SYSTEMS OF QUEUES

PARALLEL, SERIES AND NETWORK QUEUEING SYSTEMS

WITH APPLICATIONS TO

COMMUNICATIONS AND COMPUTING.

A Dissertation

Presented in partial fulfilment

of the requirements

for the degree of Master of Science.

UNIVERSITY OF WAIKATO

1972

M.D. CAMDEN.

ABSTRACT.

The need for understanding of various real-life queueing situations has caused the theory of queues to grow in various directions. One of these directions treats the class of systems where more than one service is available to, or required by, each customer. The theory for this class has progressed along a central path from which various developments branch. Multichannel poisson queues constitute the first part of the path; series and then networks of poisson stations follow. The branches from the path involve models with non-poisson or general assumptions, models with limited waiting room or limited numbers of customers, models with nonzero transit times, and others. The path leads to models for communication networks, and hence to models for computer networks. The theory of computer timesharing branches from the beginning of the path.

The dissertation attempts to outline the present state of the theory for those models which occur on this path or its branches. The derivation of probability distributions for queue length and waiting time are presented in full for models on the central path, and in brief for the other models.

CONTENTS.

SYSTEMS OF QUEUES.

CHAPTER 1.

INTRODUCTION.

This chapter defines the aims and the subject matter of the dissertation, and sets the study in its context.

CHAPTER 1 : INTRODUCTION.

1.1 AIMS.

This study is intended to be a synthesis of knowledge about certain aspects of queueing. It presents a broad view of the mathematics associated with these aspects, but does not claim to provide a full coverage. The papers and texts cited do not constitute a full survey of the literature, but do, it is hoped, indicate the general shape of the theory which has so far evolved. This study does not claim to make any original contributions to the Theory of Queueing.

1.2 CONTENT.

The aspects of queueing included were chosen in accordance with three underlying interests:

.an interest in systems more complex than those in which a customer queues once before a single server;

.an interest in applications to computer systems; and

.an interest in the construction of rigorous methods from intuitive probability concepts.

The first of these interests guided the development of the three theory-oriented chapters (chapters 2, 3, and 4.). The second gave rise to the two applications-oriented chapters (chapters 5 and 6). The third guided the selection of methods to be accorded fuller treatment. Under these influences, the content has taken the following shape:

The central topic of chapter 2 is QUEUEING SYSTEMS WITH PARALLEL CHANNELS. Results for single-channel systems are included, since series and network systems may contain

single-channel stations; and to provide coverage of systems with various arrival and service distributions. Some of chapter 2's theory is used in chapter 3, which treats SYSTEMS INVOLVING SERIES OF QUEUES. Series systems are special cases of the topic of chapter 4, SYSTEMS INVOLVING NETWORKS OF QUEUES. The network theory is extended , in chapter 5, to describe a class of real-life COMMUNICATIONS NETWORKS, and then is extended further to describe the special communication networks which involved COMPUTER NETWORKS. Chapter 6 is not another step in this progression, but develops theory from chapter 2 for a study of COMPUTER TIME-SHARING , and mentions other applications to computer systems.

Chapter 7 discusses the usefulness of the line along which the theory of this study has developed.

The third of the interests mentioned above, and the need to give fuller treatment to methods involved in later chapters, give rise to an emphasis on SYSTEMS INVOLVING POISSON QUEUES IN EQUILIBRIUM. Queueing Theory diverges in many directions. Apart from the direction which leads to networks, the direction which this study explores most fully is that involving NON-POISSON AND GENERAL DISTRIBUTIONS for arrival and service times. The most important direction not explored is that involving QUEUE SYSTEMS WITH PRIORITIES. Where they are available, simulation studies have been mentioned together with the theory.

The measures of congestion studied are the distributions of number in system, and of waiting time.

1.3 CONTEXT.

An inspection of the titles of recent papers indicates the directions in which Queueing Theory and its applications have developed. The following lists of some of these directions should set the present study in its context. Its subject matter is a very small part of the existing knowledge about Queueing.

Theoretical developments excluded from this study include:

.systems with priorities other than " first come first served ";

.busy period distributions;

.bulk arrivals and service;

.baulking and reneging;

.rate of approach to equilibrium.

Other developments receive a partial treatment:

.non-poisson systems;

.transient solutions;

.conditions for the existence of equilibrium solutions;

.effect of limited waiting room;

.jockeying;

.feedback;

.cost structures.

Applications which have prompted the development of queueing theory include much besides communications and computing. The application receiving most attention in the journals at present is the behavior of vehicular traffic. This involves:

.intersections and gap-acceptance;

.bottlenecks;

.tollgates;

.jams;

.parking spaces;

.the taxi problem;

.take-off and landing at airports;

.utilization of wharves and canals.

Traffic of calls at telephone exchanges is still receiving attention.

Industrial problems to which Queueing Theory is being applied include:

.machine interference;

.provision of spare machines, and personnel allocation;

.inventory control;

.scheduling of jobs;

.dams and other storage systems;

.conveyor systems.

Problems involving human traffic include:

.provision of doctors, hospital beds etc., for medical care;

.design of retail shops;

.design of customs checkpoint systems.

1.4 TERMINOLOGY.

Symbols and terms are defined as the need arises; and may be redefined, with different meanings.

In the theory-oriented chapters, the word " system " refers to idealized situations. In the applications-oriented chapters, real-life situations and their theoretical images are referred to respectively as " systems " and

" models ".

Bracketed numbers in the text refer to the alphabetical list of references. Where two numbers are given, the second refers to the first relevant page of the work referenced.

Functions which are Laplace transforms are denoted by the symbol for the original function, with an asterisk. Thus P* denotes the Laplace transform of P.

Kendall's A/B/c notation (46) is used throughout.

CHAPTER 2.

QUEUES WITH MULTIPLE CHANNELS.

The chapter studies queue lengths and waiting times for single-station multi-channel systems, with various arrival and service assumptions. Results are included on single-channel systems with more general service distributions and on the effects of jockeying.

CHAPTER 2 : QUEUES WITH MULTIPLE CHANNELS.

2.1 INTRODUCTION.

This chapter treats queue systems which consist of several channels in parallel. The steadystate M/M/c system is studied in detail; then follow less detailed treatments of

.time-dependent solutions

.systems with other arrival and service distributions

.systems with arrival and service distributions defined in general terms.

The queue discipline for all the systems studied is the first-come first-served discipline.

The discussion of general approaches to multichannel systems is intended to extend to the limits of the theory as it stands now. It seems that general approaches have not been a topic of great interest since the 1950's.

The M/M/c system is chosen for detailed treatment

.as an example of an easily solved case;

.because of its theoretical simplicity;

.because of its many applications;

.and- this follows from the last two reasons- because it is used as a component of more complex systems.

Exponential interarrival times are to be expected whenever the arrivals come from a large population acting independently. The exponential model is liable to fit well except where:

.arrivals depend on some time-dependent process (like the wearing out of TV sets bought in the early 1960's; or the lunch-hour rush at a coffee bar), or

.arrivals are scheduled, and the errors about the scheduled

times are small (as at a port, an airport, or a dentist's waiting room). When scheduled arrivals occur close to the appointment times, it may be appropriate to assume a regular distribution. Where only each k-th potential customer enters the system, the k-th Erlang distribution may be appropriate. Other cases involving non-exponential treatment are: bulk arrivals; arrivals by appointment at regular intervals; random arrivals at discrete time points; non-stationary arrival patterns; arrival patterns correlated with the state of the system. (23, 16)

The use of the exponential distribution for service time is harder to justify. Some services consist of many independent tasks, any of which may be omitted (as in retail shops). Others consist of a skilled operation or a man-machine interaction (as in a toll gate service, landing an aircraft, or car following). W.R. Blunden (9) found by observation that some services in the first group (e.g, service at a service station, supermarket checkout channel, pharmacy) were approximately exponential. Others in the first group and many in the second group are better modelled with a distribution of the Pearson III shape. Distributions of this shape (shifted Gamma; shifted Erlang; shifted exponential) allow for a time interval during which no services are completed. Tollgate operation, aircraft landing, and aircraft take-off fitted this type of distribution. Two services which, fortunately, have approximate exponential distributions, are telephone calls (23, 20) and computer jobs. (80).

Thus some justification exists for concentrating on exponential arrival and service distributions.

Multiple channel queue systems exist wherever several alternate channels perform the same service. Examples are: parallel telephone trunk lines; tollgates; computer systems containing several machines or a multiprogramming machine; a garage with several mechanics; a typing pool.

The three principle measures of congestion examined in the literature are queue length, waiting time, and busy periods.

Section 2.2 studies methods of discovering probability functions for queue length; section 2.3 studies probability functions for waiting time, plus a few associated results on queue length.

Queue length distributions have the advantage that they are independent of queue discipline (23, 26); whereas the aim, in constructing a priority system, is to reduce waiting times.

Section 2.4 applies some of this theory to the simplest multichannel system: the steady state system M/M/2. It compares this to related 2-channel systems.

2.2 QUEUE LENGTHS.

2.2.1 THE SYSTEM M/M/c.

In the system studied here, customers arrive in a single poisson stream. There are C identical service channels; an arrival goes to any of the idle channels, or if no channels are idle, he goes to the tail of a single queue. The customer at the head of this queue begins his service as soon as a channel becomes available.

The Birth-Death Equations.

Let $P_n(t) = \Pr[\text{at time } t, \text{ queue length} = n]$, and let

the system begin operation at $t = 0$. The probability density function for inter-arrival time t is given by

$$a(t) = \lambda e^{-\lambda t} \quad , \quad \lambda > 0, \quad t > 0;$$

and the cumulative probability function is

$$A(t) = 1 - e^{-\lambda t} = 1 - (1 - \lambda t + \lambda^2 t^2/2 - + - + -).$$

We consider the probability of an arrival occurring in the interval dt; t is the time since the previous arrival.

$$\Pr\left[\text{no arrival occurs in } (t, t+dt) / \text{ no arrival occurred in } (0,t)\right]$$

$$= \Pr\left[\text{no arrival occurs in } (0, t+dt)\right] / \Pr\left[\text{no arrival occurred in } (0,t)\right]$$

$$= (1 - A(t+dt)) / (1 - A(t))$$

$$= e^{-\lambda(t+dt)} / e^{-\lambda t}$$

$$= e^{-\lambda dt}$$

$$= 1 - \lambda dt + \underline{0}$$

where $\underline{0}$ represents a term with a factor of $(dt)^2$, and:

$\underline{0} \to 0$ as $dt \to 0$.

This probability is independent of time from the start of operation, and time from the last arrival. Let t now be the time from the start of operation. Then:

$$\Pr\left[\text{an arrival occurs at a time in } (t, t+dt)\right] = \lambda dt.$$

Similarly;

$$\Pr\left[\text{a departure occurs in } (t, t+dt)\right] = r\mu dt,$$

when r channels are occupied at t.

$[n = 0 \text{ at } t+dt]$ implies $[n = 0 \text{ at } t$, and no arrival occurred during $dt]$ or $[n = 1$ at t, and no arrival, and one departure occurred during $dt]$.

All other possibilities have probabilities which are of small order in dt.

So:

$$P_o(t+dt) = P_o(t) \cdot (1 - \lambda dt + \underline{o}) + P_1(t) \cdot (1 - \lambda dt + \underline{o}).\mu dt,$$

and

$$\frac{P_o(t+dt) - P_o(t)}{dt} = \mu P_1(t) - \lambda P_o(t) + \underline{o}/dt \quad .$$

Let $dt \to 0$; the last equation becomes:

$$\dot{P}_o(t) = \mu P_1(t) - \lambda P_o(t).$$

When queue length n is $0 < n < c$ at $t + dt$, we have three cases:

at t, queue length was n, and during dt no arrivals and no departures occurred;

at t, queue length was $n+1$, and during dt no arrivals and one departure occurred;

at t, queue length was $n-1$, and during dt one arrival and no departure occurred.

(Probabilities of more than one arrival or departure occurring in dt become part of the term $\underline{o}$). Thus:

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

On letting $dt \to 0$, this yields

$$\dot{P}_n(t) = -P_n(t)\,(\lambda + n\mu) + P_{n-1}(t)\mu(n+1) + P_{n-1}(t)\lambda$$

When $n \geq c$, the mean departure rate is $c\mu$, and the relevant equation is found similarly.

The complete set of birth-death equations (28, 454) is:

$$\dot{P}_o(t) = -\lambda P_o(t) + \mu P_1(t);$$

$$\dot{P}_n(t) = -(\lambda + n\mu)\,P_n(t) + \lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) \quad \text{for } 0 < n < c;$$

$$\dot{P}_n(t) = -(\lambda + c\mu)P_n(t) + \lambda P_{n-1}(t) + \mu c P_{n+1}(t), \text{ for } n \geq c.$$

A unique set of solutions to these equations has been shown to exist, (76, 84) when $\rho = \lambda/(c\mu) < 1$.

Time-Dependent Solution.

The solution given by Saaty (76, 112) is too lengthy to reproduce here. It commences by defining two generating functions thus:

$$P(z,t) = \sum_{n=0} P_n(t)z^n \; ; \quad \text{and}$$

$$Q(z,t) = \sum_{n=0}^{c-2} P_n(t)z^n \; .$$

When the birth-death equations are multiplied by appropriate powers of z and summed, we get differential equations for P and Q. These are transformed, and the resulting equations in the Laplace transforms, P* and Q*, eventually yield expressions for the Laplace transforms P_n^*, for $n = 0,1,\ldots$. These expressions are very complex. Since the Laplace transform is a moment-generating function, the moments of the $P_n(t)$ can be found from their transforms.

Explicit expressions have been found for the inverse transforms of the P_n^* for the M/M/1 system (76, 93) but not for the systems with $c > 1$. The steady-state probabilities P_n can be obtained from the P_n^* by using the fact that

$$\lim_{t\to\infty}[P_n(t)] = \lim_{s\to 0}[sP_n^*(s)] .$$

Instead, we shall derive them directly from the birth-death equations as follows.

Steady-State Solution

We assume that a steadystate probability distribution for n exists, when $\lambda/(c\mu) < 1$; then $P_n(t) \rightarrow P_n$ and $\dot{P}_n(t) \rightarrow 0$, as $t \rightarrow \infty$; for $n = 0,1,..$. The birth-death equations become, after a little arrangement:

$$(0+1)\mu P_1 - \lambda P_0 = 0 ;$$

$$(n+1)\mu P_{n+1} - \lambda P_n = n\mu P_n - \lambda P_{n-1}, \quad \text{for } 0 < n < c ;$$

$$c\mu P_{n+1} - \lambda P_n = c\mu P_n - \lambda P_{n-1} \quad \text{for } n \geq c ;$$

Let $R_n = n\mu P_n - \lambda P_{n-1}$, for $n = 1,2,\cdots\cdots c$;

then $R_1 = 0$;

$R_2 = R_1 = 0$

$R_c = R_{c-1} = \cdots = 0$.

So $P_1 = (\lambda/\mu)P_0$

$P_2 = (\lambda/2\mu)P_1 = (\lambda/\mu)^2 P_0/2$

and $P_n = (\lambda/c\mu)^n c^n P_0/n!$ for $n = 0,1, \cdots\cdots c$.

Thus, $P_c = (\rho^n c^n/c!)\, P_0$, where $\rho = \lambda/c\mu$.

Let $R_n = c\mu P_n - \lambda P_{n-1}$, for $n = c, c+1, c+2, \cdots\cdots$

Then $R_c = R_{c+1} = R_{c+2} = \cdots = 0$.

So $P_n = (\lambda/c\mu)\, P_{n-1}$, for $n = c+1, c+2, \cdots\cdots$

$= (\lambda/c\mu)^n P_c$.

P_0 can be found from

$$\sum_{n=0}^{\infty} P_n = 1,$$

to be $P_0 = 1/\left(\sum_{n=0}^{c-1} (c\rho)^n/n! + ((c\rho)^c/c!) \sum_{r=0}^{\infty} \rho^r \right)$

$= 1/\left(\sum_{n=0}^{c-1} (c\rho)^n/n! + (c\rho)^c/(c!\ (1-\rho)) \right)$

The steadystate solution, with P_0 as shown above, is

$P_n = ((c\rho)^n/n!)\, P_0$ for $0 < n \leq c$

$((c\rho)^c/c!)\, \rho^{n-c} P_0 = (c^c/c!)\, \rho^n P_0$ for $n \geq c$.

For the series to converge, ρ must be < 1.

Let $X = \sum_{n=0}^{c-1} (c\rho)^n/n!$

Some further results can now be stated, in terms of P_o, P_c, X :

Pr [an arrival does not have to wait for service]

= Pr [at least one channel is idle]

$= P_o + P_1 + \ldots + P_{c-1}$

$= X P_o$.

Pr [an arrival does have to wait]

$= 1 - XP_o = P_c / (1 - \rho)$.

Average queue length = E[n]

$= (0.1 + 1.c\rho/1! + \ldots + c.(c\rho)^c/c!) P_o$

$+ ((c+1)\rho^1 + (c+2)\rho^2 + \ldots) P_c$

$= c\rho XP_o + (\rho / (1 - \rho)^2) P_c$

= E [number of channels busy] + E [number waiting for service] .

Utilization factor

= E [proportion of channels busy]

$= c\rho XP_o / c = \rho XP_o$

= traffic intensity · Pr [at least 1 channel is idle] .

2.2.2 VARIANTS ON THE M/M/c SYSTEM.

Unlimited Number of Channels.

When each arrival calls into operation a service channel we can find relevant formulas by letting $c \to \infty$ in formulas for the M/M/c system. In the steady state,

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_o \quad , \quad \text{for all } n > 0 .$$

Summation gives $P_o = e^{-\lambda/\mu}$,

so $P_n = (r^n/n!).e^{-r}$,

and n has a poisson distribution, with mean $r = \lambda/\mu$.

Obviously, average number waiting = average waiting time = 0.

Saaty (76, 99) develops time-dependent solutions for $P_n(t)$

by defining the generating function: $P(z,t) = \sum_{n=0}^{\infty} P_n(t)\, z^n$, and solving its differential equation.

n-dependent Arrival and Service Rates. (1, 255)

Sometimes prospective queuers are influenced by the size of the queue they intend to join; and servers may be spurred to work faster by a lengthening row of customers. We denote the average arrival rate and average service rate when there are n in the queue, by λ_n and μ_n. The steady-state equations for the single-server system are

$$\mu_1 P_1 - \lambda_0 P_0 = 0 ;$$

$$\mu_{n+1} P_{n+1} - \lambda_n P_n = \mu_n P_n - \lambda_{n-1} P_{n-1} , \quad \text{for } n > 0 .$$

Hence $P_{n+1} = (\lambda_n / \mu_{n+1})\, P_n = \left((\lambda_n \cdots \lambda_0) / (\mu_{n+1} \cdots \mu_1) \right) P_0$.

If the series: $\frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \cdots$

converges, it converges to $1/P_0$, and a steady state exists.

This method can easily be applied to c-channel systems; since the departure rate depends on the number of channels occupied. In fact, the derivation above is a special case of n-dependent service times.

Service Rates Differing for Differing Channels. (76, 290)

The c servers perform the same function, but their average rates of service $(\mu_1, \mu_2, \cdots \mu_c)$ can differ. Besides the probabilities P_n, we need to consider the probabilities for the $2^c - 1$ states which arise according to which of the channels are occupied.

Saaty considers the system with 2 parallel channels. Five differential equations are developed. They can be used as steadystate equations if $\rho = \frac{\lambda}{\mu_1 + \mu_2} < 1$. He also uses: (total number served in long time T) = T.(mean arrival rate); to find another equation. The solution is:

$P_1 = (1-\rho)(1-P_0)$

$P_n = \rho^{n-1}(1-\rho)(1-P_0)$ for $n \geq 2$. For $n \geq 2$, the distribution is geometric as usual.

2.2.3 OTHER SYSTEMS.

The System M/D/c. (76, 154)

Let the constant service time at each channel be 1 time unit. A steady state will exist if $\lambda < c$. Let $a_n = \sum_{i=0}^{n} P_i$,

$= \Pr[\text{queue length} \leq n]$.

$P_0 = \Pr[\text{at one service-time earlier, queue length} \leq c]$

$\cdot \Pr[\text{no arrival during 1 service}]$;

$= a_c e^{-\lambda 1}$.

For $n > 0$, similarly,

$$P_n = a_c \frac{\lambda^n e^{-\lambda}}{n!} + P_{c+1} \frac{\lambda^{n-1} e^{-\lambda}}{(n-1)!} + + + P_{c+n} e^{-\lambda}.$$

The generating function is defined by:

$$P(z) = \sum_{n=0}^{\infty} P_n z^n .$$

A differential equation in P(z) is constructed from the above relationships and solved. It is interesting to note that in this case, with service times non-exponential; the P_n are not geometric.

The System $M/E_k/c$.

The remarks in 2.1 suggest that the model with exponential arrivals, Erlangian service, and several channels can be fitted well to many real life situations. A method for treating this model is presented by J.Mayhugh and R.McCormick (62):

The service at each channel is considered to consist of k stages: each of which takes a time distributed exponentially with parameter μ/k. The state of the system at any time can be defined by the symbol n: $x_k\ x_{k-1} \cdots\cdots x_1$; where .n is the number in the system;

. x_k is the total number, summed across the channels, in the first stage of service;

. x_{k-1} is the total number at the second stage;

. etc.

These states are ordered in a certain way, and indexed. Where j is the index of some state, p(j,t) is the probability that the system is in state j at time t. Differential-difference equations are formed. When the existence of equilibrium state-probabilities p_j is assumed, these equations yield a set of "initial" equations and a set of "cyclic" equations. Each of these equations is a relation between several of the p_j. Much matrix algebra produces a solution for the p_j. These can be summed to give the usual steady-state queue length probabilities.

Mayhugh and McCormack deal with the systems $M/E_3/2$, and $M/E_3/3$, then the general system $M/E_k/c$.

The System GI/M/c.

The geometric property of P_n, for $n > c$, survives when the assumption of exponential arrivals is removed. Cox and Smith consider the system as a Markov process by observing it at the instants when customers arrive. (23, 69).

Let A(x) be the arrival distribution, and at equilibrium let:

p_{ij} = Pr [(n+1)th customer finds j customers ahead of him/ nth customer found i ahead of him];

= Pr [i+1-j customers were served during the arrival interval]

$$= \int_{x=0}^{\infty} e^{-c\mu x} \frac{(c\mu x)^{i-j+1}}{(i-j+1)!} \, dA(x) = W_{i-j+1}, \text{ say; for } j \geq c+1.$$

Let h_k be the mean number of occurences of $n = k+1$, between successive occurrences of $n = k$ (where n is the queue length at a regeneration point). n can increase by no more than 1; and p_{ij} depends only on i-j. Hence if n reaches k+1, the

probability that it returns to $k+1$ before returning to some $n \leq k$ is independent of k; for $k+1 \geq c$. Thus h_k is independent of k, and $h_k = h$, say. But $h = P_{n+1}/P_n$, so $P_n = \text{const.}h^n$, for $n \geq c-1$. The writers also show that $h = \rho$.

2.3 WAITING TIMES.

This section

.treats the steady state M/M/c system fully;

.discusses the more general systems M/G/c, and GI/M/c;

.treats the formula $\bar{n} = \lambda \bar{w}$;

.then, since methods for dealing with waiting times in multi-channel systems do not go very far, discusses waiting time methods in single channel systems.

