\]

Suppose that a departing customer has waiting-time and service-time \( w \) and \( v \), respectively, and leaves \( q \) customers behind him. Then \( q \) is the number of arrivals in a total time \( w + v \) and so,

(1.7.6) \[
E[q'] = E[w'] + E[v'] / \lambda .
\]
From this, and (1.7.5), it is easily deduced that

\[ \frac{E[s^2]}{E[v^2]} = \frac{\rho}{\pi (1-\rho)} \cdot \left( 1 + \text{var} \left( \frac{v}{\mu} \right) \right) \]

(1.7.7)

If the mean service time is kept constant and the frequency of calls for service remains the same, then \( \lambda \) and \( \mu \) will be fixed and in these circumstances it follows from (1.7.7) that maximum efficiency will be obtained if, and only if, there is no variation in the service time. With a negative-exponential distribution of service times, however, the ratio (1.7.7) is equal to twice the minimum value.

For a fixed form of service time distribution the ratio is a constant multiple of \( \rho/(1-\rho) \) and consequently can only be reduced by reducing \( \rho \) (i.e., by a reduction either in mean service time or in frequency of calls for service). This implies an increase in the fraction of time \( 1-\rho \) (1.7.4) during which the counter is unused, and so an increase of efficiency in one direction leads to a decrease in another.

Now let us consider the distribution of waiting times. If

\[ E[s^2]/E[v^2] \]

is considerably less than unity an approximation to the limiting distribution can be found by performing the necessary integrations of (1.5.2)

Consider now the most important situation, when the arrival times are random.

Let \( G_1(y) \) and \( G_2(y) \) be the distribution functions of \( s \) and \( t \) respectively, so that both vanish for negative arguments and

\[ g(y) = \int g_1(y+z) \, dg_2(z) \]
If the arrivals are random, the distribution of $t$ is negative exponential and in this case,

$$G(y) = \lambda \int_0^\infty e^{-\lambda z} G_1(y+z) \, dz,$$

where $E[t]$ is $1/\lambda$.

Let $F^*(x)$ be defined by the equation

$$(1.7.8) \quad F^*(x) = \int_{y \leq x} F(x-y) \, dG(y)$$

for all $x$.

Then, in the special case of negative exponential arrivals, we have, whenever $x < 0$, in virtue of the fact that $G_1(x) = 0$, the relations

$$F^*(x) = \lambda \int_{y \leq x} \int_z^{\leq 0} e^{-\lambda z} F(x-y) \, dG_1(y+z) \, dz$$

$$(1.7.9) \quad = \lambda \int_{u \geq 0} \int_{v \geq 0} e^{-\lambda (u+v-x)} F(u) \, dG_1(v) \, du$$

$$= C e^{\lambda x}, \text{ where } C \text{ is a constant.}$$

Taking Fourier transforms of (1.7.8) we have,

$$\int_0^\infty e^{i\tau x} \, dF^*(x) = \int_0^\infty \int_0^\infty e^{i\tau x} \, dF(x-y) \, dG(y)$$

hence, since

$$\int_{-\infty}^0 dF^*(x) = F(0)$$

and

$$F^*(x) = F(x) \text{ for } x \geq 0,$$
\[
\int_{-\infty}^{0} e^{i\tau x} dF^*(x) - \int_{-\infty}^{0} dF^*(x) + \int_{0}^{\infty} e^{i\tau x} dF(x) = \int_{-\infty}^{0} e^{i\tau x} dF(x) \int_{-\infty}^{0} e^{i\tau y} dG(y)
\]

Letting \( \phi(\tau) \) and \( \psi(\tau) \) denote the characteristic functions of \( w \) and \( s \) respectively, then

\[
\frac{C \ \lambda}{\lambda + i\tau} - C + \phi(\tau) = \phi(\tau) \frac{\lambda}{\lambda + i\tau} \cdot \psi(\tau)
\]

and thus,

\[
(1.7.10) \quad \phi(\tau) = C \cdot \left[ 1 + \lambda \cdot \frac{1 - \psi(\tau)}{i\tau} \right]^{-1}
\]

Now \( \phi(0) = 1 \). Thus, as

\[
\lim_{\tau \to 0} \frac{1 - \psi(\tau)}{i\tau} = -\mathbb{E} \int_{-}\mathbf{1}
\]

it follows that, as \( 1/\lambda = \mathbb{E}[s]\),

\[
(1.7.11) \quad C = 1 - \mathbb{E}[s] / \mathbb{E}[v] \cdot \mathbf{1}
\]

Lindley \( \int_{10} \), gives tables of the values of the probability of not having to wait, mean waiting time and variance of waiting time in the cases of random arrival and regular arrival when the service time is of \( \chi^2 \) distribution with 2 or 3 degrees of freedom. He also showed that if service time were exponentially distributed and arrivals were regular then the waiting time was again exponentially distributed, and it has been known for a long time that the same is true if both arrivals and service times were exponentially distributed. W. L. Smith \( \int_{10} \), however, has shown that if the service time is distributed exponentially, then so is the waiting time, whatever the arrival-time distribution.
CHAPTER TWO
GENERAL QUEUEING SYSTEMS

2.1 Introduction.
In this chapter we will consider only the simple arrangements
of one or two queues in series, either single, or multi-server in
either case. We will discuss such topics of interest as, the out-
put; independence of interdeparture intervals and the state of the
system at those departures; mutual independence of interdeparture
intervals; and the distributions of waiting times at the various queues.

2.2 The output of a single multi-server queue.
We will consider, with P. J. Burke [37], a queue with random
arrivals and exponential service times. Let there be \( s \) servers
each with an exponential service time with mean \( 1/\mu \), and let the
average interval between arrivals be \( 1/\lambda \). Then, it is well known,
Kendall [37], that if \( s \mu > \lambda \) there is an equilibrium distribu-
tion of states of the system, where by state of the system we mean
the number of customers queueing at any particular instant, under
general independence conditions between the service and inter-arri-
val times and the state of the system at any particular time. We
shall assume that all customers remain in the system until they have
received service, otherwise the queue discipline, or order of service,
is irrelevant, as, in this section, we are only interested in the
output.
The probability of the system being in some state \( k \) when a customer departs, is the same as that of being in the state \( k \) when a customer arrives. For, the first case represents a transition from \((k+1)\) to \( k \), whilst the latter represents a transition from \( k \) to \((k+1)\). Over a period of time, the number of transitions in each direction cannot differ by more than one, but the limit of the probability in the latter case is known to be equal to the probability of the system being in the state \( k \) at an arbitrary instant. Hence, so is the former limit.

Following Feller [27], let the probability that the system is in state \( k \) at any time \( t \), be \( p_k(t) \), and consider \( p_k(t+\delta t) \).

In any small interval of time, \( \delta t \), the system is such that the probability of more than one event happening is of the order of \( \delta t^2 \) and hence is negligible. Thus, the probability of the system being in the state \( k \) at time \( t+\delta t \) can be represented by,

\[
p_0(t+\delta t) = p_0(t)(1 - \lambda \delta t) + p_1(t) \mu \delta t, \quad (k = 0)
\]

\[
p_k(t+\delta t) = p_k(t)(1 - k \mu \delta t - \lambda \delta t) + p_{k+1}(t)(k+1) \mu \delta t + p_{k-1}(t) \lambda \delta t, \quad (1 \leq k < \infty)
\]

\[
p_k(t+\delta t) = p_k(t)(1 - s \mu \delta t - \lambda \delta t) + p_{k+1}(t)s \mu \delta t + p_{k-1}(t) \mu \delta t, \quad (k \geq \infty).
\]
From these equations, we can infer that

\[
\frac{dp_0(t)}{dt} = \mu p_1(t) - \lambda p_0(t),
\]

\[
\frac{dp_k(t)}{dt} = (k+1)\mu p_{k+1}(t) - (\lambda + k\mu)p_k(t) + \lambda p_{k-1}(t),
\]

\((k+1 \leq s)\)

\[
\frac{dp_k(t)}{dt} = s\mu p_{k+1}(t) - (\lambda + s\mu)p_k(t) + \lambda p_{k-1}(t),
\]

\((k+1 > s)\).

In the steady state our probabilities are independent of time, and so we have:

\[
p_1 = \frac{\lambda}{\mu} \cdot p_0,
\]

\[(2.2.1) \quad p_{k+1} = -\frac{(\lambda + k\mu)}{(k+1)\mu} p_k + \frac{\lambda p_{k-1}}{(k+1)\mu}, \quad (k+1 \leq s).\]

\[
p_{k+1} = -\frac{(\lambda + s\mu)}{s\mu} p_k + \frac{\lambda p_{k-1}}{s\mu}, \quad (k > s).
\]

These equations are obviously satisfied by,

\[
p_k = p_o \left(\frac{\lambda}{\mu}\right)^k / k!, \quad (0 \leq k < s).
\]

\[
p_k = p_o \left(\frac{\lambda}{\mu}\right)^k / s! s^{k-s}, \quad (k > s).
\]

This solution is just a special case of the more general multi-stage solution of R. R. P. Jackson 117, which will be discussed.
later and the uniqueness of which will be demonstrated. To complete the solution, \( P_0 \) is determined by the normalising condition
\[ \sum_{k=0}^{\infty} P_k = 1. \]

Let \( L \) denote the arbitrary inter-departure interval, and \( n(t) \) the state of the system at a time \( t \) after the previous departure. Let \( F_k(t) \) be the joint probability that \( n(t) = k \) and that \( L > t \).

The marginal distributions of \( L \) and the equilibrium probabilities are given by,
\[ \sum_{k=0}^{\infty} F_k(t) = F(t) \]
the distribution function of \( L \),

and
\[ F_k(0) = P_k \]
the equilibrium probability of being in state \( k \).

For an infinitesimal interval of length \( dt \),
\[ F_0(t + dt) = F_0(t) (1 - \lambda dt) , \]
neglecting terms of \( o(dt^2) \).

Similarly,
\[ F_k(t+dt) = F_k(t)(1-\lambda dt - k \mu dt) + F_{k-1}(t) \lambda dt, \] \( (k < s) \)

and
\[ F_k(t+dt) = F_k(t)(1-\lambda dt - s\mu dt) + F_{k-1}(t) \lambda dt, \] \( (k > s) \)

These equations lead to:
\[ F_0'(t) = -\lambda F_0(t), \]
\[ F_k'(t) = \lambda F_{k-1}(t) - (\lambda + k \mu) F_k(t), \quad (k < s) \]
\[ F_k'(t) = \lambda F_{k-1}(t) - (\lambda + s \mu) F_k(t), \quad (k \geq s) \]

subject to the initial conditions, \( F_k(0) = p_k \). The latter equations yield,

\[ (2.2.2) \quad F_k(t) = p_k e^{-\lambda t} \]

as the unique solutions subject to the initial conditions.

Thus, we have that,

\[ F(t) = \sum_{k=0}^{\infty} p_k e^{-\lambda t} = e^{-\lambda t} \]

and consequently, the marginal distribution of the interdeparture interval is exponential with parameter \( \lambda \), i.e. the same as the inter-arrival processes.

2.3 The independence of the interdeparture interval and the state of the system at the departure.

We have,

\[ p(t+dt > L > t, \text{ and } n(L+o) = k) \]
\[ = \begin{cases} F_{k+1}(t) (k+1) \mu dt, & \text{for } k+1 \leq s, \\ F_{k+1}(t) s \mu dt, & \text{for } k+1 > s, \end{cases} \]

i.e.
\[ \frac{1}{k!} \left( \frac{\lambda}{\mu} \right)^k p_0 e^{-\lambda t} \lambda dt, \]

and,
\[ \frac{1}{s! s^{k-s}} \cdot \left( \frac{\lambda}{\mu} \right)^k p_0 e^{-\lambda t} \lambda dt, \]
respectively, from (2.2.1) and (2.2.2).

These probabilities can be factored into the marginal distributions of \( n(L) \) and \( L \), thus proving the independence of \( L \) and \( n(L) \).

2.4. **The independence of inter-departure intervals.**

Let \( \Lambda \) represent the set of lengths of an arbitrary number of inter-departure intervals subsequent to \( L \). Because of the randomness of the input, and because of the exponential service time distribution, the output process is Markovian with respect to the state of the system. This implies that,

\[(2.4.1) \quad p(\Lambda /n(L)) = p(\Lambda /n(L), L),\]

where to avoid ambiguity \( n(L) \) can be taken to mean \( n(L+0) \). The independence of \( n(L) \) and \( L \) gives us,

\[(2.4.2) \quad p(n(L), L) = p(n(L)) p(L).\]

We can write,

\[(2.4.3) \quad p(L, n(L), \Lambda) = p(\Lambda /L, n(L)) p(L, n(L)).\]

and on substituting from (2.4.1) and (2.4.2), we have,

\[p(L, n(L), \Lambda) = p(\Lambda /jn(L)) p(n(L)) p(L)\]

Whence,

\[p(L, \Lambda) = \sum_{n(L)=0}^{\infty} p(\Lambda, n(L)) p(L) = p(\Lambda) p(L).\]
2.5 A single server queue with random arrivals and general service time.

In this case, P. D. Finch \( \int_0^\infty \) considers a single server queue with a Poisson input process and general service time.

Let the Poisson input process have parameter \( \lambda \) and let the service time of customer number \( r \) be \( s_r \). Suppose that the \( s_r \) are identically and independently distributed with density function \( B(x) \), such that \( \int_0^\infty s^{-1} < \infty \), and suppose \( B(x) \) possesses a continuous second derivative.

Define \( n_{r+1}(t) \) as before, \( \ell_r \) as the time interval between the departures of customers number \( r \) and \( r+1 \), and \( H_{r+1}(t, j) \) as the joint frequency function for \( n_{r+1} \) and \( \ell_r \), that is:

\[
H_{r+1}(t, j) \delta t = p(n_{r+1} = j, t + \delta t > \ell_r > t), \quad (j = 0, 1, \ldots)
\]

we have for our marginal distributions of \( \ell_r \) and the steady state probabilities,

\[
H_{r+1}(t) \delta t = \sum_{j=0}^{\infty} H_{r+1}(t, j) \delta t = p(t + \delta t > \ell_r > t).
\]

\[
P_{r+1}(j) = \int_0^\infty H_{r+1}(t, j) \, dt = p(n_{r+1} = j), \quad (j = 0, 1, \ldots)
\]

We consider only the case \( \lambda \int_0^\infty s^{-1} < 1 \), for, as we know, this is the only case of practical importance which has a stationary distribution.

The number of customers left behind by the departure number \( r+1 \) will be independent of the time interval between departures numbers.
$r$ and $(r+1)$, in the limit as $r \rightarrow \infty$ if, and only if,

$$H(t, j) = H(t) p(j),$$

where, $H(t) = \lim_{r \rightarrow \infty} H_r(t)$.

