AD646233

ORC 67-1
JANUARY 1967

CLOSED NETWORKS OF QUEUES

by

Richard J. Swersey

D D C
FEB 6 1967
B

OPERATIONS RESEARCH CENTER

COLLEGE OF ENGINEERING

ARCHIVE COPY

UNIVERSITY OF CALIFORNIA - BERKELEY

CLOSED NETWORKS OF QUEUES

by

Richard J. Swersey
Operations Research Center
University of California, Berkeley

January 1967 ORC 67-1

This research was supported by the Office of Naval Research under Contract Nonr-222(83), the National Science Foundation under Grant GP-4593 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

CONTENTS

	Page
Chapter I - Introduction	
1. Queuing Networks	1
2. Analytic Framework	2
3. Optimization Problems	3
4. Definitions and Notation	4
Chapter II - Cyclic Queues	
1. Introduction	6
2. Steady-State Equations	7
3. Operating Characteristics	8
4. Cyclic Queues and Markov Chains	8
5. Customer-average Wait in Queue	13
6. Statistical Inference	18
7. Customer-dependent Service Rates	26
8. Conclusions	27
Chapter III - General Networks of Queues	
1. Introduction	28
2. Jobshop Systems	28
3. Closed Queuing Networks	29
4. Customer-average Wait in Queue	35
Chapter IV - Miscellaneous Extensions	
1. Approximations to Open Systems	37
2. Customer-dependent Service Rates	38
3. Comparison of a Single Customer and a Two Customer System	39
4. The Marginal Distribution of the Number of Customers at Service Center i	41
Chapter V - Optimization Problems	
1. Introduction	44
2. Cost-Performance Alternatives	44
3. Maximization of Production Rate	47
4. The Marginal Cost of Congestion	52
5. PCA Airlines - A Discrete Selection Problem	53
Acknowledgments	58
References	59

"CLOSED NETWORKS OF QUEUES"

ABSTRACT

RICHARD J. SWERSEY

A closed network of queues consists of a finite set of N customers, a finite set of M single-channel servers, and a set of arcs (i,j) which represent the allowed instantaneous movement from station i to station j. We also assume that all customers are identical in their stochastic behavior, that movement is governed by a set of given transition probabilities such that they form an irreducible Markov chain, that service times are governed by an exponential distribution, and that the imbedded Markov chain defined on the instants of service completions is irreducible.

Under these assumptions steady-state operating characteristics are derived through analyses of the time-average steady-state equations and of the underlying Markov chain. The general results are specialized to cyclic queues and to open networks of queues (jobshop-like queuing systems); the structure of the steady-state probabilities is the same as that of the cyclic queue. We also show that the results can be generalized to a multi-channel server problem and that the number of customers at a given service center has an IFR (increasing failure rate) distribution.

Optimal allocation of labor (servers) from a fixed pool is considered and it is shown that, in a cyclic queue, maximum output is achieved from an equal allocation of labor to each service center. A discrete optimization problem is then considered for a more general network that represents the operation of an airlines maintenance base.

Chapter I

INTRODUCTION

1. Queuing Networks

Consider the operation of a haulage system in a coal mine. Shuttle cars are loaded at various working faces and travel to a common unloading point at a belt conveyor. The shuttle car, thus, is served at two stations--a randomly selected working face where it waits if and only if another car in its loop is being loaded, and the unloading point where cars from several loops may queue up to be served. This is a special case of a queuing network--it is cyclic for a given car because the same operations are repeated continually, but has a more general network character because a randomly-chosen car unloaded at the conveyor may return to one of several faces.

A more complex structure is represented by the overhaul and maintenance base of an airline system. Here, an item is probably inspected for unseen difficulties, as well as undergoing routine maintenance, and hence may go through various movements in the system before it returns to service. We have, in effect, a closed loop system with various internal series-parallel movements, and congestion at every station within the system.

These two examples have the following things in common:

(1) The number of served units and service stations is fixed, i.e., the system is closed (where in the second example the airplane in service is part of the system).

(2) The possible internal movements from station to station provide a well-defined network.

(3) Variability of service times creates congestion in the form of random queues in front of each service station.

For these reasons we shall refer to the models developed in this thesis as closed networks of queues. We shall determine their operating characteristics under assumptions specified below, and we will contrast their behavior with other models which have been analyzed in the literature - cyclic queues, and open networks of queues (jobshop-like queuing systems)

2. Analytic Framework

We define a closed queuing network as consisting of

(1) a finite set of N customers (serviced units such as airplane, shuttle cars, machines)

(2) a finite set of M single channel servers (service stations such as machines, repairmen, etc.)

(3) a set of arcs (i,j) which represent the allowed (instantaneous) movement from station i to station j.

We will also make the following assumptions:

(1) All customers i = 1, 2,...N are identical units in terms of their stochastic behavior in movement over the network, and in selection of a service time at some station j = 1,...M.

(2) Movement over the network is governed by a set of given transition probabilities, p_{ij}.

p_{ij} = P [customer moves to station j | he has just completed service at station i]

We assume that $\Sigma p_{ij} = 1$, and the associated Markov chain is irreducible.

(3) The service time distribution $F(t;i,n)$ at station i is the same for all customers, possibly dependent on the total number of customers at this station, $n = 1,2...N$. In particular, we assume:

$$F(t;i,n) = 1 - e^{-\mu_i(n)t} \qquad t \geq 0$$

It is important to note that the model loses the identity of individual customers, and this is not suitable for special items which have unique paths through the system. For an analysis of a class of such problems see Swersey [9]. Apart from the lost identity of the customers, the most critical assumption is (3). However, eliminating the exponential property yields insurmountable difficulties because it is impossible to define appropriate imbedding points which lead to a meaningful analysis and to significant results. In principle, however, one can extend the results to the Erlang distributions by the method of stages.

3. Optimization Problems

In addition to steady-state operating characteristics, we are also interested in system economics. For example, one can associate a cost with the time a customer waits in queue to be served and another cost with the time a server waits for a customer. If one is given a specific queuing network, it is of interest to compute the trade-off between expected costs of customer idle time, and expected costs of server idle time. One can then decide whether to buy new machines to decrease service times, or whether to add servers or customers to the system. Or, if the queuing network represents a production cycle, one can compare the marginal value of extra output with the marginal costs of obtaining that output. And so on.

Thus, the objective of this dissertation is twofold--to develop closed form expressions for steady-state probabilities, expected waiting times, and other relevant operating characteristics of queuing networks, and to analyze some optimal operating policies on some special structures of these models. We will also compare these models. We will also compare these models with Koenigsberg's [6] cyclic queues and Jackson's [4, 5] open networks of waiting lines.

4. Definitions and Notation

i = index of a service center ($i = 1 \ldots M$)

μi = exponential service rate at i (independent of the number of customers).

n_i = total number of customers at i, equal to those waiting for service plus the one in service

$$\sum_{i=1}^{M} n_i = N$$

$(n_1, n_2, \ldots n_M)$ = state vector: n_1 customers at server 1, n_2 at server 2, etc.

$\mathcal{A}$ = set of all possible arrangements of $(n_1 \ldots n_M)$

$A = \text{card}\,\mathcal{A} = \frac{(N+M-1)!}{(M-1)!\,N!}$ (the dimension of $\mathcal{A}$)

$a, b, \ldots$ = elements of $\mathcal{A}$

$a = (n_1 \ldots n_M)$ a point coordinate of interest

$a(i,j)$ = the neighboring point of a, $(n_1, n_2 \ldots n_i+1, \ldots n_j-1, \ldots n_M)$

$P(a)$ = time average steady-state probability of being in state a.

$n_i(a)$ = number of customers at station i for state a.

$\mathcal{A}(\cdot)$ = set of all possible arrangements of $(n_1 \ldots n_M)$ such that property $(\cdot)$ holds.

$A(\cdot) = \text{card}\,\mathcal{A}(\cdot)$

$I[\cdot]$ = indicator function = $\begin{cases} 1 & \text{if property } (\cdot) \text{ holds} \\ 0 & \text{otherwise} \end{cases}$

q_i = steady-state average rate at which customers pass through station i.

$$q_i = \sum_{j=1}^{M} P_{j,i} q_j \qquad i = 1 \ldots M$$

where $P_{j,i}$ is defined in assumption (2), Chapter I.

$\gamma_i = \mu_i / q_i$

$X_i = \gamma_i / \gamma_1$, the relative mean service rate at i with respect to server 1.

Chapter II

CYCLIC QUEUES

1. Introduction

In 1958, Koenigsberg published the first [6] of two [7] articles on cyclic queues. In his model, M, servers are linked in series with the last one in a closed loop with the first one. (Figure 1)

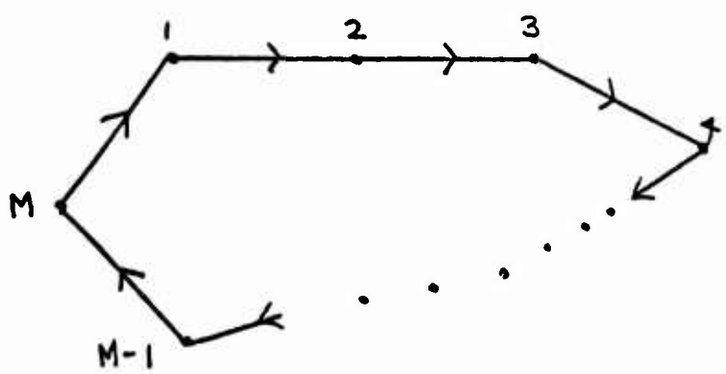


Figure 1

Cyclic Queue

In terms of the more general network model this is a special case where,

$\mu_i(n) = \mu_i$ for all $n=1,\ldots N (i=1\ldots M)$

$p_{i,i+1} = 1 \quad i=1,\ldots M (M+1$ is defined as $1)$

$p_{i,j} = 0 \quad j \neq i+1$

and defining $q_1 = 1 \geq q_i = 1 \quad i = 2\ldots M$

and hence $\gamma_i = \mu_i \quad i = 1\ldots M$

In this chapter we will discuss his results and cast them into a unified framework, correcting some minor errors. Our purpose is not just to illustrate the beginnings of research on queuing systems; in later chapters we will show <u>that even the most complicated network structure can be recast so that the results have the same form as Koenigsberg's original model</u>! It is this fact alone that makes analytical optimization of networks of queues tractable. (Chapter V)