2.3.1 THE SYSTEM M/M/c.

The steady state distribution of waiting time (including service) can be deduced from the P_n found earlier. The system is observed at the arrival of each new customer. If the new customer finds n, number waiting excluding himself, $< c$, he goes straight into service. Thus $\Pr[\text{waiting time} = 0] = \sum_{n=0}^{c-1} ((c\rho)^n/n!)P_0$; where P_0 has the value given in 2.2.1. If he finds $n = c$, he has to wait for one service to be completed. If $n < c$, he has to wait one partial service time, and n-c full service times. Since the service times are exponential, the partial service times have the same distribution as full service times.

We will show by induction that the sum of k exponential variables has an Erlangian distribution. That is; where $w = t_1 + t_2 + \cdots + t_k$, and each t_i has density function $c\mu e^{-c\mu t_i}$, that the density function of w is given by

$$g_k(w) = (c\mu (c\mu w)^{k-1} e^{-c\mu w})/(k-1)!$$

For $k = 1$, $g_k(w)$ reduces to the exponential function.

Let $S = t_1 + + + t_{k+1}$.

$= w + t_{k+1}$; then

$$g_{k+1}(s) = \int_{0}^{s} \underset{(s=w+t_{k+1})}{c\mu \frac{(c\mu w)^{k-1}}{(k-1)!}} e^{-c\mu w} c\mu e^{-c\mu t_{k+1}} dw ;$$

$$= \int_0^s \frac{c\mu (c\mu w)^{k-1}}{(k-1)!} dw \; c\mu e^{-c\mu s} ;$$

$$= c\mu \frac{(c\mu s)^k}{k!} e^{-c\mu s} .$$

The density function for $w > 0$ is given by:

$$f(w) = P_c \cdot g_1(w) + P_{c+1} \cdot g_2(w) + + +$$

$$= P_c \cdot c\mu ((c\mu w)^0/0! + (c\mu w)^1/1! + + +) e^{-c\mu w}$$

$$= e^{-c\mu(1-\rho)w} c\mu P_c , \text{ for } w > 0.$$

Some further results follow:

$$\Pr[w > W] = \int_W^\infty f(w)\, dw = e^{-c\mu(1-\rho)W} . P_c/(1-\rho) ;$$

$\Pr$ [new arrival does have to wait]

$= \Pr[w > 0] = P_c/(1-\rho)$.

$$\text{Mean waiting time} = 0.\Pr[w = 0] + \int_{0+}^{\infty} wf(w)\, dw$$

$$= P_c/(c\mu(1-\rho)^2)$$

$$= (1/\lambda) . P_c/((c\mu/\lambda)(1-\rho)^2)$$

$$= (1/\lambda) . P_c/(1-\rho)^2$$

So mean queue length $= \lambda$.mean waiting time.

2.3.2 MORE GENERAL SYSTEMS.

The System M/G/c

The integrodifferential equation for w in M/G/1 developed by Takacs and Deschamps can be extended for multichannel systems in the steady state. Saaty's treatment (76, 203) is as follows.

b(t) is the density function for service time at each channel; Let P(w) be the cumulative density function for waiting time w.

Consider a large number, N, of identical M/G/c systems. At time t, we separate out a set S of these systems : a system is

in S if a new arrival finds that his waiting time is $\geq w$. We must choose w and a time interval dt such that $w > dt > 0$. Between t and t+dt, any one of the N systems can experience either : no arrival; waiting time for a new arrival drops by dt, and the system leaves S if its waiting time at t was in (t,t+dt). We call this Case A;

or: one arrival.... we call him Alf, for reference. At t, the system was in one of these 3 states:

1. $n < c-1$. Alf's waiting time is 0: the next arrival, Bert, has no waiting time either: the system can not enter S.
2. $n = c-1$. Alf's waiting time is 0; the system joins S with probability H(w), where

   $H(y) = \Pr[$ Alf's channel is occupied for time $> y$; and the other c-1 channels finish their current service in time $> y]$.

Thus the system joins S if Bert's waiting time is $> w$.

We call this Case B.

3. All channels are occupied. The system joins S if Bert's waiting time is $\geq w$; ie, if ((Alf's waiting time) + (time from Alf's start of service to time of Bert's start of service)) $\geq w$.

   This is Case C.

or; there are several arrivals. This possibility can be ignored since we shall let $dt \to 0$.

Since we are considering the equilibrium case, (number of systems leaving S ...Case A) = (number of systems joining S... Case B, Case C) .

Case A:

$\Pr[$ at t, waiting time is in $(w, w+dt)] \to \frac{dP(w)}{dw}\, dt$, as $dt \to 0$;

so the number of systems joining S

$$\to N. \frac{dP}{dw} .dt$$

Case B:

Consider any channel apart from the one serving Alf. The mean length of service times longer than y is $\int_y^\infty (x-y)\, b(x)\, dx$,

$$= \int_y^\infty B'(x)\, dx$$

where $B'(x) = 1 - B(x)$. The mean length of service is $1/\mu$. Thus, the fraction of systems in which the channel considered is still occupied after a time y is

$$\int_y^\infty B'(x)\, dx \,/(1/\mu) \quad .$$

Then $H(y) = B'(y).(\mu \int_y^\infty B'(x)\, dx)^{c-1}$.

The number of Case B systems then is

N. (proportion with c-1 channels occupied at t).(proportion which receive an arrival in dt).H(w) ;

$= N.P_{c-1}.\lambda dt.H(w)$.

Case C:

Number of systems involved is

N. (proportion receiving an arrival during dt).(Pr [Alf's waiting time x + time y for next channel to clear $\geq$ w, and $x < w$]

$$= N.\lambda dt. \int_{\substack{0+ \\ x+y \geq w}}^{w} \frac{dP(x)}{dx} H(y)\, dx.$$

The integrodifferential equation for P(w) in the steady state is found on dividing by N dt:

$$0 = \frac{dP(w)}{dw} - \lambda P_{c-1}.H(w) - \int_0^w \frac{dP(x)}{dx} H(w-x)\, dx \quad .$$

H(w) can be found when b(x) is decided on; $P(0), P_{c-1}$ have to be found by other means . The Laplace transform of the equation gives P* as a function of H*, P_{c-1}, P(0), λ .

When $B'(x) = e^{-\mu x}$, the M/M/c solution for P(w) fits the integrodifferential equation, when H(y) has been replaced by $e^{-c\mu y}$.

Kendall has shown (46) that when this system is observed at the instants when customers arrive ($t_1, t_2, \ldots$), an imbedded Markov chain is discovered. Let $Y(t_i)$ be the number of customers found by an arrival at t_i to be ahead of him. Since the Y's depend on the memoryless property of the service-time distribution, the chain $Y(t_1)$, $Y(t_2)$ ..... is Markovian.

By using Feller's theory (28, 372) of denumerable Markov chains, Kendall finds that the geometric distribution of n (for $n > c$), and therefore the exponential distribution of w, survives when arrivals are not exponential. He states that " the limiting queue length distribution is a geometric series, save for its first (c-1) terms; the common ratio being the unique root of

$$\int_0^\infty e^{-(1-\lambda)\, c\mu x}\, dA(x) = \lambda \quad ; \quad \text{with} \quad 0 < \lambda < 1;$$

where A(x) is the cumulative distribution of interarrival times. Thus, where q is the number waiting;

$$\Pr\,[q = m / m > 0] = (1-\lambda)\lambda^{m-1}. \text{ "}$$

Since service times are exponential, the method used to find P(w) from P_n for M/M/c can be used; for $w > 0$,

$$P(w) = c\mu(1-\lambda)\, e^{-w(1-\lambda)c\mu} .$$

When $\alpha e^{-\alpha x} dx$ replaces dA(x), these equations yield the usual M/M/c results. For example; we find that Kendall's $\lambda = \rho$:

$$\alpha \int_0^\infty e^{-((1-\lambda)c\mu + \alpha)x}\, dx = \lambda$$

$$\text{ie, } \alpha = \lambda.((1-\lambda)c\mu + \alpha)$$

$$\text{ie, } \lambda = \alpha / c\mu .$$

Kendall's paper concludes with detailed results for M/M/c and D/M/c with c = 1,2,3.

2.3.3 AVERAGE WAITING TIME AND AVERAGE QUEUE LENGTH. (1, 259)

The following discussion is incorporated because the formula $\bar{n} = \lambda \bar{w}$ is just as applicable to multi-channel queues as to simpler ones. When :

$\bar{n}$ = average queue length over a time interval T,

$\bar{w}$ = average waiting time (including service time) over T, and

λ = average arrival rate during T,

the formula holds exactly for any busy period; and approximately for any long period. However, if λ = theoretical expected arrival rate, the equation will not hold exactly.

The formula is proved for the general case thus:

Consider a time interval (O,T) with n = O at both ends. Let the number of arrivals during T be N. Let $g_i(t)$

$$= \begin{cases} 1 & \text{from arrival to departure of ith customer} \\ 0 & \text{elsewhere.} \end{cases}$$

Then mean arrival rate = $n/T = \lambda$;

Mean waiting time per customer = $\bar{w} = \sum_{i=1}^{N} \int_0^T g_i(t)\, dt \,/N$;

Mean number in system = $\bar{n} = \int_0^T \sum_{i=1}^{N} g_i(t)\, dt/T$

$$= \bar{w}\, N/T = \bar{w}\lambda .$$

W.Maxwell (61) points out that if the g_i are replaced by other functions for which

$\int \Sigma = \Sigma \int$, the relation between the average values still holds.

If $g_i(t) = \begin{cases} 1 & \text{while customer i waits} \\ 0 & \text{during his service, and otherwise,} \end{cases}$

the formula becomes

(Average number waiting, not in service) = λ.(average waiting time before service).

If $g_i(t) = \begin{cases} g_i & \text{while i is in the system} \\ 0 & \text{otherwise} , \end{cases}$ (61)

we have given each customer a value or cost. $\bar{n}$ is now the mean cost content of the system, and $\bar{w}$ is the mean cost per customer.

If $g_i(t) = \begin{cases} G_i + g_i(t-r_i) & \text{from i's arrival at } r_i, \text{ to departure} \\ 0 & \text{otherwise;} \end{cases}$

we have assigned a more complex cost for each customer in the queue.

2.3.4 OTHER SYSTEMS.

The System M/G/1.

A consideration of waiting time analagous to Saaty's method above for M/G/c leads to a similar integrodifferential equation- that of Tacaks (76, 198). The Laplace transform of this leads to a differential equation of first order in time, for the transform of P(w,t). The only unknown term is P(0,t).

This system is extended by Gnedenko (76, 207) to cover systems where the customer may depart before the end of his service.

Kendall (46) points out that if we consider the number of customers left behind at a departure, we discover a Markov process. Its existence can be ascribed to the exponential input.

Another approach to the M/G/1 system yields the Pollaczek-Khintchine formulas for E[n] and E[w] (w here excludes service time); for the steady state. Again, the system is observed as customers leave:

Let n be the number left when a certain customer leaves; and let n' be the number left when the next customer leaves. Let r be the number of arrivals during this service, and let

$\delta(n) = \begin{cases} 0 & \text{if } n > 0 \\ 1 & \text{if } n = 0. \end{cases}$

Assume that $E[n]$, $E[n^2]$ are finite. Then

$$n' = n - 1 + s + r \qquad (76, 40), (84, 15).$$

We take expected values of this equation and of the equation got by squaring it. We assume that the system is in a steady state, and use the fact that arrivals in any time interval are poisson to find $E[r]$ and $E[r^2]$. The result is:

$$E[n] = \rho + (\rho^2 + \lambda^2 \operatorname{Var}[t]) / 2(1-\rho) .$$

For multichannel systems, the equation means that

$$E[\text{number in system}] = E[\text{number of channels occupied}] + E[\text{number waiting}]$$

$E[w]$ is found by considering that

$E[\text{number of arrivals during a customer's entire stay}]$

$= E[\text{number in system at his departure}]$; so:

$\lambda . (E[w] + 1/\mu) = E[n]$. Thus

$E[w] = (\rho^2 + \lambda^2 \operatorname{Var}[t]) / 2\lambda(1-\rho)$

$= \lambda E[\text{number waiting}]$. Thus

$E[\text{waiting time}] / E[\text{service time}] = (1 + V^2) \rho / 2(1-\rho)$;

where V = coefficient of variation of t. The smaller this ratio, the more efficient is the system. When service times are regular, V = 0, and efficiency is highest for a particular ρ.

These three versions of the Pollaczek-Khintchine formula are of great practical value; they indicate the effects on a system of changing its traffic intensity ρ, and the variance of its service-time distribution.

The System G/M/1.

W. Smith (83) shows that " the service time distribution exerts a strong influence over the . ..distribution of waiting time." From Lindley's integral equation he develops several theorems; we state the one which applies to the G/M/1 system: " If the service time is distributed exponentially,

so is the waiting time (which here includes service time); whatever the arrival distribution."

As shown above, Kendall extends this result to multichannel systems.

The System GI/G/1.

D. Lindley (58) considers the variable

u_i = (service time of i th customer) - (time from arrival of i th, to arrival of (i+1)th).

The distribution g(u) can be found from a(t) and b(s); the waiting time (excluding service) distribution, F(w), is found to depend only on g(u).

When $E[u] < 0$, a limiting distribution for w exists, and satisfies this Wiener-Hopf integral equation:

$$F(w) = \int_0^\infty F(y)\, g(w-y)\, dy.$$

Unfortunately this equation is not easy to solve. Lindley finds solutions for the special cases M/G/1, E_k/G/1, D/E_k/1.

Kendall (46) points out that the G/G/1 system yields a Markov process if we examine the waiting time of customers as they begin service.

The System GI/G/c

J. Kiefer and J. Wolfowitz (47) reduce this system to a random walk in c -space, find an integral equation for steady-state waiting times, and prove that a solution exists, which, for $\rho < 1$, is a probability distribution. Again, the equation is difficult to solve for the various special cases.

2.4 A SIMPLE APPLICATION.

This section takes the simplest possible multi-channel situation (poisson arrivals; two exponential servers; steady state) and

1. gives some numerical results;
2. discusses some customer strategies.

The reasons for including it are that it gives numerical meaning to a small part of the theory above; and by an elegant extension of the probability methods used above for the steady M/M/c system, it gives some surprising results for some customer strategies. Customer strategies are of special interest, since nobody spends more time designing queue systems than waiting within them.

Throughout this section, λ is the total arrival rate; 2μ is the total service rate; and $\rho = \lambda/(2\mu)$. Only steadystate solutions are considered.

2.4.1 THE M/M/2 SYSTEM COMPARED WITH TWO M/M/1 SYSTEMS.

It is important for later results to show that a poisson stream of arrivals, with rate λ, in which each arrival chooses independently, with equal probabilities (ie, 1/2 each) between 2 queues, results in 2 poisson streams, each with rate $\lambda/2$.

Let $P_n(t) = \Pr[$ n arrivals in time t $]$ for the stream with rate λ; and let $p_n(t)$ be the same probability for either of the two lesser streams. Assume that the original stream is poisson;

then $P_n(t) = (\lambda t)^n e^{-\lambda t}/n!$

Consider one of the two streams which arise after the division. n arrivals for this stream can occur in t if:

.n occur in the first stream, and they all enter the stream being considered; or

.n+1 occur in the first stream; n of these enter the stream considered;

.and so on , to ∞ .

Then $p_n(t) = P_n(t)(1/2)^n + P_{n+1}(t)(1/2)^{n+1}(n+1)$

$$+ + + \quad + P_{n+r}(t)(1/2)^{n+r}\frac{(n+r)!}{n!r!} \quad + + +$$

$$= (\lambda t/2)^n e^{-\lambda t}/n! + \cdots + (\lambda t/2)^{n+r} e^{-\lambda t}/(n!\, r!)$$

$$= \frac{(\lambda t/2)^n e^{-\lambda t}}{n!} \left(1 + \frac{(\lambda t/2)^1}{1!} + \cdots \right)$$

$$= (\lambda t/2)^n e^{-\lambda t/2} / n!$$

So $p_n(t)$ is poisson, rate $\lambda/2$.

The converse is that 2 merging poisson streams form another poisson stream. Assume $p_n(t)$ is poisson, rate $\lambda/2$. Then

$$P_n(t) = p_0(t)p_n(t) + p_1(t)p_{n-1}(t) + \cdots + p_n(t)p_0(t)$$

$$= (\lambda t/2)^n e^{-\lambda t} \left(\frac{1}{0!n!} + \frac{1}{1!(n-1)!} + \cdots + \frac{1}{n!0!} \right)$$

$$= (\lambda t/2)^n e^{-\lambda t}\, 2^n/n!$$

This identity between double and single poisson streams is essential for the comparison of two M/M/1's with one M/M/2.

Two M/M/1's

Customers arrive in two independent poisson streams at rate $\lambda/2$, to two M/M/1 systems which work independently. Each has service rate μ. For each system,

$$P_n = \rho^n(1-\rho) \; ; \; n = 0, 1, 2.....$$

where $\rho = (\lambda/2)/\mu$.

Let $P_{m,n} = \Pr[\text{m in first system, n in second}]$.

Then $P_{m,n} = \rho^m(1-\rho)\rho^n(1-\rho)$.

Let $q_r = \Pr[\text{r customers in combined system}]$;

Then $q_r = \sum_{m+n=r} P_{m,n} = (r+1)\rho^r(1-\rho)^2 \; ; \; r = 0, 1, 2.....$

Mean number in combined system

$$= E[m+n] = E[m] + E[n] = 2\rho/(1-\rho).$$

One M/M/2

The arrival distribution is the same, but this time the customers form one queue before the two channels. Putting $c = 2$ in the functions of 2.2.1 yields:

$q_1 = 2\rho q_0$,

$q_r = (2^2/2!)\rho^r q_0$ for $r \geq 2$.

Since $q_0 + q_1 + q_2 + \ldots = 1$,

$$q_0 = (1-\rho)/(1+\rho) .$$

Mean number in system

$$= 2\rho/(1-\rho^2)$$

$$= (\text{mean number for 2 M/M/1's})/(1+\rho).$$

Comparisons.

For 2 M/M/1's; and for 1 M/M/2,

$$\begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2\rho \\ 3\rho^2 \\ 4\rho^3 \end{bmatrix} . (1-\rho)^2 \quad \text{and} \quad \begin{bmatrix} 1 \\ 2\rho \\ 2\rho^2 \\ 2\rho^3 \end{bmatrix} . (1-\rho)^2/(1-\rho^2).$$

q_0, q_1, are larger for the second system; q_r remains larger for this system for $r < (1+\rho^2)/(1-\rho^2)$.

Mean number in system is smaller for the M/M/2 system by a factor between 1 and 2; hence mean waiting time is smaller by the same factor. M/M/2 clearly is preferable.

2.4.2 JOCKEYING

In studying jockeying, Koenigsberg (56) considers two channels at which service rates may differ. The four strategies considered here produce results similar to those for the two systems just discussed. The proof of his first result is outlined; the other results are stated.

Two M/M/1's; Arrivals Joining Shorter Queue.

Arrivals form a poisson stream, rate λ , and;

.if the queue lengths are equal, the arrival goes to either queue with equal probability;

.if the lengths differ, he joins the shorter queue.

Six equations (of which Koenigsberg states four) are required to cover all cases:

$m = 0, \quad n = 0$
$m = 0, \quad n = 1$
$m = 1, \quad n = 0$
$m > 1, \quad n = 0$
$m = 0, \quad n > 1$
$m > 1, \quad n > 1$

For example, the sixth is

$$\frac{d}{dt} P_{m,n} = -(\lambda + \mu_1 + \mu_2) P_{m,n} + D_1 P_{m-1,n} + D_2 P_{m,n-1} + \mu_1 P_{m+1,n} + \mu_2 P_{m,n+1}$$

where $D_1 = \begin{cases} 0 ; & m-1 > n \\ \lambda/2 ; & m-1 = n \\ \lambda ; & m-1 < n . \end{cases}$

and D_2 is defined similarly.

Generating functions are defined by

$$g_n(x) = \sum_{m=0}^{\infty} x^m P_{m,n} ;$$ and by a double summation, an equation in these is constructed. Geometric functions:

$$g_n(x) = a_n(1 + \rho x + (\rho x)^2 + \cdots) \quad ; \rho = \lambda/(\mu_1 + \mu_2)$$

satisfy this equation, when the a_n are also geometric. Thus

$$P_{m,n} = \rho^m \rho^n (1 - \rho)^2 ;$$

This is independent of μ_1/μ_2, and is the same as for two M/M/1's. However, the mean waiting times for the two queues do depend on μ_1, μ_2.

Two other strategies produce the same results:

Probabilistic Jockeying.

Arrivals again join the shorter queue; whenever one queue is longer, customers leave it at rate: (k.(difference between lengths of waiting lines)). The result is independent of k; there is no change in the equilibrium solution.

Lane Changing.

Each queue has its own input, with its own arrival rate. Probabilistic jockeying occurs as above.

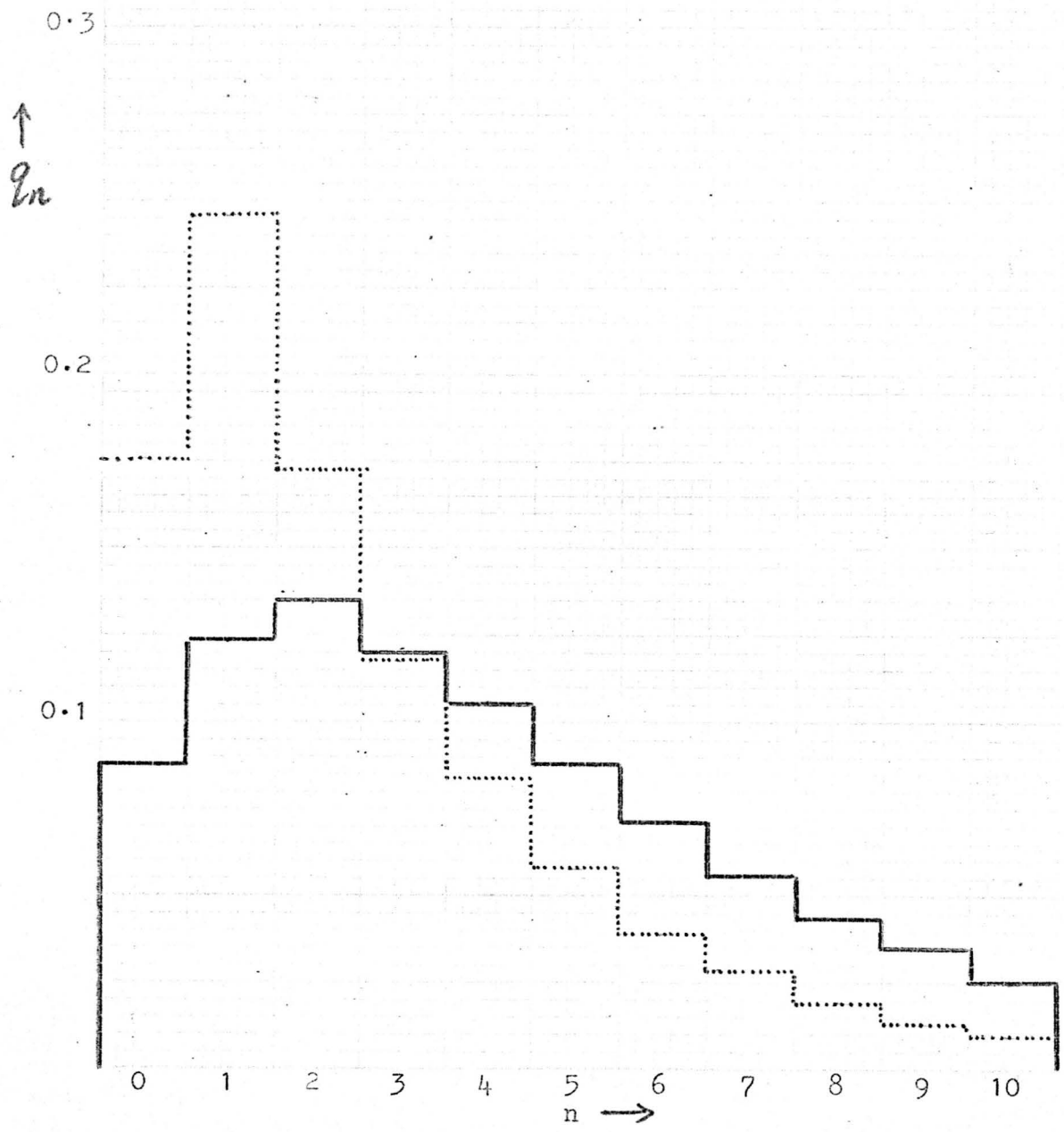
When $\mu_1 = \mu_2$, these three cases all have the same result as the case with two M/M/1's. Koenigsberg points out that in all these cases, one channel can be idle while customers

are waiting; the channels are not being fully utilised. They are fully utilised in the following case:

Instantaneous Jockeying.