Now,

$$H_{r+1}(t) = p(n_r = 0) p(t + \delta t > \ell_r > t/n_r = 0) + p(n_r > 0) p(t + \delta t > \ell_r > t/n_r = 0)$$

but the conditional distribution of $\ell_r$ given $n_r > 0$ is just that of a service time, whilst the conditional distribution of $\ell_r$ given that $n_r = 0$ is the sum of an arrival interval and a service time.

Thus,

$$H_{r+1}(t) = p_r(o) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x) + \{1 - p_r(o)\} B'(t).$$

Since, $p(o) = \lim_{r \rightarrow \infty} p_r(o)$ exists, it follows that

$$H(t) = \lim_{r \rightarrow \infty} H_r(t)$$

exists and is given by

$$(2.5.2) \quad H(t) = p(o) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x) + (1-p(o)) B'(t).$$

If $n_r = 0$, then there will be $j+1$ customers present at time $t$ after departure number $r$ where $t + \delta t > \ell_r > t$, if and only if a customer arrives in some interval $(\tau, \tau + \delta \tau)$ after departure number
r(\text{where } \tau < t), and if there are \( j \) subsequent arrivals in the time interval \((\tau, t)\) where the service time \( s_{r+1} \) extends from \( \tau \) to \( t \) but not to \( t + 8t \).

Similarly, if \( n_\tau = m \) it will be the case that \((j+1)\) customers are present at time \( t \) after the departure number \( r \), and that \( t + 8t > \langle R \rangle > t \), if and only if \( j + \langle R \rangle - m \) arrive in time \( t \) after the departure number \( r \) and \( t < s_\tau < t + 8t \).

Thus, as

\[
H_{r+1}(t, j) = p(n_{r+1} = j, t + 8t > \langle R \rangle > t) = p(n_\tau(t) = j+1, t + 8t > \langle R \rangle > t)
\]

we have,

\[
H_{r+1}(t, j) = p_r(o) \int_0^t e^{-\lambda t} \left\{ \lambda \left( \frac{t - \tau}{t} \right)^{j-1} \frac{\lambda(t-\tau)}{j!} \right\} B'(t-\tau) d\tau + \sum_{m=1}^{j+1} p_r(m) \left\{ \frac{(\lambda t)^{j+1-m}}{(j+1-m)!} \right\} e^{-\lambda t} B'(t).
\]

Since, under our assumptions the following exist,

\[
\lim_{r \to \infty} p_r(o) \quad \text{and} \quad \lim_{r \to \infty} p_r(m)
\]

then the existence of the following limit is implied.

\[
H(t, j) = \lim_{r \to \infty} H_{r}(t, j)
\]

Moreover, we can deduce that
\[ H(t, j) = p(o) \int_0^t e^{-\lambda \tau} \left( \frac{\lambda(t-\tau)^j}{j!} \right) e^{-\lambda(t-\tau)} B'(t-\tau) \, d\tau \]

\[ + \sum_{m=1}^{j+1} p(m) \left\{ \frac{(\lambda t)^{j+1-m}}{(j+1-m)!} \right\} e^{-\lambda t} B'(t) \]

In order to prove that the number of customers left by departure number \( r \) is not independent of the time interval between departures number \( r \) and \( r+1 \), in the limit as \( r \rightarrow \infty \), it is sufficient to prove that

\[ H(t, j) \neq H(t) \, p(j) \]

for some value of \( j \).

Let us consider the case \( j = 0 \).

Using equations (2.5.2) and (2.5.3) we have that a necessary condition for

\[ H(t, 0) = H(t) \, p(0) \]

is that,

\[ p'(1) B'(t) e^{-\lambda t} + p(0) B(t) \lambda e^{-\lambda t} \]

\[ = p(0) (1-p(0)) B'(t) + p(0) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} \, dB(x) \]

Differentiating both sides of (2.5.5) with respect to \( t \) we have,

\[ p(1) e^{-\lambda t} B''(t) + \lambda p(o - p(1)) B'(t) e^{-\lambda t} - \lambda^2 p(o) B(t) e^{-\lambda t} \]

\[ = p(0)(1-p(0)) B''(t) + \lambda p(0) B'(t) - \lambda^2 p(0) e^{-\lambda t} \int_0^t e^{\lambda x} \, dB(x) \]

Eliminating the integral between equations (2.5.5) and (2.5.6)
gives, after simplification,

\[ B''(t) \int p(o)(1-p(o)-p(1)e^{-\lambda t} + \lambda p(o)B'(t)(1-e^{-\lambda t}) = 0 \]

which has the solution

\[ B'(t) = C e^{\frac{-\lambda t}{1-p(o)}} \int \frac{p(o)(1-p(o)-p(1)e^{-\lambda t})}{p(o)(1-p(o)-p(1))} \]

where \( C \) is a positive normalising constant. Hence,

\[ B'(0+) = C > 0 \]

but, by letting \( t \to 0+ \) in (2.5.5) we see that, since

\[ B(0+) = 0, \]

and \( B'(0+) \neq 0 \), we must have,

\[ p(1) = p(o)(1-p(o)) \]

Equation (2.5.6) then reduces to

\[ (1-p(o)) B''(t) + \lambda B'(t) = 0 \]

whence

\[ B(t) = 1 - e^{-\lambda t/1-p(o)} \]

Thus the service time is exponential.

If we let \( P_{r+1}(n/\tau) \) be the conditional probability that customer number \( r \) leaves \( n \) customers behind him on departure, when it is given that \( I_r = \tau \), and let \( H_{r+2}(t/\tau) \) be the conditional probability density for \( I_{r+1} \), we have, by an argument similar to that used in deriving (2.5.2)
\[ H_{r+2}(t/\tau) = (1-p_{r+1}(o/\tau)) B'(t) + p_{r+1}(o/\tau) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x). \]

Thus, as \( r \to \infty \),

\[ H(t/\tau) = (1-p(o/\tau)) B'(t) + p(o/\tau) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x). \]

However, if successive intervals were independent in the limit, we would have,

\[ H(t/\tau) = H(t) \]

which would imply that,

\[ (1-p(o/\tau)) B'(t) + B(o/\tau) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x) \]

\[ = (1-p(o)) B'(t) + p(o) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x), \]

i.e. \( (p(o/\tau) - p(o)) B'(t) - (p(o/\tau) - p(o)) \lambda e^{-\lambda t} \int_0^t e^{\lambda x} dB(x) = 0. \)

It follows that either (i) \( p(o/\tau) = p(o) \),

or (ii) \( B'(t) \)

If (i) is the case it follows that

\[ H(t, o) = H(t) p(o) \]

If not (i) then (ii) must be true. Differentiating (ii) and eliminating the integral, gives \( B''(t) = 0 \) for all \( t \). Hence, \( B'(x) = C \) for all \( x \), and on substituting this back in (ii) we find that \( C = 0 \). Hence \( B(x) \) cannot be a distribution function.

Thus, (i) must be true, and we have, in conclusion, that the interdeparture intervals are independent if and only if the service time distribution is negative exponential.
2.6. The waiting time at a second multi-server queue.

E. Reich \( \sqrt{177} \), uses a different approach to attack the problem of waiting times and independence of inter-departure times. He uses the reversibility of the Markov chain to study the distribution of waiting times when the customer proceeds to a second multi-server queue after having been served at the first.

We restrict our attention to unsaturated queues in "equilibrium."

Definition:

A stationary stochastic process \( N(t) \) is said to be reversible, if \( N(t) \) and \( N(-t) \) have the same multivariate distributions.

This necessary and sufficient condition for reversibility becomes, if \( N(t) \) is a discrete or continuous parameter Markov chain with a denumerable state space,

\[
\Phi_{ij}(t) = p_i P_{ij}(t) = p_j P_{ji}(t) = \Phi_{ji}(t) \text{ for all } i, j = 0, 1, 2, ...
\]

where \( p_i \) and \( P_{ij}(t) \) are, respectively, the stationary and transition probabilities of \( N(t) \).

Kolmogorov's criterion, \( \sqrt{157} \), for the reversibility of Markov chains with a finite state space, may, in a special case, which includes Reich's, (14), "type A", be immediately generalised to the denumerable state space as follows.

Theorem.

Let \( N(k); k = 0, 1, 2, ... \) be an irreducible stationary discrete parameter Markov chain with the state space \( 0, 1, 2, ... \), let the stationary probabilities be \( u_k \), and let the single step transition
probabilities be $\pi_{ij}$. A necessary and sufficient condition for the reversibility of the Markov chain is that,

$$\pi_{12} \pi_{23} \cdots \pi_{n-1,n} \pi_{n1} = \pi_{11} \pi_{12} \cdots \pi_{12} \pi_{n1},$$

for every sequence of non-negative integers $(i_1, i_2, \ldots)$, beginning and ending with the same integer.

The proof of the theorem is to be found in Reich's paper, and as it is of no interest as far as the general theme of this account is concerned, it is omitted here.

By considering the special case he calls his "type A", Reich shows easily that the stationary birth and death process is reversible. Then, by noting its equivalence, with constant birth probabilities independent of the state of the system, and that of a stationary unsaturated queue with Poisson input, exponential service time, first come-first served queue discipline and $s$ servers, he shows three things. If customers arrivals correspond to births and departures to deaths then,

a) the sequence of departure times form a Poisson process,

b) the value of $n(t)$ is independent of all past departure times, and,

c) if $t_0$ is a departure time, then $n(t_0+\delta)$ is independent of all past departure times.

These results are equivalent to the theorem of Burke $\sqrt{37}$, proved.
earlier by a different method.

Suppose the customers, upon departure from the first queue, immediately enter a second multi-server queue, where once again service is first come-first served, with independent exponential service times. It follows from result (b), that if \( n_1(t) \) and \( n_2(t) \) refer, respectively, to the first and second queues then \( n_1(t) \) and \( n_2(t) \) are independent, \( \tau \leq t \). This result was first proved in the special case \( s = 1, \tau = t \) by R. R. P. Jackson [117], by a different method which will be demonstrated later.

Here, and throughout the rest of this account, stage time will be used to refer to the time spent by the customer between arrival and departure, i.e. the sum of the waiting time and service time. Let \( \tau_1 \) and \( \tau_2 \) represent a customer's stage times at the first and second queues respectively, and let \( n_1 \) and \( n_2 \) denote the number of customers this customer leaves behind him at the first queue, and finds at the second queue, respectively (customers being served included). As a corollary of (c) above, \( n_1 \) and \( n_2 \) are independent.

Let \( A(t, k) = P(\tau_1 < t/n_2 = k) \). If \( \lambda \) is the parameter of the Poisson input process, then \( n_1 \) is the number of Poisson events which occurred during the waiting period \( \tau_1 \). Therefore, we have that

\[
P(n_1 = j/\tau_1 = t, n_2 = k) = e^{-\lambda t} \frac{(\lambda t)^j}{j!}.
\]
Hence,

\[ E \left[ z_{1/n_2} \right] = E_t \int E_j (z_{1/n_2} = k, \tau_1 = t) \]

\[ = E_t \int e^{\lambda t z} e^{-\lambda t} \]

\[ = \int_0^\infty e^{\lambda t z} e^{-\lambda t} \, dA(t, k). \]

But the left hand side of this last equation is independent of \( k \) (as \( n_1, n_2 \) are independent). Therefore, \( A(t, k) \) does not depend on \( k \), and, thus, \( n_2 \), and consequently \( \tau_2 \), are independent of \( \tau_1 \). In other words, the numbers of customers waiting in the first queue and the second queue at any time are independent, as are also a customer's stage times in the first and second queues. We are, of course, assuming infinite queues as a possibility at either queue.
CHAPTER THREE
MULTI-STAGE QUEUEING SYSTEMS

3.1. Introduction.

In this chapter we define a series of queues, and then investigate the ergodicity of one such series. After establishing the conditions under which a steady state exists, we study the series in its steady state, both as a whole and in its single queue elements. Finally, we study the concept of a network of queues.

3.2. A series of queues.

A series of queues is a system whereby the input to any queue is the output of the preceding one in the series, except for the initial queue whose input is the input to the system as a whole.

3.3. The ergodicity of queues in series.

This topic has been treated thoroughly by J. Sachs [187]. Following him, let us consider the case where we have $s$ queues, each with one server.

For $\sigma = 1, 2, \ldots, s$, let $R_n^\sigma$ be the service time of customer number $n$ in queue $\sigma$. Denote by $R_{n+1}^0$ the time $\tau_{n+1} - \tau_n$, where $\tau_n$ is the time at which customer number $n$ enters the system, i.e., enters the first queue. Let $W_n^\sigma$ denote the waiting time of the $n$-th person in queue $\sigma$. It is immediately obvious, by induction, that

\begin{equation}
W_{n+1}^\sigma = \max \left\{ \omega, W_n^\sigma + t_n^\sigma + \sum_{\sigma=1}^{p-1} (t_n^\sigma + W_n^\sigma - W_{n+1}^\sigma) \right\}.
\end{equation}
where \( t^\sigma_n = R^\sigma_n - R^\sigma_{n+1} \).

Let \( T^\sigma_k = \sum_{i=1}^k t^\sigma_i \) and let \( D^\sigma_k = \max^* \sum_{j=1}^p -T^\sigma_{j+1} \cdots -T^\sigma_{j+p} \),

where \( \max^* \) implies maximisation over all \( 0 \leq j_1 \leq \cdots \leq j_p \leq k \).