2. Steady-State Equations

Because the service times are exponential, we can immediately write down the steady-state equations (after Koenigsberg [6]):

$$P[a]\sum_{i=1}^{M}\gamma_i I[n_i>0|a] = \sum_{i=1}^{M} P[a(i;i+1)]\gamma_i I[n_{i+1}>0|a(i;i+1)]$$

The solutions take the form,

$$P(a) = P(N,0,\ldots,0)\,\frac{\gamma_1^{N-n_1(a)}}{\prod_{i=2}\gamma_i^{n_i(a)}}$$

or

$$\boxed{P(a) = P(N,0,\ldots,0)\prod_{i=1}^{M} X_i^{n_i(a)}} \tag{2.1}$$

and

$$Z_M^N \equiv \frac{1}{P(N,0,\ldots,0)} = \sum_{a\in\mathcal{A}}\prod_{i=1}^{M} X_i^{n_i(a)}$$

$$\sum_{a\in\mathcal{A}}\prod_{i=1}^{M} X_i^{n_i(a)} = \sum_{i=1}^{M}\frac{X_i^{N+M-1}}{\prod_{\substack{j=1\\ j\neq i}}^{M}(X_i-X_j)} \tag{2.2}$$

Equation (2.2) has been specifically evaluated for M = 1,...5. See [6] .

3. Operating Characteristics

We will repeat here some operating characteristics of cyclic queues. For details of the development, see [6].

Let $D_i = P$ [server i is idle],

$$\text{Then } D_i = \frac{{}_iZ^N_{M-1}}{Z^N_M} \qquad (i = 1 \ldots M)$$

where ${}_iZ^N_M$ indicates that X_i is omitted from the summation in (2.2).

Let θ_i the average number of customers served per unit times at station i; then,

$$\theta_i = (1-D_i)\mu_i \tag{2.3}$$

and $\theta_i = \theta_j = \theta$ for all i and j.

Let $E[L(i)]$ = expected number of customers at server i. Then

$$E[L(i)] = \frac{X_i \, \partial Z^N_M / \partial X_i}{Z^N_M} \tag{2.4}$$

and $E[L_q(i)]$ = mean number of customers waiting at server i

$$= E[L(i)] - (1-D_i) \, . \tag{2.5}$$

4. Cyclic Queues and Markov Chains

In order to find the average waiting-time per customer at each server we will need to know the steady-state probabilities of the imbedded Markov chain. After our own development of these probabilities we can correct Koenigsberg's result for average waiting-time.

Because of assumptions (3) and (4) of Chapter I, a cyclic queue with exponentially distributed service times is a continuous time Markov process with a discrete state space. Since the customers are identical (assumption

(1), Chapter I) the state space of the Markov chain imbedded on the instants of service completions at any service center, consists of all arrangements, $\mathcal{A}$, and its dimensionality is A. Number all possible states a,b = 1,2...A and let,

$\pi(a,b)$ = Markov transition probabilities between states (a) and (b) on the instant of a service completion

$\pi(a)$ = stationary Markov chain probability of being in state a.

Theorem II-1: The Markov transition probabilities between states (arrangements) a and b (a,b = 1,2...A) are:

$$(2.6) \qquad \pi(a,b) = \begin{cases} \dfrac{\sum_{i=1}^{M} \mu_i \, I[n_i(b) = n_i(a) - 1]}{\sum_{i=1}^{M} \mu_i \, I[n_i(a) > 0]} & \text{if a to b is possible} \\ 0 & \text{otherwise} \end{cases}$$

Proof: First consider the special case, M = 3 and N = 2. The transition matrix is:

$$[\pi] = \begin{array}{c|cccccc} & (200) & (002) & (020) & (101) & (110) & (011) \\ \hline (200) & 0 & 0 & 0 & 0 & 1 & 0 \\ (002) & 0 & 0 & 0 & 1 & 0 & 0 \\ (020) & 0 & 0 & 0 & 0 & 0 & 1 \\ (101) & \frac{\mu_3}{\mu_1+\mu_3} & 0 & 0 & 0 & 0 & \frac{\mu_1}{\mu_1+\mu_3} \\ (110) & 0 & 0 & \frac{\mu_1}{\mu_1+\mu_2} & \frac{\mu_2}{\mu_1+\mu_2} & 0 & 0 \\ (011) & 0 & \frac{\mu_2}{\mu_2+\mu_3} & 0 & 0 & \frac{\mu_3}{\mu_2+\mu_3} & 0 \end{array}$$

Consider the transition (110)→(020) and call this event G.

P(G) = P [service completion at Center 1 before a completion at Center 2]

$$= \mu_1 \int e^{-(\mu_1+\mu_2)t} \, dt = \mu_1/\mu_1+\mu_2$$

since the only other competing transition is (110) → (101), which requires a service completion at Center 2. The general result (2.6) follows directly from these observations.

The formula (2.6) can also be derived from Billingsley's[2] "intensity functions," where we interpret a term in the numerator as the intensity of a jump into the state it represents, and the denominator as the intensity of jumps out of the current state.

Theorem II-2: The $\pi(a,b)$ of (2.6) are transition probabilities of a closed recurrent class of states and furthermore there exists a unique set of stationary probabilities of the imbedded Markov chain given by the solution to

$$\pi(b) = \sum_{a \in \mathcal{A}} \pi(a)\, \pi(a,b) \tag{2.7}$$

$$\sum_{a \in \mathcal{A}} \pi(a) = 1 \ .$$

Proof: It is obvious that

$$\pi(a,b) \geq 0 \text{ and}$$

$$\sum_{b \in \mathcal{A}} \pi(a,b) = 1 \text{ for all } a \in \mathcal{A}$$

Since the p_{ij} form an irreducible Markov chain, all service centers communicate. That is, a customer can move from any center i to any center j, with non-zero probability, in a finite number of steps ($\leq M$).

Now states a and b differ from each other only in the placement of a finite number ($< N$) of customers. By moving one customer at a time, any state, a, can be changed to any other state, b, in a finite number of steps ($< NM$) and this set of changes occurs with non-zero probability. Hence, some finite power of $[\pi(a,b)]$ is eventually non-zero, implying all states communicate; communication implies that all states of the chain belong to a single irreducible class. Furthermore, since at least one state is persistent (A is finite) the chain is irreducible ergodic.

Theorem II-3: The stationary probabilities of the associated Markov chain are given by:

$$\pi(b) = \frac{\prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[r_i > 0 | b]}{\sum_{b \in \mathcal{A}} \prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[n_i > 0 | b]} \tag{2.8}$$

and furthermore they are related to the time-average probabilities by,

$$\pi(b) = \frac{P(b) \sum_{i=1}^{M} \mu_i \, I[n_i>0|b]}{\sum_{b\epsilon\mathcal{B}} P(b) \sum_{i=1}^{M} \mu_i \, I[n_i>0|b]} \tag{2.9}$$

Proof: Formula (2.8) is obtained after much tedious manipulation of (2.7).

We know that the $\pi(b)$ are asymptotically equivalent to the ratio:

$$\frac{\text{Expected number of transitions into state } b}{\text{Expected number of transitions in the system}}$$

The number of transitions out of state b differs by, at most, one from the number of transitions into b for some time period T. The expected number of transitions out of state b in a long time T is,

$$P(b) \sum_{i=1}^{M} \mu_i \, I[n_i>0|b] \cdot T$$

where $P(b)$ is a time-average probability, and the total number of transitions in the system during time T is

$$\sum_{b\epsilon\mathcal{B}} P(b) \sum_{i=1}^{M} \mu_i \, I[n_i>0|b] \cdot T$$

Hence, as $T \to \infty$, $\pi(b)$ is given by (2.9).

Corollary (II-3): The time-average probabilities, $P(b)$, Theorem II-3 are given by formula (2.1).

Proof: From an examination of (2.8) and (2.9) we conclude that

$$P(b) = C \prod_{i=1}^{M} X_i^{n_i(b)}$$

where C is a constant.

Let

$$C = 1/Z_M^N$$

Then

$$P(b) = 1/Z_M^N \prod_{i=1}^{M} X_i^{n_i(b)}$$

which is (2.1).

From equation (2.8) we can derive all of the $\pi(b)$ and use them to solve for the P(b) in equations (2.9). No claim is made that computing P(b) from the $\pi(b)$ is easier than using equation (2.1); however, if we consider a system that is a variation on the cyclic queue (see[7]) the construction of the state space and solution for the $\pi(b)$ might prove to be an easier task than setting up the balance equations and solving for the P(b). Koenigsberg's approach, however, yields the transient solutions if they are of interest. Equation (2.9) establishes the link between the theory of Markov chains and the time-oriented queuing formulation for cyclic congestion systems.

Equation (2.8) can be changed to a less cumbersome form, and we can do some cancellation:

$$\pi(b) = \frac{\prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[n_i>0|b]}{\sum_{b\in\mathcal{R}} \prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[n_i>0|b]} \tag{2.10}$$

As in [6] define

$$B_M^N = \sum_{b\in\mathcal{R}} \prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[n_i>0|b] \tag{2.11}$$

After much algebra one obtains,

$$B_2^N = 2\mu_1 \sum_{i=0}^{N-1} X_2^{\,i}$$

$$B_3^N = 3\mu_1 \left[1 + \sum_{j=1}^{N-1} \sum_{k=2}^{3} X_k^{\,j} + \sum_{n_2=1}^{N-2} \sum_{n_3=1}^{N-n_2-1} X_2^{\,n_2} X_3^{\,n_3} \right]$$

and in general, $(M > 2)$ (2.11) becomes

$$(2.12) \quad B_M^N = M\mu_1 \left(1 + \sum_{j=1}^{N-1} \sum_{k=2}^{M} X_k^{\,j} + \sum_{n_2=1}^{N-2} \cdots \sum_{n_M=1}^{N+M-4-\sum n_i} \prod_{j=2}^{M} X_j^{\,n_j}\right)$$

Therefore, an alternate form for (2.10) is,

$$(2.13) \qquad \pi(b) = 1/B_M^N \left(\prod_{i=1}^{M} X_i^{\,n_i(b)} \sum_{i=1}^{M} \mu_i \; I[\bar{n}_i > 0 \,|\, b]\right.$$

We will use (2.8) directly in the next section in order to compute the average wait at each server, and we will use the Markov chain analysis in a later section on statistical inference.