Again customers join the shorter of two M/M/1's. The last customer on either queue jockeys to the shorter queue on the instant it becomes shorter by more than 1. The results for mean number in system and mean number waiting, are identical with those for the M/M/2 system; which means that they are optimal.

Probabilities for Number in System.



——— an M/M/2 with arrival rate λ,
service rate μ at each channel, $\rho = \cdot 7$.

----- 2 M/M/1's in parallel with arrival rate $\lambda/2$,
service rate μ, at each, $\rho = \cdot 7$.

CHAPTER 3.

QUEUES IN SERIES.

The chapter studies output distributions, then series systems with various assumptions as to arrival, service, size of waiting room, number of customers.

CHAPTER 3 : QUEUES IN SERIES.

3.1 INTRODUCTION.

Some Terminology.

A set of parallel service channels, each providing the same service, will be referred to as a "service station" ; "queue length" will mean the total number of customers waiting before a station, plus the number in service there.

Outline Of the Chapter.

This introductory section mentions some features of the theory of series systems. The study of output distributions in Section 2 is essential for much of what follows. The study of the two-stage systems in Section 3 exemplifies some of the approaches available. Section 4 studies methods for many stations in series. The simple system treated in Section 5 should give numerical meaning to some of the foregoing theory.

The Development of Series Queue Theory.

In many real-life situations, a customer is served at one station, and departs only to queue before another station. The customer may have to proceed along a series of many stations.

A series system is more prone to congestion than a single-stage system; but it is more sensitive to changes in design; and there are more opportunities for changes. The theoretician can suggest little which is not obvious to improve the performance of a single-stage system; with a series of stations he has more scope. For example, he can pinpoint where waiting room is most needed, or find the most efficient ordering for the stations. Most of the papers cited in this chapter arose from some practical situation.

This practical bias shows its influence in several ways. In the late 1950's there appeared the elegant theorems on

poisson series systems. Later (in the 1960's) we find treatments of systems with more realistic service distributions and with limited waiting room between stages. Work on systems with general distributions has not been carried far. Though transient solutions and rates of reaching equilibrium are of interest to the practical worker, his first interest will be in the more manageable steadystate solutions. Thus little work has appeared on transient solutions.

A few of the results in this chapter can be deduced from the results for networks. These results are presented here because they lead up to the more general results, and because series systems have considerable practical importance themselves. Theoretically, a series system is distinguished from the rest of the network systems by the fact that there is the same arrival point, routing, and departure point for all customers.

Applications.

Serial arrangements of processors occur in several different guises within computing systems. Other real-life situations which have provoked theoretical development are:

.industrial production lines, where several different operations are required;

.inspection or repair of vehicles or other machinery;

.flow of vehicles through a series of intersections;

.movement of freight cars in a railway yard;

.message traffic in a relay system;

.several-stage service in retail shops;

.the receptionist's room and the doctor's surgery;

.hospitals.

3.2 OUTPUT DISTRIBUTIONS.

When customers leaving a service station immediately join the queue before another station, the output distribution of the first station becomes the arrival distribution for the second. In studying series and nets of queues two items are of key importance:

.the nature of the output distribution from each service station;

.the conditions for two adjacent stations to operate, in the limit, independently.

This section studies methods of finding output distributions; and conditions under which independence is attained. The theory for systems more general than M/M/c seems restricted to cases for which $c = 1$. Some results for these cases are given.

3.2.1 THE SYSTEM M/M/c.

Bourke's Results.

Before Bourke's paper (10) of 1956, O'Brien (67) and others had stated or assumed that the output of an M/M/1 system was exponential, In the paper cited, Bourke shows that for the steadystate M/M/c system with no defections and any queue discipline, the number of departures in an arbitrary time interval has the same distribution as the number of arrivals in the same interval. In other words, each departure interval is exponential, and independent of all earlier departure intervals.

Bourke points out that this fact simplifies the theory of series systems; each station can be treated separately. His paper treats, as an example, the system comprising a set of identical channels (sales clerks) from which customers must go to a second set of channels (cashiers).

His proof follows: it deals with an M/M/c system.

T is chosen arbitrarily and will represent an interval between two departures.

Let $k(t)$ = queue length at time t .

Let $F_n(t) = \Pr\left[t \leq T, \text{ and } k(t) = n\right]$.

The initial conditions on the F_n are

$$F_n(0) = P_n \ ;$$

where P_n is as defined in 2.2.1 above.

The events associated with the probability densities $F_n(t)$ exclude events involving a departure, so:

$$F_o(t+dt) = (1 - \lambda dt)\, F_o(t) \ ;$$

$$F_n(t+dt) = (1 - \lambda dt)(1 - n\mu dt)\, F_n(t) + \lambda dt(1 - (n-1)\mu dt)\, F_{n-1}(t) \ , \ n \leq c;$$

$$F_n(t+dt) = (1 - \lambda dt)(1 - c\mu dt)\, F_n(t) + \lambda dt\,(1 - c\mu t)\, F_{n-1}(t) \ , \ n > c.$$

Then $\dot{F}_o(t) = -\lambda F_o(t)$; ①

$$\dot{F}_n(t) = -(\lambda + n\mu)\, F_n(t) + \lambda F_{n-1}(t) \ , \ n \leq c; \quad ②$$

$$\dot{F}_n(t) = -(\lambda + c\mu)\, F_n(t) + \lambda F_{n-1}(t) \ , \ n > c. \quad ③$$

From ① ;

$$F_o(t) = F_o(0)e^{-\lambda t} = P_o\, e^{-\lambda t} \ .$$

In ② and ③ , we try solutioms of the form

$$F_n(t) = K_n e^{-\lambda t} \ ;$$

and get :

$$-\lambda K_n = -\lambda K_n - n\mu K_n + \lambda K_{n-1} \ , \ n \leq c$$

ie, $K_n = (\lambda/n\mu)\, K_{n-1} \ , \ n \leq c;$

and similarly,

$$K_n = (\lambda/c\mu)\, K_{n-1} \ , \ n > c.$$

Thus $F_n(t) = (\lambda/\mu)(\lambda/2\mu).....(\lambda/n\mu)\, P_o e^{-\lambda t}$

$$= P_n\, e^{-\lambda t} \ , \ n \leq c.$$

The same holds for $n \geq c$.

When departure time T has density function $f(T)$,

$$\int_t^\infty f(T)dT = Pr\left[T > t\right] = Pr\left[t > T, \text{ and } n = 0 \text{ or } 1 \text{ or } 2\ldots\right]$$

$$= \sum_{n=0}^{\infty} F_n(t)$$

$$= e^{-\lambda t}$$

so $f(T) = \lambda e^{-\lambda T}$;

which is the distribution for interarrival times.

Bourke also proves that the state of the system just after a departure is independent of the preceding departure time. We factorize the following probability:

$$Pr\left[k(T+0) = n-1, \text{ and } T \in (t, t+dt)\right]$$

$$= \begin{cases} F_n(t).n\mu dt & , n = 1,2,\ldots.c \\ F_n(t).c\mu dt & , n \geq c ; \end{cases}$$

$$= \begin{cases} \lambda e^{-\lambda t} dt\ (n\mu/\lambda)\ P_n & , \quad n \leq c; \\ \lambda e^{-\lambda t} dt\ (c\mu/\lambda)\ P_n & , \quad n > c \end{cases}$$

$$= \lambda e^{-\lambda t}.P_{n-1} .$$

Hence the departure interval beginning at time t depends only on the state of the system at $t+0$. This state is independent of the departure interval which finished at t, and therefore of all earlier departures.

Reich's Results.

By a different argument Reich (73) reaches the same results; then proves some further results.

He first finds a condition for a certain class of markov processes to be reversible, and hence shows that a birth-death process with state-dependent rates is reversible. If the input to the process is not state-dependent, its output is a poisson process.

His theorem for an M/M/c system is:

(a) the sequence of departure times forms a poisson process;

(b) queue length is independent of all past departure times;

(c) if t_o is a departure time, $k(t_o + 0)$ is independent of all past departure times.

Reich then finds two limitations on generalisations of the theorem. The first is that though the output of an $E_j/E_j/c$ system is E_j for $j = 1$ (i.e., E_j is equivalent to M) or $j = \infty$ (i.e., E_j is equivalent to D); the same does not hold for all j. His counter-example is the $E_2/E_2/1$ system. The second is that " if the arrival and departure epochs of a single-counter queue are both poisson, then the service time is exponential, or a step function at 0 ."

3.2.2 SINGLE-CHANNEL SYSTEMS

The System M/G/1.

Finch (29) finds the same results as Reich. These two conditions on an M/G/1 system:

.service times are exponential, and

.infinite waiting room is available;

are necessary for the independence of departure time and queue length at the end of the departure time; and for two adjacent departure intervals to be independent.

M.G.Fs of Output Distributions.

The output distributions of systems more general than M/M/1 will not be exponential, but their MGFs can be found simply in some cases. Makino (59) finds MGFs for the systems M/G/1, $E_k/M/1$, and $E_2/E_2/1$. His methods are illustrated by the following treatment of the M/G/1 case.

Let f(t), g(t), h(t) be the density functions for arrival times, service times and departure times respectively.

Let their MGFs be $M_f(z)$, $M_g(z)$, $M_h(z)$. λ and μ are arrival and service rates; $\rho = \lambda/\mu$.

$P_o = \Pr[\text{queue length} = 0]$.

$f^*g(t)$ = convolute of $f(t)$ and $g(t)$.

At equilibrium, arrival rate = departure rate. MGFs have these properties:

. $M_{f^*g}(z) = M_f(z).M_g(z)$;

. $M_f(0) = 1$;

. $M'f(0) = E[t]$;

. when f is exponential, rate λ , $M_f(z) = \lambda/(\lambda - z)$.

A departure leaves the system empty with probability P_o; time to next departure is the sum of an arrival time and a service time. A departure leaves the channel occupied with probability $(1 - P_o)$; time to next departure is a service time.

Thus $h(t) = P_o.f^*g(t) + (1 - P_o).g(t)$.

The MGFs must therefore obey this relationship:

$$M_h(z) = P_o.M_f(z)\, M_g(z) + (1 - P_o)M_g(z).$$

Differentiate, and let $z = 0$:

$$1/\lambda = P_o(1.(1/\lambda) + (1/\mu).1) + (1 - P_o)(1/\mu).$$

This yields

$$P_o = 1 - \rho.$$

With f exponential,

$$M_h(z) = ((1 - \rho)\lambda/(\lambda - z) + \rho)\, M_g(z)$$

$$= \frac{\lambda}{\mu}\,\frac{\mu - z}{\lambda - z}\, M_g(z) .$$

<u>The System M/M/1</u>.

This result is appended for its simplicity:

$$h(t) = \rho f^*g(t) + (1 - \rho)g(t)$$

$$= \lambda e^{-\lambda t}.$$

3.3. TWO QUEUES IN SERIES.

Much of the theory for two stations in series follows from the theory of output distributions. The first part of this section discusses the steadystate behaviour of two multi-channel poisson systems in series. Other multichannel results are special cases of results presented in the next section, 3.4. The second part of this section outlines the development of the theory for two single-channel stations in series, along several different lines. The concepts of limited waiting space, and limited numbers of customers, are of considerable practical importance.

3.3.1 TWO M/M/c SYSTEMS.

Arrivals occur exponentially at rate λ, and are served at a station with c_1 parallel channels, each with exponential service rate μ_1. Departing customers immediately queue before a second station, of c_2 channels. Each of these has exponential service, rate μ_2.

$\rho_i = (\lambda / c_i \mu_i)$, for $i = 1, 2$.

Reich uses his theorem (3.2.1 above) to show that the two queue lengths are independent. Let $m(t)$, $n(t)$ be the two lengths at time t. $n(t)$ depends on departure intervals which finished at $T \leq t$; but $m(t)$ is independent of these. Thus $m(t)$, $n(T)$ are independent for $T \leq t$.

Since the theory for two multichannel queues in series with infinite waiting room has not been developed past poisson queues, the full derivation is given here for the joint distribution of equilibrium queue lengths in the $M/M/c_1 \ldots /M/c_2$ system. This was treated first by R.R.P.Jackson (41) for $c_1 = c_2 = 1$.

Let $P_{m,n}$ = Pr [length of first queue = m, and length of second = n; in equilibrium].

We will allow ourselves to suspect independence, and so will try for a solution factorizable into functions of m and n:

$$P_{m,n} = M_m \cdot N_n \quad .$$

In a small time interval dt, three transitions are possible: an arrival; the transfer of a customer from the first to the second queue; a departure from the system. These possible events give rise to 9 equations. For (2) .... (9), only the coefficients printed in round brackets for (1) are given. The rest of (2) .. (9) is identical with (1). When $P_{m,n}$ has negative arguments, its coefficient turns out to be zero.

m	n	$\{\lambda(1)+\mu_1(c_1)+\mu_2(c_2)\}P_{m,n} =$	$\lambda(1)P_{m-1,n}$	$+\mu_1(c_1)P_{m+1,n-1}$	$+\mu_2(c_2)P_{m,n+1}$	
$\geq c_1$	$\geq c_2$	1	m_1	c_2	1	$m+1$ $n-1$ c_2 (1)
$>0, <c_1$	$\geq c_2$	1	c_1	n	1	c_1 $n+1$ (2)
$\geq c_1$	$>0, <c_2$	1	m	n	1	c_1 c_2 (3)
$>0, <c_1$	$>0, <c_2$	1	0	0	0	0 1 (4)
$=0$	$=0$	1	0	n	0	1 $n+1$ (5)
$=0$	$>0, <c_2$	1	0	c_2	0	1 c_2 (6)
$=0$	$\geq c_2$	1	m	0	1	0 1 (7)
$>0, <c_1$	$=0$	1	c_1	0	1	0 1 (8)
$\geq c_1$	$=0$	1				(9)

Replacing $P_{m,n}$ by $M_m \cdot N_n$ in (5) gives :

$$N_1 = (\lambda / \mu_2)\, N_0;$$

or $\mu_2 N_1 - \lambda N_0 = 0;$ (10)

and in (8), with $n = 1$;

$$(\lambda + \mu_1)\, M_1 N_0 = \lambda M_0 N_0 + \mu_2 M_1 N_1 \qquad (11)$$

(10) and (11) yield : $N_1 = (\lambda / \mu_2)\, M_0$. (12)

(6) becomes, for $0 < n < c_2$;

$$(\lambda + n\mu_2)\, M_0 N_n = \mu_1 M_1 N_{n-1} + (n+1)\mu_2 M_0 N_{n+1} \cdot$$

(12) eliminates M_0 and M_1 :

$$(\lambda + n\mu_2)\, N_n = \lambda N_{n-1} + (n+1)\mu_2 N_{n+1} .$$

This rearranges to:

$$n\mu_2 N_n - \lambda N_{n-1} = (n+1)\mu_2 N_{n+1} - \lambda N_n . \quad (13)$$

Similar treatment of (7) gives another familiar equation.

For $n \geq c_2$;

$$c_2\mu_2 N_n - \lambda N_{n-1} = c_2\mu_2 N_{n+1} - \lambda N_n . \quad (14)$$

(10) , (13) , and (14) combined indicate that the left sides of these three equations are all zero; and so

$$N_n = \begin{cases} ((c_2 \rho_2)^n / n!)\, N_o & \text{for } n = 0, 1 \ldots\ldots c \\ ((c_2)^{c_2} / c_2!)\, N_o\, \rho_2^n & \text{for } n = c, c+1, \ldots\ldots \end{cases} \quad (15)$$

A similar treatment of (8) and (9) produces the analogous solution for M_m; M, m and 1 replace N, n and 2 in (15) .

The solution;

$$P_{m,n} = M_m . N_n$$

also satisfies (1) , (2) , (3) and (8) . M_o, N_o can be found by summing $P_{m,n}$ over m and n. Thus the solution for the joint equilibrium queue length distribution is

$$P_{m,n} = P_m . P_n$$

where P_n is as defined in 2.2.1 .

The mean number in the system is

$$\sum_{m=o}^{\infty} \sum_{n=o}^{\infty} (m+n)\, P_{m,n} = \sum_{m=o}^{\infty} mP_m + \sum_{n=o}^{\infty} nP_n$$

which is simply the sum of the mean numbers which would occur if the two stations worked in complete independence, with arrival rate λ at each. It makes no difference whether the faster or slower station comes first.

3.3.2 SINGLE-CHANNEL SYSTEMS WITH INFINITE STORAGE.

The M/M/1...../M/1 System.

When $c_1 = c_2 = 1$, the results of 3.3.1 become:

$$P_{m,n} = \rho_1^m \rho_2^n (1-\rho_1)(1-\rho_2) .$$

This system has recieved extensive study, by R.R.P Jackson (41), Hunt (38), Cox (23, 138), Saaty (76, 259) and others. A result of Reich is of importance: (73)

Let w_1, w_2 denote the two waiting times (including service time). Let m be the number of customers left behind in the first queue by some customer, and let n be the number this customer finds in the second queue. n is independent of w_1, and so w_2 is independent of w_1.

Bourke (12) proves that this independence fails if w_1, w_2 are the waiting times excluding service times. He shows that $\Pr[w_2 = 0 / w_1 = 0] < 1 - \lambda/\mu_2 = \Pr[w_2 = 0]$.

Non-Poisson Systems.

Pearce (69) outlines a method of handling two-station single-channel systems, and treats the system M/D/1...../D/1 as an example. For two queues to form in such a system, the first service time must be shorter than the second. He defines a double generating function:

$$g(x,y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} x^m y^n P_{m,n}$$

and by summing the steadystate difference equations, expresses $g(x,y)$ in terms of the marginal generating function $g(0,y)$. This is determined by contour integration.

3.3.3 SYSTEMS WITH LIMITED NUMBERS OF CUSTOMERS.

R.R.P Jackson (42) studies the M/M/1...../M/1 system for which only N customers are available. The input is exponential, except when there are N in the system. Then no further arrivals can occur.

There are now seven steadystate difference equations

governing $P_{m,n}$. Their solution:

$$P_{m,n} = \rho_1^{\,m} \rho_2^{\,n} P_{0,0}$$

is similar to the solution for the unrestricted case; but this time

$$1/\ P_{0,0} = \sum_{\substack{m=0 \\ (m+n \leq N)}}^{N} \sum_{n=0}^{N} \rho_1^{\,m} \rho_2^{\,n} \ .$$

Expressions for mean number in system; mean number waiting, mean number in service are easily found; but lengthy. Again, the order in which the stations occur makes no difference. This suggests that the system behaves like a cyclic arrangement of two queues.

3.3.4 A TWO-QUEUE-CYCLIC SYSTEM.

A fixed number N of customers circulate around two stations; each station has a single exponential channel. A customer leaving either channel immediately joins the queue before the other channel. The steady-state situation is simple, so is analysed here in full. N+1 states are possible;

let $P_n = \Pr\left[\text{length of "first" queue is } n\right]$, $n = 0, 1, \ldots N$,

$= \Pr\left[\text{length of "second" queue is } N-n\right]$.

The equations are:

$$0 = -\mu_2 P_0 + \mu_1 P_1 \ ;$$

$$0 = -\mu_1 P_N + \mu_2 P_{N-1} \ ;$$

$$\mu_1 P_n - \mu_2 P_{n-1} = \mu_1 P_{n+1} - \mu_2 P_n \ , \quad n = 1, 2, \ldots N-1 \ .$$

(Only N of these N+1 equations are independent.)

The equations yield:

$$\mu_1 P_n = \mu_2 P_{n-1} \quad , \quad n = 1, \ldots N \ .$$

The solution is

$$P_n = (\mu_2 / \mu_1)^n P_0 \ ;$$

where P_0 is found by summing the probabilities.

When $\mu_1 = \mu_2$, each state has equal probability $1/(N+1)$.

3.3.5 SYSTEMS WITH LIMITED WAITING ROOM.

Let the maximum allowable number of customers in the second of two series queues be N. Waiting room for the first queue is not restricted. A customer who finishes service at the first station while the second queue has length N is forced to wait until a departure occurs from the second station. In this situation, the first station is said to be blocked.

Maximum Utilization.

The first channel is actually serving for only a fraction of the time during which customers are waiting at it. Thus as ρ_1 increases towards 1, saturation may be reached when ρ_1 takes some value less than 1. This value, ρ_{max}, is known as the " maximum utilization " of the system.

Hunt (38) studies the M/M/1... ../M/1 system with $N = 1$ (ie: no waiting room between stations); and finds ρ_{max} in terms of μ_1, μ_2. For $\mu_1 = \mu_2$, $\rho_{max} = 2/3$.
The system with $N > 1$ but finite is given, by Hunt, N+2 state probabilities. For $\mu_1 = \mu_2$, they yield

$$\rho_{max} = (N+1)/(N+2).$$

Naturally this approaches 1 as N increases.

Some Approaches to State Probabilities.

Avi-Itzhak and Yadin (5) study the system M/G/1...../G/1 where $N = 1$, in the steady state. They regard the first station as an M/G/1 where arrivals who initiate a busy period have a different service distribution.

Let s_1, s_2 be the two service times. Let $T = \max(s_1, s_2)$; This is the time a customer spends in the first station if he arrives in it to find it busy or blocked. If he finds it idle,

he spends s_1 there. The writers use these ideas to construct the MGF of the variable:

(Time waiting before first station)+

(Time spent at first station; T or s_1)+

(Time spent in second station; s_2);

and the MGF of the variable

(number left behind by a customer leaving first channel).

They show that the concept of maximum utilization is not restricted to systems with exponential service. In M/G/1.... /G/1 , equilibrium can exist only if

$$\lambda E[T] < 1 ;$$

and for i = 1, 2

$$\rho_i = \lambda E[s_i] \leq \lambda E[T] .$$

Finally, the writers treat the special cases: M/D/1..../D/1 and M/M/1..../M/1.

Prabhu (72) studies the transient M/G/1...../G/1 system, and finds Laplace-Stieltjes transforms of the state probabilities. In the limit as $t \to \infty$, his results are in accord with those of Avi-Itzak and Yadin.

Makino (59) finds the MGF for the output distribution of an M/M/1..../M/1 system with finite N, and with $\mu_1 = \mu_2$. He takes a weighted sum of the MGFs for the three situations in which a departure interval consists of:

.an arrival interval plus two service times (one at each channel.)

.two service times

.one service time.

He equates the MGF of the output distribution with the MGF of the output distribution from a one-stage M/G/1 system.