Also, let

\[
H^P_n = \sum_{\sigma=1}^p \left( t^\sigma_n + w^\sigma_n - w^\sigma_{n+1} \right),
\]

then, for all \( n \), it is obvious that

\[
H^1_n = D^1_{n-1} - D^1_n.
\]

Assume now that

\[
H^P_n = D^P_{n-1} - D^P_n \quad \text{for all} \quad n.
\]

Then, from (3.3.1) and the induction hypothesis,

\[
W^{p+1}_{k+1} = \max \left( \omega, W^p_k + t^{p+1}_k + H^P_k \right)
\]

\[
= \max \left( \omega, W^p_k + t^{p+1}_k + D^P_{k-1} - D^P_k \right)
\]

(3.3.2)

\[
= \max \left( T^{p+1}_k - T^{p+1}_j + D^P_j - D^P_k \right) \quad 0 \leq j \leq k
\]

Upon using this expression for \( k = n - 1 \) and \( k = n \) we obtain

\[
t^{p+1}_n + W^{p+1}_n - W^p_n = t^{p+1}_n + \max_{0 \leq j \leq n-1} \left( (T^{p+1}_{n-1} - T^{p+1}_j + D^P_j - D^P_{n-1}) \right)
\]

\[
= \max_{0 \leq j \leq n} \left( (T^{p+1}_n - T^{p+1}_j + D^P_j - D^P_n) \right)
\]
\[ = \max_{0 < j < n-1} (-T_{j}^{P+1} + D_{j}^{P}) - \max_{0 < j < n} (-T_{j}^{P+1} + D_{j}^{P}) \]

\[ + D_{n}^{P} - D_{n-1}^{P} \]

\[ = D_{n-1}^{P+1} - D_{n}^{P+1} + D_{n}^{P} - D_{n-1}^{P} \]

Thus,

\[ H_{n}^{P+1} = t_{n}^{P+1} + W_{n}^{P+1} - W_{n+1}^{P+1} + R_{n}^{P} \]

\[ = D_{n-1}^{P+1} - D_{n}^{P+1} \]

and hence,

\[ (3.3.3) \quad H_{n}^{P} = D_{n-1}^{P} - D_{n}^{P} \]

for all \( n \) and all \( 0 \leq p \leq s \).

\[ z_{n} = (W_{n}^{1}, ..., W_{n}^{s}) \] is not a Markov process, but it is obvious that \( Y_{n} = (W_{n}^{1}, ..., W_{n}^{s}, R_{n}^{1}, ..., R_{n}^{s-1}) \) is the \( n \)-th random variable in a stationary Markov process, for we have, \( (3.3.1) \),

\[ W_{n+1}^{P} = \max (o, W_{n}^{P} + t_{n}^{P} + \sum_{\sigma=1}^{p-1} (t_{n}^{\sigma} + W_{n}^{\sigma} - W_{n+1}^{\sigma})) \]

\[ = \max (o, W_{n}^{P} + t_{n}^{P} + H_{n}^{P-1}) \]

but \( t_{n}^{P} \) and \( H_{n}^{P-1} \) are both functions of \( Y_{n} \) only. This leads us to:-

Lemma 1.

Let \( t = (t_{1}, t_{2}, ..., t_{s}) \), \( x = (x_{1}, ..., x_{s}) \) and let

\[ r = (r_{1}, ..., r_{s-1}), \quad R_{n} = (R_{n}^{1}, ..., R_{n}^{s-1}) \]
where \( t_i \geq 0 \), \( i = 1, 2, \ldots s \)
and \( x_j \geq 0 \), \( j = 1, 2, \ldots s \).

Then,

\[
p(z_n \leq t/z_1 = x, R_1 = r) \leq p(z_n \leq t/z_1 = 0, R_1 = r)
\]

for all \( n, x, t, r \).

**Proof of Lemma 1.**

Fix a point \( w \) in the sample space of \( R_2, \ldots, R_n \), \( R_2^0, \ldots, R_n^0, \ldots, R_1^s, \ldots, R_n^s \) and let

\[
W_{1,1}^1(w, x) = x_1,
W_{1,2}^2(w, x) = x_2,
\vdots
\vdots
\vdots
W_{1,s}^s(w, x) = x_s.
\]

Then,

\[
W_{j}^p(w, x) = \max (0, W_{j-1}^p(w, x) + t_{j-1}^p(w) + H_{j-1}^{p-1}(w, x)),
\]

for \( 2 \leq j \leq n \).

It is clear that \( W_{j}^1(w, 0) \leq W_{j}^1(w, x) \) for each \( j \). We then observe that,

\[
t_{n}^{p-1}(w) + W_{n}^{p-1}(w, x) - W_{n+1}^{p-1}(w, x)
= t_{n}^{p-1}(w) + W_{n}^{p-1}(w, x) - \max (0, W_{n}^{p-1}(w, x) + t_{n}^{p-1}(w)
+ H_{n}^{p-2}(w, x))
\]
\[ = \min \left( w_{n}^{P-1}(w, x) + t_{n}^{P-1}(w), - H_{n}^{P-2}(w, x) \right). \]

Thus, we have

\[ W_{j}^{P} = \max (0, W_{j-1}^{P}(w, x) + t_{j-1}^{P}(w) + H_{j-1}^{P-2}(w, x)) \]

\[ + \min (W_{j-1}^{P-1}(w, x) + t_{j-1}^{P-1}(w), - H_{j-1}^{P-2}(w, x)) \]

and it follows easily that \( W_{j}^{P}(w, 0) \leq W_{j}^{P}(w, x) \) for all \( 2 \leq j \leq n \), which completes the proof of Lemma 1.

**Lemma 2.**

\[ p(z_{n} \leq t/z_{l} = 0) \to F(t) \text{ as } n \to \infty, \] where \( F \) is an \( s \) dimensional distribution function whose variation over \( s \)-dimensional space may be less than 1. i.e. \( F \) may not be a probability distribution function.

**Proof of Lemma 2.**

Let \( H(x, r) = p(t_{2} \leq x, R_{2} \leq r/z_{l} = 0) \)

Then,

\[ (3.3.4) \quad p(z_{n+1} \leq t/z_{l} = 0) = \int p(z_{n+1} \leq t/z_{2} = x, R_{2} = r, z_{l} = 0) dH(x, r) \]

Since \( \{ Y_n \} \) is a stationary Markov process,

\[ p(z_{n+1} \leq t/z_{2} = x, R_{2} = r, z_{l} = 0) = p(z_{n} \leq t/z_{l} = x, R_{l} = r) \]

\[ \leq p(z_{n} \leq t/z_{l} = 0, R_{l} = r) \] by Lemma 1.
Let \( H^* \) be the distribution function of \( R_1 \), and therefore of \( R_2 \). Then, using (3.3.5) in (3.3.4) we have,

\[
P(z_{n+1} \leq t/z_1 = 0) \leq \int P(z_n \leq t/z_1 = 0, R_1 = r) \, dH(x, r)
\]

\[
= \int P(z_n \leq t/z_1 = 0, R_1 = r) \, dH^*(x, r)
\]

\[
= P(z_n \leq t/x_1 = 0).
\]

Thus, \( P(z_n \leq t/z_1 = 0) \) is a monotone sequence and therefore converges to a limit which we call \( F(t) \). Replacing \( z_1 \) by \( \alpha_1 \), we have,

\[
P(t \leq \alpha_1, \ldots, W_n \leq \alpha_s / W_1^\sigma = 0, \sigma = 1, 2, \ldots s) \longrightarrow F(\alpha_1, \ldots, \alpha_s).
\]

**Theorem.**

Let \( \mu_\sigma = E[R_1^\sigma] \), for \( \sigma = 0, 1, \ldots, s \). Then, if \( \max_{1 \leq \sigma \leq s} \mu_\sigma < \mu_0 \), \( F \) is a probability distribution function.

**Proof of Theorem 1.**

Because of Lemma 2, we need only show that, under the conditions stated here, \( \{z_n\} \) is bounded in probability, i.e., for all \( n \),

\[
P(z_n \leq t/z_1 = 0) \geq 1 - \eta(t)
\]

where \( \eta(t) \to 0 \) as \( t \to \infty \).

This is equivalent to proving that \( W_n^1, \ldots, W_n^s \) are each bounded in probability. Lindley \( \int_{16}^{16} \) showed that \( W_n^1 \) was bounded in probability and Sachs proves that \( W_n^2 \) is bounded in probability, but the proof is long and of mainly intrinsic value only, and is omitted here. This
leaves us then, \( W_n^1, \ldots, W_n^s \) and a trivial induction argument makes it obvious that we need only consider the \( \{ W_n^s \} \). Actually, the argument given is legitimate when \( s \) is replaced by any \( 1 \leq p \leq s \).

We observe that

\[
-D_n^{s-1} \leq T_n^1 + \cdots + T_n^{s-1}.
\]

Thus, from (3.3.2) with \( k = n, p = s - 1 \),

\[
W_{n+1}^s \leq \max_{0 \leq j \leq n} \left( T_n^1 + \cdots + T_n^s - T_n^s + D_j^{s-1} \right)
\]

(3.3.6)

\[
= \max_{0 \leq j_1 \leq \cdots \leq j_s \leq n} \left( T_n^s - T_j^s + \cdots + T_n^1 - T_j^1 \right)
\]

Let \( s_0 = s \), and define \( s_i \) to be the largest \( \sigma < s_{i-1} \), \( (\sigma \geq \sigma) \),

with the property that

(3.3.7)

\[
\mu_{\sigma} - \mu_{s_{i-1}} > 0.
\]

Notice that \( \sigma = 0 \) satisfies (3.3.7), so that \( s_1 \) is well defined. Let \( k \) be the first \( i \) such that \( s_i = 0 \). Then it is easy to check that

\[
\mu_0 = \mu_{s_k} > \cdots > \mu_{s_0} = \mu_s,
\]

and that, for \( s_1 < \sigma < s_{i-1} \),

(3.3.8)

\[
\mu_\sigma - \mu_{s_{i-1}} < \mu_{s_i}.
\]

Define, for \( i = 1, \ldots, k \),
\[ U_n^i = \max_{0 \leq j_1 \leq \cdots \leq j_g \leq n} \left( \sum_{\sigma = s_i+1}^{s_i-1} (T^\sigma_n - T^\sigma_j) \right) \]
\[ \leq \max_{0 \leq j_{s_i+1} \leq \cdots \leq j_{s_{i-1}} \leq n} \left( \sum_{\sigma = s_i+1}^{s_i-1} (T^\sigma_n - T^\sigma_j) \right). \]

Because of (3.3.6), we have,

(3.3.9) \[ W_{n+1}^g \leq U_n^1 + \cdots + U_n^g \]

and, therefore, in order to show that \{W_n\} is bounded in probability, we have only to show that each \( U_n^i \), \((i = 1, \ldots)\), is bounded in probability, which can be summarized in the following lemma.

**Lemma 3**

For \( m = 1, \ldots, M \), let \( \lambda_i^m \), \( i = 1, \ldots \) be a sequence of identically and independently distributed random variables with

(3.3.10) \[ E/\sqrt{\lambda_i^m} = \lambda_m - \lambda_{m-1} \]

where \( \lambda_0 > \lambda_M > \max_0 < a < M \lambda_a > \min_0 < a < M \lambda_a \geq 0 \). (It is not assumed that \( \lambda_i^m \) and \( \lambda_{i+1}^m \) are independent of one another).

Also, let \( S_k^m = \sum_{i=1}^{k} \lambda_i^m \), and let \( \psi_n = \max_{m=1}^{M} (\sum_{m=1}^{M} (S_n^m - S_j^m)) \).

Then, \( \psi_n \) is bounded in probability.

It is easily shown that the \( U_n^i \) can be taken as the \( \psi_n \) and satisfy the conditions of the lemma, so that the proof of Lemma 3 gives us our proof of the theorem.

**Proof of Lemma 3**

Let \( \gamma = \min'(\lambda_i - \lambda_j) \), where \( \min' \) implies minimization over all
\( o \leq i, j \leq M \) with \( \lambda_i - \lambda_j > 0 \). Let \( \delta = \gamma/M \). \( \delta \) is, of course, strictly positive. For \( 2 \leq m \leq M \), define \( \epsilon_m = \lambda_M - \lambda_{m-1} + (M-m+1)\delta \), and let \( \epsilon_{M+1} = 0 \), and \( \epsilon_1 = 0 \).

Then,

(3.3.11) \( \epsilon_2, \ldots, \epsilon_M \) are positive.

(3.3.12) \( \lambda_m - \lambda_{m-1} + \epsilon_{m+1} - \epsilon_m = -\delta \), \( (2 \leq m \leq M) \).

(3.3.13) \( \lambda_1 - \lambda_0 + \epsilon_2 = \lambda_M - \lambda_0 + \int (M-1)/M \gamma \)

\[ < (-1/M)(\lambda_0 - \lambda_M) \]

\[ < 0 \]

Letting \( i_{M+1} = n \), and taking note of the fact that \( \epsilon_{M+1} = 0 \), we have

(3.3.14) \( \sum_{m=1}^{M} (n-j_m + 1) \epsilon_{m+1} - (n-j_m) \epsilon_m = 0 \),

where \( o \leq j_1 \leq \ldots \leq i_{M+1} = n \).

Upon using (3.3.14), we find

\[ \psi_n = \max \sum_{m=1}^{M} (s_m^n - s_{j_m}^n) + (n-j_{m+1}) \epsilon_{m+1} - (n-j_m) \epsilon_m \]

\[ \leq \sum_{m=1}^{M} \max (s_m^n - s_{j_m}^n) + (n-j_{m+1}) \epsilon_{m+1} - (n-j_m) \epsilon_m \]

(3.3.15) \( \leq \sum_{m=1}^{M} \max_{o \leq j_m \leq n} \int s_m^n - s_{j_m}^n + (n-j_m)(\epsilon_{m+1} - \epsilon_m) \)

Each of the terms in the summation of the right hand side of (3.3.15) is bounded in probability, since

$$\max_{0 < j_m < n} \sqrt{S_n - S_j} + (n-j_m)(\varepsilon_{m+1} - \varepsilon_m) \gamma = \max_{0 < k < n} \sqrt{S_n - S_k} \gamma,$$

where, $$S_k = \sum_{i=1}^{k} (x_i^m + \varepsilon_{m+1} - \varepsilon_m),$$

and $$E \sqrt{X_i^m + \varepsilon_{m+1} - \varepsilon_m} \gamma < c,$$ owing to (3.3.12) and (3.3.13).

This concludes the proof of Lemma 3 and hence of Theorem 1.

**Theorem 2**

If $$\max_{1 \leq \sigma \leq s} \mu_\sigma > \mu_0,$$ then $$F \equiv 0.$$

**Proof of Theorem 2**

Let $$p$$ be the first $$\sigma > 1$$ with $$\mu_\sigma > \mu_0.$$ We need only show that $$W_n^p \to \infty$$ in probability.

Using (3.3.2), we find

$$W_n^p = \max_{0 < j < n} \sqrt{T_n - T_j} - \frac{D^p - D^{p-1}}{c_n} \gamma$$

$$= \max_{0 < j_1 < \ldots < j_n} \sqrt{T_n - T_{j_1} - T_{j_2} - \ldots - T_{j_n}} - \frac{D^p - D^{p-1}}{c_n} \gamma$$

(3.3.16)

$$= \max_{0 < j_1 < \ldots < j_n} \sqrt{T_n + \ldots + T_{j_n} - T_{j_1} - \ldots - T_{j_{n-1}}} \gamma - \max_{0 < j_1 < \ldots < j_n} \sqrt{T_n + \ldots + T_{j_{n-1}} - T_{j_1} - \ldots - T_{j_{n-2}}} \gamma.$$
The last term on the right hand side of (3.3.16) is bounded in probability, because \( \max_{1 \leq \sigma \leq p-1} \mu_\sigma < \mu_0 \), by Theorem 1.