5. Customer-average Wait in Queue

In order to find the average wait in queue at any station we must distinguish between the average number in a queue, at an arbitrary instant of time and the average number found by an arbitrary arrival.

Define the random variables,

$R(i)$ = total number of customers at server i just after some arrival,

$R_q(i)$ = number of customers in queue at server i just after some arrival,

$\mathcal{A}_i(n_i \geq 1)$ = set of all possible arrangements of $(n_1 \ldots, n_M)$ for which $n_i \geq 1$ for some service center i

Define:

$\pi_i(b)$ = P [system in state b after the next arrival at station i]

$\pi_i(b)$ = P [system in state b after the next transition | next transition is an arrival to station i]

<u>Theorem II-4</u>:

$$(2.14) \qquad \pi_i(b) = \begin{cases} \dfrac{M\mu_i \prod_{i=1}^{M} X_i^{n_i(b)}}{B_M^N} & \text{if } n_i(b) \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$$(2.15) \qquad \frac{1}{M} = \frac{\sum_{b \in \mathcal{A}_i(n_i \geq 1)} \pi(b(i-1,i))\, \mu_{i-1}}{\sum_{i=1}^{M} \mu_i \; I[n_i > 0 \mid b(i-1,1)]}$$

<u>Proof</u>:

From the definition of $\pi_i(b)$,

$$\pi_i(b) = \frac{\pi(b(i-1,i))\, \mu_{i-1} / \sum_{i=1}^{M} \mu_i \; I[n_i > 0 \mid b(i-1,i)]}{P[\text{next transition is an arrival to } i]}$$

We have already seen that in the steady state all service centers do the same amount of work; that is, they all have the same average transition rate θ . Therefore,

P [next transition is an arrival to i] = $\frac{1}{M}$,

which by definition, is given by (2.15). We note that $\pi_i(b)$ is a proper probability sequence over for all i .

<u>Corollary (II-4)</u>:

$$(2.16) \qquad E[R_i] = \frac{M\mu_i}{B_M^N} \sum_{b \in \mathcal{A}_i(n_i \geq 1)} \prod_{i=1}^{M} X_i^{n_i(b)} \qquad i=1,\ldots,M$$

Proof: Trivially shown by the definition of the expected value and the use of (2.14).

We can also express (2.16) in a form similar to (2.5) by defining:

Define:

Q_i = number of customers at i after the next transition in the system

Then,

$E(R_i)$ = 1 + mean number of customers at i (except arrivals at i), averaged over all transitions.

Recall that the probability a transition is an arrival at i is just $\frac{1}{M}$. Then the mean number at i (except arrivals) averaged over all transitions is just,

$E(Q_i)$ when the transition is not an arrival at i

$E(Q_i) - 1$ when an arrival at i occurs.

Thus, $E(R_i) = 1 + E(Q_i)\left(1-\frac{1}{M}\right) + \left(E(Q_i)-1\right)\frac{1}{M}$ $\qquad i = 1\ldots M$

(2.16a) $\qquad E(R_i) = E(Q_i) + 1 - \frac{1}{M}$

Noting, that

$$\partial B_M^N/\partial\mu_i = -\frac{1}{\mu_i}\sum_{b\in} n_i(b)\prod_{k=1}^{M} X_k^{n_k(b)}\sum_{i=1}^{M}\mu_i\, I[n_i>0|b] + \sum_{b\in\mathcal{A}_i(n_i\geq 1)}\prod_{k=1}^{M} X_k^{n_k(b)} \qquad (i=2\ldots M)$$

we find

$$E(Q_i) = -\frac{(\partial B_M^N/\partial\mu_i)\,\mu_i}{B_M^N} + \frac{1}{M}$$

Thus, an alternate form for (2.16) is:

(2.16b)
$$E(R_i) = \frac{-(\partial B_M^N/\partial \mu_i)\,\mu_i}{B_M^N} + 1 \qquad i=2,\ldots,M$$

By similar arguments,

(2.16b)
$$E(R_1) = \frac{-(\partial B_M^N/\partial \mu_i)\,\mu_i}{B_M^N} + N + 1$$

Define:

$W(i)$ = time spent at server i by a customer

$W_q(i)$ = queuing time at server i by a customer

Then,

$$E(W(i)) = E[E(W(i)|n_i \text{ at } i \text{ after last arrival})]$$
$$= E(\,n_i/\mu_i\,) = 1/\mu_i\; E(n_i)$$

But $n_i = R_i$, so:

(2.17)
$$E(W(i)) = E(R_i)/\mu_i = \frac{M}{B_M^N} \sum_{b\in\mathcal{R}_i(n_i\geq 1)} n_i \prod_{i=1}^{M} X_k^{n_k(b)}$$

and

(2.17a)
$$E(W_q(i)) = E(W(i)) - \frac{1}{\mu_i}$$

We would like to correct here a result of Koenigsberg's [6] (formulas (17), (18), p. 28). He claims that the mean wait in queue is just

$$E(L_q(i))/\mu_i$$

where $E(L_q(i))$ is the time average number in the queue; but this result would be correct only if the arrival stream at each service center was Poisson.

For the special case where all the service centers have the same rate, 1 say,

$$E(W_q(i)) = \frac{N-1}{M}$$

and Koenigsberg gives $$E(W_q(i)) = \frac{N(N-1)}{M(N+M-1)}$$

A comparison of this discrepancy for various values of N is given below: (M=3)

N	$E(W_q(i))$	$E(W_q$ (Koenigsberg))
3	.667	.4
4	1	.667
10	3	2.5
100	33	32.3
1000	333	332

Koenigsberg's formula understates the mean wait in queue and its use might cause difficulties in analyzing a real world situation. However, we will show later that as the number of customers in the system gets very large, that (2.12a) approaches Koenigsberg's result. His other major results, formulas (2.1) - (2.6), here, are stated correctly in [6]. One other operating characteristic will be useful in Chapter IV.

Define:

T = time for an arbitrary customer to complete one cycle of the service systems.

Then, $$T = \sum_{i=1}^{M} W(i)$$

$$E(T) = \sum_{i=1}^{M} E(W(i)) \tag{2.18}$$

6. Statistical Inference

The previous sections all dealt with a system for which the service rates were known; however, in many real-world situations, the service rates are unknown. One method of estimating these rates would be to observe the system and note the transitions that actually occurred and the length of time between transitions. From methods developed by Billingsley [2] we could then estimate the service rates and test hypotheses concerning these estimates.

The method used here (see Swersey [9]) parallels the development by Wolff [10] for birth and death processes.

If the system is observed continuously for a fixed time T, a sequence of S (where S is a random variable) transitions will be observed. We can then estimate the service rates by the maximum likelihood method. Denote the likelihood function for the observed sequence by $L_S(\underline{\theta})$ where $\underline{\theta} = (\mu_1 \ldots \mu_M)$.

Let,

$$U_{Si} = \partial/\partial\theta \ \ln L_S(\underline{\theta})$$

$$\underline{U_S} = (U_{S1}, U_{S2}, \ldots U_{SM})$$

We are interested in the asymptotic properties of the function $\underline{U_S}$ and of the maximum likelihood estimators, $\underline{\theta}$, obtained by setting. $\underline{U_S} = 0$.

As $T \to \infty$, $S \to \infty$ a.s.

As $T \to \infty$

$$\frac{U_S}{\sqrt{S}} \xrightarrow{\mathcal{L}} N(0, \sigma(\underline{\theta})) \tag{2.19}$$

where $\sigma(\underline{\theta})$ is the variance-covariance matrix of $\underline{\theta}$. Also,

$$\boxed{\frac{(\hat{\underline{\theta}} - \theta)}{\sqrt{S}} \xrightarrow{\mathcal{L}} N(0, \sigma^{-1}(\theta))} \tag{2.20}$$

The proofs for (2.19) and (2.20) are given by Billingsley [2].

Moreover, if we assume that

$$H_o: \ \underline{\theta} = \underline{\theta}^o$$

and H_0 is true, then

$$(2.21) \qquad 2[\underset{\underline{\theta}}{\text{Max}} \ln L_s(\underline{\theta}) - \ln L_s(\underline{\theta}^o)] \xrightarrow{\mathcal{L}} \chi^2_k$$

A statistical inference analysis of cyclic queues follows; note, however, that we are making statements only about asymptotic properties of estimators.

Let $\varphi(n,a,b)$ = likelihood due to the nth transition, where the system moved from state a to state b.