This MGF is known in terms of the MGF of the service time distribution. In other words, the series system is reduced to a one-stage system. First and second service times and waiting time in the second queue are regarded as being the service time in the one-stage system.

Limited Waiting-Room and Cyclic Systems.

Kleinrock (49) develops the theory of queues in series to show that the time for n jobs to be processed at each of P processors (computers) is less than the time for n jobs to go through a single processor P times. For P = 2, he investigates the ratio of these times by the following method.

Let maximum queue length before the second processor be N-1. When the length of the second queue reaches N (this includes the customer in service) no further customers can enter the system. Otherwise, it is assumed, there is a non-zero supply of customers waiting before the first processor. The system is equivalent to the two-queue cyclic system discussed above. In the cyclic system, the first queue has no waiting customers if and only if there are N-1 customers in the second queue. This limited space system behaves as if there were no customers before the first processor if and only if there are N-1 customers in the second queue.

The optimal arrangement is $\mu_1 = \mu_2$; the ratio

$$\frac{\text{time for n jobs on 2 processors}}{\text{time for n jobs on 1 processor}}$$

$$= (N+1)/(2N)$$

3.4 MANY QUEUES IN SERIES.

As the number of stations in series increases, the effect of the arrival distribution diminishes, and the effect of the

service distributions increases in importance. The work on series systems seems restricted to systems where each station has the same service-time distribution. Hence the theory of this section has been classified according to service distributions. The theory for systems with exponential service is treated first. It develops from the work of Bourke and Reich on independence. For systems with other distributions, or limited waiting room, the independence disappears, and other methods must be used. Discussion follows, on systems with regular, then Erlangian service - both these distributions are good approximations for many real-life services. Notes on some other topics follow.

3.4.1 SYSTEMS WITH EXPONENTIAL SERVICE, AND UNLIMITED WAITING ROOM.

Extension of Results for Multichannel Systems.

Exponential arrivals occur at the first of k multichannel systems in series. For $i = 1, 2, \ldots, k$, the ith station has c_i channels, each with exponential service at rate μ_i, and unlimited waiting room. This system, with $k=2$, was treated in 3.3.1. The output from the second and subsequent stations is still exponential, rate λ ; so we could expect the independence of queue lengths at fixed time to survive where $k > 2$. This system was studied first by R.R.P Jackson (41), (42). His proof is outlined by Saaty (76 , 261):

Let n_j, $j=1, \ldots, k$, be the queue length at the ith station; let the equilibrium probabilities for the states of the system be given by $P_{n_1, n_2, \ldots n_k}$; $n_j = 1, 2, \ldots$ for $j = 1, \ldots, k$. The solution to the steadystate equations (which will not be written out here) is:

$$P_{n_1, n_2, \ldots, n_k} = P_{o, \ldots, o} \prod_{j=1}^{k} b_{n_j, j}$$

where $b_{n_j,j} = \begin{cases} (c_j \rho_j)^{n_j} / (n_j)! & , n_j < c_j \\ (c_j)^{c_j} (\rho_j)^{n_j} / (c_j)! & , n_j \geq c_j; \end{cases}$

and $\rho_j = \lambda / (c_j \mu_j)$.

$P_{o,\ldots,o}$ is found by a k-fold summation.

The marginal probability that there are n_j customers at the jth station is $b_{n_j,j} / \left(\sum_{n_j=o}^{\infty} b_{n_j,j} \right)$; i.e., the jth station has the same equilibrium queue length distribution that it would have if it was operating alone, with exponential arrivals at rate λ . The mean queue length at the jth station, mean number waiting, and mean number in service there, are given by the expressions of 3.3.1 . The mean total number in system, etc, are found by adding the means for each station.

The distribution function for waiting time (excluding service), w_j, at the jth station, is given by

$$\Pr \left[w_j > T \right] = \frac{P_{c_j}}{1-\rho_j} e^{-c_j \mu_j (1 - \rho_j) T} .$$

Nelson, (65) , shows that if the distribution of waiting time (excluding service), w_j, at each of k stations, is given by : $\Pr \left[w_j > T \right] = K e^{r_j T}$,

then the distribution of total waiting time (excluding service), t, is given by

$$\Pr \left[t > T \right] = \sum_{j=1}^{k} A_{jk} e^{r_j T} ,$$

where $A_{jk} = K_j \prod_{\substack{i=1 \\ i \neq j}}^{k} ((1 - K_i r_j)/(r_j - r_i))$, $j = 1,..,k;$

and $r_j \neq r_i$ for $i \neq j$.

If $r_j = r_i$ for $i \neq j$, the A_{jk} can be computed by perturbing r_i, r_j.

The proof is by induction on k, and involves considering separately the cases $t = 0$, $t > 0$; and $w_k = 0$, $w_k > 0$.

Extension of Results on Independence of Single-Channel Waiting-Times.

Reich (74) extends his earlier results (73) to show that, for a series of k single-server stations, the waiting times (including service) for a particular customer are independent. His results are as follows:

Let w_{ij} be the waiting time (including service) for the ith customer at the jth station. The distribution of w_{ij} is the convolution:

$$\Pr\left[w_{ij} \leqslant x\right] = W(x) * S(x) \quad , i = 1, \ldots\ldots k;$$

where $$W(x) = \begin{cases} 1 - \rho_j e^{(\lambda - \mu_j)x} & , x \geqslant 0 \\ 0 & , x < 0 \end{cases};$$

$$S(x) = \begin{cases} 1 - e^{-\mu_j x} & , x \geqslant 0 \\ 0 & , x < 0 . \end{cases}$$

Also,

$$\Pr\left[w_{i1} \leqslant x_1,\ w_{i2} \leqslant x_2,\ \ldots\ldots,\ w_{ik} \leqslant x_k\right] = \prod_{j=1}^{k} \Pr\left[w_{ij} \leqslant x_j\right];$$

so that for each fixed i, the variables w_{ij}, $j = 1, \ldots\ldots, k$, are independent.

3.4.2 SYSTEMS WITH EXPONENTIAL SERVICE, AND LIMITED WAITING ROOM.

Several different approaches, most of them approximate or numerical, have been developed for these systems. A full analytical solution has not been found; for the limitations on waiting room remove the property of independence.

Extension of Kleinrock's Cyclic Model.

A two-station system is equivalent to a two-station cyclic system with a limited number of customers. This equivalence does not hold for a larger system, as the customers have more freedom in distributing themselves throughout the system. However, Kleinrock (49) approximates the behaviour of a four-

station system (with the same waiting room at each station), by considering it as a series of two joint stations, each with exponential service, and with limited waiting room between them. This pair of joint stations is equivalent to a cyclic system. Each of the joint stations consists of two exponential stations with limited room between them, and the service time for the joint station consists of two waiting times and two service times. The approximation is good. Series of 8, 16, 32, 64,...stations can be treated similarly. Again Kleinrock examines the performance measure R, where

$$R = \frac{E[\text{time to process n customers in the series system}]}{E[\text{time for a single processor to perform the equivalent task}]}.$$

R declines steadily as the number of stations rises.

A Numerical Approach.

Hillier and Boling (37) extend the work of Hunt (38) by two computer-based procedures. The first is exact, and deals with Erlang service as well as exponential. The second is approximate, and deals with exponential service only.

There are k single-channel stations in series, in the steady state. The first queue is never empty. The maximum queue length at the jth station is N_j ($j = 2, 3, \ldots, k$), and the service distribution there is Erlang, with shape parameter s_j, and mean rate μ_j. There are hundreds of possible states, so the transition matrix for them is generated within the computer. It starts with a certain state, and allows a service or (for $s_j > 1$) a partial service to occur at each station. Other states result; transition probabilities μ_j are introduced into the matrix; and the process continues till no

new states result. Equilibrium state probabilities P_n are found by solving the equation:

[vector of P_n's][transition matrix] = [vector of P_n's] .

Let R be the mean output rate; since the first queue is never empty, R is the maximum throughput rate for the system. When all service rates are the same, $\rho_{max} = R/\mu$. R is found by summing P_n over states from which a departure from the system can occur, and multiplying this probability by μ_k s_k.

In the second, approximate, procedure, each station is considered as a single channel with exponential arrival and service distributions. The independence results can then be applied to the system. The effective arrival and service rates are found by an iterative procedure.

The numerical results of Hillier and Boling for a special case are graphed here. The special case has four stations, each with exponential service at the same rate; $N_2 = N_3 = N_4 = N$. The loss of efficiency caused by limited waiting room is clear, and the agreement between the two procedures is remarkably good. The second procedure can be used to study the effects of unbalancing a production line; i.e., introducing unequal service rates. This can improve efficiency.

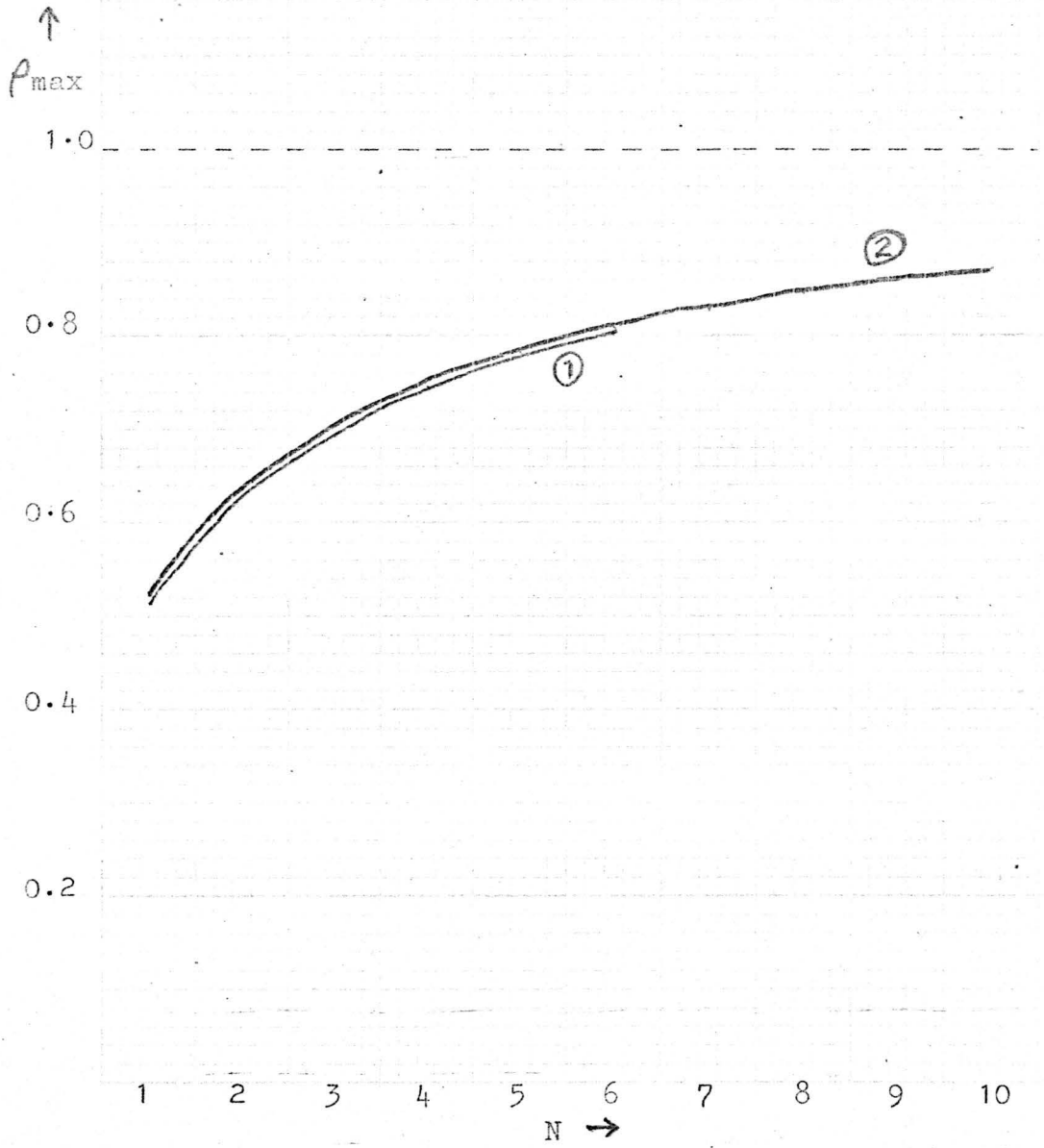
Optimal Ordering of Stations.

Patterson (68) also takes a Markovian approach to the problem of k stations : exponential arrivals occur at the first station, each has exponential service with its own rate; and its own finite waiting room. Stochastic processes are defined on subsets of the set of states; state probabilities are evaluated. The mean rate of departure from the system is

Maximum Utilization as

Function of Waiting Room,

For a Series of 4 Stations.



Calculated by ① exact procedure

② approximate procedure.

influenced by the order in which the stations are placed. Patterson's conclusions follow:

When the arrival rate is high, i.e.,the first station is usually occupied, fast stations should be placed between the slow ones, so that they can serve as extra storage spaces. If each station has unlimited waiting room, the departure distribution is exponential, and its mean rate cannot be larger than the smallest mean service rate, μ_{min}. If limited waiting space occurs before one or more stations, the departure rate is less than μ_{min}. Departure rate is maximized when blocking is minimized; this is achieved by putting fast stations directly before slow stations.

3.4.3 SYSTEMS WITH REGULAR SERVICE TIMES.

Two studies of series systems with regular service times appeared at about the same time; their results are similar, and centre on the intuitively obvious fact that the slowest station will dominate the behaviour of the system (as it does in systems with exponential service) . The less general study is mentioned here first.

Single-Channel Systems.

Avi-Itzhak (4) studies the series of k single-server stations, with arbitrary arrival distribution ; regular service distributions (with a possibly different rate μ_i at each station); limited waiting room except at the first station.

The system is equivalent to one with maximum queue lengths of 1 before each station except the first; we can insert stations with zero service time before each space in which a single customer can be held. Avi-Itzhak shows that D_n (the departure time of nth customer) depends only on D_{n-1}, the

arrival time of the nth customer, μ_{max} (the longest service), and S (the sum of all the services). As D_1 is independent of the order of the stations, so is D_n for all n. Thus the system is equivalent to the system obtained by placing the slowest station first. No blocking can then occur at any of the stations. The system behaves like the system comprising the queue that would result before a single channel with regular service of length $\mu_{max.}$, together with a service time of S.

Avi-Itzhak extends these results to series where each station has the same number of multiple channels.

Multiple-Channel Systems.

Friedman (3/) considers two systems. Model I differs from the systems studied by Avi-Itzhak in having a different number of channels at each station; and no restrictions on waiting room. Model II differs from Model I in that one station is a single channel, with variable service time which is always longer than (service time)/(number of channels) at the other stations.

Friedman defines the relation:

A dominates B , (A, B are stations) ;

which is nearly equivalent to

A is slower than B.

He uses dominance to show that mean waiting time and other problems can be reduced to corresponding problems for a system of fewer stations. Any single-channel system, or any Model II system, can be reduced to a single-station system.

3.44 SYSTEMS WITH GENERAL SERVICE DISTRIBUTION.

An infinite number of identical single-channel stations

are connected in series; the arrival distribution and the common service distribution are arbitrary. Let $d_{i,j}$ be the time between the ith and the (i+1)th customer's departures from the jth station. Masterson and Sherman (60) study the problem of whether $d_{i,j}$ reaches equilibrium as $j \to \infty$. They find that it does not for i = 1 or 2; though it may, as $i \to \infty$. As might be expected from 3.4.2, in the special case of constant service, equilibrium is achieved at j = 1, and maintained for j > 1.

3.5 A COMPARISON.

Some basic difference in behaviour between single-station and series systems, and between series systems with and without waiting room, are demonstrated by the following comparison of three simple systems.

Let A represent an M/M/1 system, with service rate $\mu/2$. Let B and C both represent M/M/1..../M/1 systems, with service rate μ at all channels. B has unlimited waiting room; C has room only for the customer in service at the second station, and unlimited room at the first station. Each system has arrival rate λ ; and $\rho = \lambda/\mu$. The mean total service time is the same for each, and is $2/\mu$.

State Probabilities.

Let $P_n = \Pr[\text{total number in system} = n]$. These state probabilities (or sums of them) are presented only for A and B. P_n for C can be found (5) from the expansion of a very complex generating function. A is more congested than B; its P_n decrease more slowly with n than that of B:

n	P_n for A	P_n for B
0	$1 \cdot (1 - \rho/2)$	$1 \cdot (1 - \rho^2)$
1	2ρ	2ρ
2	$4\rho^2$	$3\rho^2$
3	$8\rho^3$	$4\rho^3$

For $2^n/(n+1) < (1-\rho)^2/(1-2\rho)$, P_n for A is smaller. Since the left side is 1 for $n = 0, 1$; and since the right side is greater than 1 for all ρ ; A has the smaller P_0 and P_1 always.

Mean Number in System.

Let $E[n]_K$ be the mean number in system K, where K = A, B, C. Then:

$$E[n]_A = 2\rho/(1-2\rho)$$

$$E[n]_B = 2\rho/(1-\rho).$$

Avi-Itzhak and Yadin treat C as a special case of the problem they study (5), and find that:

$$E[n]_C = (2\rho(2-\rho^2))/((2+\rho)(2-3\rho)).$$

Evidently the upper bounds on ρ for $E[n]_K$ to be finite and non-negative are 1/2, 1, 2/3; for A, B, C. The value for C agrees with Hunt's results.

Again we see that A is always more congested than B, since:

$$E[n]_A > E[n]_B, \quad \text{for } \rho > 0.$$

Apparently it is more efficient to provide two short exponential services (and cause two short waits) than to provide one exponential service of twice the mean length.

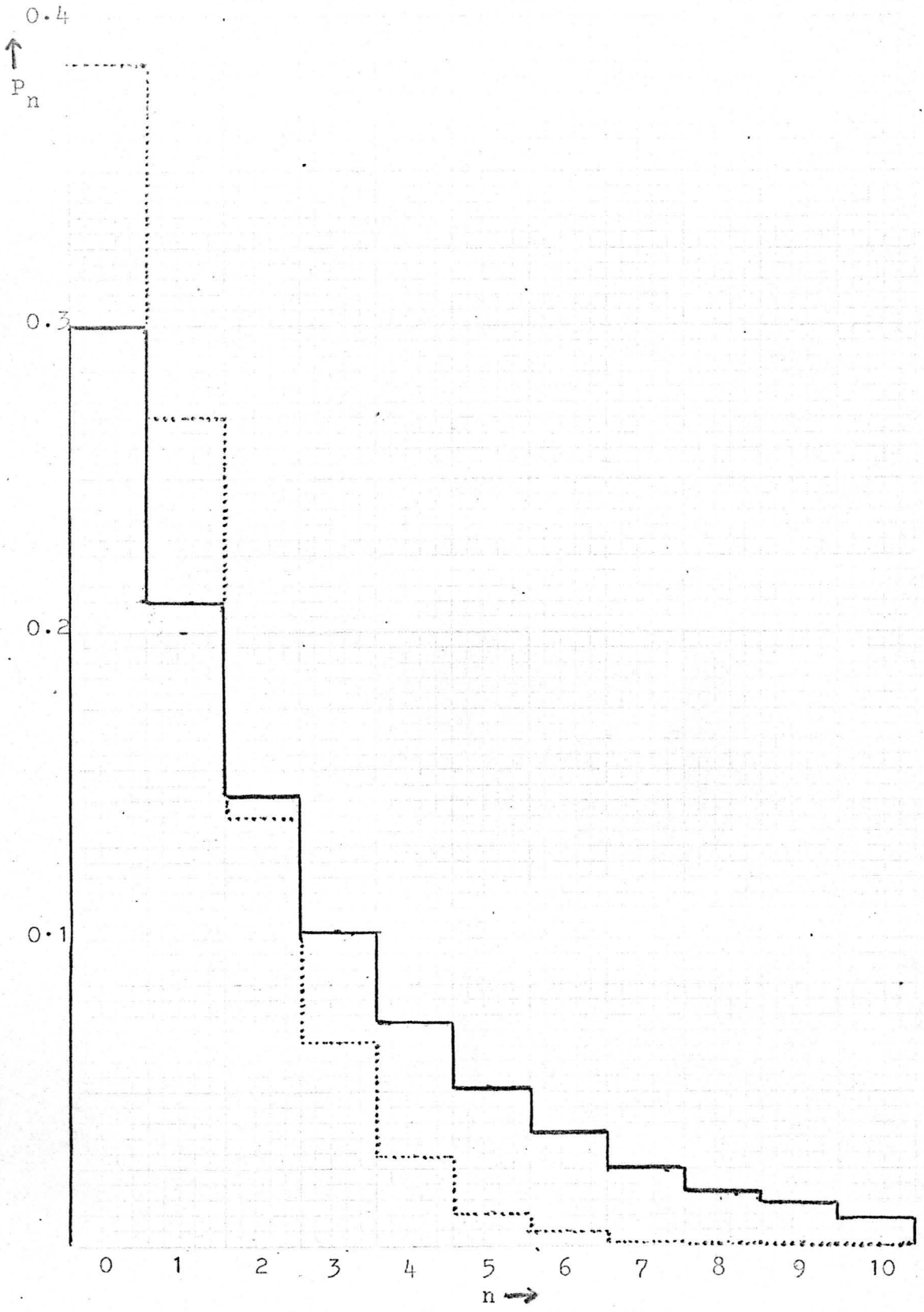
B is more congested than C for $0 < \rho < r$, where r is the root of $E[n]_B = E[n]_C$.

The value of r is about 0·6.

Apparently, at low traffic intensity ($\rho < r$) the 2 stations are utilized better if all waiting is done at the first station.

A is always more congested than C.

Probabilities for Number in System.



——— M/M/1 with arrival rate λ, mean service time μ, $\lambda/\mu = .7$.

--- M/M/1../M/1 with: arrival rate λ, mean service $1/(2\mu)$ at each station, $\lambda/\mu = .7$.

CHAPTER 4.

NETWORKS OF QUEUES.

The chapter studies open and closed networks of poisson queues, with an extension for nonzero transit times. It includes a comparison of simple examples studied in chapters 2, 3 and 4.

CHAPTER 4 : NETWORKS OF QUEUES.

4.1 INTRODUCTION.

Queue Networks and Their Applications.

Mathematics, Physics, and Operations Research use the word " network " to cover various classes of phenomena which are quite different both in real life and in theoretical treatment. The subject of interest here is networks of queues. These networks consist of nodes and links. Each node is a service station with its line of waiting customers; the link joining each pair of nodes is a path along which customers may travel in either direction. The flow of customers along these links is discrete.

Most of the theory presented here was engendered by applications of the job-shop type. These include production, repair, and customer service processes. The earliest applications, however, were to communications systems. This type of application probably generates more interest now than applications of the job-shop type. Related to this field is the field of electronic data processing.

Strengths and Weaknesses of the Existing Theory.

" Many real systems one wishes to study seem to be networks ", according to Evans (26); yet in 1968 Saaty (78) could say that " although the subject of queues has been pushed far....., the general problem of a net of queues remains in its infancy without substantial change." The theory presented in 4.2, 4.3, and 4.4 together with the comments in this section, is intended to indicate the extent of the theory at the present time. Results developed specifically for communications networks will follow in

Chapter 5.

The strength of the extant theory is that it explores to the limit the type of model in which arrivals and services are exponential. From the state probabilities it is possible to determine the optimal arrival and service rates, number of channels, and routing weights. There are many weaknesses: little or nothing is known about transient behaviour; about systems with non-exponential service; or about systems where rates and routing depend on the identity of the customers.