Looking at the first term on the right hand side, we have

\[
\max_{0 \leq j_1 \leq \ldots \leq j_n \leq n} \left[ T_n^p + \ldots + T_n^{j_p} - T_{j_p}^{j_{p-1}} - \ldots - T_{j_1}^1 \right]
\]

\[
= \max_{0 \leq k \leq n} \left[ \sum_{j=k+1}^{n} \Sigma_{\sigma=1}^{p} (R_{j}^\sigma - R_{j+1}^\sigma) \right]
\]

\[
= \max_{0 \leq k \leq n} \left[ \sum_{j=k+1}^{n} (R_{j}^p - R_{j}^0) + \Sigma_{\sigma=1}^{p} (R_{k+1}^{\sigma-1} - R_{n+1}^{\sigma-1}) \right]
\]

\[
\geq \max_{0 \leq k \leq n} \left[ \sum_{j=k+1}^{n} (R_{j}^p - R_{j}^0) - \Sigma_{\sigma=1}^{p} R_{n+1}^{\sigma-1} \right].
\]

The last term written is bounded in probability, whilst the preceding term tends to infinity in probability as \( n \to \infty \) because \( \mu_p \geq \mu_0 \).

It is then quite clear that \( W_{n+1}^p \) must tend to infinity in probability and hence \( F = 0 \).

To summarize, we have proved that if and only if \( \max_{1 \leq \sigma \leq s} \mu_\sigma < \mu_0 \), then the set of waiting times at the \( s \) stages tends to a limit which has a probability distribution function.

3.4. The steady state solution.

Suppose that we have a sequence of queues. Suppose, further, as R. R. P. Jackson does, that at queue \( i \) there are \( r_i \) identical servers, each of which has associated with it an exponential service time distribution with mean \( 1/\mu_i \) units of time. If a customer
arrives at the queue to find all $r_i$ servers busy, we assume that he takes his place in line behind any previous arrivals and waits until a server becomes empty, service being first come-first served.

Suppose, further, that new customers arrive at the first queue at random, with a mean rate of arrival $\lambda$, and leave when they finish service at the $k$-th queue. We assume that $\lambda < r_i \mu_i$ for all $i$. Hence from the previous section, the system has a stationary probability distribution function. We have shown in 2.2 that the input to each queue will be random with a mean rate $\lambda$.

Let $P(n_1, n_2, \ldots, n_k, t)$ be the probability that at time $t$ there are $n_1$ customers in the first stage, $n_2$ in the second, ..., and $n_k$ in the $k$-th. If we consider the changes in the system over a small time interval $\delta t$, we can obtain a set of differential difference equations in the usual manner. Then, as before, we obtain a set of steady state equations in which, $P(n_1, \ldots, n_k, t) = P(n_1, \ldots, n_k)$ for all $t$. Define,

$$a(n_j) = \begin{cases} n_j & n_j < r_j \\ r_j & n_j \geq r_j \end{cases} \quad j = 1, 2, \ldots, k$$

$$b(n_j) = \begin{cases} 1 & n_j \neq 0 \\ 0 & n_j = 0 \end{cases} \quad j = 1, 2, \ldots, k$$

and $b(n_k + 1) = 1$. 
Then the steady state equations are,

\[
(\lambda + \sum_{j=1}^{k} \sigma(n_j) a(n_j) \mu_j) P(n_1, \ldots, n_k) = \sum_{j=1}^{k} \sigma(n_j + 1) a(n_j + 1) \mu_j P(n_1, \ldots, n_j + 1, n_{j+1}, \ldots, n_k) + \lambda P(n_1 - 1, n_2, \ldots, n_k),
\]

where all P's containing a negative value of n are automatically taken to be zero. The last term in the summation on the right hand side is

\[
\sigma(n_k + 1) a(n_k + 1) \mu_k P(n_1, n_2, \ldots, n_{k-1}, n_k + 1).
\]

It can be shown that \( P(n_1, n_2, \ldots, n_k) = P(0) \prod_{j=1}^{k} b(n_j) \) is a solution of the set of equations, where

\[
b(n_j) = \frac{1}{n_j!} \left( \frac{\lambda}{\mu_j} \right)^{n_j}, \quad n_j < r_j,
\]

\[
= \frac{1}{r_j^n} \left( \frac{\lambda}{\mu_j} \right)^{r_j} \left( \frac{\lambda}{\mu_j - r_j} \right)^{n_j - r_j}, \quad n_j \geq r_j.
\]

\( P(0, \ldots, 0) = P(0) \) can be found by utilising the usual normalising equation, \( \sum_{j=1}^{\infty} P(n_1, \ldots, n_k) = 1 \), in the following manner.

The \( b(n_j) \) are all positive and \( \sum_{n_j=0}^{\infty} b(n_j) \) is a convergent series, equal to
\[
\frac{r_j}{r_j} \sum_{n_j=0}^{\infty} \left( \frac{\lambda}{\mu_j r_j} \right)^n_j = \frac{r_j}{r_j} \frac{1}{1 - \left( \frac{\lambda}{\mu_j r_j} \right)}
\]

where \( \frac{\lambda}{\mu_j r_j} < 1 \).

The positivity of the terms ensures that

\[
\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_k=0}^{\infty} \left( \frac{1}{\prod_{j=1}^{k} \mu_j n_j} \right) = \prod_{j=1}^{k} \sum_{n_j=0}^{\infty} \frac{1}{\mu_j n_j} = \prod_{j=1}^{k} A_j^{-1}.
\]

and so, on writing \( \sum_{n_j=0}^{\infty} b(n_j) = A_j \), \( j = 1, \ldots, k \), it follows that

\[
P(o) = \prod_{j=1}^{k} A_j^{-1}.
\]

The solution we have given can be verified by substitution. It remains to show that it is a unique solution.

Proof of Uniqueness of solution

Consider the associated Markov process generated by the states \((n_1, \ldots, n_k)\) of the real system, but with the transitions now supposed to occur at discrete times separated by time intervals, all of length \( h \).

If we choose the value of \( h \) such that the quantities \( \lambda h, 1 - \lambda h, k_i \mu_i h, 1 - k_i \mu_i h; (k_i < r_i) \) lie between \( 0 \) and \( 1 \), they can be regarded as transition probabilities. This process obviously differs from the true process.

If we let \( A \) be the transition matrix of the associated process, we can show that it is reducible, by the following argument. It is clearly possible to reach the null state from any other state, merely
by successive departures without arrivals, for instance. The probability of this occurring is clearly positive, as it is the product of such numbers as \((1-\lambda h), k_i \mu_i h, (1 - k_i \mu_i h)\). Similarly, it is possible to reach any other state from the null state, again with positive probability. Combining these two procedures, it is apparent that we can reach any state from any other state, with positive probability. Thus, the matrix \(A\) is irreducible (Feller [5], p. 318), and hence a unique solution exists to the equation \(x A = x\), where \(x\) is the row vector of state probabilities. In other words, there exists a unique solution to the equation \(x (A - I) = 0\).

The diagonal elements of \(A\) are of the form \(1 - a_n h\) where \(a_n\) is a function of \((\lambda; \mu_1, \ldots, \mu_k; k_1, \ldots, k_k)\) whilst all other elements have the factor \(h\). Thus, we may write \((A - I)/h = B\), and it follows that there exists a unique solution to \(y B = 0\).

This is the required result, for \(B\) is, in fact, the matrix of the steady state equations of the true process.

**Distributions for any stage**

Let the marginal probability that there are \(n\) customers at stage \(j\) of service, i.e. in queue \(j\), be given by \(P_j(n)\), where

\[
P_j(n) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_{j-1}=0}^{\infty} \sum_{n_{j+1}=0}^{\infty} \sum_{n_k=0}^{\infty} P(n_1, n_2, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k)\]

\(n, n_{j+1}, \ldots, n_k\)
In the same way as we evaluated \( P(o) \) it can be shown that 
\[ P_j(n) = n_j^{-1/n_j} = n_j^{-1/A_j}. \]
Thus, the probability distribution of customers present in stage \( j \) is dependent only on \( \mu_j, r_j, \lambda \), and is the same as that obtained from the system with one queue; \( r_j \) servers; random input and a negative-exponential distribution of service times.

In the case where we have a single server system, we obtain

\[
(3.4.1) \quad P(n_1, \ldots, n_k) = P(o) \sum_{j=1}^{\infty} \frac{n_j^k}{\mu_j^k} \frac{\lambda^k}{\mu_j^k} 
\]

where \( P(o) = \prod_{j=1}^{k} \left( 1 - \lambda / \mu_j \right) \).

We note that \( (3.4.1) \) factorises, so that the distributions \( P_j(n), (j = 1, 2, \ldots, k) \), are independent and we may write immediately, putting \( x_j = \lambda / \mu_j \).

\[ P_j(n) = n_j^k (1 - x_j). \]

Thus, the average number of customers in stage \( j \) is

\[
\sum_{n=0}^{\infty} n x_j^n (1 - x_j) = x_j (1 - x_j) \sum_{n=1}^{\infty} n x_j^{n-1} 
\]

\[
= x_j (1 - x_j) (1 - x_j)^{-2} 
= x_j (1 - x_j)^{-1}.
\]

The average number of customers being services is equal to the probability of one or more customers being in the stage.
\[ = \sum_{n=1}^{\infty} x_j^n (1 - x_j) \]

(3.4.3)

\[ = x_j. \]

Thus, from (3.4.2) - (3.4.3), the average number of customers waiting in line is

(3.4.4)

\[ x_j^2 (1 - x_j)^{-1}. \]

To obtain the average values for the system as a whole, we must sum the averages for the individual stages.

Let us now consider a customer arriving at stage \( j \) to find \( n \) customers already there. This event has probability \( x_j^n (1 - x_j) \) of occurring. Now, knowing that the distribution of service times is negative exponential with mean \( 1/\mu_j \), we have,

\[ P(1 \text{ departure in } \xi, \xi + d\xi) = \mu_j e^{-\mu_j \xi} d\xi. \]

Therefore,

\[ P(\text{second departure is in } \xi, \xi + d\xi) = d\xi \int_{\xi}^{\infty} \mu_j e^{-\mu_j x} - \mu_j (\xi - x) dx = \xi e^{-\mu_j \xi} d\xi. \]

At this point let us assume that the probability of departure number \( n-1 \) occurring in the interval \( \xi, \xi + d\xi \) is,

\[ \frac{\mu_j^{n-1} \xi^{n-2}}{(n-2)!} e^{-\mu_j \xi} d\xi. \]
Then the probability of departure number \( n \) occurring in this interval is given by

\[
\frac{d\xi}{\xi} \int_0^\xi \frac{\mu_j^{n-1} x^{n-2}}{(n-2)!} e^{-\mu_j x} \mu_j e^{-\mu_j (\xi-x)} \, dx = \frac{\mu_j^n \xi^{n-1}}{(n-1)!} e^{-\mu_j \xi} d\xi
\]

which is true for \( n = 1, 2, \ldots \) by induction.

Hence, the waiting time distribution is given by the probability density function,

\[
(3.4.5) \quad r_j(\xi) = \sum_{n=1}^{\infty} (1 - x_j) \frac{x^n \mu_j^n \xi^{n-1}}{(n-1)!} e^{-\mu_j \xi}
\]

and the probability of not writing at all is,

\[
(3.4.6) \quad (1 - x_j).
\]

The conditional probability density function of waiting time, given that there is some waiting, is \((\mu_j - \lambda) e^{-\mu_j \xi} \), \(-(\mu_j - \lambda)\).

In addition, it is easy to show that

a) the mean waiting time is \( x_j / \{\mu_j (1 - x_j)\} \),

b) given that the mean service time is \( 1/\mu_j \), a measure of efficiency defined to be \( \frac{\text{mean waiting time}}{\text{mean service time}} = x_j / (1 - x_j) \),

and that

c) the average time spent waiting in the system as a whole is \( \sum_j \frac{\lambda}{\mu_j (\mu_j - \lambda)} \), whilst the average service time is \( \sum_j \frac{1}{\mu_j} \).
When the service times are all equal the system is somewhat simplified, and it is easily seen that

\[ P(n) = \sum_{n_1 + \ldots + n_k = n} \binom{n}{\frac{\lambda}{\mu}} P(0) \]

\[ = \binom{n+k-1}{n} \left( \frac{\lambda}{\mu} \right)^n (1 - \frac{\lambda}{\mu})^k. \]

3.5. A network of queues.

Let us consider now a more complicated arrangement, as discussed by J. R. Jackson /10/, in which we have, once again, a number of queues. However, in this case, we assume that a customer finishing service at queue \( k \) transfers to queue \( j \) according to a fixed probability associated with the queue it is leaving. Call this probability \( \theta_{jk} \). In addition, customers arrive randomly from outside the system and enter queue \( j \) at a mean rate \( \lambda_j \).

Specifically, our assumptions are these:

Let there be stages \( 1,2,\ldots,M \) (each stage is a queue). Then, for \( m = 1, 2, \ldots, M \),

1) Stage \( m \) contains \( r_m \) servers.

2) Customers from outside the system arrive in stage \( m \) in a Poisson type series at a mean rate \( \lambda_m \).

3) Customers arriving in stage \( m \), from inside or outside the system, are served on a first come-first served basis with an exponential distribution of service times mean \( 1/\mu_m \).
Once served in stage $m$, a customer goes instantaneously to stage $k$, $(k = 1, 2, \ldots, M)$, with probability $\theta_{km}$. He leaves the system with probability $1 - \sum_k \theta_{km}$. It is on the basis of assumption (4) that this system is termed a 'network' of waiting lines.

For $m = 1, 2, \ldots, M$, let $\Gamma_m$ be the mean arrival rate of customers at stage $m$ from any source. It is obvious that in a steady state we must have

\[(3.5.1) \quad \Gamma_m = \lambda_m + \sum_k \theta_{mk} \Gamma_k.\]

Let $n_m$ denote the number of customers waiting or in service in stage $m$, and define the state of the system to be the vector $(n_1, \ldots, n_M)$.

Define $P_m(n)$, $(m = 1, 2, \ldots, M; n = 1, 2, \ldots)$ by the following equations.

\[(3.5.2) \quad P_m(n) = P_m(0) \left(\frac{\Gamma_m}{\mu_m}\right)^n n^! / n! , \quad n = 0, \ldots, \Gamma_m,\]

\[P_m(n) = P_m(0) \left(\frac{\Gamma_m}{\mu_m}\right)^{\Gamma_m} \gamma_{\Gamma_m}! \gamma_{\Gamma_m}^{n-\Gamma_m} , \quad n \geq \Gamma_m ,\]

where the $P_m(0)$ are determined by the equations $\sum_{n=0}^{\infty} P_m(n) = 1$.

We then have:

**Theorem.**

A steady state distribution of the state of the present system is given by the products.

\[P(n_1, \ldots, n_M) = P_1(n_1) P_2(n_2) \ldots P_M(n_M) ,\]

Provided $\Gamma_m < \mu_m r_m$ for $m = 1, 2, \ldots, M$, (the usual steady state condition).
Proof.