$\tau(n,a,b)$ = time spent in state a on transition $n = (t_n - t_{n-1})$, $n = 1 \ldots S$

Assuming exponential service times, it is easily shown that the density function of $\tau(n,a,b)$ is given by

$$(2.22) \qquad f(\tau(n,a,b) = \sum_{i=1}^{M} \mu_i \ I_n[n_i(a)>0]$$

$$\exp\left[-\tau(n,a,b) \sum_{i=1}^{M} \mu_i \ I_n[n_i(a)>0] \right]$$

where the indicator function now depends on n. The transition probability for the nth transition is obtained from (2.4).

$$(2.23) \qquad \pi(n,a,b) = \frac{\sum_{i=1}^{M} \mu_i \ I_n[n_i(b) = n_i(a) - 1]}{\sum_{i=1}^{M} \mu_i \ I_n[n_i(a)>0]}$$

Then $\varphi(n,a,b)$ is just the product of (2.22) and (2.23).

$$\varphi(n,a,b) = \sum_{i=1}^{M} \mu_i \, I_n[n_i(b) = n_i(a)-1]$$

$$\exp\left[-\tau(n,a,b) \sum_{i=1}^{M} \mu_i \, I_n[n_i(a)>0] \right]$$

We assumed S transitions occurred in time T. The <u>probability</u> that the $(S+1)^{st}$ transition occurred after T is $(t_S < T)$

$$P[\tau_{S+1}>t-t_S] = \exp[-(t-t_S) \sum_{i=1}^{M} \mu_i \, I_S[n_i(a)>0]$$

Assume, also, that a stationary distribution in time exists. Then the observations start when the system is in state P(0), say. The likelihood function is asymptotically equivalent to

$$L_S(\underline{\theta}) = P(0)\exp[-(t-t_S) \sum_{i=1}^{M} \mu_i \, I_S[n_i(a)>0] \prod_{n=1}^{S} \varphi(n,a,b)$$

or more conveniently,

$$\ln L_S(\underline{\theta}) = \ln P(0)-(t-t_S) \sum_{i=1}^{M} \mu_i \, I_S[n_i(a)>0] + \sum_{n=1}^{S} \ln \varphi(n,a,b)$$

In large sample theory, the starting conditions can be neglected.

$$\ln L_S(\underline{\theta}) = -(t-t_S) \sum_{i=1}^{M} \mu_i \; I_S[n_i(a)>0] + \sum_{n=1}^{S} \ln \varphi(n,a,b)$$

and ignoring end effects, we get

$$\ln L_S(\underline{\theta}) = \sum_{n=1}^{S} \ln \sum_{i=1}^{M} \mu_i \; I_n[n_i(b) = n_i(a) - 1]$$

$$- \sum_{n=1}^{S} \tau(n,a,b) \sum_{i=1}^{M} \mu_i \; I_n[n_i(a)>0]$$

In order to estimate the M parameters, μ_i, it will be convenient to redefine the summation indices.

Let d_i = the number of times during T that the completion was at stage i.

$$= \sum_{n=1}^{S} I_n[n_i(b) = n_i(a) - 1]$$

Then,

$$\sum_{n=1}^{S} \ln \sum_{i=1}^{M} \mu_i \; I_n[n_i(b) = n_i(a) - 1] = \sum_{i=1}^{M} d_i \ln \mu_i \; .$$

Let γ_i = the amount of time during T that $n_i(a) > 0$

$$= \sum_{n=1}^{S} \tau(n,a,b) \; I_n[n_i(a)>0]$$

Then,

$$\sum_{n=1}^{S} \tau(n,a,b) \sum_{i=1}^{M} \mu_i \; I[n_i>0|a] = \sum_{i=1}^{M} \gamma_i \; \mu_i$$

and

$$\ln L_S(\underline{\mu}) = \sum_{i=1}^{M} d_i \ln \mu_i - \sum_{i=1}^{M} \gamma_i \; \mu_i$$

Let

$$\partial \ln L_S(\underline{\mu})/\partial\mu_i = 0 \quad , \qquad \text{then}$$

$$d_i/\mu_i - \gamma_i = 0$$

and the maximum likelihood estimator of μ_i is,

$$\hat{\mu}_i = d_i/\gamma_i \qquad i=1,\ldots.M$$

In order to obtain the variance-covariance matrix, we use a procedure of conditioning on a particular transition, then on the state, and finally removing the conditioning.

Let

$$G(\mu_i) = \partial/\partial\mu_i \; \ln \varphi(n,a,b)$$

$$\ln \varphi(n,a,b) = \ln \sum_{i=1}^{M} \mu_i \; I_n[n_i(b) = n_i(a)-1]$$

$$- \tau(n,a,b) \sum_{i=1}^{M} \mu_i \; I_n[n_i>0|a]$$

Consider a particular station i, say,

$$G(\mu_i) = \begin{cases} -1/\mu_i - \tau(n,a,b) & \text{if } n_i|b = n_i|a - 1 \\ -\tau(n,a,b) & \text{if } n_i|a>0 \\ 0 & \text{otherwise} \end{cases}$$

$$G(\mu_i\mu_j) = \begin{cases} \partial^2/\partial\mu_i\partial\mu_j \ln \varphi(n,a,b) = -1/\mu_i^2 & \text{if } i=j \text{ and } n_i|b = n_i|a - 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, under suitable regularity conditions on $\varphi(n,a,b)$ (See Billingsley [2] for conditions on interchanging order of integration and differentiation)

$$V(\, G(\mu_i)|n,a,n_i>0|a\,) = -E(\, G(\mu_i\mu_i)|n,a,n_i\geq 0|a\,)$$

$$= \frac{1}{\mu_i^2} \frac{\mu_i}{\sum_{i=1}^{M} \mu_i\, I_n[n_i>0|a]}$$

$$= 1/\mu_i \sum_{i=1}^{M} \mu_i\, I_n[n_i>0|a]$$

Also,

$$V(\, G(\mu_i)\,) = E(\, V(G(\mu_i)|a\,)\,) = \sum_{a=1}^{A} V(G(\mu_i)|a)\, \pi(a)$$

$$= \sum_{a\in\mathcal{A}_i(n_i\geq 1)} \frac{\pi(a)}{\mu_i \sum_{i=1}^{M} \mu_i\, I[n_i>0|a]}$$

Substituting (2.9) for $\pi(a)$

$$(2.24) \qquad V(G(\mu_i)) = \frac{\sum_{a \in \Omega_i(n_i \geq 1)} P(a)}{\mu_i \sum_{a=1}^{A} P(a) \sum_{i=1}^{M} \mu_i \, I[n_i > 0 | a]}$$

We have already defined $(1-D_i)$ as the fraction of time the ith stage is working; the numerator of (2.24) is precisely $(1-D_i)$. If we expand terms in the denominator we have,

$$V(\, G(\mu_i)\,) = \frac{1 - D_i}{\mu_i M R}$$

where $R = (1-D_i)\mu_i$ is the output of the system and is the same for each stage i. Then,

$$V(\, G(\mu_i)\,) = 1/M\mu_i^{\,2} \qquad \doteq \quad \sigma(\underline{\theta})$$

is a diagonal matrix.

We are interested in the inverse of $\sigma(\underline{\theta})$.

$$(2.24a) \qquad V_{asymp}(\sqrt{S}(\hat{\mu}_i - \mu_i)) = M\mu_i^{\,2}$$

Note also that the asymptotic variance has a slightly different form if we normalize on T instead of on S. The change is equivalent to multiplying (2.9) by the limiting form of T/S which is clearly $1/M\mu_i(1-D_i)$. The variance becomes

$$(2.24b) \qquad V_{asymp}(\sqrt{T}(\hat{\mu}_i - \mu_i)) = \mu_i/(1-D_i)$$

Now the dependence on the time average probabilities appears formally. In order to test hypotheses on the μ_i , we make use of equation (2.21) Noting that the maximum over $\underline{\mu}$ of the likelihood function is obtained when we substitute the maximum likelihood estimators, the test that, $\mu^o=(\mu_1^o,\ldots\mu_M^o)$ is given by

$$2[\sum_{i=1}^{M} d_i \ln (d_i/\gamma_i\mu_i^o) + \sum_{i=1}^{M} \gamma_i\mu_i - S] \xrightarrow{\mathcal{L}} \chi_M^2$$

We can define an appropriate constant for any level of significance and check that the test quantity does not exceed the constant.

In particular, we might want to test the nested hypothesis that all of the μ_i are equal. The likelihood equation reduces to,

$$\ln L_S(\underline{\mu}) = S \ln \mu - \mu \sum_{i=1}^{M} \gamma_i$$

and $$\hat{\mu} = S/\sum_{i=1}^{M} \gamma_i$$

The appropriate test quantity becomes,

$$2[\sum_{i=1}^{M} d_i \ln (d_i/\gamma_i) - S \ln(S/\sum_{i=1}^{M} \gamma_i)] \xrightarrow{\mathcal{L}} \chi_{M-1}^2$$

7. Customer-Dependent Service Rates

The probability structure of the cyclic queue model can easily be extended to the more general model in which the service rates are occupancy-dependent; i.e.:

$$F(t;i,n) = 1 - e^{-\mu_i(n)t}$$

For example, in certain production systems it is possible that the service rate varies as the number of customers gets larger or smaller at each station: the steady-state probabilities turn out to have a similar form to (2.1).

Theorem II-5: For a cyclic queue with customer-dependent service rates, the steady-state probabilities are given by,

$$(2.25) \qquad P(a) = P(N,0,\ldots,0) \frac{\prod_{k=1}^{N-n_1(a)} \mu_1(N-k+1)}{\prod_{i=2}^{M} \prod_{k=1}^{n_i(a)} \mu_i(k)}$$

where a product is interpreted as unity if the upper limit is zero.

Proof: The steady state equations are given by,

$$P(a) \sum_{i=1}^{M} \mu_i(n_i(a))\, I[n_i>0|a] = \sum_{i=1}^{M} \mu_{i-1}(n_{i-1}(a)+1)\, P(a(i-1,i))\, I[n_i>0|a(i-1,i)]$$

From (2.25),

$$P(a(i-1,i)) = P(N,0,\ldots,0) \frac{\prod_{k=1}^{N-n_1(a)} \mu_1(N-k+1)}{\prod_{\substack{j=2 \\ j\neq i,i-1}}^{M} \prod_{k=1}^{n_j(a)} \mu_j(k) \prod_{k=1}^{n_{i-1}(a)+1} \mu_{i-1}(k) \prod_{k=1}^{n_i(a)-1} \mu_i(k)}$$

$$P(a(i-1,i)) = P(N,0,\ldots,0) \frac{\prod_{k=1}^{N-n_1(a)} \mu_1(N-k+1)\, \mu_i(n_i(a))}{\prod_{j=2}^{M} \prod_{k=1}^{n_j(a)} \mu_j(k)\, \mu_{i-1}(n_{i-1}(a)+1)}$$

Therefore the right hand side reduces to

$$\sum_{i=1}^{M} \mu_i(n_i(a))I[n_i>o|a]\,P(a)$$

by assuming (2.25) is true.