Evans (26) sums this up: " The work supporting J.R. Jackson's paper is really a terminal point for the study of individual queues as opposed to a starting point in network analysis." He suggests that a network system should be studied as a whole: " The designer..... is not interested in trying to provide conditions under which queues may operate independently. He is interested in how to introduce dependencies so that the system will operate in a desirable way. "

Alternative Approaches.

It seems that a quite different approach will be needed to advance the theory further. Saaty (78) suggests the use of graph theory. This has been explored for the single-server queue by Benes (7). Each state of the system is equivalent to a pattern in some abstract space. An ordering can be imposed on the patterns. Possible transitions between states are equivalent to certain types of movement in the " lattice " of patterns.

Saaty also suggests developing the analogy between queue networks, and continuous flow networks. Channel capacity

and traffic intensity correspond to pipe diameter and fluid pressure.

Linear programming may be used in optimizing the flow through a queue network.

One powerful line of approach to which resort must and has been made, is simulation. An example is Nelson's work on " Queueing Network Experiments with Varying Arrival and Service Parameters."(66) He simulates the time-dependent behaviour of a two-station network where arrival and service processes are each assigned distributions of three types; exponential, erlang with shape parameter 2, and regular. The aim is to discover and isolate dependencies among the variables describing the state of the system. The results of the simulation are compared to those of an analytical approximation model.

Evans (27) suggests an alternative concept of network capacity. The usual concept is the maximum of (arrival rate/ service rate) for stable behaviour; the alternative is the expected output rate of a system operated under a saturation load. This concept has been used by Kleinrock (49) and others.

Outline of the Chapter.

The only systems treated in the rest of this chapter are those with exponential arrival and services. 4.2 treats the work of J.R.Jackson on a very general form of the open network model. 4.3 presents the work of Gordon and Newell on a less general closed model. 4.4 more briefly treats the work of Pogner and Bernholtz: this concerns a more general closed model than that of 4.3 ; and involves the possibility of

nonzero transit times between stations. The topic of 4.3 is a special case of the topic of 4.4 .

Only equilibrium solutions are involved. FCFS priority is assumed throughout, though some of the results are independent of priority. Conway and Maxwell (22) conjecture that if a priority discipline is such that it leaves the output of an M/M/c system still poisson, then the independence results for networks will hold under this discipline.

4.5 analyses a very simple network, then collects together for comparison the simple systems analysed at the ends of the chapters 2, 3 and 4.

4.2 OPEN SYSTEMS WITH EXPONENTIAL ARRIVALS AND SERVICE.

The model studied by J.R.Jackson (40) consists of a group of N single-channel stations. A single poisson stream feeds the system, and divides into N streams for the N stations according to pre-assigned probabilities. The stream of customers leaving each station divides into N+1 streams. N of these go back to the stations; the other leaves the system. The rate of the arrival stream depends on the total number in the system; the rate of service at each station depends on the queue length there. The state-dependence of these rates is almost arbitrary.

It is proved above in 2.4.1 that a poisson stream split in halves becomes two poisson streams. Apparently, under a simple condition (arrival rate is constant) the poisson character of the streams within the network survives all the splittings and confluences which the streams undergo. This allows the stations to function as if independent.

Jackson's study has great generality; but serious limitations. Obvious limitations are the restrictions to exponential distributions, and unlimited waiting room. A less obvious limitation is that all customers are assumed identical: rates and routing are assigned to them probabilistically. The lengths of service received by one customer are independent of one another.

Since the results of J.R.Jackson are of central significance in this dissertation, they are given here almost in full. The network of stations with several identical channels arises as a special case of the general model. A result on waiting-times concludes this section.

4.2.1 EQUILIBRIUM STATE PROBABILITIES FOR THE GENERAL MODEL.

Notation, and Restrictions on the Model.

There are N service stations. Let k_n be the queue length at the station n. The states of the system are N-dimensional " state vectors " $\underline{k}$, where

$$\underline{k} = (k_1, k_2, \ldots, k_N).$$

Let $P(\underline{k}) = \Pr$ [system is in state $\underline{k}$ /a steady state has been reached] .

Let $S(\underline{k})$ be the total number in the system when it is in state $\underline{k}$. Then

$$S(\underline{k}) = k_1 + k_2 + \cdots + k_N .$$

The arrival rate when the system is in state $\underline{k}$ is $\lambda(S(\underline{k}))$; so $\lambda(K)$ must be defined for $K = 0, 1, \ldots$. The following restriction is set on $\lambda(K)$:

ASSUMPTION: Either the $\lambda(K)$ are all positive; or there exists a K_o, with $\lambda(K) > 0$ if $K \leq K_o$,

but $\lambda(K) = 0$ if $K > K_o$.

The service rate at station n , for state $\underline{k}$, is $\mu(n,k_n)$. Thus $\mu(n,k_n)$ must be defined for $n = 1,\ldots,N$; $k_n = 0, 1,\ldots$

ASSUMPTION : The $\mu(n,k_n)$ are all positive, except that each $\mu(n,0) = 0$.

Let the "outside world ", in its function of supplying customers to the network, be thought of as station 0. In receiving customers leaving the network, it can be thought of as station $N+1$. The customer flow is defined by the $\lambda(K)$, $\mu(n,k_n)$, and a matrix R of probabilities:

$$R = \left\{ r(m,n) \;/\; m = 0, 1,\ldots,N;\ n = 1, 2,..,N+1 \right\}.$$

$r(m,n)$ is the probability that a customer leaving station m goes to station n. Hence we have:

ASSUMPTION: for $m = 0,\ldots,N$; $\left\{ r(m,n)/n = 1,\ldots,N+1 \right\}$ is a probability distribution.

A customer should have zero probability of wandering forever in the net. This requirement is equivalent to:

ASSUMPTION: The equations

$$e(n) = r(0,n) + \sum_{m=1}^{N} e(m)\, r(m,n) \quad , n = 1, 2..N;$$

have a unique set of non-negative solutions $\left\{ e(n)/n = 1,\ldots,N. \right\}$. $e(n)$ can be interpreted as the expected number of times a customer visits station n.

<u>The Steady State Equation.</u>

We consider

Pr [system is in state $\underline{j}$ at $t+h$/system is in state $\underline{k}$ at t]

for various cases, and replace terms in h^2 by 0:

No movements occur :

$$1 - h\,\lambda(S(\underline{k})) \sum_{n=1}^{N} r(0,n) - h \sum_{n=1}^{N} \mu(n,k_n)(1 - r(n,n)); \quad \text{if } \underline{j} = \underline{k}.$$

A customer arrives: $(x = 1,\ldots,N)$

$h\,\lambda(S(\underline{k}))\ r(0,x)$, if $\underline{j} = \underline{k}$ except that $j_x = k_x + 1$.

A customer departs: $(x = 1,\ldots,N)$

$h\mu(x,k_x)\ r(x,N+1)$, if $\underline{j} = \underline{k}$ except that $j_x = k_x - 1$.

A customer moves from one station to another:

$(\ x = 1,\ldots,N;\ y = 1,\ldots,N;$ and $x \neq y\ )$

$h\mu(x,k_x)\ r(x,y)$, if $\underline{j} = \underline{k}$ except that $j_x = k_x - 1$,

$j_y = k_y + 1$.

More than one transition occurs:

0, for all other $\underline{j}$.

These $N^2 + N + 1$ probabilities lead, in the usual fashion, to this steady-state difference equation:

$$0 = -\left(\lambda(S(\underline{k}) + \Sigma\mu(n,k_n)\ (1 - r(n,n))\right) P(\underline{k})$$
$$+ \Sigma\lambda(S(\underline{k}) - 1)\ r(0,n)\ P(\underline{h}_n)$$
$$+ \Sigma\mu(n,k_n + 1)\ r(n,N+1)\ P(\underline{l}_n)$$
$$+ \Sigma\Sigma\mu(n,k_n + 1)\ r(n,m)\ P(\underline{j}_{mn})$$

The single sums are for $n = 1,..,N$; the double sum is for $m = 1,..,N$; $n = 1,\ldots,N$; $m \neq n$.

$P(\underline{k})$ is zero when $\underline{k}$ has a negative component. $\underline{h}_n = \underline{k}$ except that its nth component is $k_n - 1$; $\underline{l}_n = \underline{k}$ except that its nth component is $k_n + 1$; $\underline{j}_{mn} = \underline{k}$ except that its mth component is $k_m - 1$, and its nth component is $k_n + 1$.

The Solution.

Jackson introduces these notations: (an empty product takes the value 1)

$$W(K) = \prod_{L=0}^{K-1} \lambda(i), \quad \text{for } K = 0, 1,\ldots ;$$

a product of arrival rates.

$$w(\underline{k}) = \prod_{n=1}^{N} \prod_{i=1}^{k_n} \left(e(n)/\mu(n,i) \right) ,$$

a product of products of effective service times.

$T(K) = \sum w(\underline{k})$; summed over $\underline{k}$ with $S(\underline{k}) = K$; for $K = 0, 1, \ldots$.

The solution is given (when it exists) by

$P(\underline{k}) = \gamma\, w(\underline{k})\, W(S(\underline{k}))$,

where γ is found by using the fact that the $P(\underline{k})$ sum to 1. It exists when the sum for $(1/\gamma)$ converges.

Every nontransient state of the system communicates with every other, and hence if a limiting probability distribution exists, it is unique.

Routine substitution shows that the solution satisfies the difference equation.

The solution resembles earlier results for series of .../M/c stations; in being of the form

(product of arrival rates)/(product of products of effective service rates).

4.2.2 COROLLARIES, EXTENSIONS, SPECIAL CASES.

Corollaries.

By summing the $P(\underline{k})$, we find that

$\Pr[\text{number in system} = K] = \gamma\, T(K) W(K)$.

The conditional probabilities

$\Pr[\text{system is in state } \underline{k} \;/\; S(\underline{k}) = K]$

do not depend on the arrival rate .

The routing matrix R influences the steady state only through the e(n).

An Extension.

Jackson extends his theorem to cover systems which must include at least K* customers, and which have a maximum queue length k_n^* at each station.

Special Cases.

A case of particular interest is that in which the arrival

rate is constant; i.e.,

$$\lambda(K) = \lambda .$$

This condition is sufficient for $P(\underline{k})$ to factor into N separate probabilities. Let

$$w_n(k) = \prod_{i=1}^{k} (\lambda e(n)/\mu(n,i)), \text{ for } k = 0, 1, \ldots;$$

$$P_n(k) = \begin{cases} w_n(k) \Big/ \sum_{i=0}^{\infty} w_n(i), & \text{if the sum converges} \\ 0, & \text{otherwise; for } k = 0, 1 \ldots . \end{cases}$$

Then when $P(\underline{k}) = \pi \left(\prod_{n=1}^{N} \prod_{i=1}^{k_n} (e(n)/\mu(n,i)) \right) \lambda^{S(\underline{k})}$,

$$P(\underline{k}) = P_1(k_1) \ldots P_N(k_N) .$$

$P_n(k)$ is the equilibrium distribution of a system with exponential arrivals at rate $\lambda e(n)$, and with service identical to that at station n. Since e(n) is the mean number of visits to station n, $\lambda e(n)$ is the mean arrival rate there. Hence we can say that the equilibrium queue length distributions are independent.

A further specialization of the constant arrival rate case is of interest. Let

$$\mu(n,k) = \mu_n \min(k, c_n); \quad \text{for } n = 1, \ldots, N.$$

Station n now operates as a set of c_n parallel channels each with rate μ_n . The outcome is that the $P_n(k)$ are identical with the familiar solution for an M/M/c system with arrival rate $\lambda e(n)$, and c_n channels with rate μ_n. J.R.Jackson proved this result in an earlier paper (39).

Jackson also considers:

the effect of fixing service rates; and

the closed system which results when

$$\lambda(K) = 0 \text{ for } K \geq K^*$$

in the extension above.

4.2.3 WAITING TIMES.

An M/M/c system with arrival rate $\lambda\, e(n)$, and service as usual, has this distribution for waiting time (excluding service) w_n: (see 2.3.1)

$$\Pr[w_n > T] = A_n e^{b_n T} \qquad (1),$$

where $A_n = \Pr[\text{queue length} = c_n] / (1 - \rho_n)$

$$= \Pr[w_n > 0];$$

$$b_n = -c_n \mu_n (1 - \rho_n); \text{ and}$$

$$\rho_n = \lambda\, e(n)/(c_n \mu_n).$$

Nelson (65) asserts that the results of Bourke, Reich and J.R.Jackson " establish the validity of a stage-by-stage analysis " for the network of N exponential (μ_n, c_n)-stations with exponential arrivals, rate λ. The distribution of w_n, the waiting time at station n, is then given by (1). The result of Nelson treated in 3.4.1 can be applied to the system. The probability density function for total waiting time (excluding service) w then is again a weighted sum of exponentials; with finite probability at $w = 0$.

4.3 CLOSED SYSTEMS WITH EXPONENTIAL SERVICE.

Gordon and Newell (34) study the equilibrium distribution of customers in an N-station network, where station n consists of c_n channels each with exponential service at rate μ_n. No arrivals to or departures from the system occur; the number in the system, K, is fixed. This system differs from Jackson's, in that the k_n are not independent since

$$k_1 + + + + k_n = K .$$

4.3.1 EQUIVALENCE OF OPEN AND CLOSED SYSTEMS.

This N-station closed system is equivalent to an (N-1)-station open system, for which:

the total number in system has an upper limit K_o, and for which the arrival rate is as follows:

Let $k_2, k_3, .., k_N$ be the queue lengths at the N-1 stations, and let

$$k_1 = K_o - k_2 - - - k_N .$$

c_1 is some positive integer; μ_1 is some positive number. Customers arrive exponentially at rate $\mu_1 \min(k_1, c_1)$.

The equivalence between the open and closed systems follows from the equivalence between the " outside world " and another station, with c_1 channels and rate μ_1.

When $c_1 = 1$, the input process is exponential, except that no new customers are accepted when the system is " full ".

4.3.2 EQUILIBRIUM STATE PROBABILITIES.

In the closed system, the equilibrium difference equation is

$$\left(\sum_{n=1}^{N} d(k_n)\, a_n(k_n)\, \mu_n \right) P(\underline{k}) = \left(\sum_{m=1}^{N} \sum_{n=1}^{N} d(k_n)\, a_m(k_m + 1)\, \mu_m\, r(m,n) \right) P(\underline{h})$$

where $\underline{k}$, $P(\underline{k})$, $r(m,n)$ are as in 4.2.1;

$$d(k) = \begin{cases} 0 & \text{if } k \leqslant c_n \\ 1 & \text{if } k \geqslant c_n ; \end{cases}$$

$$a_n(k) = \begin{cases} k & \text{if } k \leqslant c_n \\ c_n & \text{if } k \geqslant c_n ; \end{cases} \quad \text{and}$$

$$\underline{h} = \underline{k} \text{ except that } h_m = k_m + 1$$

$$h_n = k_n - 1.$$

The writers use a separation-of-variables technique to find this solution:

$$P(\underline{k}) = c_o \prod_{n=1}^{N} \left(x_n^{k_n} / B_n(k_n) \right)$$

where c_o is a normalising factor;

$$B_n(k) = \begin{cases} k! & \text{if } k \leq c_n \\ c_n!\, c_n^{k-c_n} & \text{if } k \geq c_n ; \end{cases} \quad \text{and}$$

x_n, $n = 1,\ldots,N$, are the solutions to the N equations

$$\sum_{m=1}^{N} r(m,n)\mu_m x_m = \mu_n x_n .$$

The x_n are unique except for a common factor. They can be interpreted as:

E [total amount of service time required by a customer at the nth station, during some arbitrary time period] .

4.3.3 EXTENSIONS, SPECIAL CASES.

An Extension.

Gordon and Newell (34) continue, to consider the effect of allowing K to increase to infinity. If one station is more congested (i.e., x_n/c_n is larger) than the others, its queue length grows to infinity, and the other stations function independently of one another.

Special Cases.

When the routing probabilities r(m,n) form a doubly stochastic matrix, (i.e., columns as well as rows sum to 1), the x_n are equal. Then

$$P(\underline{k}) = c_o \prod_{n=1}^{N} (1/B_n(k_n)).$$

When $r(m,n) = \begin{cases} 1 & \text{for } n = m+1 \\ 0 & \text{otherwise} \end{cases}$,

$\{R(m,n)\}$ is a doubly stochastic matrix, and the routing is cyclic. These cyclic queue systems are a special case of closed network systems.

4.4 CLOSED EXPONENTIAL SYSTEMS WITH TIME LAGS.

The models described above assume that it takes zero time to move from one station to another; this assumption is a poor approximation in many real situations. Posner and Bernholtz (70) treat a system with generally distributed transit times between the stations. Their model is a closed one with N single-channel stations; each channel is exponential, with rate depending on the queue length there. The N^2 transit times all have their own distributions. Again, routing is governed by a matrix of probabilities.

4.4.1 EQUILIBRIUM STATE PROBABILITIES.

Complete specification of the state of the system at equilibrium now involves three arrays: $\underline{k}$, an N-dimensional vector giving the N queue lengths; $A = \{a(m,n)\}$, an $N \times N$ matrix where a(m,n)is the number of customers in transit from m to n; $X = \{x(m,n)\}$, an $N \times N$ array of vectors where each x(m,n) is an a(m,n)-dimensional vector. The l-th component of x(m,n) will be written as $x_{l,m,n}$; and is the elapsed time-in-transit for the l-th of the a(m,n) customers on their way from m to n.

Integrodifferential difference equations are developed, together with boundary and continuity equations. The steadystate solution to these is the following density function:

$$f(\underline{k}, A, X) = c_o \cdot \left(\prod_{m=1}^{N} (e(m)^{k_m} / \prod_{j=0}^{k_m} \mu_m(j) \right) \cdot \prod_{m=1}^{N} \prod_{n=1}^{N} \prod_{i=0}^{a(m,n)} \left((r(m,n)(1 - G_{m,n}(x_{l,m,n}))) \right);$$

where the e(n) are the solutions to

$$e(n) = \sum_{m=1}^{N} e(m)\, r(m,n) \quad , \; n = 1,..,N;$$

$G_{m,n}(x) = \Pr[\text{transit time from } m \text{ to } n \leq x]$; and

c_o, $\mu_m(j)$, $r(m,n)$ are as in earlier models.

The distribution of queue lengths is found by integrating out all the $x_{l,m,n}$ and summing over all the $a(m,n)$. The distribution $P(\underline{k})$ is found to depend on the means only of the transit time distributions:

$$P(\underline{k}) = c_o \left(\prod_{m=1}^{N} (e(m)^{k_m} / \prod_{j=0}^{k_m} \mu_m(j)) \right) h^L / L! \ ;$$

where $h = \sum_{m=1}^{N} \sum_{n=1}^{N} \left(r(m,n).(\text{mean transit time from m to n}) \right)$

= mean transit time for a single transfer;

$L = K - k_1 - k_2 - \ - \ - k_N$

= number of customers in transit.

On taking the limit as $h \to 0$, a more general form of Gordon and Newell's result is found. $P(\underline{k}) > 0$ only when $L = 0$.

4.4.2 AN EXTENSION INVOLVING SEVERAL CLASSES OF UNITS.

In a later paper (7/) Posner and Bernholtz attack the problem of the customer-independence of the parameters in all the models studied so far. They study the closed network of their previous paper, which this time holds customers from several classes. Each class has possibly different service rates, routing matrix, and transit-time distributions. They find the steadystate distribution, and the marginal joint steady state distributions for each class.

4.5 SOME EXAMPLES.

This section analyses one of the simplest possible networks; then presents comparisons and graphs involving this network, and the simple systems analysed at the ends of chapters 2 and 3.

4.5.1 A SIMPLE NETWORK .

The system to be considered is an open one of two single-channel stations, with exponential arrivals and service, and with zero transit times. We retain three parameters:

the arrival rate λ ;

the service rate μ , common to both channels;

the parameter η which governs the routing.

Arrivals go to either station with equal probability; departures from either station go to the other with probability $1 - \eta$, or leave the system with probability η .

Thus the routing matrix is:

$$R = \begin{array}{c} \begin{matrix} (\,m=0 & 1 & 2\,) \end{matrix} \\ \begin{bmatrix} 1/2 & 0 & 1-\eta \\ 1/2 & 1-\eta & 0 \\ 0 & \eta & \eta \end{bmatrix} \end{array} \begin{matrix} (\,n=1 \\ 2 \\ 3\,). \end{matrix}$$

Then

$$e(1) = 1/2 + e(1).0 + e(2).(1 - \eta) ,$$

$$e(2) = 1/2 + e(1).(1 - \eta) + e(2).0 \quad ; \text{ and so}$$

$$e(1) = e(2) = 1/(2\eta) .$$

Thus the effective arrival rate at each station is $\lambda/(2\eta)$, and the effective traffic intensities are

$$\rho_1 = \rho_2 = \lambda/(\mu 2\eta) = \rho .$$

Where m, n are the queue lengths at stations 1 and 2, the joint equilibrium queue length distribution is given by:

$$P_{m,n} = (1 - \rho)^2 \rho^{m+n} .$$

The distribution for number in system N is given by:

$$P_N = (N + 1)\rho^N (1 - \rho)^2 .$$

Hence $E[N] = 2\rho/(1 - \rho)$.

All these results are identical with those for two channels in series (which system is a special case of the two-

station network) except for the different meaning taken by ρ.

The value of λ/μ below which equilibrium is possible, is 2γ. If $\gamma = 1$, we have two parallel stations. As $\gamma \to 0$, we have a system which becomes congested very early; as the customers require an average of $1/\gamma$ services each.

4.5.2 COMPARISON OF SIMPLE PARALLEL, SERIES, AND NETWORK SYSTEMS.

The graphs following the end of this chapter provide a visual comparison of the congestion occuring in the six simple systems treated earlier. They show the effects on this congestion of providing channels in parallel, series and network. For each system, the expected number in system is plotted against ρ ; where ρ is the ratio:

E [total service time required by one customer] / E [inter-arrival time] .

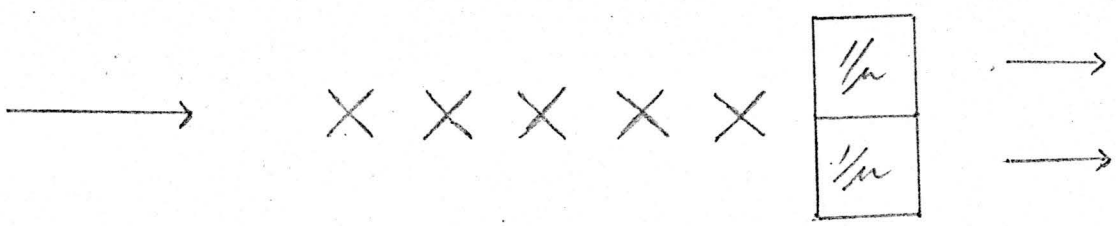
Arrivals to each system are poisson, with the same rate for all. Service at each channel is exponential; the service channels are shown as rectangles, with their mean service time written within. These times have been arranged so that in each system,

E [total service time required per customer] = $1/\mu$.