The general approach of Feller \[\text{(57)}\] gives us the steady state equations,

\[
\left( \sum_{i} \lambda_i + \sum_{i} \alpha_i(n_i) \mu_i \right) P(n_1, \ldots, n_M) \\
= \sum_{i} \alpha_i(n_{i+1}) \mu_i \Phi^*_i P(n_1, \ldots, n_{i+1}, \ldots, n_M) \\
+ \sum_{i} \lambda_i \xi_i P(n_1, \ldots, n_i-1, \ldots, n_M) \\
+ \sum_{i} \sum_{j} \alpha_j(n_{j+1}) \mu_j \Theta_{ij} P(n_1, \ldots, n_j+1, \ldots, n_i-1, \ldots, n_M),
\]

where \( \alpha_i(n_i) = \min(n_i, r_i) \),

\( \Phi^*_i = 1 - \sum_k \Phi_{ki} \),

\( \xi_i = \min(n_i, 1) \).

The theorem can be proved by verifying that the given probabilities satisfy the steady-state equations. Thus, we have shown that at least insofar as the steady states are concerned, the system we are discussing behaves as if each of its stages was an independent system. This conclusion is hardly surprising in view of the earlier work of Burke \[\text{(57)}\], and Reich \[\text{(177)}\].

Koenigsburg \[\text{(177)}\] independently discusses a cyclic queueing system, i.e. one in which operations are carried out in a fixed cycle, of \( m \) stages. This, of course, is just a special case of the above in
which $\lambda_i = 0$ for all $i$, and

$$\theta_{jk} = 1, \quad j = k+1 \pmod m$$

$$\theta_{jk} = 0, \quad \text{otherwise}.$$

Koenigsburg's steady state distributions are the appropriate special cases of Jackson's results, and his calculation of the amount of time each server is busy (though in his case they are machines), and the output of each machine (obviously equal) are covered by the work of (3,4). He gives tables of mean numbers of customers being served and waiting; the utilisation factor of the machines and the time to complete a full circuit, as well as discussing the application of his results in the coal-mining industry.
CHAPTER FOUR
THE EFFECT OF LIMITING THE SIZE OF THE WAITING ROOM

4.1. Introduction.

We have assumed, hitherto, that at any stage in a network, we can have an infinite queue length. However, in most practical cases, owing to lack of storage space or inventory costs, this assumption is unrealistic. In the case where we have a finite queue, a new customer who, on arrival, finds the queue is already full, may depart, never to return, or he may stop in the previous stage and thus cause a blockage at that particular point, stopping production there. The effects of this limitation of the queue size have been discussed in the literature, notably by P. D. Finch [6] and G. C. Hunt [7].

4.2. A waiting room of size $N$.

Suppose that we have a waiting room which will hold $N$ customers, excluding the one being served, and that if a customer arrives to find the room full, he leaves, never to return. We suppose that the arrivals are random, with mean rate $\lambda$ and we let the distribution function of the service time of customer number $r$, $s_r$ say, be $B(s)$. Then $k_r$, the probability of $r$ arrivals in a service time, is given by

$$k_r = \frac{1}{r!} \int_0^\infty e^{-\lambda s} (\lambda s)^r d B(s).$$
Let $p_n^r$ denote the probability that customer number $r$ leaves $n$ customers behind him on completion of his service. Then,

$$p_n^r = 0 \quad \text{if } n > N$$

and,

$$(4.2.2) \begin{cases} p_{o}^{r+1} = (p_1^r + p_0^r) k_o \\ p_{n}^{r+1} = p_{n+1}^r k_o + p_{n}^r k_1 + \dotsc + p_2^r k_{n-1} + (p_1^r + p_0^r) k_n, \\ (n < N) \\ p_{N}^{r+1} = p_N^r k_1 + p_{N-1}^r k_2 + \dotsc + p_2^r k_{N-1} + (p_1^r + p_0^r) k_N, \end{cases}$$

where $K_j = k_j + k_{j+1} + \dotsc$

The existence of a limiting distribution follows from the fact that the process is a finite, irreducible, aperiodic Markov chain.

The equations which determine the limiting distribution are

$$(4.2.3) \begin{cases} P_n = P_{n+1} k_o + P_{n} k_1 + \dotsc + P_2 k_{n-1} + (p_1^r + p_0^r) k_n, \quad (n = 0, 1, \dotsc, N-1), \\ P_N = P_N^r k_1 + P_{N-1}^r k_2 + \dotsc + P_2 k_{N-1} + (p_1^r + p_0^r) k_N. \end{cases}$$

Write $P(x) = P_0 + P_1 x + P_2 x^2 + \dotsc + P_N x^{N-1}$,

$$k(x) = k_o + k_1 x + \dotsc + k_r x^r + \dotsc$$

Thus,
\[ P(x)k(x) = (P_0 + P_1)k_0 + \int (P_0 + P_1)k_1 + P_2k_0 \int x \]
\[ + \int (P_0 + P_1)k_2 + P_2k_1 + P_3k_0 \int x^2 + \ldots \]
\[ + \int (P_0 + P_1)k_{N-1} + \ldots + P_Nk_0 \int x^{N-1} \]
\[ + \int (P_0 + P_1)k_{N+j} + \ldots + P_Nk_{j+1} \int x^{N+j} , \]
\[ j = 0, 1, 2, \ldots \]

On substituting from (4.2.3), we have

\[ P(x)k(x) = P_0 + P_1x + P_2x^2 + \ldots + P_Nx^{N-1} + P_Nx^N \]
\[ + \int P_N - P_N + P_{N+1}x + \ldots + P_N'x^j + \ldots \int x^N , \]

where \( P_N' = \int (P_0 + P_1)k_{N+j} + \ldots + P_Nk_{j+1} \int . \)

Hence,

\[ P(x)k(x) = P_0(1 - x) + xP(x) + x^N \int P_N - P_N + \sum_{j=1}^{\infty} P_{N+j}x^j \int . \]

and so,

\[ (4.2.4) \quad P(x) = \frac{P_0(1 - x)}{k(x) - x} + \frac{x^N \int P_N - P_N + \sum_{j=1}^{\infty} P_{N+j}x^j \int }{k(x) - x} \]

If \( (1 - x)/(k(x) - x) \) can be expanded as a power series in \( x \) for some suitable region of convergence, only the first term in (4.2.4) contributes to the coefficient of \( x^n \) for \( n < N \). The \( P_n \) can then be determined in terms of \( P_0' \), which in turn can be found from the normalizing equation \( \sum_{j=1}^{N} P_j = 1 \).
For example, in the case when the service time has a negative exponential distribution with mean value $1/\mu$, $k_x = \alpha^x \beta$ where $\alpha = \lambda/(\lambda + \mu)$, $\beta = \mu/(\lambda + \mu)$. We then have,

$$k(x) = \beta(1 + \alpha x + \alpha^2 x^2 + \ldots)$$

$$= \frac{\beta}{1-\alpha x} .$$

Thus,

$$\frac{1 - x}{k(x) - x} = \frac{(1 - x)(1 - \alpha x)}{\beta - x + \alpha x^2} ,$$

and we find

$$\begin{align*}
(4.2.5) \quad P_n &= \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+1}} \cdot \left(\frac{\lambda}{\mu}\right)^n , \quad (n = 0, 1, \ldots, N; \lambda \neq \mu) , \\
&= \frac{1}{N+1} , \quad (n = 0, 1, \ldots, N; \lambda = \mu) .
\end{align*}$$

Therefore,

$$\begin{align*}
(4.2.6) \quad \lim_{N \to \infty} P_n &= \begin{cases} 
\left(1 - \frac{\lambda}{\mu}\right)^n, & \lambda < \mu \\
0, & \lambda = \mu
\end{cases}.
\end{align*}$$

These values agree with the well known solutions for the familiar case $N = \infty$.

However,

$$\begin{align*}
(4.2.7) \quad \lim_{N \to \infty} P_n &= \begin{cases} 
1 - \frac{\mu}{\lambda}, & \text{if } \lambda > \mu , \\
0, & \text{if } \lambda \leq \mu ,
\end{cases}
\end{align*}$$

which elucidates the difference between the limiting solutions in the
two cases $\lambda \leq \mu$, $\lambda > \mu$.

If we confine ourselves to the usual case where both the input process and the service mechanism have a negative exponential distribution function, it is easy to find the distribution function for the waiting time of a customer. Denoting by $Q_n$ the probability that a customer arrives to find $n$ customers before him, including the one (if any) being served, and by $\ell_r$ the probability that $r$ customers leave the system between consecutive arrivals, we have, for $r = 0, 1, \ldots, N$,

$$Q_{n+1-r} = Q_{n-r} \ell_0 + Q_{n-r+1} \ell_1 + \cdots + Q_{n-1-r} \ell_{r-1} + (Q_N + Q_{N+1}) \ell_r$$

We can determine the $Q_n$ in a similar manner to the $P_n$ above and we find that $Q_{n+1-r}$ is the coefficient of $x^{r-1}$ in the expansion of $Q_{n+1}(1-x)/\{\ell(x) - x\}$, where $\ell(x) = \ell_0 + \ell_1 x + \ell_2 x^2 + \cdots$. In the case of random arrivals we find,

$$Q_n = \begin{cases} \frac{1 - \mu/\lambda}{1 - (\mu/\lambda)^{N+2}}, & \lambda \neq \mu, \\ \frac{1}{N+2}, & \lambda = \mu. \end{cases} \quad (4.2.3)$$

The first equation of (4.2.3) can be rearranged to give,

$$Q_n = \frac{(\mu/\lambda)^{N+1-n}}{1 + \mu/\lambda + \cdots + (\mu/\lambda)^{N+1}} = \frac{(\lambda/\mu)^n}{1 + \lambda/\mu + \cdots + (\lambda/\mu)^{N+1}} = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+2}} \cdot (\lambda/\mu)^n.$$
Thus,
\[
\lim_{N \to \infty} Q_n = \lim_{N \to \infty} P_n ,
\]
a fact which we have assumed implicitly in 2.2.

If a customer finds \( m \) customers before him on arrival \((m < N+1)\) then, he waits a time \( s_1 + \ldots + s_m \), where \( s_j \) is a random variable with negative exponential distribution. If \( S_n(x) \) is the distribution function of the sum of \( n \) service times the distribution function for waiting times, \( W_N(x) \), is given for \( x > 0 \) by
\[
W_N(x) = Q_0 + Q_1 S_1(x) + \ldots + Q_n S_n(x) + Q_{N+1} S_{N+1}(x).
\]
We have, easily,
\[
S_n(x) = 1 - e^{-\mu x} \int 1 + \mu x + (\lambda x)^2/2! + \ldots + (\mu x)^{n-1}/(n-1)! \, dx.
\]
Therefore, for \( \lambda \neq \mu \),
\[
W_N(x) = 1 - \frac{(\lambda/\mu) e^{-\mu x}}{1 - (\lambda/\mu)^{N+2}} \cdot \int 1 + \lambda x + \ldots + \frac{(\lambda x)^N}{N!} \cdot \frac{(\lambda)^{N+1}}{\mu} \left\{ 1 + \mu x + \ldots + \frac{(\mu x)^N}{N!} \right\} \, dx.
\]
Let us denote the mean waiting time for a waiting room of size \( N \) by \( E_N(w) \). We have then that for \( \lambda \neq \mu \),
\[
E_N(w) = \frac{(\lambda/\mu)}{\mu - \lambda} \int \frac{1 - (N+2)(\lambda/\mu)^{N+1} + (N+1)(\lambda/\mu)^{N+2}}{1 - (\lambda/\mu)^{N+2}} \, dx.
\]
and, for $\lambda = \mu$,

$$E_n(w) = \frac{1}{2\mu} \cdot (N+1).$$

In the case $\lambda < \mu$ and $N = \infty$, when the limiting distribution exists, we have

$$E_\infty(w) = \frac{\lambda^N}{\mu^{N-\lambda}},$$ as before.

It was proved earlier that if we have

(i) an infinite waiting room,

(ii) a Poisson input process, parameter $\lambda$,

(iii) a distribution function $B(x)$ for service time which possesses a continuous 2nd order derivative,

(iv) $\lambda E/\bar{s}/ < 1$

then

a) The queue size left by a departing customer is independent (in the limit) of the duration of the interval since the previous departure.

b) Two successive departure intervals are independent (in the limit).

It is of interest to see that these two properties do not hold when we have a waiting room of size $N < \infty$ and $\lambda E/\bar{s}/ < \infty$. For, we proved in 2.5 that the queue size is independent of the interdeparture interval if, and only if,

$$E(t) = 1 - e^{-\lambda t/(1-P(0))},$$
and we have just shown (4.2.5), that, when \( N < \infty \),

\[
P(0) = \frac{1 - \rho}{1 - \rho^{N+1}},
\]

where \( \rho = \lambda/\mu = \lambda \mathbb{E}[s] / \) is the traffic intensity. But (4.2.9) implies that \( \mathbb{E}[s] = (1 - P(0))/\lambda \), i.e. that \( P(0) = 1 - \lambda \mathbb{E}[s] = 1 - \rho \).

Thus, for \( N < \infty \), it cannot be true that the queue size is independent of the interdeparture interval.

We also showed that (b) is the case only when \( N = \infty \) and the service time distribution was negative exponential. Thus, (b) does not hold when \( N < \infty \).

To sum up, when we have a waiting room of size \( N < \infty \), the queue size is not independent of the duration of the interval since the previous departure, nor are two successive departure intervals independent.

4.3. Several special cases.

Statement of the cases.

G. C. Hunt [9] considers a sequence of single-server queues with exponential service times and four different types of waiting room where,

(i) infinite queues are allowed in front of each server,

(ii) no queues are allowed in front of any server, except the first which may have an infinite queue,

(iii) finite queues are allowed in front of each server, with the exception of the first which may have an infinite queue, and

(iv) no queues and no vacant facilities are allowed, with the exception of the queue before the first server which may be
infinite, the line moves all at once, as a unit.

Case (1).

This we have already discussed extensively and so we will just reiterate some of the results.

1) The steady state exists if the mean service rate is greater than the mean arrival rate.

2) Assuming Poisson arrivals at the first stage, which with (1) we recall, ensures Poisson arrivals at all subsequent stages, we have that the steady state solution for stage \( j \) is given by,

\[
P_n = (1 - \rho) \rho^n,
\]

where \( \rho = \lambda/\mu_j \) is the utilisation factor or traffic intensity, \( \lambda \) the mean arrival rate and \( \mu_j \) the mean service rate of stage \( j \).

3) We can define the mean number of units in the system as

\[
L = \sum_{j} \sum_{n=0}^{\infty} n(1 - \rho_j)\rho_j^n = \sum_{j} \rho_j(1 - \rho_j)^{-1}.
\]

Case 2. No queues allowed.

For the two stage problem, in this case, we can define the following time dependent state probabilities.