Unfortunately, no further reduction of (2.25) is possible for further analysis of operating parameters. Particular problems of interest, however, shall be handled directly.

For example, the multiple server problem can be analyzed directly using (2.25).

Let $$\mu_i(k) = \begin{cases} k\mu_i & \text{if } 1 \leq k \leq S_i \\ S_i\mu_i & \text{if } k > S_i \end{cases}$$

where S_i is the number of parallel channels at service center i. As a special case let $S_j = S$, $S_i = 1 (i \neq j)$. The steady-state probabilities reduce to,

$$(2.26) \qquad P(a) = \begin{cases} P(N,0,\ldots 0)\dfrac{\prod_{k=1}^{M} X_k^{n_k(a)}}{n_j!} & \text{if } n_j(a) < S \\[2ex] P(N,0,\ldots 0)\dfrac{\prod_{k=1}^{M} X_k^{n_k(a)}}{S!\,S^{n_j(a)-S}} & \text{if } n_j(a) \geq S \end{cases}$$

8. Conclusions

We have stated Koenigsberg's key results for cyclic queues, and have indicated some extensions through analysis of the related Markov process. With these fundamentals as the base, we will next investigate more general queuing networks, and show that the structure of the state probabilities is similar to that of the simple cycle. Then, the results derived here will be used to analyze these more general structures.

Chapter III

GENERAL NETWORKS OF QUEUES

1. Introduction

We now investigate general closed networks of queues. The relationship between these networks and Koenigsberg's cyclic queue model will be further developed and we will compare our system to the so-called jobshop system of Jackson [4] and [5].

2. Jobshop Systems

Jackson [4] and [5] has described what he calls a jobshop-like queuing system. Such a system is defined as obeying assumptions (1) - (3) of Chapter 1 (where $N = \infty$) plus the following:

(4) customers from outside the system arrive at server i in a Poisson process with parameter β_i

(5) once served in department i , a customer goes to department j with probability p_{ij} ; he leaves the system with probability $1 - \sum_{i=1}^{M} p_{ij}$

Obviously, these assumptions mean that the probabilities of movements of a given customer are independent of his previous history. As Wolff [11] has pointed out, this is precisely not a jobshop model, as it allows a customer to cycle in the system forever without ever leaving it. In a typical jobshop, if an item returns to a machine for re-working, there is a strong likelihood that either the service time distribution is different than on its first pass, or the movement probabilities are changed or both.

Instead we will refer to Jackson's system as an open network of queues. His results give insight into the closed or cyclic queuing networks, and they are also useful in the analysis of appropriate systems.

Jackson's main result is as follows:

Define:

$$P_{n_i} = P_0^{\,i} \, (\Gamma_i/\mu_i)^{n_i} \qquad n_i = 1,\ldots. \tag{3.0}$$

where $P_0^{\,i} = P[\text{ Server i is idle }]$ and

$$\Gamma_i = \beta_i + \sum_{k=1}^{M} p_{ki} \, \Gamma_k \qquad = \text{average arrival}$$

rate of customers to server i from any source. Then, (iff $\Gamma_i < \mu_i$)

$$P(n_1,\ldots.n_M) = P_{n_1} P_{n_2} \ldots\ldots P_{n_M} \tag{3.1}$$

Equation (3.1) is a powerful result. The steady-state probabilities can be treated as though each service center were independent. The result is similar in form to that for queues in tandem.

Based on equation (3.1), we might expect that for the cyclic queuing network the steady-state probabilities take a form similar to the series cyclic queuing formula (2.1). Jackson [4] has extended his model to include the cases where the service rates depend upon the number present; however, he has not eliminated the assumptions that belie the jobshop description. We will now define such a system in detail and prove that our intuitive feeling is correct.

3. Closed Queuing Networks

A closed queuing network was defined in Chapter I. We will repeat it here for clarity.

We define a closed queuing network as consisting of

(1) a finite set of N <u>customers</u> (serviced units such as airplane, shuttle cars, machines)

(2) a finite set of M single channel <u>servers</u> (service stations such as machines, repairmen, etc.)

(3) a set of arcs (i,j) which represent the allowed (instantaneous) movement from station i to station j.

We will also make the following assumptions:

(1) All customers i = 1,2,...N are <u>identical</u> units in terms of their stochastic behavior in movement over the network, and in selection of a service time at some station j = 1,...M.

(2) Movement over the network is governed by a set of given transition probabilities, p_{ij}.

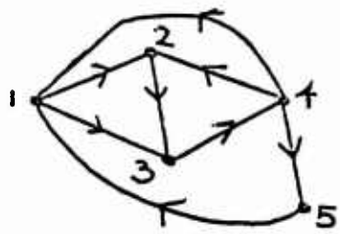
$$p_{ij} = P\ [\text{customer moves to station } j \mid \text{he has just completed service at station } i\ ]$$

We assume that $\sum_j p_{ij} = 1$, and the associated Markov chain is irreducible.

(3) The service time distribution F(t;i,n) at station i is the same for all customers, possibly dependent on the total number of customers at this station, n = 1,2...N. In particular, we assume:

$$F(t;i,n) = 1 - e^{-\mu_i(n)t}$$

For clarity consider a typical network as illustrated in Figure 2a.



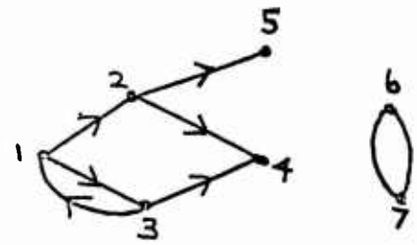


Figure 2a Figure 2b

Closed Queing Network and Improper Network

Note that we do not preclude subloops anywhere in the network. If we consider station 1 as being in a main loop, note that stations 2, 3, and 4, and 4, 5 form subloops. This implies that any customer could cycle indefinitely within the network without ever returning to station 1; however, from assumption (2) this event has probability zero. Figure 2b represents an improper network; stations 4 and 5 and 6, 7 do not communicate with the rest of the network, thereby voiding assumption (2).

We defined q_i in Chapter I; they may be found from

$$q_1 = 1$$

$$(3.2) \qquad q_i = \sum_{j=1}^{M} p_{ji}\, q_j \qquad i=1,\ldots,M$$

Note that if they were normalized by using (3.2) and

$$\sum_{i=1}^{M} q_i = 1$$

they would represent the steady-state probabilities of finding a given customer (N = 1) at station i, of the Markov chain.

Note also that the absence of any subloops other than those containing station 1 makes the solution of equations (3.2) almost trivial; this is

caused by the near-singularity of the set of equations.

Using the same notation as in Chapter II, the time average steady-state equations are,

$$(3.3) \qquad P(b) \sum_{i=1}^{M} \mu_i \; I[n_i>0|b] = \sum_{j=1}^{M} \sum_{i=1}^{M} \mu_j \; p_{ji} \; P(b(j,i)) \; I[n_i>0|b(j,i)]$$

We note that these equations cover all of the possible transitions in the imbedded irreducible Markov chain for if a particular center j does not connect directly to center i, $p_{j,i} = 0$, and that term is eliminated from the equations.

<u>Theorem III-1</u>: The solution to system (3.3) is given by,

$$(3.4) \qquad \boxed{P(b) = P(N,0,\ldots 0) \prod_{i=1}^{M} X_i^{\,n_i(b)}}$$

<u>Proof</u>: From all of the previous assumptions about the process it is enough to show that (3.4) satisfies (3.3).

Define:
$$Y(b) = \prod_{i=2}^{M} q_i^{\,n_i(b)} / q_1^{\,n_1(b)}$$

where b represents the usual point coordinates of . From this definition, it is clear that (3.4) is equivalent to,

$$P(b) = P(N,0,\ldots 0) \prod_{i=1}^{M} (\mu_1/\mu_i)^{n_i(b)} \; Y(b)$$

Also, from physical reasoning,

$$Y(b(j,i)) = Y(b) \; q_j/q_i \qquad (j \neq i)$$

If (3.4) is correct,

$$\mu_j \; p_{ji} \; P(b(j,i)) = P(N,0,\ldots 0) \left[\prod_{i=1}^{M} (\mu_1/\mu_i)^{n_i(b)} \right] \mu_i \; p_{ji} \; Y(b(j,i))$$

For $j = 1$,

$$P(b(1,i)) = P(b)\ \gamma_i/\gamma_1$$

and $$\mu_1\ p_{1i}\ P(b)\ \gamma_i/\gamma_1 = P(b)\ q_1\ p_{1i}\ \gamma_i$$

Therefore, equations (3.3) reduce to,

$$P(b) \sum_{i=1}^{M} \mu_i\ I[n_i>0|b] = P(b)[\sum_{i=1}^{M} (p_{1i}\ \gamma_i\ q_1 + \sum_{j=2}^{M} \frac{\mu_i p_{ji} q_j}{q_1})I[n_i>0|b]\ ]$$

but $$\mu_i/q_i \sum_{j=2}^{M} p_{ji}\ q_j = \mu_i/q_i(q_i - p_{1i}\ q_1)$$
$$= \mu_i - \gamma_i\ p_{1i}\ q_1$$

and the right hand side has been reduced to

$$P(b) \sum_{i=1}^{M} \mu_i\ I[n_i>0|b]$$

so relations (3.3) and (3.4) are equivalent.

Theorem III-1 is a powerful tool because the forms of the time-average results of Chapter II can be used here with suitable modification.

Let:

θ_i = average number of customers served at center i per unit time.