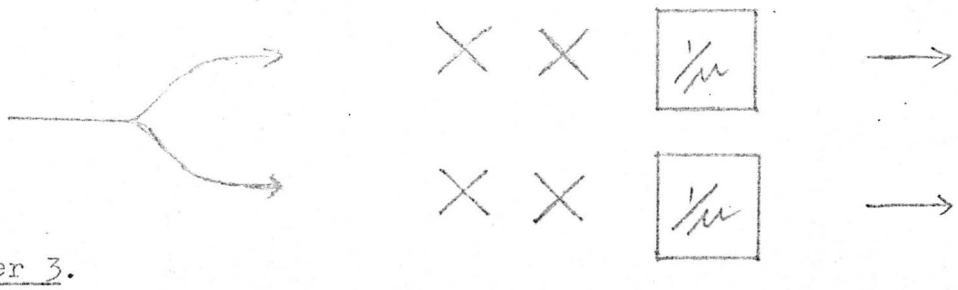
Waiting customers are shown as X's.

Chapter 2.

System 1; an M/M/2 system:



System 2: two M/M/1's:

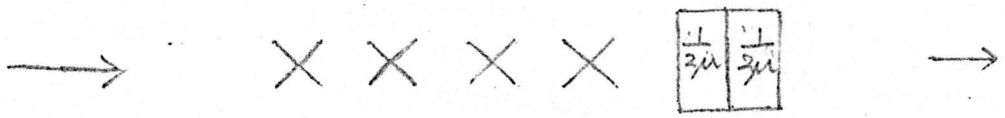


Chapter 3.

System 3; an M/M/1 :

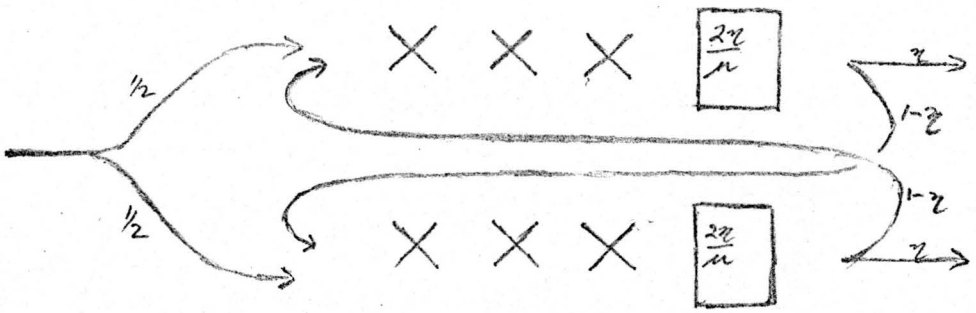
System 4: a series of two M/M/1's:

System 5: the same, with one waiting space:



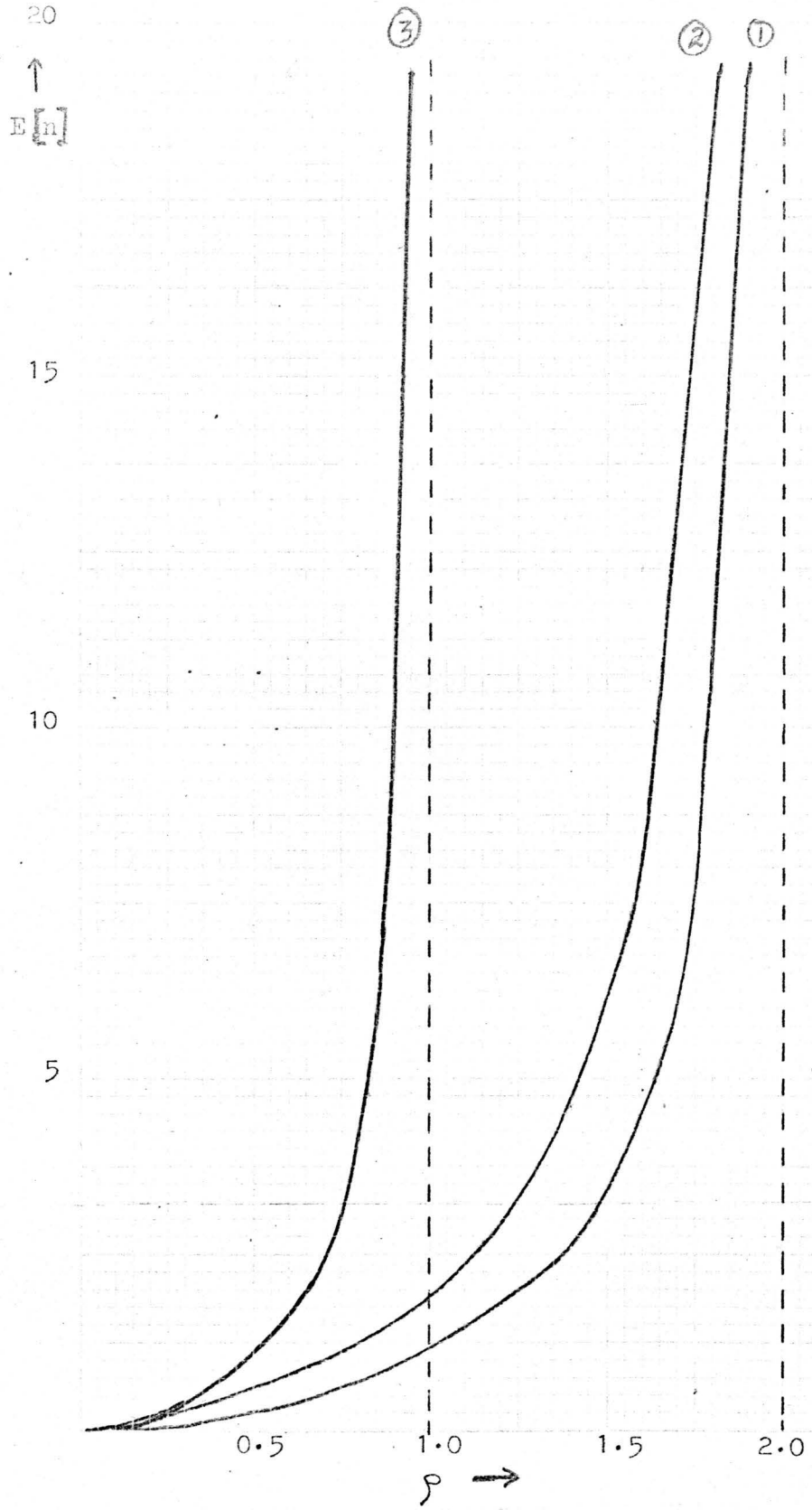
Chapter 4.

System 6; a network:



Expected Number in System

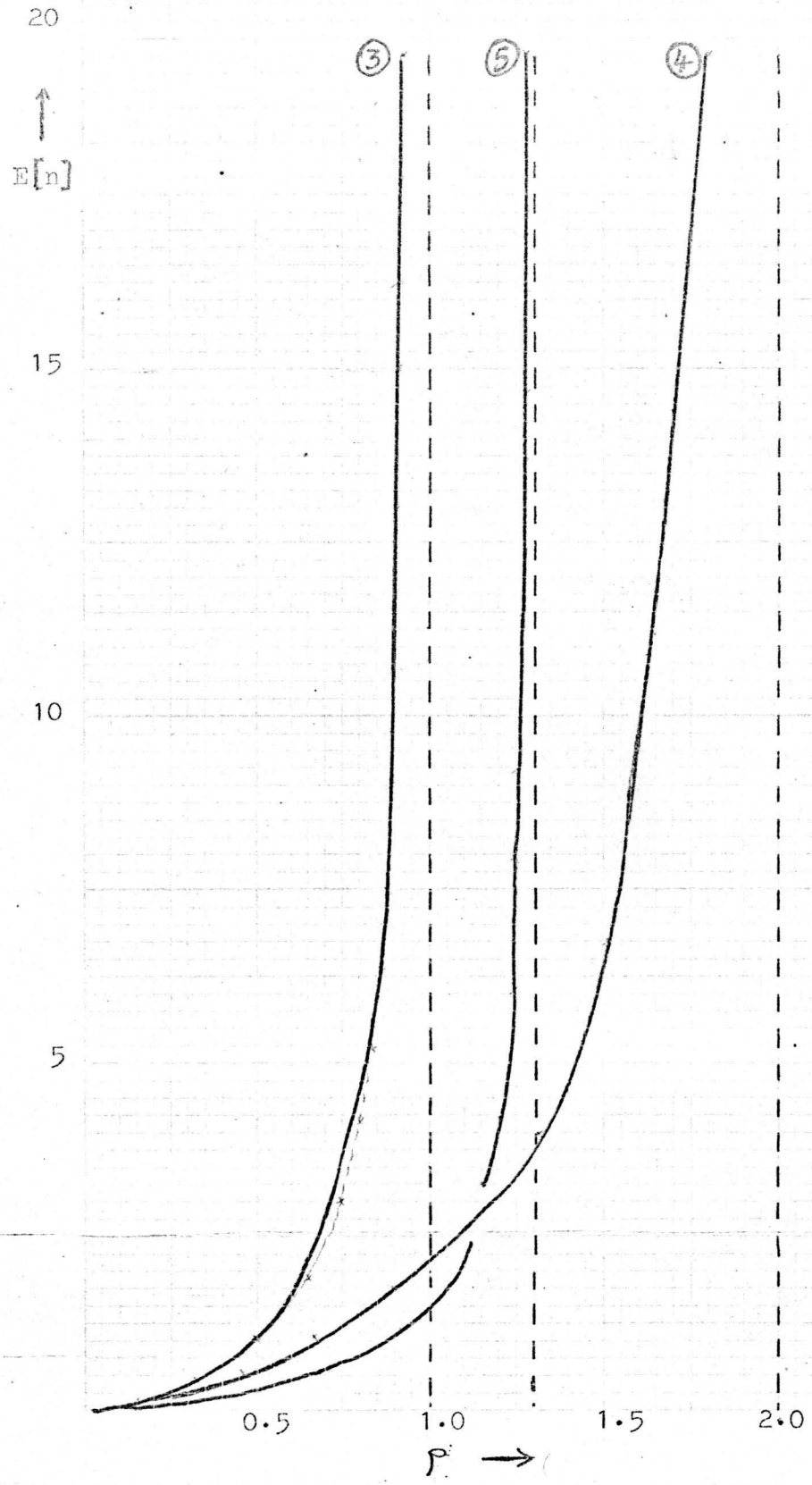
as Function of Traffic Intensity.



for ① an M/M/2. ② two M/M/1's in parallel.
③ one M/M/1.

Expected Number in System

as Function of Traffic Intensity.



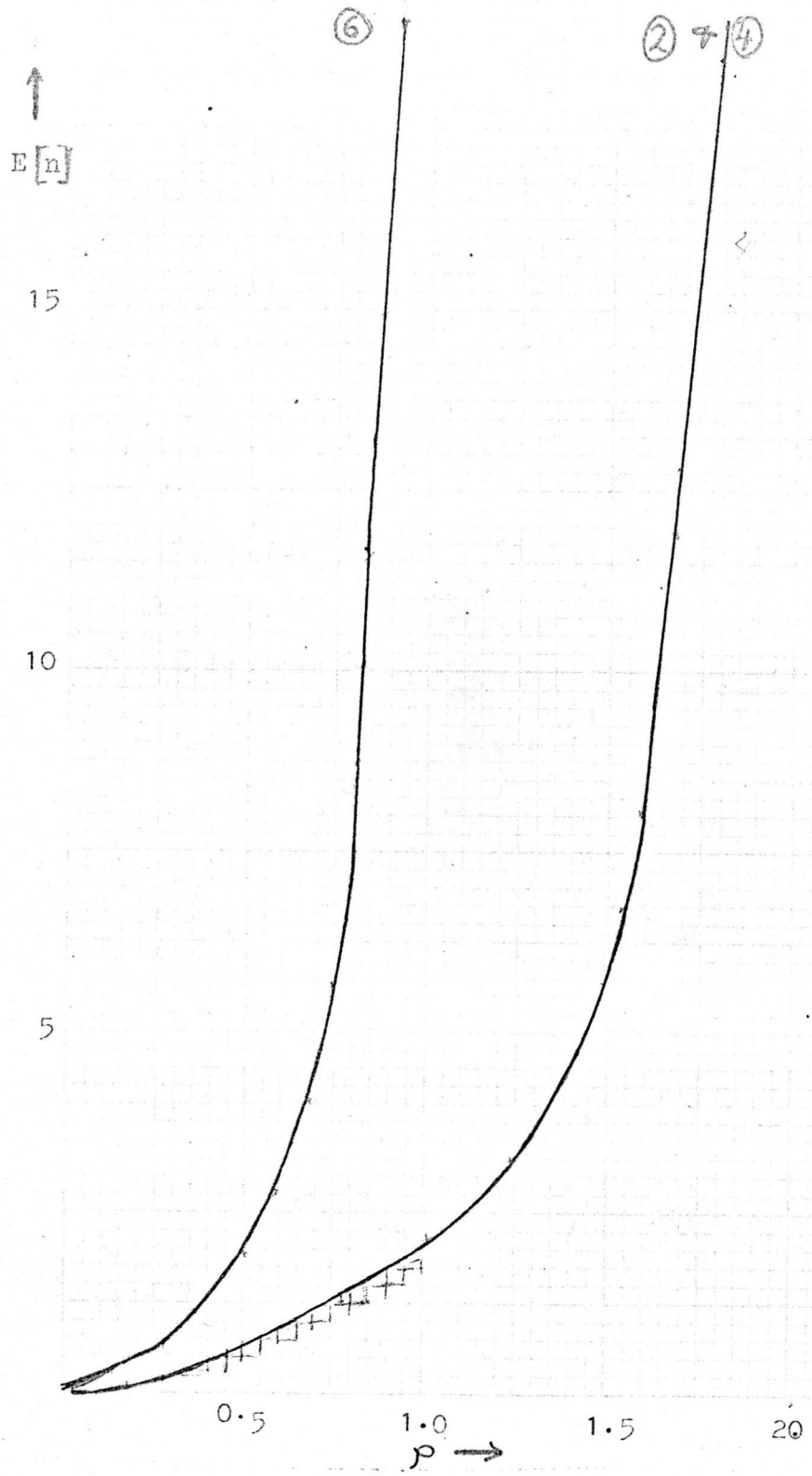
for ③ an M/M/1.

④ an M/M/1.../M/1 with unlimited room.

⑤ an M/M/1.../M/1 without waiting room.

Expected Number in System

as Function of Traffic Intensity.



for ② two M/M/1's in parallel;

④ two M/M/1's in series;

⑥ a network of two .../M/1's.

CHAPTER 5.

COMMUNICATION NETWORKS.

The chapter applies the theory for open networks of poisson queues to a type of communication network, then extends the model by applying it to a network for communication among computers. A model for a satellite network is described.

CHAPTER 5 : COMMUNICATION NETWORKS.

5.1 INTRODUCTION.

Networks (hereafter referred to as " nets ") which involve the flow of information around a set of nodes have always been a significant feature of community life. Over recent centuries and years their significance has increased rapidly, along with the growth of

.mail systems;

.radio communications;

.voice telephone systems;

.digital telegraph systems like Gentex and Telex;

and, in the last few years

.satellite nets;

.systems for communications between computing machines.

Other systems resemble these communication nets, like parcel flow in a railway system, or vehicle flow in a road net.

Two main classes of models have been explored. This chapter deals principally with the class of stochastic flow nets; but first a mention will be made of the other class: Continuous Flow Networks.

The theory of continuous flow networks does not allow for a randomly varying workload, which will fluctuate above and below the system's capacity. Nevertheless, some of the work in this field was generated specifically for the design of communications nets.

Elias, Feinstein , and Shannon (25) present a proof of the max-flow min-cut theorem. Chien (18) gives a method of relating branch capacities to terminal capacities.

Hakimi (35) generalizes the above theorem on maximum flow rates. Gomory and Hu (33) use generalized linear programming to assign branch capacities required by minimum flow rates.

Stochastic Flow Networks.

This body of theory may not have the defect mentioned above for continuous flow theory, but it does suffer from serious limitations. These are inherited from the theory of queue nets. The theory almost compels us to assume exponential distributions and unlimited waiting room - or to resort to simulation.

Analysis Versus Simulation.

This chapter should show the value of combining different approaches to a queueing problem. Kleinrock (52) lists 5 approaches.

1. Pure Analysis.

This approach reveals at once the effects of varying any of the parameters. The model, however, must be simple.

2. Approximate Analysis.

Approximate solutions are found for exact equations derived from the model.

3. Iterative Solution.

Solutions to exact equations are found by numerical analysis. This can be clumsy.

4. Simulation.

This approach can retain great detail in the model, but is expensive and slow. If there are several parameters to be varied, the desirable number of simulations can be very large.

5. Measurement.

This is the only way of verifying the results of the other approaches. It can not be performed during the design stage.

Scope of the Chapter.

The chapter does not claim to cover all the existing theory of communication nets; it should indicate along what lines and roughly to what extent the theory has developed, for nets of a certain type; the store-and-foreward type. In the simpler models, messages may have to wait, but are not lost; and delay consists only of waiting time, and transmission time - which corresponds to service time.

Outline of the Chapter

5.2 studies Kleinrock's model for a communication net; 5.3 applies and extends this to a computer net. 5.4 is a brief comment on a satellite net model.

5.2 A MODEL FOR A COMMUNICATION NETWORK.

This section outlines most of the study of a store-and-foreward communication net which Kleinrock presents in " Communication Networks; Stochastic Message Flow and Delay " (48) . The work is a significant advance in, and application of the theory of poisson queues, so it is treated here at some length.

The rest of this section

.shows how a communication net can be reduced to a net of independent exponential single-server stations with unlimited storage;

.presents Kleinrock's theorem on the optimal assignment of channel capacities;

.discusses the validity and value of the " independence

assumption ";

.compares three routing procedures.

Kleinrock's study of the effect of priority on delay in the steady M/M/1 and M/G/1 systems is omitted.

The terminology has been altered, to make clear the difference between communication centres and transmission channels.

5.2.1 THE NETWORK AND THE MODELS.

Kleinrock (48) describes a simple net to which many real-life nets are very similar. It consists of several "communication centres "; each has tributaries from which it receives and to which it passes messages. Thus messages enter and leave the net from each centre. Each centre can pass messages to some or all of the others. When a message arrives at a centre, it waits till a suitable channel, i.e., a transmitter, is free. Each of these channels leads to one other centre. Kleinrock does not clearly state whether there is one queue within each centre, or one before each channel. We choose the latter, and so transform the net of centres, " model 1," into a net of channels," model 2 ". This new net resembles the model of J.R.Jackson.

The Network of Channels.

Before each channel there is a waiting space (an electronic memory; or paper tape), assumed unlimited in capacity. A message does not join the queue in this space until it has arrived in its entirety. The time spent in arriving has already been accounted for as the service time at the sending channel. On arrival at a centre, a message either leaves the system, or joins a queue before one of the channels leading

out of the centre. In terms of our net of channels; the message leaves one channel, and either disappears from the system, or is routed to the queue before another channel. The only channels to which it can go are those of the centre at which it is arriving.

At each transmission, the service time depends on the message length, and the capacity (in bits/second) of the channel.

The message lengths are assumed exponential, mean μ; arrivals from outside are poisson. A capacity, C_i (in bits/second) is assigned to each channel i.

Two Problems.

Model 2 resembles J.R. Jackson's model, with single-channel stations and state-independent rates; call this "model 3." Two problems prevent us from identifying model 2 with model 3:

1. Each message in model 2 has a definite origin and destination. Each customer has a definite origin, but wanders among the stations at random until it chances to leave the system.
2. Each message has a fixed length; whereas in model 3, the services for a particular customer are independent of each other.

The first problem is not too serious. If the output from each channel of model 2 is exponential, the input to each will be exponential. The nett arrival rate, λ_i, at channel i, will depend on the routing procedure; but the nature of the distribution will not.

Kleinrock deals with the second problem by assigning each message a new length, each time it arrives at a channel, from an exponential distribution with mean

($1/\mu$) bits.

Thus the lengths assigned to a particular message are independent of each other; this is the "independence assumption".

Delay at an M/M/1 System.

We consider a single transmitter, receiving a poisson stream, rate λ_i, of messages with exponential length, mean $1/\mu$; and the capacity of the transmitter is C_i. Let

$T_i = E[\text{delay (=waiting time plus service time)}]$,

for a message in the system. Then (see 2.3.1)

$$T_i = \frac{1}{\mu C_i} \cdot \frac{1}{(1 - \lambda_i / (C_i \mu))}$$

Mean Delays in the Networks.

Consider the time elapsing for a message in model 1, between arrival from a tributary at a centre, to departure from a centre to a tributary. This time is identical with the time elapsing for a message in model 2, between arrival from outside before a channel, to departure from a channel to the outside. If we impose the independence assumption on model 2, it becomes identical with model 3, and the channels of model 2 are identified with the single-channel " stations " of model 3.

The delays at the stations of model 3 are independent of each other. Station i of model 3 has the same equilibrium behaviour as the M/M/1 system, so the delay at station i of model 3 is T_i. Therefore the delay at channel i of model 2 is T_i, and the T_i's of model 2 are independent.

Since the λ_i represent the nett arrival rate at the channels of model 2, the mean delay for a message passing through model 2 is

$$T = \left(\sum_{i=1}^{N} \lambda_i T_i \right) / \gamma$$

where

N = number of channels in model 2; and

γ = total arrival rate of messages from outside.

Let

$$\lambda = \sum_{i=1}^{N} \lambda_i ;$$

then the mean number of channels which process a message is $\bar{n}$, where $\bar{n} = \lambda / \gamma$.

5.2.2 OPTIMAL CAPACITY ASSIGNMENT.

Let d_i be the cost of supplying one unit of capacity to channel i of model 2. The total cost of the network is D, where $D = \sum_{i=1}^{N} d_i C_i$.

This costing model assumes that: the whole cost of the net lies in the transmitters (channels) and transmission lines; and that the cost-capacity relationship is linear.

The first theorem of Kleinrock (48, 144) below gives the C_i which minimize T for fixed D. The second gives the λ_i which minimize T under the optimal assignment of the C_i. The proof is given of the first only.

First Theorem.

(All sums in rest of this section are for $j = 1, \ldots, N$)

Let $G = T + a\,(\Sigma d_j C_j - D)$,

where a is a Lagrange multiplier. The N equations: $dG/dC_i = 0$;

yield: $C_i = \lambda_i/\mu + \frac{1}{\sqrt{\lambda a}}\sqrt{\frac{\lambda_i}{d_i \mu}}$, $i = 1,\ldots,N$.

To find a, we multiply each by d_i, and sum:

$D = \Sigma d_j C_j = \Sigma\left((d_j \lambda_j)/\mu\right) + 1/(\sqrt{\lambda a}) \cdot \Sigma\sqrt{\lambda_j/(d_j \mu)}$. Then

$$C_i = \lambda_i/\mu + \left(\left(\sqrt{\lambda_i/d_i}\right)/\left(\Sigma\sqrt{\lambda_j/d_j}\right)\right).\left(D - \Sigma\left(d_j \lambda_j/\mu\right)\right) .$$

In this optimal assignment, each channel receives enough capacity to reduce its traffic intensity to 1 or less:

$C_i \geq \lambda_i/\mu$; so:

traffic intensity $= \lambda_i/(C_i \mu) \leq 1$.

The assignment : $C_i = \lambda_i/\mu$: leaves a part D_e of the total cost D unfulfilled, where $D_e = D - \Sigma d_j \lambda_j/\mu$.

This part of the cost is allocated in proportion to $\{$(arrival rate)/(cost per unit of capacity)$\}^{1/2}$ Under this optimal allocation,

$$T = (\bar{n}/\lambda)\left(\Sigma\sqrt{\lambda_j d_j}\right)^2/D_e .$$

T depends strongly on $\bar{n}$ and D_e.

Second Theorem.

In the special case where cost is related directly to capacity, we can take

$d_i = d = 1$, for $i = 1,..,N$; and

$D = C$.