\( p_1(n) = p(n, 0, 0) \): (n-1) units in the initial queue and one unit in service at the first stage (unless \( n = 0 \), which means the system is empty),
\[ \mathbf{p}_2(n) = p(n, o, l): \text{(n-1) units in the initial queue, one unit in service in the first stage, and one unit in service in the second (unless n=0, when we have just one unit, in the second stage)}, \]

\[ \mathbf{p}_3(n) = p(n, l, l): \text{n units in the initial queue, l unit in the first stage with service completed, and one unit in service in the second stage}, \]

with the corresponding steady state probabilities, \( p(n, o, o) \), \( p(n, o, l) \), \( p(n, l, l) \). We also define three new state probabilities,

\[ x_i = \sum_{n=0}^{\infty} p_i(n) , \quad (i = 1, 2, 3) , \]

These states are obviously unchanged by new arrivals. The only transitions between these states occur when some customer completes his service. The equations for these \( x_i \), following Feller [57], are given by

\[
\begin{align*}
\dot{x}_1 &= -\mu x_1 + \mu_2 x_2 + \mu_1 p(o,o,o) , \\
\dot{x}_2 &= -(\mu_1 + \mu_2) x_2 + \mu_1 x_1 + \mu_2 x_3 - \mu_1 p(o,o,o) + \mu_1 p(o,o,1) , \\
\dot{x}_3 &= -\mu_2 x_3 + \mu_1 x_2 - \mu_1 p(o,o,1) ,
\end{align*}
\]

where \( \mu_1, \mu_2 \) are the mean service rates of the first and second stages respectively.

Let \( \mu_1 = \mu_2 \) for the moment. Then, for the steady state, we have:-
\[ -x_1 + x_2 + P(o,o,o) = 0 \]

\[(4.3.2) \quad x_1 - 2x_2 + x_3 - P(o,o,o) + P(o,o,1) = 0 \]

\[ x_2 - x_3 - P(o,o,1) = 0 \]

As the utilisation increases, \( P(o,o,o) \) and \( P(o,o,1) \) become smaller and smaller and eventually can be neglected. In the limit,

\[ x_1 = x_2 = x_3 \]

Thus, the fraction of time that the first stage is free to service incoming units equals the fraction of time it is unblocked, which, in the limit, is

\[ \frac{x_1 + x_2}{x_1 + x_2 + x_3} = \frac{2}{3} \]

Thus, the maximum effective service rate of the first stage is \( 2\mu/3 \) and its maximum possible utilisation is \( 2/3 \).

It is obvious that, in the general \( N \)-stage system with no queues, blocking monotonically increases as the number of the stage decreases, and the blocking is greater in the first stage than in any other. Hence, the maximum utilisation of the first stage is the maximum utilisation of the system as a whole.

The general steady state equations can be obtained, by the usual method, from which we can deduce the mean number of units in the system, which is used for comparison purposes, as

\[ L = \sum_{n=0}^{\infty} \left( n P_1(n) + (n+1) P_2(n) + (n+2) P_3(n) \right) L \]
Case 3. Finite queues allowed.

If we have a waiting room of size \( N \) in front of the second stage, by defining \( N+3 \) state probabilities as above, we obtain by a similar method the maximum possible utilisation for different service rates \( \mu_1, \mu_2 \) as

\[
\rho_{\text{max}} = \frac{\mu_2 (\mu_1^{N+2} - \mu_2^{N+2})}{\mu_1^{N+3} - \mu_2^{N+3}},
\]

which reduces to

\[
\rho_{\text{max}} = \frac{N + 2}{N + 3},
\]

for the same service rate in both stages.

Note that \( \rho_{\text{max}} \to \frac{\mu_2}{\mu_1} \) as \( N \to \infty \), for \( \mu_1 > \mu_2 \),

as is well known.

Case 4. No queues, no vacancies.

This case can be treated exactly as the two previous ones. However, there is an easier method.

For the steady state operation, the mean throughput rate must be greater than the mean arrival rate. Thus, if \( \tau_N \) is the mean throughput time for \( N \) stages, the maximum possible utilisation is given by \( \rho_{\text{max}} = (\mu_2 \tau_N)^{-1} \). Now, let \( R_N(t) \) be the probability that service in all \( N \) stages has been completed in a time \( t \) after the line moves. Then,

\[
\tau_N = \int_0^\infty t \cdot R_N(t) \, dt.
\]
We must now find an expression for $R_n(t)$. For two stages, $R_2(t)$ is equal to the probability that service is completed in the first stage in a time less than $t$ multiplied by the probability that the service is completed in the second stage between $t$ and $t + dt$, plus a similar quantity for the second stage. Thus, the mean throughput time for two stages with equal service rates is

$$\tau_2 = 2\int_0^\infty \mu e^{-\mu t}(1 - e^{-\mu t}) t \, dt = \frac{3}{2\mu}.$$

We can therefore deduce that the maximum possible utilisation is $\rho_{\text{max}} = \frac{2}{3}$.

Hunt then gives a table of $L$ for cases 1, 2 and 4 for two stages with equal service rates in which $L$ case 1 < $L$ case 2 < $L$ case 3 for different $\rho$'s, and also a table for the calculated values of $\rho_{\text{max}}$ for different numbers of stages. He also showed that

$$\rho_{\text{max}_1} > \rho_{\text{max}_2} > \rho_{\text{max}_3} > \rho_{\text{max}_4}.$$
CHAPTER FIVE
FEEDBACK

5.1. Introduction.

Two particular cases of J. R. Jackson's \( \sqrt{10} \) network model are worthy of special consideration and have received such from P. D. Finch \( \sqrt{7} \). They are both cases of 'feedback', namely terminal and single-stage feedback.

Feedback occurs when a customer at some stage in a sequence of queues completes his service at that stage and then rejoins the queue at some previous stage or, possibly, the same one.

Let there be \( m \) stages, each with a single server. With terminal feedback, there is feedback only from the \( m \)-th server to one of the other servers with a finite probability \( p_j (j = 1,\ldots, m) \), where \( p = \sum_{j=1}^{m} p_j \) and \( q = 1 - p \) is the probability (supposed non-zero) of the customer not returning to the system. In the second case, that of single stage feedback, the customer, upon completion of his service at stage \( j \) rejoins the queue at that stage with probability \( p_j \) (not the same \( p_j \) as in the terminal feedback) where \( 0 < p_j < 1 \) for \( j = 1,\ldots, m \). The \( p_j \) are supposed to be independent of the state of the system at the moment of return and of the customer who has just completed service. In each case, customers are assumed to arrive at the first stage at random, with a mean rate \( \lambda \). The service time of server \( j \) is further assumed to be negative exponential with
mean value $1/\mu_j$. No assumptions are made regarding the point of
the queue at which a returning customer joins it as we are interested
only in the number of customers in the different stages.

The constraint is imposed on the system that there must be no
more than $N$ customers present. A new customer, arriving to find
$N$ already in the system, departs, never to return. The existence of
a limiting distribution for the number of customers present follows
because the process is an irreducible, aperiodic Markov chain.

5.2. Terminal feedback.

Write $P(n_1, \ldots, n_m)$ for the steady state probability of having
$n_j$ customers in stage $j$, including the one (if any) then being
served, subject to the condition $\sum_{j=1}^{m} n_j \leq N$. We then have, by the
usual method, the steady state equations.

Writing,

$$
e_j = \begin{cases} 1, & n_j > 0, \\ 0, & n_j = 0, \end{cases}$$

$$k = \begin{cases} 1, & \sum_{j=1}^{m} n_j < N, \\ 0, & \sum_{j=1}^{m} n_j = N, \end{cases}$$

these are:
\[ o = - \left( \lambda k + \sum_{j=1}^{m} \mu_j \epsilon_j \right) P(n_1, \ldots, n_m) + \lambda \epsilon_1 P(n_1-1, n_2, \ldots, n_m) \]
\[ + \sum_{j=1}^{m-1} \mu_j \epsilon_j p_j P(n_1, n_2, \ldots, n_{j-1}, n_{j-1}, n_{j+1}, \ldots, n_{m-1}, n_{m+1}) \]
\[ + \mu_m p_m \epsilon_m P(n_1, \ldots, n_m) + \sum_{j=1}^{m-1} \epsilon_j \mu_j P(n_1, \ldots, n_{j-1}, n_{j+1}, n_{j+1}, \ldots, n_m) \]
\[ + (1 - p) \mu_m k P(n_1, n_2, \ldots, n_{m-1}, n_m + 1). \]

It can be verified by direct substitution, that this set of equations is satisfied by

\[ P(n_1, \ldots, n_m) = (x_1)^{n_1}(x_2)^{n_2} \cdots (x_m)^{n_m} P(o, o, \ldots, o) \]

where \( x_j = \lambda(1 - p + p_1 + p_2 + \cdots + p_j)/\mu_j(1 - p), (j = 1, 2, \ldots, m), \)

and \( p = \sum_{j=1}^{m} p_j. \)

\( P(o, o, \ldots, o) \) can be determined by normalisation.

The uniqueness of the solution can be demonstrated in a manner similar to that of R. R. P. Jackson, explained in section 3.3.

Indeed, when each \( p_j \) vanishes, each \( \lambda/\mu_j < 1 \), and \( N = \infty \). Then this is, exactly, Jackson's solution.

If each \( x_j < 1 \), \( P(o, \ldots, o) \) tends to a finite non-zero limit as \( N \to \infty \), but when any \( x_j > 1 \) then \( P(o, \ldots, o) \), and hence \( P(n_1, \ldots, n_m) \) for any \( \{n_j\} \), tends to zero as \( N \to \infty \).
5.3. Single stage feedback.

In this case, the steady state equations are:

\[
o = - \left( \lambda K + \mu_1 q_1 e_1 + \cdots + \mu_m q_m e_m \right) P(n_1, \ldots, n_m)
+ \lambda e_1 P(n_1 - 1, n_2, \ldots, n_m)
+ \sum_{j=1}^{m-1} \mu_j q_j e_{j+1} P(n_1, \ldots, n_{j+1}, n_{j+1} - 1, \ldots, n_m)
+ \mu_m q_m K P(n_1, n_2, \ldots, n_m + 1),
\]

where \( \sum_{j=1}^{n} n_j \leq N \) and \( q_j = 1 - p_j \).

It can be shown, as before, that \( P(n_1, \ldots, n_m) = (x_1)^{n_1} \cdots (x_m)^{n_m} \times P(o, \ldots, o) \), where \( x_j = \lambda/\mu_j q_j \), is the unique solution to these equations. \( P(o, \ldots, o) \) is again given by normalisation.

As in the previous section, it is necessary and sufficient, for \( P(n_1, \ldots, n_m) \) to tend to a finite non-zero limit as \( N \) tends to infinity, that each \( x_j \) be less than unity.
CHAPTER SIX

APPLICATIONS

6.1. Introduction.

It is reasonable to ask, "What practical situations can be met by applications of the theory in this account?"

The answer is that there are many and varied industrial and mechanical systems which satisfy the limitations of this theory or yield to slight variations in the application of the theory.

For instance, Koenigsburg [147] developed his theory whilst working on the organisation of coal-mining equipment and personnel; R. R. P. Jackson [117] was concerned with a study of the operations of a machine shop repairing aircraft engines; likewise J. R. Jackson [107] also received his stimulus whilst studying machine shops, in research for the British Navy and Burke [137] uses as an example the derivation of the optimum number of sales-clerks and cashiers in a shop. Further applications by Koenigsburg [147], Karush [127] and Benson and Gregory [27] are discussed, at length, and the solution of two common problems is deduced.

6.2. The spare and standby machine problem.

Koenigsburg [147] discusses what may be termed "the spare and standby machine problem" in which we have a closed system with two stages, one a repair and the other a working stage. There are N machines, of which a maximum of A can be working at any time (A
operators), and $M$ repairmen. Breakdowns are assumed to occur randomly with a mean rate $\mu_2$ whereupon the machines enter the repair stage, each repairman having an exponential distribution of service times with mean rate $\mu_1$. The numbers $N, A, M$ and $\mu_2/\mu_1$ completely determine the output of the system.

The problem, of course, is just the simplest example of a multi-server cyclic queue. Denoting by $P(n_1, n_2)$ the steady state probability of the state $(n_1, n_2)$, where $n_1$ and $n_2$ are, respectively, the number of machines in stages one and two ('repair' and 'service and waiting'), we obtain, by the usual procedure, the steady state solution

$$(6.2.1) \begin{cases}
P(n_1, N-n_1) &= (\frac{n_1}{n_1}) (\frac{\mu_2}{\mu_1})^{n_1} P(o, N), \quad (n_1 < M \leq N-A), \\
P(n_1, N-n_1) &= (\frac{n_1}{M!M^{-1}}) (\frac{\mu_2}{\mu_1})^{n_1-M} P(o, N), \quad (M \leq n_1 \leq N-A), \\
P(n_1, N-n_1) &= (\frac{n_1}{N-A!/(N-n_1)!M!M^{-1}}) (\frac{\mu_2}{\mu_1})^{n_1-M} P(o, N), \quad (N-A < n_1).
\end{cases}$$

or

$$(6.2.2) \begin{cases}
P(n_1, N-n_1) &= (\frac{n_1}{n_1}) (\frac{\mu_2}{\mu_1})^{n_1} P(o, N), \quad (n_1 \leq N-A < M), \\
P(n_1, N-n_1) &= (\frac{n_1}{N-A!/(N-n_1)!}) (\frac{\mu_2}{\mu_1})^{n_1} P(o, N), \\
P(n_1, N-n_1) &= (\frac{n_1}{M!M^{-1}}) (\frac{\mu_2}{\mu_1})^{n_1-M} P(o, N), \quad (N-A \leq n_1 < M),
\end{cases}$$

$$(N-A < M \leq n_1).$$
When $N = A$ we have a special case of $N - A < M$ and the results are identical to those for the Swedish Machine problem. We then have:

\[
P(n_1, N - n_1) = \frac{N!}{(N-n_1)!n_1!} \left( \frac{\mu_2}{\mu_1} \right)^{n_1} P(o, N), (n_1 < M),
\]

\[
P(n_1, N - n_1) = \frac{N!}{(N-n_1)!M!N^{n_1-M}} \left( \frac{\mu_2}{\mu_1} \right)^{n_1} P(o, N), (n_1 \geq M).
\]

From these results, it is easy to derive such quantities of interest as mean number of units; mean length of the waiting line and the mean number of units being served, at either stage, or the probabilities that the servers are fully occupied or that machines must wait at either stage.

Koenigsgberg \cite{H} gives tables of these values and also the utilisations of machines, operators and repair facilities for various values of $N, A, M$ and $\frac{\mu_2}{\mu_1}$ and shows how these can be used to throw light upon an organisational problem in which five alternatives for improving the system are considered.

(i) Adding one repairman,

(ii) adding one machine,

(iii) adding one machine and one operator,

(iv) adding one machine and one repairman, and

(v) speeding up repair time.