Then obviously,

(3.5) $$\theta_i = (1 - D_i)\mu_i \quad \text{where } (1 - D_i) = P[\text{server i is busy}]$$

Note that

$$\theta_i \neq \theta_j \qquad j=1,\ldots M$$

i.e., the number of customers served per unit time is not identical at each state as in Chapter II. To show this, (3.5) can be expanded so that,

$$\theta_i = \gamma_1\ q_i\ Z^N_{M-1} \qquad (i = 1\ldots M)$$

In Chapter II, $q_i = q_j$ $(i,j = 1\ldots M)$ and hence $\theta_i = \theta_j$.

But in our general network, $q_i \neq q_j$ and

$$\theta i = \theta j \frac{qi}{qj},$$

or the normalized service rate at station i equals the normalized arrival rate as we would expect.

We can relate all of the output rates to θ_1, say,

$$\theta_i = \theta_1 \, q_i/q_1 \quad , \tag{3.5a}$$

and the mean number of customers served in the system per unit time is,

$$K = \sum_{i=1}^{M} \theta_i = \theta_1/q_1 \sum_{i=1}^{M} q_i \, .$$

Then,

$$P\,[\text{a given transition is out of state } i] = \theta_i/K = q_i/\sum_{i=1}^{M} q_i \tag{3.6}$$

We see that in the case of a cyclic queue where $qi = qj$ $(i,j = 1...M)$, (3.6) reduces to,

$$P\,[\text{transition is out of state } i\,] = \frac{1}{M}$$

Before we investigate the mean wait, it will be necessary to find the steady-state probabilities, $\pi(b)$, of the associated Markov chain, and show that they retain the properties of the cyclic queue.

<u>Theorem III-2</u>: The Markov chain steady-state probabilities are given by,

$$\pi(b) = 1/B_M^N \prod_{i=1}^{M} X_i^{n_i(b)} \sum_{i=1}^{M} \mu_i \, I[n_i>0|b] \tag{3.7}$$

We recognize (3.7) as the same form as (2.13), where the general network structure is incorporated into the definition of X_i. i.e.,

$$X_i = \gamma_1/\gamma_i = (\mu_1/q_1)/(\mu_i/q_i)$$

Our assumptions here are the same as those of <u>Theorem II-1</u> and hence the are the probabilities of a single irreducible Markov chain. Therefore it is enough to show that (3.7) solves

(3.8) $$\pi(b) = \sum_{a \in \mathcal{A}} \pi(b,a)\ \pi(a) \qquad (b \in \mathcal{A})$$

Equations (3.8) can be written as,

$$\pi(b) = \sum_{i=1}^{M} \sum_{j=1}^{M} \pi(b(i,j))\ \mu_i p_{ij}\ I[n_j>0|b] \Big/ \sum_{i=1}^{M} \mu_i\ I[n_i>0|b]$$

If (3.7) is true, then after some cancellation

$$\pi(b(i,j)) = (1/B_M^N)\ X_j \prod_{k=1}^{M} X_k^{n_k(b)}\ X_i \sum_{i=1}^{M} \mu_i\ I[n_i>0|b]$$

Then,

$$\pi(b) = (1/B_M^N) \sum_{i=1}^{M} \sum_{j=1}^{M} 1/q_j \prod_{k=1}^{M} X_k^{n_k(b)}\ \mu_j\ q_i\ p_{ij}\ I[n_j>0|b]$$

Now interchange the order of summation and,

$$\pi(b) = (1/B_M^N) \prod_{k=1}^{M} X_k^{n_k(b)} \sum_{j=1}^{M} (1/q_j)\mu_j\ I[n_j>0|b] \sum_{i=1}^{M} p_{ij}\ q_i$$

but $\sum_i p_{ij}\ q_i = q_j$ by definition, and the theorem is proved.

Thus, the Markov chain probabilities also retain the same form as in the series cyclic queue. We should note, however, that B_M^N cannot be determined by using formula (2.12).

4. <u>Customer-average Wait in Queue</u>

Analogously to Section 5 of Chapter II, define:

$\pi_j(b)$ = P [system in state b just after the next arrival to station j]

Then,

$$\pi_j(b) = \frac{\sum_{i=1}^{M} \pi(b(i,j))\ \mu_i\ p_{ij}\ I[n_j>0|b]}{P[\text{next transition is an arrival to } j] \sum_{i=1}^{M} \mu_i I[n_i>0|b]}$$

We know from (3.6) that,

$$P\ [\text{ next transition is an arrival to } j\ ] = q_j \Big/ \sum_{i=1}^{M} q_i$$

Therefore, from (3.6) and (3.7)

(3.9)
$$\pi_j(b) = \begin{cases} (\gamma_j/B_M^N)\left(\sum_{i=1}^{M} q_i\right) \prod_{k=1}^{M} X_k^{n_k(b)} & \text{if } n_j \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Define the random variable:

R_i = number at station i just after an arrival at i.

Then,

(3.10)
$$E(R_i) = (\gamma_i/B_M^N) \sum_{j=1}^{M} q_j \sum_{b \in \Omega_i(n_i \geq 1)} n_i \prod_{k=1}^{M} X_k^{n_k(b)}$$

and the mean wait is given by,

(3.11)
$$E(W_i) = E(R_i)/\mu_i$$

Chapter IV

MISCELLANEOUS EXTENSIONS

1. Approximations to Open Systems

Gordon [3] has also formulated a general closed network model. He has left the solution to the steady-state equations in a more cumbersome form than (2.1) and he has not shown the reduction in form to the cyclic queue case. His formulation is used to analyze the following two problems:

a - Asymptotic results for the marginal distribution of the number of customers at a service center as $N \to \infty$, M finite; and as $M \to \infty$, N finite.

b - Cyclic queues with limited storage space for customers between service centers.

His results on the first of these problems are relevant here. He has shown that as N becomes unbounded one can make two meaningful observations:

(1) the number of customers at the slowest service center, S, also becomes unbounded and hence the center is always busy;

(2) the marginal probability distribution for the number of customers at stage i ($\mu_i > \mu_S$), becomes geometric in the single channel case. That is,

$$P(n_i) = P_o^{\,i} X_i^{\,n_i}$$

Suppose we treat the slowest center as a Poisson input stream to the rest of the network. Using Jackson's definitions from Section 2, Chapter III we get,

$$\Gamma_S = \mu_S \,, \qquad \Gamma_i = \sum_{k=1}^{M} p_{ki}\, \Gamma_k \qquad (i \neq S)$$

It can easily be shown that Γ_i reduces to

$$\Gamma_i = \mu_S q_i$$

Hence,

$$P(n_i) = P_o^{\,i} (\Gamma_i/\mu_i)^{n_i} = P_o^{\,i} X_i^{\,n_i}$$

which is Gordon's result. Thus, the closed network system is asymptotically equivalent to an open network system. If more than one service center has the slowest service rate, Gordon's result holds only for the remaining centers with faster service rates.

2. Customer-Dependent Service Rates

For a general network structure we can easily handle a service distribution which is of the form

$$F(t;i,n) = 1 - e^{-\mu_i(n)t}$$

We assume that the internal movement probabilities remain fixed with respect to the number of customers at a service center. Then, it is easily shown that,

$$P(a) = P(N,0,\ldots 0) \frac{\prod_{k=1}^{N-n_1(a)} \mu_1(N-k+1) \prod_{i=2}^{M} q_i^{\,n_i(a)}}{\prod_{i=2}^{M} \prod_{k=1}^{n_i(a)} \mu_i(k)\, q_1^{\,N-n_1(a)}}$$

where an upper limit of zero in a product is interpreted as unity. We can specialize this result to,

$$\mu_i(k) = \begin{cases} k\mu_i & \text{if } 1 \le k \le S_i \\ S_i \mu_i & \text{if } k > S_i \end{cases}$$

where S_i = number of parallel channels at service center i. Letting $S_i = S$ and $S_j = 1$ $(j \neq i)$ we get the same result as (2.26), i.e.,

$$P(a) = \begin{cases} \dfrac{P(N,0,\ldots 0) \prod_{k=1}^{M} X_k^{\,n_k(a)}}{n_i!} & \text{if } n_i(a) < S \\[2ex] P(N,0,\ldots 0)\,(1/S!\, S^{n_i(a)-S}) \prod_{k=1}^{M} X_k^{\,n_k(a)} & \text{if } n_i(a) \ge S \end{cases}$$

3. Comparison of a Single Customer (N=1) and a Two Customer (N=2) System

For a single customer system there is no interference anywhere in the system. The steady-state results can be summarized as follows:

$$P(a) = \begin{cases} X_i / \sum_{i=1}^{M} X_i & \text{if } n_i(a) = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi(a) = \begin{cases} q_i / \sum_{i=1}^{M} q_i & \text{if } n_i(a) = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_i(a) = \begin{cases} 1 & \text{if } n_i(a) = 1 \\ 0 & \text{otherwise} \end{cases}$$

$E(R_i) = 1$

$E(W_i) = 1/\mu_i$, and $E(W_q(i)) = 0$.

We note that the Markov chain probabilities are independent of the service rates whereas the time average probabilities depend upon both service rates and normalized arrival rates (q_i). If the q_i were normalized by

$$\sum_{i=1}^{M} q_i = 1 ,$$

the Markov chain probabilities of the process would be identical to the Markov chain probabilities of a random walk through the network, i.e.,

$$\pi(a) = \begin{cases} q_i & \text{if } n_i(a) = 1 \\ 0 & \text{otherwise} \end{cases}$$

For purposes of comparison with a two-customer system the most meaningful parameter that shows interference is $E(W_q(i))$ the mean wait in queue. Omitting the details we get,

$$E(W_i) = q_i(2X_i^2 + \sum_{j\neq i} X_i X_j)/\sum_{i=1}^{M} q_i B_M^2$$

$$E(W_q(i)) = q_i X_i^2/\sum_{i=1}^{M} q_i B_M^2$$

Let L_i be the ratio of $\dfrac{E(W_i)}{E(W_q(i))}$

For N = 1, we get $L_i = \infty$, no interference.