It is to be expected that under any routing procedure, some or all channels will have a non-zero minimum arrival rate; let this be k_i for channel i.

Under the special-case optimal assignment, and under these constraints:

$\lambda_i \geq k_i$; $i = 1,..,N$; and

$\Sigma \lambda_i = \lambda$;

the assignment of the λ_i which minimizes T is:

$$\lambda_1 = \lambda - k_2 - k_3 - \cdots - k_N$$

$$\lambda_i = k_i \text{ , for } i = 2, 3, .., N .$$

Thus the best routing procedures concentrate as much traffic as possible into a few; preferably 1, channels. This channel will receive as much as possible of the capacity.

5.2.3 THE INDEPENDENCE ASSUMPTION.

The idea of a message changing its length is obviously unrealistic; but Kleinrock makes use of the independence assumption for these three reasons:

1. The mathematics for even a 2-station series " net " is intractable without it. Permanent message lengths produce a dependency between inter-arrival times and lengths of adjacent messages as they move through a net.
2. Messages arriving at a model 1 centre usually have several channels out of the centre. Dependencies should be reduced when adjacent messages take different channels.
3. Experimental evidence backs up the guess of 2. Kleinrock simulated nets consisting of a central node, with

(A) one channel leading in, one channel leading out;

(B) several channels leading in, one channel leading out;

(C) one channel leading in, several channels leading out;

(D) several channels leading in, several channels leading out.

The graphs of mean message delay in (C) and (D), even with only two channels leading out, were close to exponential; and quite different from the graphs for (A) and (B). Graphs of mean message delay in more complex nets with the messages retaining a permanent length, were very similar to graphs for

the same nets with independent message lengths.

Kleinrock concludes that the introduction of the independence assumption produces an easily analysed model, which behaves in essentially the same way as the more realistic model.

5.2.4 ROUTING PROCEDURES.

Kleinrock discusses the relative merits of three routing procedures. The theoretical work above:

.yields the optimal capacity assignment;

.makes possible an analysis of random routing;

.suggests the general principles that traffic should be concentrated in as few channels as possible, and that $\bar{n}$ should be kept as small as possible.

Simulation methods have to take over at this point.

Description of the Three Procedures.

Under random routing, the message receives directions to its next destination each time it leaves a centre; according to a probability matrix. Only circulant matrices are tested - these are special cases of doubly stochastic matrices.

Under fixed routing, the message is assigned a destination before it enters the net. Given its present location and its destination, its next location is determined uniquely by an " incidence matrix ".

Alternate routing is less rigid, but still deterministic. The entries of the incidence matrix are now lists; the message goes to the first node on the list a channel to which is idle.

Advantages and Disadvantages.

Random routing has the obvious disadvantage that a message

may take a long random walk among the nodes before chancing on its destination. The mean and variance of the number of steps required are large; average delay is large; the net becomes overloaded at comparatively low rates of input. The advantages, apart from theoretical simplicity, are that no directory information is required, and that a net with random routing is little affected by failure of some of the nodes or links.

Alternate and fixed procedures both tend to concentrate the traffic better; and so make best use of a good capacity assignment. Alternate routing is worse than fixed, in that it spreads the traffic; and in times of congestion, it sends messages via longer routes, thus making the traffic load even heavier . It is better than fixed in that often a good capacity assignment cannot be made. The traffic matrix may not be known clearly, or may vary with time.

5.3 A MODEL FOR A COMPUTER NETWORK.

One of the most significant developments in computing during the 1960's was the advent of time-sharing. With the help of their time-sharing machines, various institutions around the world have developed highly specialized resources such as programs, data files, hardware and human talent. The desire to make these localized resources available to distant users led to the concept of computer networks. According to Kleinrock (52), this concept " represents the next major breakthrough in the use of computers."

The membership of the net described below consists of university and research laboratory systems. The future of

computer nets is not limited to such systems; there are already commercial systems in existence, like the Bank Data Processing System in New Zealand.

The Advanced Research Projects Agency (A.R.P.A.) network is treated in some detail here, for these reasons:

.A vast amount of thought, queueing theory, and queue simulation have been devoted to it.

.These studies follow on from the studies of nets and communication nets in this and earlier chapters.

.Little if any information on other computer nets is available in the journals of Operations Research and Computer Systems.

.The store-and-foreward system described here is typical of many present and future systems, not all of which are computer nets.

This section describes the A.R.P.A. net, then summarizes some of the studies done during its planning and experimental stages. These studies are the work of various A.R.P.A. personnel (13), (30), (36),(75) ; much of the work and the reporting comes from Kleinrock (52), (53), (54).

In a private communication (55), Kleinrock states that " the study of communication networks has recently become of central importance in the design of computer networks."

5.3.1 A DESCRIPTION OF THE ARPA NETWORK.

In 1968, after preliminary investigations, the Advanced Research Projects Agency of the U.S. Department of Defence decided to proceed with the plan of linking a number of computing centres, scattered throughout the U.S. .

Previous attempts at linking distant computers had not

been very successful. Roberts and Wessler (7[illegible]) point out that the problems resulted from the inadequacy of, or inefficient use of, the available communication services. Such services would be either too slow or too expensive; and the advantages of linking several systems are small unless a considerable number of systems are involved. Such a number would require extremely expensive communication links.

The solution to these problems lay in connecting the member systems only via a store-and-foreward net of small special-purpose machines. The member systems are called Hosts; the small machines are Imps (for Interface Message Processor).

The Hosts.

The network was designed to grow till it contained 19 or more Hosts. It has now attained this state. Each Host is a research centre, and consists of one or more machines; program and data files; and the resident human talent.

The choice of the 19 members was made so that the net would include numerous computer researchers; a variety of specialized facilities; workers in a variety of disciplines; machines with incompatible hardware structures. The experiment was intended to demonstrate the possibility of co-operation among extremely different elements.

The Hosts communicate with the rest of the net only through their Imps; and should otherwise function essentially as they did before connection into the net.

The Imps.

As Hosts contact the net only through their Imps, the system reduces to a network of 19 Imps.

The essential function of the Imp net is to convey messages reliably and quickly. " Real " messages originate only from the Hosts. If Host A has a message for Host B, the message enters the net-after being broken down into 1 kilobit " packets " - at Imp A. The packets are relayed separately across the Imp net to Imp B, where they are reassembled, and passed to Host B.

Viewed without the Hosts, the Imp net looks very much like the model 1 net of the previous section: messages arrive in and depart from the net at each Imp, and are passed from Imp to Imp. However, this net has certain special features.

The Message Traffic.

The requirement of reliability is met by various hard- and software precautions. The strategy which concerns us here is the acknowledgements procedure. Each packet of a message is passed from Imp to Imp separately. The receiving Imp performs a cyclic error check, then sends an acknowledgement. The sending Imp keeps a copy of the packet until it receives an acknowledgement. If the acknowledgement does not arrive (because of detection of a transmission error, or full buffers at the receiving Imp), the sending Imp transmits the message again.

Various other short non-" real " messages go from Imp to Imp; these concern the traffic density throughout the system.

The Routing Procedure.

The procedure used is a development from alternate-routing. Each Imp keeps a matrix which tells it the average delay which a message for any destination will suffer, if sent out along any channel. These matrices are updated often.

Special Features of the Imp Network.

The Imp net differs from the model 1 communication net in these ways.

1. The Imp's storage is limited- to about 70 packet-sized buffers.
2. Extra traffic arises from retransmission after errors or buffer overflow.
3. There are two distinct types of message:
   "real" messages, i.e., Host-Host messages, whose mean length is $1/\mu'$, = 560 bits; and Imp-Imp messages, which are shorter, and have a distribution which approximates the exponential, with the overall mean length $1/\mu$, = 350 bits.
4. A request from Host A to Host B usually produces a reply from B to A. The reply will usually be longer than the request.
5. The length of error checks and other processes at the Imps are significant when compared with the transmission time delays.
6. Channel costs are more complex.
7. Channel capacities are restricted to a few discrete values.
8. The time for a signal to travel the length L_i of the ith channel is non zero; let P be such that this time is PL_i.

5.3.2 STUDIES ON THE ARPA NETWORK.

The realization of a net of this nature is an expensive process. Consequently, much analysis and simulation was done during the designing phase. The aim of these studies was to optimize the performance of the net for variable topological structure, capacity assignment, routing procedure, and priority discipline.

Analysis of Average Message Delay.

The assumption that each message length comes from the same exponential distribution produces T-values which compare very badly with values reached by simulation. The T we wish to know is not the overall delay, but the delay for " real " messages. The travel time; the short (assumed constant) processing time at Imp and Host; and the message type problem are all accomodated by this model:

$$T = \sum_{i=1}^{N} \frac{\lambda_i}{\gamma}\left(\frac{1}{\mu C_i} + \frac{\lambda_i/(\mu C_i)}{\mu C_i - \lambda_i} + PL_i + 10^{-3}\right) + 10^{-3}$$

... in seconds.

Special features 3, 5 and 8 have been accounted for. The mean delay has been replaced by mean service for " real " messages, plus mean wait for all messages.

The assumption of a mix of two exponential distributions for message lengths may not be accurate, but the results produced by this model are in close agreement with simulation results. (The model also uses the independence assumption). The analysis can be extended further, to include the effect on T of priorities.

The Pollaczek-Khintchine formula can be used for the mean waiting times of more general distributions; however, the independence property is lost.

Channel Cost Functions.

Data indicate that the most realistic relation between channel cost and length is either:

$$d_i = A\, L_i^{\,a} \quad \text{where } 0 \leq a \leq 1;$$

or:

$$d_i = A \ln(aL_i) .$$

Kleinrock (54) finds the optimal capacity assignments for

these two relations.

Routing Procedures.

Kleinrock (64) reports the results of simulations for three variants of dynamic alternate routing. These are:

1. Synchronous updating- the matrices are updated every half second;
2. Synchronous updating with loops suppressed: packets are prevented from oscillating between two Imps;
3. Asynchronous updating; the matrices are updated only when traffic densities change markedly.

The performance improves from 1. through 2. to 3.

Topology.

Perhaps the first decision in designing a net is to select which pairs of Imps are to receive direct links. Frank, Frisch and Chou (30) use a mixture of iteration and analysis first to create new feasible topologies, then to calculate their performance measures.

Other Problems.

Analytic methods for coping with special features 1, 2 and 4 are not available. Their effect can be judged by comparison of simulation results.

5.4 A MODEL FOR A SATELLITE NETWORK.

A satellite network requires a model somewhat different from those discussed above. El-Bardai (24) presents a model, and finds an expression for the expected delay along each possible route.

The Model.

There are m ground stations g_i, and n satellites s_j.

All are at rest relative to one another. Each g_i can communicate only with certain " near " s_j; and similarly each s_j can communicate only with certain g_i. Only messages to be relayed from g_i, say, via one satellite, to g_k, are considered. Messages for g_k arrive at g_i in a poisson stream of rate λ_{ik}, and queue there for transmission. The net is synchronized so that transmission of a message can begin only at the discrete time points 0, d, 2d, 3d... . Transmission time is less than d; but the distances are such that the transit times (the times for a signal to pass from one node to another) are important. The g_i have infinite storage; the s_j can only receive and retransmit one message; they have no storage. If more than one message is transmitted to an s_i at the same time, only one message is retransmitted; the others are lost.

A message destined to go from g_i to g_k will become lost with a certain probability; and will reach g_k with the complement of this probability. If it reaches g_k, its total delay will consist of its waiting time at g_i, plus two transit times. El-Bordai finds the probabilities of loss, and the expected total delays, then presents an approximate method for treating satellites with finite memories.

CHAPTER 6.

COMPUTER TIME-SHARING.

The chapter applies some of the theory for single-server queues, extended to describe feedback, to a simple time-sharing system; then outlines the development of more complex models from this simple model.

CHAPTER 6 : COMPUTER TIME-SHARING.

6.1 INTRODUCTION.

Computer systems function under fluctuating workloads, and so are candidates for study as queue systems. This chapter mentions some of the computer applications of queues, then concentrates on the application which has received most attention.

The computers of Chapter 5 provide input to and accept output from the queue network, without being part of it. This chapter does not develop from the immediately previous chapters, but returns to the theory of Chapter 2. This theory has been applied and extended in the study of computer systems.

Computer Systems and Queues.

Some computer systems or subsystems can be studied via the existing single-station theory. Chang (17) presents a summary of relevant results on single-server queues for use by systems analysts; together with a long list of applications. For other subsystems, special models have been developed. Examples include:

.The study by Boudreau and Kac (11) of a closed two-station system;

.Chang's model (14) for the queue of requests for access to main storage- the requests come from the input channels and the processor, and are serviced in bulk;

.Coffman's study (20) of the queues of requests for access to drum storage- as the drum rotates, each queue is served in order; producing a set of systems which resemble M/D/1 systems;

.Chang's model (16) for real-time processing- each job consists of a high-priority communications task, plus a longer processing task.

Most of these applications deal with the problem of matching slow in/out devices with the fast processor.

A field which has received more thorough treatment is that of priorities. External priorities can be fixed or delay-dependent , with or without pre-emption, with or without resumption of a rejected job. Kleinrock's book (48,71) devotes a chapter to these. External priorities will not be considered further here.

There remains the problem of providing shorter waiting-times for shorter jobs. The optimal strategy is to apply the shortest-first discipline; but this is impossible if job lengths are unknown until the end of the job. The technique employed is time-sharing.

The Aims of Time-Sharing.

Time-sharing is designed to make best use of both the machine and its human users, when the machine-user relationship is an interactive or conversational one. It is not designed for the machine with a few long jobs to process, but for the machine with many short jobs and many users.

Krishnamoorthi and Wood (57) list the following desirable properties.

.simultaneity: each user has the impression that he has full use of the machine - or at least of as much of its potential that he can keep up with;

.independence: the programs handled do not interfere with each other;

.immediacy: all jobs receive some attention within seconds of being submitted.

To achieve these goals, each job receives a small quantum (from a few milliseconds up to about 10 seconds) in the processor; then is returned to the user; or waits. To provide models for such systems, the theory of queues with feedback has developed rapidly.

Scope of the Existing Theory.

Since the end of the 1950's, numerous working time-sharing systems have appeared, together with numerous theoretical papers, systems analysis papers, and some simulation studies. Coffman (19) summarized the theory available by 1967, and a further summary was compiled by McKinney (63) in 1969.

It is interesting to note that this branch of Queue Theory, like the theory for networks, is largely restricted to models with exponential arrival and service times. Besides these, models with exponential arrivals from a finite number of sources have been developed to treat systems where the number of users is restricted to the number of terminals.

The independence property of the exponential distribution is shared by its discrete analog : the geometric distribution. One of the earliest time-sharing models used the assumption that job lengths were restricted to discrete values:

(quantum size).n ; n = 1, 2,...

and geometrically distributed; with arrivals occuring as Bernoulli trials at the end of each quantum.

The M/G/1 system with feedback has also received attention.

There seems to be little published work on simulation;

the major work in this field is the monograph of A. Scherr (80).

Outline of the Chapter.

This section has presented some general ideas on computer applications of queue theory, the time-sharing application in particular. 6.2 describes and analyses one basic model in detail. 6.3 lists the many possible variations on this model, and indicates briefly the mathematical methods used.

6.2 A SIMPLE TIME-SHARING SYSTEM AND MODEL.

The system treated here is chosen as a typical example of time-sharing systems; and because analysis of the model applies and extends the theory of poisson queues presented in earlier chapters. Again, the simplicity results largely from the poisson assumptions.

6.2.1 THE SYSTEM AND THE MODEL DESCRIBED.

The M/M/1 system with feedback occurs in other situations apart from computing, but since computing systems motivated the development of the theory, we describe one time-sharing system here before describing the model designed for it.

The Computer System.

A stream of jobs enters the system, and forms a queue. Usually the job does not arrive at the time-sharing part of the system until it has passed through an input device, and is stored on disc or otherwise. After waiting for its turn, the job receives one quantum or slice of computer time. If it finishes within this time, it departs - this may involve queueing again before an output device. If it does not finish, and other jobs are waiting, the processing stops,

and the job is stored (in core, if there is room; or by rolling it out onto disc). The next job enters the processor. The time involved in this transfer is the swap time. If stored, the job rejoins the queue, at the back. After a further wait it receives another quantum of processing; and so on until it is completed.

The Model.

The arriving stream of jobs is poisson, rate λ. Each job requires a certain processing time t; and t is exponential with mean $1/\mu$. Quantum length is a constant, q. Swap time, T, is taken to be zero. The space for storage of new and incomplete jobs is taken to be infinite.

The assumption of poisson input is likely to be valid for systems with many independent users; many systems do have exponential job lengths; and the memoryless property means that a job which survives a number of quanta still has an exponential life ahead of it. The system can be designed to hold quantum size almost constant. The assumption of zero swap time introduces systematic differences between system and model.

6.2.2 ANALYSIS AND RESULTS.

The system designer and manager will be interested in throughput; and " congestion ", i.e., the distribution of number in system. The latter affects the amount of storage to be provided for waiting jobs. The customers will be interested in the turnaround time. Both groups are concerned with queues before in/out devices; but only the subsystem between these is analysed here.

Queue Length.

In any interval dt, an arrival occurs with probability $\lambda\,dt$, and a departure with probability $\mu\,dt$. The occurrence of a swap within the interval makes no difference, as we have assumed zero swap time, and the swap does not alter the number in the system. Hence the distribution of number in system is identical with that for the simple M/M/1 system:

$$P_n = \rho^n(1 - \rho) ,$$

where

$$\rho = \lambda/\mu .$$

As for M/M/1, the upper limit on throughput is μ jobs/second.

Turnaround Time.

Since time-sharing aims to give preference to short jobs, we examine the expected value of the turnaround time or response time, R, for a job of given length, t. The analysis is the work of J. Shemer (82).

The job requires N passes through the processor, where

$$(N - 1)q < t \leq Nq .$$

Let

w_i = E [waiting time(without service) between (i-1)th and ith pass / t] , i = 1,..,N.

The turnaround time consists of N waits and the processing time; so

$$E[R/t] = \sum_{i=1}^{N} w_i + t .$$

However much processing a job has already received, the probability that a job still requires more than a quantum of processing is

$$\int_q^\infty \mu e^{-\mu t}\, dt ,$$

$$= e^{-\mu q} .$$

Thus the expected length of each pass is Q,

where $$Q = \int_0^q t.\mu e^{-\mu t}\, dt + q.e^{-\mu q}$$

$$= (1/\mu)(1 - e^{-\mu q}).$$

(As $q \to 0$, $Q \to 1/\mu$.)

When job J arrives, the expected number ahead of it is $E[n]$, and

$$E[n] = \rho/(1 - \rho).$$

The expected wait before J's first pass is w_1, where

$$w_1 = Q\rho/(1 - \rho).$$

During J's first wait and pass, jobs join the queue behind J from two sources:

. jobs arrive from outside at rate λ;

. $(e^{-\mu q})$ of the jobs before J require more processing.

Thus
$$w_2 = Q(\ (w_1 + q)\lambda + E[n]\ e^{-\mu q}\ ).$$

Similarly,

$$w_i = Q(\ (w_{i-1} + q)\lambda + (w_{i-1}/Q)\ e^{-\mu q}\ )$$

$$= g w_{i-1} + \lambda Qq \quad , i = 2, 3...$$

where

$$g = \lambda Q + e^{-\mu q}.$$

Since

$$w_3 = g\, w_2 + \lambda Q q = gg\, w_1 + \lambda Qq\, g + \lambda Qq$$

etc, to

$$w_N = g^{N-1} w_1 + \lambda Qq(1 + g + \cdots + g^{N-2})$$

$$= g^{N-1} w_1 + \lambda Qq(1 - g^{N-1})/(1 - g)$$

$$= g^{N-1}(w_1 - \lambda Qq/(1 - g)) + \lambda Qq/(1 - g);$$

The expected total waiting time is W, where

$$W = \sum_{i=1}^{N} w_i$$

$$\frac{(1 - g^N)}{(1 - g)} \left(w_1 - \frac{\lambda Q q}{1 - g} \right) + \frac{N \lambda Q q}{1 - g} \cdot$$

Expected turnaround time for a job of length t then is given by:

$$E[R/t] = \frac{(1 - g^N)}{1 - g} \left(\frac{Q\rho}{1-\rho} - \frac{Q q \lambda}{1-g} \right) + \frac{N \lambda Q q}{1 - g} + t ;$$

$$Q = (1/\mu)(1 - e^{-\mu q}) ;$$

$$g = \lambda Q + e^{-\mu q} ;$$

$$t/q \leq N < t/q + 1 .$$

An Extension.

In the limit as $q \to 0$, the time-sharing model becomes a " processor-sharing " model. Every job in the system is processed for the whole time it is in the system; at rate:

μ /(number in system).

Let
$q \to 0$

Then $Q \to q$,

$g \to 1 - (\mu - \lambda)q$;

$q N \to t$,

and so

$E[R/t] \to t/(1 - \rho)$.

The P_n remain unchanged.

This processor-sharing model is obviously unrealistic, as it does not account for swap times. However, it demonstrates the limit to which the time-sharing model tends.

Graphs at the end of this chapter display $E[R/t]$ for the time-sharing model (with $q = \mu/2$), and for the two limiting cases;

$q \to \infty$; the simple M/M/1 model;

$q \to 0$; the processor-shared model.

As is apparent for the cases graphed, jobs with

$$t < 1/\mu$$

have their turnaround time reduced.

6.3 OTHER MODELS.

This section is not a summary of time-sharing theory, but an outline of the models which have received study plus a few examples illustrating the methods involved in the theory.

6.3.1 POINTS OF VARIATION.

The many possible time-sharing models can all be considered as variants or developments from the simple model of 6.2. The points within the model at which variations can be introduced are listed here; together with the variants which have received study.

Number of Processes.

Though systems exist with more than one processor in parallel, the theory has not developed past single-processor models.

Number of Queues.

The Round Robin model (RR) with one queue leads to the Foreground-Background model with N queues (FB_N). In the latter, arrivals join queue 1. After their first pass they join queue 2 or leave the system; and so on to queue N, which is an RR queue. Queue i receives processing only when queues 1, 2, ..i-1 are empty. FB_N systems give an even faster service for short jobs than RR systems, and do not suffer from overloading by long jobs.

Models studied include RR, FB_2, FB_N, and FB_∞. Their

relative merits are discussed by Adiri and Avi-Itzhak (3), in their study of the FB_N model.

Quantum Size.

Quantum size can be treated as a constant, or as a random variable. It can be dependent on the state of the system, on the amount of processing already granted, or on external priorities. The aim is to reduce the time lost in swapping, and so to improve performance under heavy loading. Waiting times for short jobs will be increased. Mullery and Driscoll (64) present an algorithm in which quantum size is constant; but swaps are permitted only when necessary.

Many of the models have been transformed into processor-sharing models, by allowing q to approach zero.

Swap Time.

Swap time can be assumed zero, or treated as a positive constant or random variable.

Arrival Process.

The interarrival distributions studied are restricted entirely to geometric, exponential, and exponential from a finite number of sources. There is probably no need for models with other arrival assumptions.

When arrivals occur by Bernoulli trial at constant intervals, a geometric inter-arrival distribution results.

Job Lengths.

Most of the papers assume either exponential or geometric job lengths. Chang (15) and recently Sakata, Noguchi and Oizumi (79) have studied models with general job length distributions.

Other Studies.

Besides $E[R/t]$ and P_n, other performance measures have been studied; these include cycle time and busy period distribution. Coffman and Kleinrock (21) study tactics which the user can employ to " beat " the scheduling system. A little work has been contributed on cost structures.