6.3. An inventory problem.

W. Karush applies the methods of this thesis to an inventory problem in which customer demand for a given commodity is assumed to
be Poisson with a mean rate $\lambda$. The constant inventory is $n$ and at any moment this may be split up into an in-stock amount $n_o$ and an in-replenishment amount $n-n_o$. If a customer arrives when $n_o > 0$ then the result is a unit sale and the initiation of replenishment of that unit, but if he arrives when $n_o = 0$ then he leaves, never to return, a lost sale is recorded and no replenishment is initiated.

Successive replenishment times are assumed to be independent and random. The replenishment cycle assumed is best explained by a simple diagram.

The $\alpha_i$'s represent stages $E_{ij}$ in the replenishment path $i$ which have mean service times $1/\mu_{ij}$. The $r$ paths are taken with probabilities $\alpha_i$. Each path consists of a finite chain of exponential stages. When an exponential stage is completed, the next is immediately begun. The mean replenishment time for path $i$ is thus

$$\frac{1}{\mu_i} = \frac{1}{\mu_{i1}} + \frac{1}{\mu_{i2}} + \ldots + \frac{1}{\mu_{ik_i}}$$

and the mean replenishment time $1/\mu$ for the whole system is given by

$$\frac{1}{\mu} = \frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2} + \ldots + \frac{\alpha_r}{\mu_r}.$$
Using this model and our past theory for finding the steady state probabilities of the system, we obtain the following steady state solution,

\[(6.3.1) \quad p(n_0/n) = \frac{1}{c} \cdot \frac{\rho^q}{q!} ,\]

where \( q = n - n_0 \), \( \rho = \lambda/\mu \) and \( C \) is the normalising factor given by:

\[ C = 1 + \rho + \rho^2/2! + \ldots + \rho^n/n! .\]

In our usual form:

\[ p(n_0/n) = \frac{\rho^q}{q!} \cdot p(n/n) ,\]

and, in particular, we have

\[(6.3.2) \quad p(0/n) = \frac{\rho^n/n!}{1+\rho+\ldots+\rho^n/n!} \]

for the probability of being out of stock, or the ratio of lost sales to total demand.

We can also obtain the probability that there are exactly \( n_{11} \) replenishments in stage \( E_{11} \); \( n_{12} \) in stage \( E_{12} \); \ldots; \( n_{ij} \) in stage \( E_{ij} \); \ldots. This is given by the product

\[(6.3.3) \quad \frac{\lambda^q}{c} \sum_{i,j} \frac{\alpha_{ij}^{n_{ij}}}{n_{ij}! \mu_{ij}^{n_{ij}}} , \quad \text{where} \quad q = \sum_{i,j} n_{ij} \]

and summing these products over all sets \( \{n_{ij}\} \) with fixed \( q \) yields \((6.3.1)\).
If, instead of assuming that the mean arrival rate \( \lambda \) is a constant, we assume that it is \( \lambda(n_o) \), a function of the stock on hand, the formula (6.3.3) is modified by replacing the factor \( \lambda^q \) by \( \lambda(n)\lambda(n-1)\ldots \lambda(n_o+1) \) for \( n_o = 1, \ldots, n-1 \) and 1 for \( n_o = n \). Thus, formula (6.3.1) becomes

\[
p(n_o/n) = \frac{1}{c} \cdot \frac{(\lambda(n)/\mu)(\lambda(n-1)/\mu) \ldots (\lambda(n_o + 1)/\mu)}{q!},
\]

and, of course, requires a modification of the normalising constant \( C \).

By showing that \( L(N) = p(o,n) \) is a convex function of \( n \), Karush goes on to determine the optimum allocation of inventory dollars among several commodities, each with a replenishment cycle like the above, so as to minimize the total lost sales for a constant inventory \( N \). This, however, entails methods totally irrelevant to the present discussion and will be omitted. The point is that by means of the theory presented here, he was able to advance to a stage when the solution of the inventory problem reduced to the standard problem of 'distribution of effort', which has been treated in the literature.

6.4. A generalisation of the machine interference model.

Benson and Gregory \( /27 \) carry the analysis of the cyclic queue a stage further when they consider a closed loop of queues in a series, each with a single server, the customers of which travel between consecutive service gates with a transit time which is a random variable. Each server may also have customers from sources outside
the system which join the queue and are served with those from the closed loop, on a first come-first served basis. On completion of their service, the customers from outside depart to a destination outside the system. Benson and Gregory quote as an example an airline, which operates a service between A and B. The two queues are those for service facilities, which may also have to cater to a certain amount of casual service from other airports.

In the general case, suppose that the circuit has r service gates in series, the service time of which and the transit times between which are assumed to have negative exponential distributions, as have the inter-arrival times of successive customers from the 'open' or 'from outside' system. Suppose further that these times at gate i have means $1/\mu_i$, $1/\lambda_i$ and $1/\nu_i$ respectively, where $1/\lambda_i$ is the mean transit time from gate i to gate (i+1), and $1/\nu_r$ refers to the transit time from gate r to gate 1. Each open queue has a traffic intensity $\rho_i = \nu_i \mu_i^{-1}$, which is assumed to be less than unity.

Let $N$ be the total number of customers belonging solely to the closed system (a constant) and denote by $y_{ij}$ the customer j-th in line at gate i, including the one (if any) in service, where $y_{ij} = 0$ if the customer is from the open system and $y_{ij} = 1$ if he is from the closed system. Then, if $s_i$ is the total number of customers in transit from gate i to gate i+1, we have

$$\sum_{i=1}^{r} s_i + \sum_{i=1}^{r} \sum_{j=1}^{s_i} y_{ij} = N.$$
The state of the system may be described at any time by the values of \( k_1, \ldots, k_r \) together with the \( r \) sequences \( y_{11}, \ldots, y_{is_1} \).

The steady state equations can be written down in the usual manner, and can be shown to have the solution:

\[
p \left( k_1, k_2, \ldots, k_r; Y(1, s_1); \ldots; Y(1, s_1); \ldots; Y(r, s_r) \right) = C \prod_{i=1}^{r} \frac{\frac{m_i}{v_i}}{k_i^{s_i} \mu_i \lambda_i},
\]

where \( Y(i, t) \) represents the row vector \( y_{11} \ldots, y_{it} \); \( m_i = s_i - n_i \) is the number of customers from the open queue at gate \( i \);

\[ n_i = \sum_{j=1}^{s_i} y_{ij} \] is the number of customers from the closed queue at gate \( i \). It follows, because of the distinguishable arrangements of customers, that

\[
p(k_1, \ldots, k_r; m_1, \ldots, m_r; n_1, \ldots, n_r) = C \prod_{i=1}^{r} \frac{(m_i + n_i)!}{m_i! n_i! \frac{\frac{m_i}{v_i}}{k_i^{s_i} \mu_i \lambda_i}}
\]

from which we can derive, by summing over all arrangements of \( n_2, \ldots, n_r; k_1, \ldots, k_r \), and all values of \( m_1 \), that the probability of exactly \( u \) customers from the closed queue at gate 1 is given by:

\[
p_1(u) = C (1 - \frac{1}{v_1^{u_1}})^{-1} \frac{1}{u} \prod_{i=1}^{r} (1 - \frac{1}{v_i^{u_i} \mu_i^{u_i}})^{-1} \sum_{n_2+\ldots+n_r=0}^{N-u} \left( \frac{x}{(N-u-n_2-\ldots-n_r)!} \right) \left( \frac{(\mu_2^{u_2})^{-n_2} \cdots (\mu_r^{u_r})^{-n_r}}{(N-u-n_2-\ldots-n_r)!} \right) \left( \frac{(\lambda_1^{u_1}+\ldots+\lambda_r^{u_1})^{-N-u-n_2-\ldots-n_r}}{(\lambda_1^{u_1}+\ldots+\lambda_r^{u_1})!} \right).
\]
In their 1951 paper on machine interference, Benson and Cox introduced the function

\[ F(x, N) = \sum_{n=0}^{N} \frac{N! x^n}{(N-n)!} \]

generalising this, they have

\[ F(a_1, \ldots, a_r; n) = \sum_{n_1+\ldots+n_r = n} \frac{n!}{n_1! \cdots n_r!} \cdot a_1^{n_1} a_2^{n_2} \ldots a_r^{n_r} \cdot \frac{1}{(n-n_1-\ldots-n_r)!} \]

In fact, \( F(a_1, \ldots, a_r; n) \) is generated by the identity,

\[ F(a_1, \ldots, a_r; n) = F(a_1, \ldots, a_r; n-1) + a_r F(a_1, \ldots, a_r; n-1), \]

and is symmetrical in the variables \( a_1, \ldots, a_r \). If we let

\[ \Lambda^{-1} = \lambda_1^{-1} + \ldots + \lambda_r^{-1} \]

the average total transit time per circuit and write \( x_i = \Lambda (\mu_i - \nu_i)^{-1} \), the servicing factor for the closed queue at gate \( i \) then, we obtain

\[ p_1(u) = C \Lambda^{-N} x_1^u F(x_2, \ldots, x_r; N-u) \left[ (N-u)! \right]^{-1} \prod_{i=1}^{r} (1 - \nu_i \mu_i)^{-1} \]

Since \( \sum_{u=0}^{N} p_1(u) = 1 \), we have

\[ 1 = C \Lambda^{-N} \prod_{i=1}^{r} (1 - \nu_i \mu_i)^{-1} \sum_{u=0}^{N} \frac{x_1^u F(x_2, \ldots, x_r; N-u)}{(N-u)!} \]

However,

\[ \sum_{u=0}^{N} \frac{x_1^u F(x_2, \ldots, x_r; N-u)}{(N-u)!} = \sum_{u=0}^{N} \frac{x_1^u}{(N-u)!} \sum_{n_2+\ldots+n_r = n} \frac{n_1! n_2! \ldots n_r!}{(N-u-n_2-\ldots-n_r)!} \]
\[
= \frac{1}{N!} F(x_1, x_2, \ldots, x_r; N) .
\]

Thus,
\[
C = \bigwedge_{i=1}^{N-r} \left( 1 - \mathcal{N}(x_i, x_{i+1}, \ldots, x_r; N) \right)^{-1} \prod_{i=1}^{r} (1 - \mathcal{N}(x_i, x_{i+1}))^{-1}
\]

and, therefore,
\[
p_1(u) = \frac{\binom{N}{u} x_1^u F(x_2, \ldots, x_r; N-u)}{(N-u)! F(x_1, \ldots, x_r; N)} ,
\]

with corresponding expressions for \( p_1(u) \).

The generating function of \( p_1(u) \) is
\[
\mathcal{G}(z) = \sum_{u=0}^{N} \frac{\binom{N}{u} (z x_1)^u F(x_2, \ldots, x_r; N-u)}{(N-u)! F(x_1, \ldots, x_r; N)} ,
\]
\[
= \frac{F(z x_1, x_2, \ldots, x_r; N)}{F(x_1, x_2, \ldots, x_r; N)} .
\]

Therefore,
\[
\mathbb{E}[u^L] = x_1 \left( F(x_2, \ldots, x_r; N) \right)^{-1} \frac{\partial}{\partial x_1} F(x_1, x_2, \ldots, x_r, N) .
\]

If \( L = \mathcal{L}_1 + \mathcal{L}_2 + \ldots + \mathcal{L}_r \), the total number of customers in transit, then
\[
\mathbb{E}[L^r] = N - \left\{ F(x_1, \ldots, x_r; N) \right\}^{-1} \sum_{i=1}^{r} x_i \frac{\partial}{\partial x_i} F(x_1, \ldots, x_r; N) ,
\]
\[
= N \cdot \frac{F(x_1, \ldots, x_r; N-1)}{F(x_1, \ldots, x_r; N)} ,
\]

after some reduction.
If \( a \) is the average time required to make a full circuit of the \( r \) gates and transits, then any customer spends a proportion \( \Lambda^{-1} a^{-1} \) of his time in transit. Thus, as this equals the fraction of customers in transit,

\[
a = N \Lambda^{-1} \left( B \left[ \frac{1}{r} \right] \right)^{-1}
\]

\[
= \Lambda^{-1} F(x_1, \ldots, x_r; N) \left( F(x_1, \ldots, x_r; N-1) \right)^{-1}.
\]

The rate of arrival of closed system customers at any gate is \( N a^{-1} \).

Denote the probability that there are no customers from either system at gate \( 1 \) by \( q_1(0) \). We then have:

\[
q_1(0) = C \sum_{\ell_1 + \ldots + \ell_r + n_2 + \ldots + n_r = N} \prod_{i=2}^{r} \frac{\mu_i}{\ell_i!} \left( 1 - \nu_i \right)^{-1} \left( 1 - \nu_{r+1} \right)^{-1} \left( \ell_i - 1 \right) \lambda_i \left( \ell_i \right) \left( \ell_i + 1 \right) \left( \ell_i + 2 \right) \ldots \left( \ell_i + n_i \right) \lambda_1 \left( \ell_1 \right) \left( \ell_1 + 1 \right) \left( \ell_1 + 2 \right) \ldots \left( \ell_1 + n_1 \right)
\]

\[
= C \prod_{i=2}^{r} \left( 1 - \nu_i \right)^{-1} \sum_{n_2 + \ldots + n_r = 0}^{N} \prod_{i=2}^{r} \left( \mu_i - \nu_i \right)^{-n_i} \left( \mu_r - \nu_r \right)^{-n_r} \left( \ell_i - 1 \right) \lambda_i \left( \ell_i \right) \left( \ell_i + 1 \right) \left( \ell_i + 2 \right) \ldots \left( \ell_i + n_i \right) \lambda_1 \left( \ell_1 \right) \left( \ell_1 + 1 \right) \left( \ell_1 + 2 \right) \ldots \left( \ell_1 + n_1 \right)
\]

\[
= C \Lambda^{-N} \left( N! \right) \prod_{i=2}^{r} \left( 1 - \nu_i \right)^{-1} \left( 1 - \nu_{r+1} \right)^{-1} \left( F(x_2, \ldots, x_r; N) \right)^{-1}
\]

\[
= (1 - \nu_1 \mu_1^{-1}) F(x_2, \ldots, x_r; N) \left( F(x_1, \ldots, x_r; N) \right)^{-1},
\]

a product of the following two factors:

1) The probability that there are no customers from an open queue with the same input and service mechanism as the open queue here.
ii) The probability $p_1(o)$ that there are no customers from the closed system. Alternatively, $p_1(o)$ can be interpreted as the probability that there are no customers from a closed system with the same conditions as the one here, neglecting the effect of the open queue, after reducing the mean rate of service from $\mu_i$ to $\mu_i - v_i$.

Benson and Gregory point out that, because the distribution of the sum of $k$ variables, each with the same negative exponential distribution, is known to be $\chi^2_{2k}$, then, if we introduce service gates with zero average service times, these results hold for transit times with $\chi^2_{2k}$ distributions and, in particular, for constant transit times.