For N = 2, we get,

$$L_i = 2 + (1/X_i) \sum_{j\neq i} X_i$$

$$= 2 + (\gamma_i/\gamma_1) \sum_{j\neq i} X_j$$

As γ_i gets large, L_i increases linearly. One can increase γ_i by increasing the service rate or by decreasing the probabilities of reaching service center i. Thus, for a center with a fast service rate and small probability of being reached L_i is large and we have little or no interference; for a center with a slow service rate which can be reached with high probability, L_i decreases and we have more interference. Notice that if μ_i is large but finite we always get some interference. In fact for N = 2,

$$\pi(a) = X_i^2 \mu_i / B_M^2 \qquad \text{if } n_i(a) = 2$$

$$= \gamma_1^2 \, q_i / \gamma_i \, B_M^2$$

The probability is inversely proportional to μ_i and hence is small; it can be neglected if γ_i is large enough.

As we increase N, the interference terms get larger and it is not apparent that even for fast service centers the mean delay in queue will be close to zero.

4. The Marginal Distribution of the Number of Customers at Service Center i

Reference has been made to Gordon's [3] work on marginal distributions of the number of customers at service center i. He has concluded that not much can be said in general about these distributions and he has developed asymptotic results for them. We can, however, prove a property of these distributions that will be of value in one of the optimization problems to be considered in Chapter V.

Define:

$p_i(k;N) = P[$ k customers at service center i when there are N in the system $]$

Theorem IV-1: $p_i(k;N)$ has an IFR (increasing failure rate distribution).

Proof: The failure rate function is defined as,

$$r_i(k;N) = p_i(k;N) / \sum_{j=k}^{N} p_i(j;N)$$

It is necessary to show,

$$r_i(k;N) \leq r_i(k+1;N) \qquad \text{for } k = 1,\ldots.N-1 \tag{4.1}$$

where N is arbitrary .

We proceed by induction on N.

N=1: $r_i(1;1) = 1$ by definition
$r_i(0;1) = p_i(0;1) < 1$ implying $r_i(0;1) \leq r_i(1;1)$

Now assume $r_i(k;N) \le r_i(k+1;N)$ for $k = 1,\ldots.N-1$

and we will show (4.1) for N+1. From the definition of $r_i(k;N)$, (4.5)

reduces to

$$p_i(k;N)/p_i(k+1;N) \le 1 + \Big(\sum_{a\in\mathcal{A}_i(n_i=k)} \prod_{k=1}^{M} X_k^{n_k(a)}\Big) \Big/ \sum_{j=k+1}^{N} \sum_{a\in\mathcal{A}_i(n_i=j)} X_k^{n_k(a)}$$

or

$$(4.2) \qquad p_i(k;N)/p_i(k+1;N) \le 1 + (X_i^k Z_{M-1}^{N-k}) \Big/ \sum_{j=k+1}^{N} X_i^j Z_{M-1}^{N-j}$$

where Z_M^N is defined by (2.2) and $Z_M^0 = 1$. We need to show

$$(4.3) \qquad p_i(k;N+1)/p_i(k+1;N+1) \le 1 + (X_i^k Z_{M-1}^{N-k+1}) \Big/ \sum_{j=k+1}^{N+1} X_i^j Z_{M-1}^{N-j+1}$$

Now,

$$p_i(k;N+1)/p_i(k+1;N+1) = X_i^k Z_{M-1}^{N-k+1} / X_i^{k+1} Z_{M-1}^{N-k} = Z_{M-1}^{N-k+1} / X_i Z_{M-1}^{N-k}$$
$$= p_i(k-1;N)/p_i(k;N)$$

But we have assumed the result is true for N. Therefore,

$$p_i(k;N+1)/p_i(k+1;N+1) \le 1 + (X_i^{k-1} Z_{M-1}^{N-k+1}) \Big/ \sum_{j=k}^{N} X_i^j Z_{M-1}^{N-j}$$

But,

$$\sum_{j=k}^{N} X_i^j Z_{M-1}^{N-j} = (1/X_i) \sum_{j=k+1}^{N+1} X_i^j Z_{M-1}^{N-j+1}$$

Therefore,

$$p_i(k;N+1)/p_i(k+1;N+1) \le 1 + (X_i^k Z_{M-1}^{N-k+1})/\sum_{j=k+1}^{N+1} X_i^j Z_{M-1}^{N-j+1}$$

establishing the induction. Hence $p_i(k;N)$ is an IFR distribution.

Chapter V

OPTIMIZATION PROBLEMS

1. Introduction

Managers of service systems are concerned with more than an analytic description of steady-state operating characteristics; they face economic problems, such as maximizing output, subject to resource constraints. For example, service rates in a coal mine may be linearly related to the number of workers doing a specific job; given a limited total number of workers in the mine, the problem is to allocate them to the work stations in order to maximize output. Or, a manager of an engine overhaul facility is interested in adding parallel production lines to increase output, or to reduce the probability of having no engines in service, and wishes to know if the benefits obtained will offset the extra cost.

These two examples offer a contrast in approaches to optimization problems. The former is more or less a continuous decision problem in the control variables, while the latter is a matter of costing and comparing specific alternatives using the results of Chapter III. In this chapter we shall examine the continuous decision problems in order to see if we can infer some general properties of the operating parameters of such systems. An example of discrete selection can also be found in Section 5.

2. Cost-Performance Alternatives

The two examples of optimization problems stated above provide a natural focus for two major classes of problems. These are:

1 - Allocation of resources to an existing queuing network.

2 - Design of a new queuing network and/or alteration of the structure of an existing network.

In both classes of problems the decision maker is usually interested in optimizing cost or performance, or in evaluating tradeoffs between cost and performance. Let us consider some possible forms for these problems.

For allocation problems we can identify the following major elements:

1 - Objective function - minimum cost, maximum production, etc.

2 - Technology - the necessary work stations in a network to process the customers, the arrangement of the work stations, the rules that define movement in the network, the relationships between resource inputs and service rates.

3 - Resources - labor, machines, capital, equipment that can change the technology, etc.

A typical mathematical form for such a problem is,

$$\text{(V.1)} \qquad \max_{M} F(\mu) \quad \text{subject to} \quad \sum_{i=1}^{M} K_i \leq L, \quad K_i \geq 0$$

where $\mu = (\mu_1, \ldots \mu_M)$

and $\mu_i = f(K_i)$

Here, k_i is the amount of resource (such as labor or capital) allocated to service center i, L is the maximum amount of the resource, and the service rate depends upon the allocation of the resource; $F(\mu)$ might be production rate, or cost rate, and its form depends upon the technology of the system.

We might also consider the possibility that the technology represented by the network can be altered by changing the internal movement probabilities. For example, suppose that revenue is generated each time a customer is served at service center i and that a cost, c_{ij} (per hour of operation), is associated with his movement to service center j. We can program the p_{ij} by the following problem:

$$\text{(V.2)} \qquad \text{Max } G(1-D_i) - \sum_{j=1}^{M} c_{ij}p_{ij} \qquad \text{subject to}$$

$$\sum_{j=1}^{M} p_{ij} = 1 . \qquad p_{ij} \geq 0$$

where G = revenue per customer

The first term in the objective function is the average revenue per hour and the second term is the average transportation cost per hour. Note that the p_{ij} appear implicitly in the first term in the definition of D_i. In solving problem (V.2) one must be sure that the programmed p_{ij}'s do not violate the assumption of irreducibility of the associated Markov chain; such a violation could occur if one or more of them had a zero value.

Next, consider a cyclic queue model of airline maintenance operation where a plane is served at two maintenance stations and then goes into flight. (Figure 3)

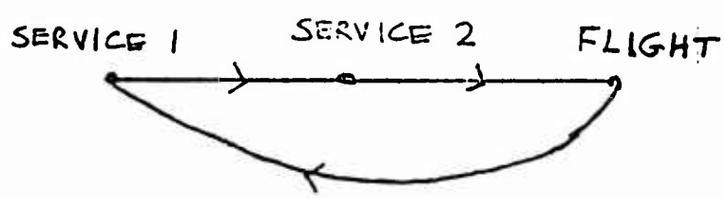


Figure 3

Airline Maintenance - Cyclic Queue

A plane earns \$1 for every hour in service and it costs \$ r for every hour of waiting time in queue. It would be of interest to describe the function of net revenue versus the number of planes, N, in the system in order to determine the marginal cost of congestion. On every cycle through the system the net revenue is, (expected value)

$$1 - r \sum_{i=1}^{3} E(W_q(i))$$

This type of problem is a typical example of a design problem. Another type of design problem would be the measurement of increased system output by increasing the number of service channels at a station and comparing the cost of increased service to the value of extra output. See [6] and [7] for example.

3. Maximization of Production Rate

Let us reconsider problem (V.1) as applied to a cyclic queue where the service rates are proportional to the resource and the objective is maximization of production rate, i.e.,

$$\text{Max } F(\mu) = (1-D_i)\,\mu_i \qquad \text{subject to}$$

(V.1a)
$$\sum_{i=1}^{M} \mu_i \le L$$
$$\mu_i \ge 0 \qquad \text{for all } i=1,\ldots,M$$

where $\mu_i = CK_i$, $q_i = 1$ for all i,

and C is a constant.

The problem can be reformulated as,

(5.1)
$$\text{Max } \Phi(\mu) = F(\mu) - \alpha\left(\sum_{i=1}^{M} \mu_i - L\right)$$

where α is the well known Lagrangian multiplier. If $F(\mu)$ was a concave function the solution of (5.1) would be at hand for the condition for a stationary point,

$$\partial\Phi/\partial\mu_i = 0 \qquad i=1,\ldots,M$$

is given by,

$$(5.2) \qquad (1-D_i) - \mu_i \, \partial D_i/\partial\mu_i = (1-D_j) - \mu_j \, \partial D_j/\partial\mu_j \qquad i,j = 1,\ldots,M$$

Concavity of $F(\mu)$ could not be shown because of the difficulty of analyzing ratios of polynomials. For example, consider the elements of the Hessian matrix of $F(\mu)$:

$$(5.3) \qquad \mu_i \, \partial^2 F/\partial\mu_i^2 = \mathrm{Var}(L(i))D_i - (\,E(L(i))\,)^2 D_i - E(L(i))D_i$$

$$(5.4) \qquad \mu_j \, \partial^2 F/\partial\mu_i\partial\mu_j = \mathrm{Covar}(L(i),L(j))D_i - E(L(i))E(L(j))D_i - E(L(j))D_i + (\,E(L(j))+1) \sum_{a\in\Omega_i(n_i=0)} n_j(a) \prod_k X_k^{n_k(a)}$$

These terms have proven too difficult for analysis in general and hence we cannot conclude that $F(\mu)$ is or is not concave. However, we can solve problem (V.1a) without requiring concavity.