6.3.2 MATHEMATICAL METHODS FOR SOME OTHER MODELS.

The methods involved in analysis of three of the more significant models are outlined briefly here.

The Discrete-Time Model.

The model with geometric arrivals and job lengths is significant in being one of the simplest, and one of the first to be studied. The arrivals can occur just before or just after the ends of the quanta. The " late arrivals " case is considered here.

Feedback makes no difference to queue length, so the equilibrium probabilities for the instants after the end of a quantum and before an arrival are given by:

$$P_n = (1 - a)a^n$$

where $a = \rho\sigma/(1 - \lambda q)$;

$\rho = \lambda q/(1 - \sigma)$;

$q = \Pr[\text{an arrival occurs at the end of a quantum}]$;

and job length distribution is given by

$$s_n = (1 - \sigma)\sigma^{n-1}, \quad n = 1, 2..$$

In various publications (48), (50), (51), Kleinrock finds $E[R/t]$ (where R is turnaround time, t is job length), by a method similar to that used by Shemer (see 6.2.1) for the exponential model. The expected waits before each pass are found from a recurrence relation, and summed.

Finite-Source Model.

The model with exponential arrivals from a finite number of sources is treated in numerous papers; its significance comes from its applications to systems with a finite number, N, of input terminals, each of which can have only one job at a time in the system.

Adiri (2) shows that the sequence of departure epochs is a sequence of regeneration points: the time between departures has the same distribution as the total processing time required, since queue length is unchanged by a swap not involving a departure. There is a finite number of states (N + 1), so the model behaves as a markov chain. The equilibrium queue length probabilities are found from the matrix of transition probabilities. $E[R/t]$ can then be found by the familiar recurrence and summation procedure.

The RR M/G/1 Model.

The significance of this model comes from the non-exponential job lengths in some systems. Sakata, Noguchi and Oizumi (79) present a derivation of $E[R/t]$ which is similar to the derivations above. Here it is necessary to divide the jobs into classes: class i contains the jobs which will need i phases. Instead of considering the expected number of jobs which join the queue behimd job J, M_{ij} has to be found; where

$M_{ij} = E$ [number of class j jobs in queue when a class i job begins a pass] .

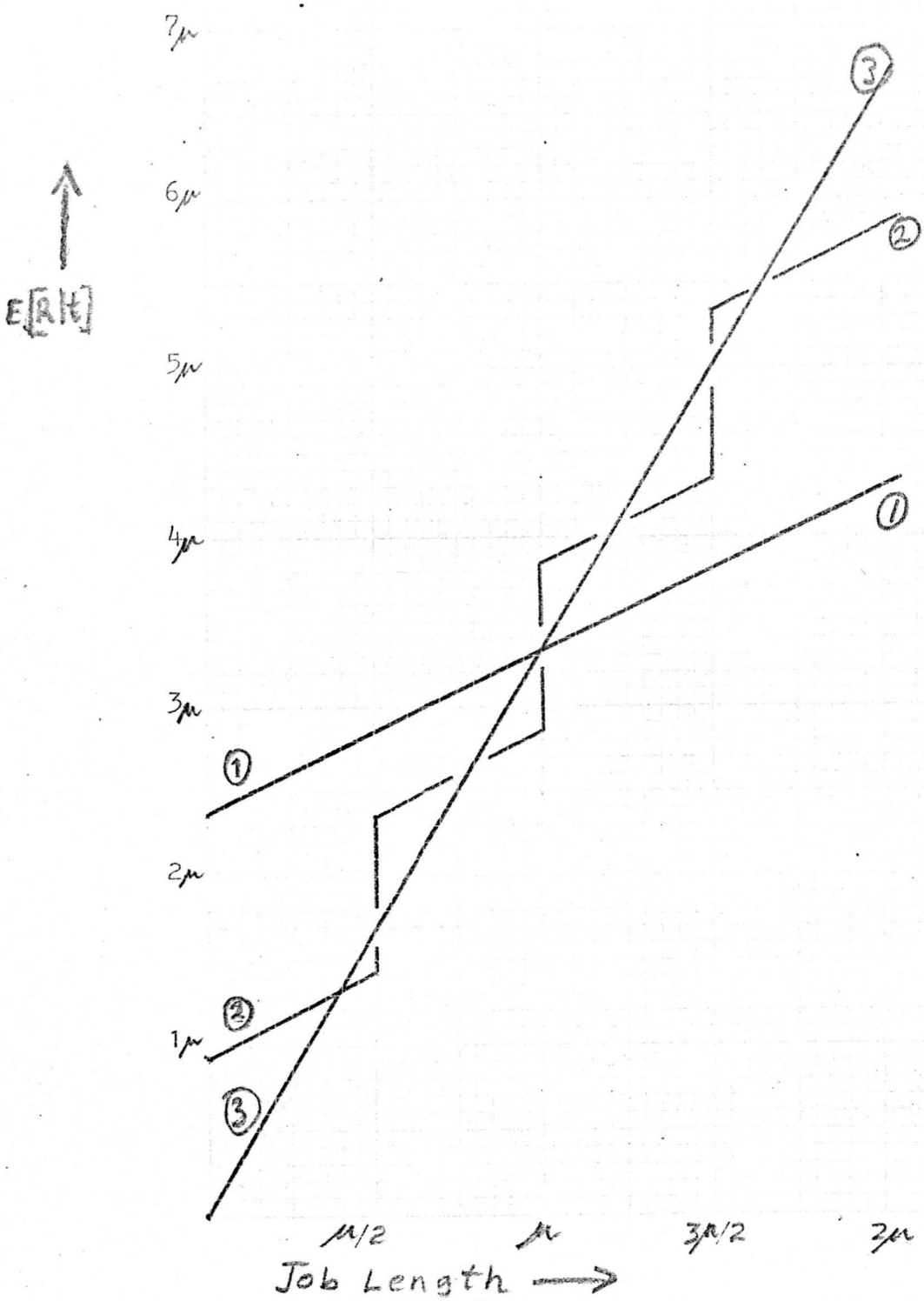
Matrix recurrence relations involving vectors of the M_{ij} are found and solved.

If the job length probability function does not reach zero

for finite job lengths, an approximate solution only can be found.

Expected Turnaround Time

as Function of Job Length.



① simple M/M/1 model

$\lambda = .7$ throughout ② time-sharing model; $q = \mu/2$

③ processor-sharing model

CHAPTER 7.

CONCLUSION.

This chapter attempts to evaluate the coverage achieved for each major topic, and the practical value of the theory which has evolved.

7.1 COVERAGE.

7.2 CRITIQUE.

CHAPTER 7 : CONCLUSION.

7.1 COVERAGE.

Saaty's " Queueing Theory " (76), published in 1961, contains 910 references; the bibliography of his " A Lament and a Bibliography " (77) consists of about 500 more; and the flow of papers is continuing still. However, many of these works are irrelevant to the present study, and some degree of coverage can be claimed for each chapter.

Chapter 2 achieves the poorest coverage. Numerous papers on multichannel systems have appeared, particularly in the early 1960's; few recent papers treat multichannel systems devoid of other complications which are outside the scope of the chapter. Three such papers which escaped inclusion are Arora's, on the Laplace transform of $P_n(t)$ for the system M/GI/2 (6), that of R.R.P Jackson and Henderson on an alternative method of finding the transforms of $P_n(t)$ for the system M/M/c (45), and Shapiro's on P_n for the system $M/E_2/c$ (81). There are probably many other works on A/B/c systems. However, the treatment of the steady state M/M/c system is complete, and the rest of the chapter indicates directions taken in the study of other systems. R.R.P Jackson's three " classical " approaches (43) are included: reduction to a markov process; the imbedded markov chain; and the probabilistic derivation of the Wiener-Hopf equation for waiting time.

The entirely separate approach of Beneš has not been included; his book deals with single-server queues (7). There may be, or at least should be, other fresh approaches,

such as the cybernetic approach called for by Ghosal (32).

Chapter 3 achieves a better coverage; for the number of relevant papers is much smaller. The study of poisson series systems, the reduction methods for use with regular service-time systems, and the MGF methods are treated. The work on cyclic queues is only touched upon.

Chapter 4 achieves an even better coverage, as the field is even more specialized. Few writers seem to have ventured into it since J.R. Jackson presented his generalized poisson model (40). There are, however, the quite different approaches mentioned at the start of chapter 5. Some of these approaches have nothing to do with queues; others do not seem to have been developed significantly.

Chapter 5 deals with a specialized model which has been studied extensively by a small group. The coverage for this model is good; but there may be other types of communication networks which have received study.

Chapter 6 deals briefly with a field which has received much attention in the last decade. The coverage of the theory and literature is not intended to be thorough, but indicative.

There are various papers on the topics of chapters 5 and 6 which are irrelevant in that they study the design of computer software, rather than the queueing models which are the real subject-matter of this study.

7.2 CRITIQUE.

As we proceed from single-station models, to series models, to network models, and then to applications of network models, the theory clearly becomes more and more limited to that of

poisson queues. In their survey papers, Bhat (8) and R.R.P. Jackson with Adelson (43) take almost contradictory stands on the merits of the similar development which occurs throughout the " classical " theory of queues, and its applications.

Bhat concludes his survey with these comments:
" It is our firm belief that the applied researcher should give up the naive idea that the steady state behaviour of a poisson arrival, exponential service queueing system is all that needs to be considered in real-life situations." He adds: "... what we need in the solution of design problems... are not simplifying assumptions that reduce the problem into familiar patterns.. (but)... more sophisticated procedures which can take care of the complexities of the situation."

Jackson and Adelson state:
" On perusing the literature, it soon became obvious that only the relatively simple problems.. are likely to yield usable solutions by analytic methods. The complexity of many of these solutions is such as to frighten off all but the most intrepid applied scientists... ."
The writers devote most of the second part of their survey (44) to systems with exponential or Erlang distributions.

A substantial piece of evidence inclines the present writer to side with Jackson and Adelson, as far as the applications of this study are concerned. The evidence is the simulation studies of Kleinrock (48), Scherr (80) and others. The poisson models can be modified in various ways which do not destroy the simplicity of their analysis, and the analysis of these

models produces results in close agreement with the results of simulations performed with much more complex and lifelike models.

REFERENCES.

The year given for each reference is the year of publication; many of the works were written a year or two before publication.

Abbreviations.

A.C.M. : the Association for Computing Machinery.

A.F.I.P.S. : American Federation of Information Processing Societies; and S.J.C.C. : Spring Joint Computer Conferences.

An. Math. Stat. : Annals of Mathematical Statistics.

I.E.E.E. : the Institute of Electrical and Electronic Engineers.

I.R.E. : the Institute of Radio Engineers.

Man. Sci. : Management Science.

N.R.L.Q. : Naval Research Logistics Quarterly.

O.R. : the journal Operations Research.

O.R. Quart. : Operations Research Quarterly.

Roy. Stat. Soc. : The Royal Statistical Society.

Soc. Ind. Ap. Math. : The Society of Industrial and Applied Mathematics.

List of References.

1. R.L. ACKOFF and M.W. SASIENI,
   Fundamentals of Operations Research.
   1968. Wiley International.
2. I. ADIRI and B. AVI-ITZHAK.
   A Time-Sharing Queue With a Finite Number of Customers.
   1969. Journal of A.C.M., 16, 315...323.
3. I. ADIRI and B. AVI-ITZHAK.
   A Time-Sharing Model with Many Queues.
   1969. O.R., 17, 1077.... 1089.
4. B. AVI-ITZHAK; A Sequence of Service Stations with
   Arbitrary Input and Regular Service Times.
   1965. Man. Sci., 11, 565... 571.
5. B. AVI-ITZHAK and M. YADIN.
   A Sequence of Two Servers with No Intermediate Queue.
   1965. Man. Sci., 11, 553... 564.
6. K.L. ARORA.
   Time-Dependent Solution of the Two-Server Queue Fed by
   General Arrival and Exponential Service Time Distributions.
   1962. O.R., 10, 327... 334.
7. V.E. BENEŠ.
   General Stochastic Processes in the Theory of Queues.
   1963. Addison-Wesley.
8. U.N. BHAT.
   Sixty Years of Queueing Theory.
   1969. Man. Sci., 15, B-280... B-294.

9. W.R. BLUNDEN.

On the Theory of Traffic Flow.

1962. Australian Road Research Board Proceedings, 1, 131.. 140.

10. P.J. BURKE.

The Output of a Queueing System.

1956. O.R., 4, 699... 704.

11. P.E. BOUDREAU and M. KAC.

Analysis of a Basic Queueing Problem Arising in Computer Systems.

1961. I.B.M. Journal of Research and Development, 5, 132... 140.

12. P.J. BURKE.

The Dependence of Delays in Tandem Queues.

1964. An. Math. Stat., 35, 874... 1875.

13. C.S. CARR, S.D. CROCKER and V.G. CERF.

Host-Host Communication in the ARPA Network.

1970. A.F.I.P.S., Proceedings of S.J.C.C., 589... 597.

14. W. CHANG.

Congestion Analysis of a Computer Core Storage System.

1967. N.R.L.Q., 14, 367... 378.

15. W. CHANG.

Queues with Feedback for Time-Sharing Computer System Analysis.

1968. O.R., 16, 613... 627.

16. W. CHANG.

Queueing Analysis of Real-Time Computer Processing.

1969. Man. Sci., 15, 658... 671.

17. W. CHANG.

Single-Server Queueing Processes in Computing Systems.

1970. I.B.M. Systems Journal, , 36... 71.

18. R.T. CHIEN.

Synthesis of a Communication Net.

1960. I.B.M. Journal 4, 311... 320.

19. E.G. COFFMAN.

Studying Multiprogramming Systems with the Queueing Theory.

1967. Datamation, 13, 47... 54.

20. E.G. COFFMAN.

Analysis of a Drum Input/Output Queue Under Scheduled Operation in a Paged Computer System.

1969. Journal of A.C.M., 16., 73... 90.

21. E.G. COFFMAN and L. KLEINROCK.

Computer Scheduling Methods and their Countermeasures.

1968. Proceedings of A.F.I.P.S., 32, 11... 21.

22. R.W. CONWAY and E.L. MAXWELL.

Network Dispatching by the Shortest Operation Discipline.

1962. O.R., 10, 51... 73.

23. D.R. COX and W.L. SMITH.

Queues.

1961. Chapman and Hall.

24. M.T. EL-BARDAI.

Queueing Analysis of Satellite Networks.

1970. O.R., 18, 654... 664.

25. P. ELIAS, A. FEINSTEIN, and C.E. SHANNON.

A Note on the Maximum Flow Through a Network.

1956. I.R.E. Transactions on Information Theory, 117..119.

26. R.V. EVANS.

—— (Discussion of Saaty's paper.)

1968. Chapter 4 of Symposium on Congestion Theory, 102..104.

27. R.V. EVANS.

Capacity of Queueing Networks.

1967. O.R., 15, 530... 536.

28. W. FELLER.

An Introduction to Probability Theory and its Applications.

Volume 1.

1950. Wiley International.

29. P.D. FINCH.

The Output Process of the Queueing System M/G/1.

1959. Journal of Roy.Stat. Soc; B, 21, 375... 379.

30. H. FRANK, I.T. FRISCH and W. CHOU.

Topological Considerations in the Design of the ARPA Computer Network.

1970. A.F.I.P.S., Proceedings of SJCC, 581... 587.

31. H.D. FRIEDMAN.

Reduction Methods for Tandem Queueing Systems.

1965. O.R., 13, 121... 130.

32. A. GHOSAL.

Cybernetic Queues.

1969. Man. Sci., 16, B-14... B-15.

33. R.E. GOMORY and T.C. HU.

An Application of Generalized Linear Programming to Network Flows.

1962. Journal of Soc. Ind. Ap. Math., 10, 260... 283.

34. W.J. GORDON andG.F. NEWELL.

Closed Queueing Systems with Exponential Servers.

1967. O.R., 15, 254... 265.

35. S.L. HAKIMI.

Simultaneous Flows Through a Communication Network.

1962. I.R.E. Transactions on Circuit Theory, CT-9(2)

169... 175.

36. F.E. HEART, R.E. KAHN, S.M. ORNSTEIN, W.R. CROWTHER and D.C. WALDSEN.

The Interface Message Processor for the ARPA Network.

1970. A.F.I.P.S.,Proceedings of S.J.C.C., 551... 567.

37. F.S. HILLIER and R.W. BOLING.

Finite Queues in Series with Exponential or Erlang Service Times- a Numerical Approach.

1967. O.R., 15, 286... 303.

38. G.C. HUNT.

Sequential Arrays of Waiting Lines.

1956. O.R., 4, 674... 683.

39. J.R. JACKSON.

Network of Waiting Lines.

1957. O.R., 5, 518... 521.

40. J.R. JACKSON.

Jobshop-Like Queueing Systems.

1963. Man. Sci., 10, 131... 142.

41. R.R.P. JACKSON.

Queueing Systems with Phase-Type Service.

1954. O R.Quart, 5, 109... 120.

42. R.R.P. JACKSON.

Queueing Processes with Phase-Type Service.

1956. Journal of Roy. Stat. Soc., B, 18, 129... 132.

43. R.R.P. JACKSON and R.M. ADELSON.

A Critical Survey of Queueing Theory, Part 1.

1962. O.R. Quart. 13, 13... 23.

44. (Part 2 of Reference 43.

Same volume, 293... 307.)

45. R.R.P. JACKSON and J.C. HENDERSON.

The Time-Dependent Solution to the Many-Server Poisson Queue .

1966. O.R., 14, 720... 722.

46. D.G. KENDALL.

Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain.

1953. An. Math. Stat., 24, 338... 354.

47. J. KIEFER and J. WOLFOWITZ.

On the Theory of Queues with Many Servers.

1955. Am. Math. Soc. Transactions. 78, 1... 18.

48. L. KLEINROCK.

Communication Nets; Stochastic Message Flow and Delay.

1964. McGraw-Hill.

49. L. KLEINROCK.

Sequential Processing Machines Analysed with a Queueing Theory Model.

1966. Journal of A.C.M., 13, 179... 193.

50. L. KLEINROCK.

Analysis of a Time-Shared Processor.

1964. N.R.L.Q., 11, 59... 72.

51. L. KLEINROCK.

Time-Shared Systems : A Theoretical Treatment.

1967. Journal of A.C.M., 14, 242... 261.

52. L. KLEINROCK.

Comparison of Solution Methods for Computer Network Models.

1969. I.E.E.E., Proceedings of Computing and Communications Conference; Rome, N.Y. 295... 303.

53. L. KLEINROCK.

Models for Computer Networks.

1969. I.E.E.E. International Conference on Communications; Boulder, Colorado. 219... 21-16.

54. L. KLEINROCK.

Analytic and Simulation Methods in Computer Network Design.

1970. A.F.I.P.S., Proceedings of S.J.C.C., 569... 579.

55. L. KLEINROCK.

——— (Private communication to Mr. J.C. Turner).

1970.

56. E. KOENIGSBERG.

On Jockeying in Queues.

1966. Man. Sci. , 12, 412... 436.

57. B. KRISHNAMOORTHI and R.C. WOOD.

Time-Shared Computer Operations with both Interarrival and Service Times Exponential.

1966. Journal of A.C.M., 13, 317... 338.

58. D.V. LINDLEY.

The Theory of Queues with a Single Server.

1952. Proceedings of Cambridge Philosophical Society, 48, 277... 289.

59. T. MAKINO.

On a Study of Output Distribution.

1966. Journal of Operations Research Society of Japan, 8, 109... 132.

60. G.E. MASTERSON and S. SHERMAN.

On Queues in Tandem.

1963. An. Math. Stat., 34, 300... 307.

61. W.L. MAXWELL.

On the Generality of the Equation $L = \lambda W$.

1970. O.R., 18, 172... 173.

62. J.O. MAYHUGH and R.E. McCORMICK.

Steady State Solution of the Queue $M/E_k/r$.

1968. Man. Sci., 14, 692... 712.

63. J.M. McKINNEY.

A Survey of Analytical Time-Sharing Models.

1969. Computing Surveys, 1, 105... 116.

64. A.P. MULLERY.and G.C. DRISCOLL.

A Processor Allocation Method for Time-Sharing.

1970. Communications of A.C.M., 13, 10... 14.

65. R.T. NELSON.

Waiting-Time Distributions for Applications to a Series of Service Centres.

1958. O.R., 6, 856... 862.

66. R.T. NELSON.

Queueing Network Experiments with Varying Arrival and Service Processes.

1966. N.R.L.Q., 13, 321... 346.

67. G.C. O'BRIEN.

The Solution of Some Queueing Problems.

1954. Journal of Soc. Ind. Ap. Math., 2, 133... 142.

68. R.L. PATTERSON.

Markov Processes Occuring in the Theory of Traffic Flow through an N-Stage Stochastic Service System.

1964. Journal of Industrial Engineering, 10, 188.. 193.

69. C. PEARCE.

On the Joint Equilibrium Queue Length Distribution in a Series Queue.

1967. Canadian Journal of Operations Research,-, 96..100.

70. M. POSNER and B. BERNHOLTZ.

Closed Finite Queueing Networks with Time Lags.

1968. O.R., 16, 962... 976.

71. M. POSNER and B. BERNHOLTZ.

Closed Finite Queueing Networks with Time Lags and Several Classes of Units.

1968. O.R. ., 16,977... 985.

72. N.U. PRABHU.

Transient Behaviour of a Tandem Queue.

1967. Man.Sci., 13, 631... 639.

73. E. REICH.

Waiting Times when Queues are in Tandem.

1957. An. Math. Stat., 28, 768... 773.

74. E. REICH.

Note on Queues in Tandem.

1963. An. Math. Stat., 34, 338... 341.

75. L.G. ROBERTS and B.D. WESSLER.

Computer Network Development to Achieve Resource Sharing.

1970. A.F.I.P.S., Proceedings of S.J.C.C., 543.. 549.

76 T.L. SAATY.

Elements of Queueing Theory with Applications.

1961. McGraw-Hill.

77. T.L. SAATY.

Seven More Years of Queues: A Lament and a Bibliography.

1966. N.R.L.Q., 13, 447... 476.

78. T.L. SAATY.

Stochastic Network Flows: Advances in Networks of Queues.

1964. Editors W.L. Smith and W.E.Wilkinson. Uni. North Carolina Press. Chapter 4 of Symposium on Congestion Theory. 86... 99.

79. M. SAKATA, S. NOGUCHI and J. OIZUMI.

An Analysis of the M/G/1 Queue Under Round-Robin Scheduling.

1971. O.R., 19, 371... 385.

80. A.L. SCHERR.

An Analysis of Time-Shared Computer Systems.

1967. M.I.T. Research Monograph No. 36.

81. S. SHAPIRO.

The M-Server Queue with Poisson Input and Gamma-Distributed Service of Order Two. 1966. O.R., 14, 685.. 695.

82. J.E. SHEMER.

Some Mathematical Considerations of Time-Sharing Scheduling Algorithms.

1967. Journal of A.C.M., 14, 262... 272.

83. W.L. SMITH.

On the Distribution of Queueing Times.

1953. Proceedings of Cambridge Philosophical Society, 49, 449... 461.

84. B. VAN DER VEEN.

Introduction to the Theory of Operational Research.

1967. Phillips-Springer.

I must record my thanks to the many people who have helped me during this undertaking . In particular I wish to thank Mr. J.C.Turner ; the librarians of the D.S.I.R. Applied Mathematics Laboratory and of Waikato and Victoria Universities ; and Miss M.A.Belsey.

M.D.Camden.