**Limits of the system.**

Benson and Gregory show that as $N$ tends to infinity with $\lambda$ constant, the queue at stage $i$ acts like an open queue with traffic intensity $x_i x^{-1}_M$, where $x_M = \max(x_i)$, $(i = 1, \ldots, r)$. Queues for which $x_i = x_M$ expand without limit. Also, only at the gate(s) at which the average service time is a maximum will reducing the service time affect the number of customers in transit, where the effect will be to increase the number. In the limit, the rate of arrival at any gate is equal to the minimum adjusted service rate.

When $N$ tends to infinity and $\lambda$ tends to zero, i.e. the average transit times between two consecutive gates tends to infinity, we find that:

i) The arrival rate of closed queue customers is $\lambda_r N$, where $\lambda_r \rightarrow \infty$ in such a way that $\lambda_r N = c$, where $c < \min(\mu_i - v_i)$. 
ii) Gate $i$ behaves, as far as the closed queue customers
are concerned, as an open queue with traffic intensity $c\xi_i^{-1}$ for all
$i$, where $\xi_i = \mu_i = \nu_i$.

iii) $Q_i(o) = (\xi_i - c)\mu_i^{-1}$, where $Q_i(o)$ is the probability
that there are no customers in stage $i$.

6.5. A consideration of some practical problems.

Suppose we have a network subject to one restriction, that
$\sum_{m=1}^{M} \mu_m = \mu$. Then, it is of practical interest to find the optimum
distribution of our service facilities in terms of output or inven-
tory, i.e. how can we maximize output or minimize our total inventory.

Suppose that we have only one server at each stage; then we have,
from (3.5.2),

\[
(6.5.1) \quad p_m(n) = p_m(o) \left( \frac{\Gamma_m}{\mu_m} \right)^n .
\]

We know that, in the steady state, the output from any queue must
be random, the same as the input. Thus, for stage $m$, output is ran-
dom, parameter $\Gamma_m$. However, only a proportion $(1 - p_m)$ of this
output actually leaves the system, the rest goes back into some other
stage of the system. Hence, actually output of the stage is random
with parameter $\Gamma_m(1-p_m)$. Thus, total output $= \sum_{m=1}^{M} \Gamma_m(1-p_m)$
which is independent of $\mu_m$, for all $m$. Thus, as long as $\mu_m > \Gamma_m$
for all $m$, ensuring a steady state distribution, the output is un-
changed by any arrangement of the $\mu_m$'s.

The mean number of customers in stage $m$, $\bar{N}_m$, is given by,
\[
\tilde{\eta}_m = \sum_{n=1}^{\infty} n P_m(n) = P_m(0) \sum_{n=1}^{\infty} n(\Gamma_m/\mu_m)^n, \text{ from (6.5.1)}.
\]

Now \( P_m(0) \) is given by \( \sum_{n=0}^{\infty} P_m(n) = 1 \)

Hence,

\[
P_m(0) (1 + \Gamma_m/\mu_m + (\Gamma_m/\mu_m)^2 + \ldots ) = 1
\]

i.e.

\[
P_m(0) = 1 - \Gamma_m/\mu_m.
\]

Thus, the mean number of customers in stage \( m \) is given by,

\[
\tilde{\eta}_m = (1 - \Gamma_m/\mu_m)(\Gamma_m/\mu_m) \sum_{n=1}^{\infty} n(\Gamma_m/\mu_m)^{n-1}
\]

\[
= \frac{\Gamma_m}{\mu_m - \Gamma_m}.
\]

Thus, the average number of customers in the system is,

\[
(6.5.2) \quad \sum_{m=1}^{M} \frac{\Gamma_m}{\mu_m - \Gamma_m}.
\]

We now attempt to minimize (6.5.2) subject to the restriction

\[
M \mu = \mu.
\]

Let

\[
F = \sum_{m=1}^{M} \frac{\Gamma_m}{(\mu_m - \Gamma_m)} + \lambda_1 \left( \sum_{m=1}^{M} \mu_m - \mu \right).
\]

where \( \lambda_1 \) is a Lagrange multiplier.
Thus, we have, in the usual way,

$$\frac{\partial F}{\partial \mu_m} = -\frac{\Gamma_m}{(\mu_m - \Gamma_m)^2} + \lambda_1$$

and

$$\frac{\partial^2 F}{\partial \mu_m^2} = \frac{2 \Gamma_m}{(\mu_m - \Gamma_m)^3}$$

Thus, \( \frac{\partial F}{\partial \mu_m} = 0 \), for all \( m \), when

$$\frac{\Gamma_m}{(\mu_m - \Gamma_m)^2} = \lambda_1$$

i.e. \( (\mu_m - \Gamma_m)^2 = \Gamma_1/\lambda_1 \)

or

$$\mu_m = \Gamma_m + \frac{\Gamma_m}{\lambda_1}$$

This root is chosen so that \( \mu_m > \Gamma_m \).

But, we require \( \sum_{m=1}^{M} \mu_m = \mu \), and so we must have

$$\mu = \sum_{m=1}^{M} \Gamma_m + \left( \sum_{m=1}^{M} \Gamma_m^{1/2} \right)^{1/2}/\lambda_1^{1/2}$$

Hence,

$$\lambda_1^{1/2} = \left( \sum_{m=1}^{M} \Gamma_m^{1/2} \right)/\left( \mu - \sum_{m=1}^{M} \Gamma_m \right).$$
Recall that $\mu_m > \Gamma_m$ for all $m$, and we have that $\frac{\partial^2 F}{\partial \mu_m^2} > 0$ for all $m$. Consequently our optimum arrangement is

$$\mu_m = \Gamma_m + \sum_{m=1}^{M} \frac{1}{\Gamma_m} \left( \sum_{m=1}^{M} \Gamma_m \right)^{1/2}$$

This is the distribution of the $\mu_m$'s which minimizes the average number of customers in the system, and, consequently, the inventory.

Let us now consider a variation on this problem. Consider the case where we have $R$ servers, each with mean service time $1/\mu$. Queue $m$ is assumed to have $r_m$ servers. The steady state solution for stage $m$ is given by (3.5.2) as

$$P_m(n) = P_m(o) \frac{(\Gamma_m/\mu_m)^n}{n!}, \ (n = 0, \ldots, r_m),$$

$$P_m(n) = P_m(o) \frac{(\Gamma_m/\mu_m)^n}{r_m! r_m^{n-r_m}}, \ (n > r_m),$$

where $P_m(o)$ are determined by the usual normalising equations

$$\sum_{n=0}^{\infty} P_m(n) = 1, \ (m = 1, \ldots, M).$$

Now,

$$\mu_m = \frac{r_m \mu}{r_m^m}.$$

Hence

$$P_m(o) = \sum_{n=0}^{r_m} \frac{(\Gamma_m/\mu_m)^n}{n!} + \sum_{n=r_m+1}^{\infty} \frac{(\Gamma_m/\mu_m)^n}{r_m! r_m^{n-r_m}} r_m^{n-r_m}.$$  

The average number in stage $m$ is given by
\[ \sum_{n=1}^{\infty} nP_m(n) = P_m(0) \sum_{n=1}^{r_m} n(\Gamma_m/r_m \mu)^n/n! + \sum_{r_{m+1}}^{\infty} n(\Gamma_m/r_m \mu)^n/r_m ! r_m^{n-r_m} \]

Hence, the average total number is,

\[
(6.5.3) \quad \sum_{m=1}^{M} \left( \sum_{n=0}^{r_m} n(\Gamma_m/r_m \sigma)^n/n! + \sum_{r_{m+1}}^{\infty} n(\Gamma_m/r_m \mu)^n/r_m ! r_m^{n-r_m} \right)
\]

Having solved the equations (3.5.1) for the \( \Gamma_m \), we have to find the arrangement of \( r_m \)'s, subject to \( \sum_{m=1}^{M} r_m = R \) and the steady state conditions \( \Gamma_m < r_m \mu \) for \( m = 1, \ldots, M \), (which, incidentally, assure us once again that the output is a constant, whatever the arrangement), which minimize (6.5.3).

The easiest way to do this calculation is, probably, a graphical method. Starting with the smallest \( r_m \) such that \( r_m \mu > \Gamma_m \), the individual average numbers of each stage can be calculated for increasing \( r_m \) and plotted as a straight line graph. Or, if the calculations are made on a modern computer, we can get the graphs directly as the output. As each \( r_m \) increases, the average number of customers in its stage will decrease. This decrease will be a convex function for increasing \( r_m \). Now, we consider our starting values on the graphs, i.e. the smallest \( r_m \)'s such that \( r_m \mu > \Gamma_m \), and find the graph which has the steepest descent for an increase of one server. We then move along this graph to the point which corres-
ponds to adding one server to the stage and repeat the process, each time picking the graph with the steepest line of descent and moving along it, until our resulting number of servers is \( R \), whereupon we note our resulting number of servers for each stage. This is then our optimum distribution. Obviously, without a computer, this leads to a lot of work. However, the work can be kept to a minimum by first calculating the mean number of customers in each stage for \( r_m \) and \( r_{m+1} \) servers only \((m = 1, \ldots, M)\), and noting the differences. Then, picking the largest of these differences, in queue \( j \) say, we calculate the average number for \( r_{j+2} \), and so on, until our total number of servers is \( R \).

The following problem frequently occurs both in industry and in businesses where a service is given, subject to a waiting room of limited size, e.g. a doctor's surgery or a hairdressers.

We have a service mechanism with a negative exponentially distributed service time \( s \), which has mean \( 1/\mu \). There is a waiting room of size \( N \) before it which is fed by a random input from another source with parameter \( \lambda \). If we consider the industrial case, the input and service mechanisms will be two machines. Now, the second machine can have its mean service time reduced, but only at the cost of producing more defectives, or, the waiting room can be enlarged, again at a certain expense. Further, when the waiting room is full, and another customer arrives, he blocks the first machine, which again proves costly.
Let the cost of defectives per unit of working time be distributed as \( f(x) \), where \( x \) is the traffic intensity, and let the cost of storage space be distributed as \( g(n) \). Further, let the cost per unit of time that the first machine is stopped be \( C \). The problem is to find the optimum working arrangement in terms of cost.

We have from 4.2 that, for a given \( N \) and \( \mu \), the probability that there will be \( N \) customers in the waiting room is

\[
P_N(1 - K_1) = P_{N-1}K_2 + P_{N-2}K_3 + \ldots + P_2K_{N-1} + (P_1 + P_0)K_N,
\]

where

\[
P_n = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{n+1}}, \quad (\frac{\lambda}{\mu})^n, \quad \text{for} \quad n = 0, 1, \ldots, N-1,
\]

and, writing \( \alpha = \lambda/\lambda + \mu \) and \( \beta = \mu/\lambda + \mu \),

\[
K_r = \beta(\alpha^r + \alpha^{r+1} + \ldots)
\]

\[
= \beta \frac{\alpha^r}{1 - \alpha} = \alpha^r.
\]

Thus,

\[
P_N(1 - \alpha) = P_N \beta
\]

\[
= \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+1}} \cdot \sum \alpha^2(\lambda/\mu)^{N-1} + \alpha^3(\lambda/\mu)^{N-2} + \ldots
\]

\[
+ \alpha^N(1 + \lambda/\mu)
\]

\[
= \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+1}} \cdot \frac{\lambda}{\lambda + \mu} \cdot (\frac{\lambda}{\mu})^{N-1},
\]
after some reduction.

Thus,
\[ P_N = \frac{1 - \frac{\lambda}{\mu} x}{1 - (\frac{\lambda}{\mu} x)^{N+1}} \left( \frac{\lambda}{\mu} x \right)^N \]

Let \( \frac{\lambda}{\mu} = x \), and assume that we have \( \mu > \lambda \), so that \( x < 1 \) (as in 4.2). Then,
\[ P_N = \frac{1 - x}{1 - x^{N+1}} \cdot x^N. \]

Now, if a customer arrives to find the waiting room full, he blocks the previous machine. This is the same as having a waiting room of size \( N+1 \). Thus, the probability that the first machine is blocked, given \( N \) and \( x \), \( P_B(N, x) \) is given by

\[ (6.5.4) \quad P_B(N, x) = \frac{1 - x}{1 - x^{N+2}} \cdot x^{N+1} \]

Hence,
\[ P_B(N, x) - P_B(N+1, x) = (1-x) \frac{x^{N+1}}{1-x^{N+2}} - \frac{x}{1-x^{N+3}} \]
\[ = \frac{(1-x)^2 x^{N+1}}{(1-x^{N+2})(1-x^{N+3})} \]
\[ = D(x, N) \text{ say.} \]

Thus, for a given \( x \), as long as we have
\[ c \cdot \frac{(1-x)^2 x^{N+1}}{(1-x^{N+2})(1-x^{N+3})} > g(N+1) - g(N), \]

our system is improving as we increase \( N \).
Consider the special case $g(N) = kN$, where $k$ is a constant. Then, we have an improvement as long as $D(x,N) > \frac{k}{C}$.

We can represent this on the following graph.

"Graph to show the 'improvement region' with respect to unit increases in the size of the waiting room."

where the shaded region is the region in which an improvement can be made.

We also have,

$$\frac{\partial P_B(N,x)}{\partial x} = \frac{x^N(1-x^{N+2})/(N+1) - (N+2)x^N}{(1 - x^{N+2})^2} + x^N(x-x^2)(N+2)x^{N+1}$$

which is positive in $0 < x < 1$ and singular at $x = 1$.

At $x = \lambda/\mu$, $\partial x/\partial \mu = -\lambda/\mu^2$ and we have:
\[
\frac{\partial P_B(N, x)}{\partial \mu} = -\frac{\lambda}{\mu} \frac{\partial P_B(N, x)}{\partial x},
\]

which is negative in \(0 < x < 1\).

Thus, as long as

\[
-\frac{C}{\partial \mu} \frac{\partial P_B(N, x)}{\partial \mu} > \frac{\partial f(x)}{\partial \mu}
\]

our system improves with increasing \(\mu\).

Let us consider the special case, \(f(\mu) = e^{-k_1 \mu}\). Then we can represent our arrangement on the following graph.

"Graph to display the 'improvement region' with respect to changes in \(\mu\)."

where the stippled region is the 'improvement region'.

To find the optimum arrangement, we find an \(x\) and an \(N\) which are in the improvement regions of both graphs. Then, on the first
graph, we find all greater $N$ such that this $x$ is still in the improvement region. For each of these $N$'s we check on both graphs to find the maximum $\mu$ for which we are still in the improvement region and take the minimum of these in each case. Now, for each particular case we increase $\mu$ until we find that

$$\frac{k}{C} - D(x, N) = \frac{\partial P_B(N, x)}{\partial \mu} - \frac{k}{C} e^{\frac{1}{\mu}}$$

and stop there, noting our values of $x$ and $N$. We calculate our profits for each of these points and find the maximum. The corresponding values of $x$ and $N$ are our optimum arrangement.
BIBLIOGRAPHY