<u>Theorem V-1</u>: The optimal allocation of resources for problem (V.1a) is

$$\mu_i = L/M \quad \text{for all } i=1,\ldots,M$$

<u>Proof</u>:

We state the obvious fact that $F(\mu)$ is homogeneous of order one. Therefore the solution lies in the plane,

$$\sum_{i=1}^{M} \mu_i = L$$

Next we prove,

<u>Lemma (5.1)</u> - $F(\mu)$ is concave in its orthogonal directions, i.e.,

$$\partial^2 F/\partial\mu_i^2 < 0$$

Proof: From equation (5.3) it is only necessary to show,

(5.5) $$\mathrm{Var}(L(i)) < (E(L(i)))^2 + E(L(i))$$

Define K^2 = coefficient of variation of the distribution of L_i.

Since $E(L(i)) > 0$, (5.5) reduces to

$$K^2 < 1 + 1/E(L(i))$$

By Theorem (IV-1) the distribution of L_i is IFR; Barlow and Proschan [1] have proven that if a distribution is IFR,

$$K^2 \leq 1 \quad \text{and hence,} \quad K^2 < 1 + 1/E(L(i))$$

proving the Lemma.

We also note that $F(\mu)$ is symmetric with respect to the service center. i.e. If μ^o is a permutation of μ, then $F(\mu^o) = F(\mu)$

Now let,

$$G = \{ \mu \mid \sum_{i=1}^{M} \mu_i = L \} = \text{set of all vectors, } \mu, \text{ in the plane.}$$

Consider some $\mu \in G$, and let

$$G_1 = \{\mu_1, \ldots, \mu_k\}$$
$$G_2 = \{\mu_{k+1}, \ldots \mu_M\}$$

such that if

$\mu_i \in G_1$ and $\mu_j \in G_2$, $\quad \mu_i \geq \mu_j$ for all i, j

Now consider some other vector $\mu^o \in G$, such that

$$\mu_i^o = \mu_i + m_i \epsilon \quad \text{for all } \mu_i \in G_1$$

$$\mu_i^o = \mu_i - n_i \epsilon \quad \text{for all } \mu_i \in G_2$$

where $\epsilon > 0$ is arbitrarily small, $m_i \geq 0$, $n_i \geq 0$.

and $$\sum_{i=1}^{k} m_i = 1 = \sum_{i=k+1}^{M} n_i$$

Because the solution plane cuts the space of μ symmetrically, one can move to any vector μ^o for which the service rates are more out of balance than in μ, by the indicated operation. The symmetry of F with respect to the service centers allows us to consider vectors only on one side of the balance point. The directional derivation of F from μ to μ^o is given by,

$$\partial F/\partial\mu = \epsilon \left(\sum_{i=1}^{k} m_i \, \partial F/\partial\mu_i - \sum_{i=k+1}^{M} n_i \, \partial F/\partial\mu_i \right)$$

Let $\mu_r = \underset{i\epsilon G_1}{\text{Min}} \; \mu_i$ and $\mu_s = \underset{i\epsilon G_2}{\text{Max}} \; \mu_i$

As a consequence of Lemma (V.1), if $\mu_i \leq \mu_j$

$$\partial F/\partial\mu_i \geq \partial F/\partial\mu_j .$$

Then,

$$\partial F/\partial\mu < \epsilon \left(\partial F/\partial\mu_r \sum_{i=1}^{k} m_i - \partial F/\partial\mu_s \sum_{i=k+1}^{M} n_i \right)$$

or

$$\partial F/\partial\mu < \epsilon \left(\partial F/\partial\mu_r - \partial F/\partial\mu_s \right)$$

But $\mu_r \geq \mu_s$ by previous assumptions implying

$$\partial F/\partial\mu_r \leq \partial F/\partial\mu_s$$

Therefore,

$$\partial F/\partial\mu < 0 .$$

Now if $\mu = (L/M,..L/M)$ then for every vector $\mu^o \epsilon G$,

$\partial F/\partial\mu \leq 0$ Which implies $F(\mu^o) \leq F(L/M,...L/M)$

Hence, production rate in a cyclic queue model can be maximized by equalizing the service rates. Without much difficulty, we will show that this result can be extended to a more general class of networks.

<u>Corollary (V-1)</u>: Given a network for which,

$$q_i = q_j \quad \text{for all } i,j$$

the optimal solution to problem (V.1a) is,

$$\mu_i = L/M \qquad i=1,...M$$

<u>Proof</u>: The proof of Theorem (V-1) rests on two major characteristics of cyclic queues. These are:

(1) $(1-D_i)\,\mu_i = \theta_i = \theta \qquad i=1,....M$

(2) L_i has an IFR distribution.

We have previously shown (Chapter III) that if $q_i = q_j$,

$$(1-D_i)\,\mu_i = \theta_i = \theta \qquad i=1,\ldots,M$$

and that L_i has an IFR distribution for any network. Therefore, Theorem (V-1) is valid here.

This type of network corresponds to one in which there is completely random internal movement such that each service center is equally likely to be visited by a customer.

The treatment of problem (V.1a) for a network where the q_i are arbitrary is analytically untractable for the following reasons:

1 - $F(\mu) = (1-D_i)\,\mu_i$ is not a constant for all $i = 1\ldots M$.

2 - $F(\mu)$ may not be concave.

3 - $F(\mu)$ is not symmetric with respect to i except in the special case noted in Corollary (V-1).

Let us suppose that the production rate is measured at a single service center, i, whose service rate is fixed; our problem reduces to allocating the resource to the M-1 other centers in order to maximize $(1-D_i)\,\mu_i$. Let us fix our attention on the single-customer network (N = 1).

<u>Theorem V-2</u>: For a single-customer (N = 1) closed network of queues and the following problem:

$$\text{Max } (1-D_i)\,\mu_i \qquad \text{subject to,}$$

$$\sum_{j\neq i} \mu_j \leq L, \qquad \mu_j \geq 0$$

the optimal solution is,

$$\mu_j = L(\sqrt{q_j}/\sum_{j\neq i}\sqrt{q_j} \tag{5.6}$$

Proof: Differentiating the Lagrangian function,

$$(1-D_i)\,\mu_i - \alpha(\sum_{j\neq i} \mu_j - L)$$

yields (5.6) as the stationary point. It is only necessary to show that this point is a maximum. Omitting the details, it is easily shown that the principal minors of the determinant of the Hessian matrix, of $(1-D_i)\,\mu_i$, alternate in sign and hence (5.6) represents an optimal solution.

For a system with an arbitrary number of customers, we can start out with (5.6) as a solution and then check to see how close we are to the stationary point. One can then iterate a few times in an attempt to get closer to the stationary point. It has been our experience that this procedure is computationally feasible, and furthermore, it produces a near-optimal solution.

4. The Marginal Cost of Congestion

We stated earlier the problem of marginal cost of congestion in an airlines maintenance system. Net revenue is given by,

$$C = 1 - r\sum_{i=1}^{3} E(W_q(i))$$

and we wish to find C as a function of the number of customers, N. We have already shown in Chapter IV that for

$$N = 1,\quad E(W_q(i)) = 0 \qquad \text{and for}$$
$$N = 2,\quad E(W_q(i)) = q_i X_i^2 / B_M^2 \sum_{i=1}^{M} q_i$$

Let $X_1 = 1$ $\qquad q_1 = 1/3$

$X_2 = 2$ $\qquad q_2 = 1/3$

$X_3 = 10$ $\qquad q_3 = 1/3$

After much calculation, for N = 3, we get,

N	C(N)
1	0
2	1 - .89
3	1 - 1.89

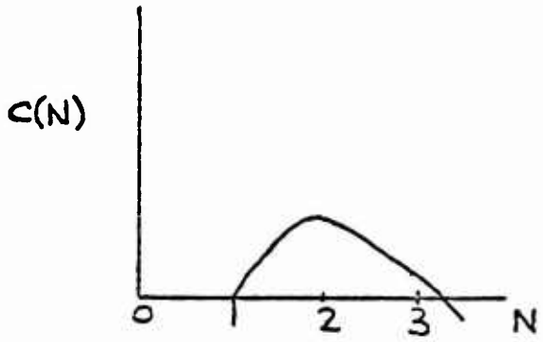


Figure 4

Marginal Cost of Congestion

If we treat N as a continuous variable, the cost curve has the form that is illustrated in Figure 4. With such a curve, a manager can decide how many customers (planes) can be handled profitably in the system. If the revenues and costs are marginal values, then the system is saturated at the value of N for which $C(N) = 0$ $(N \geq 1)$.

5. PCA Airlines - A Discrete Selection Problem

PCA Airlines is a commuter service between San Francisco and Los Angeles. Their fleet of planes consists of 5 Starstream Propjets; four of the planes run the regular service and the fifth one is held as a spare to be used when one of the others breaks down. PCA has heavy competition on its route and whenever they do not have four available planes, customers are lost to other airlines. Presently the planes have a mean failure time of 150 hours which accounts for planned services as well as random breakdowns. The planes are serviced in a two stage service station

where in the first stage they are inspected, diagnosed and cleaned and in the second stage they are repaired or overhauled; the mean service times are 2 and 10 hours respectively. Management has noticed that there are various times when they do not have four available planes and hence there is a loss of revenue. They wish to consider several possible improvements in the maintenance system in order to cut lost revenues.

The present system is illustrated in Figure 5.

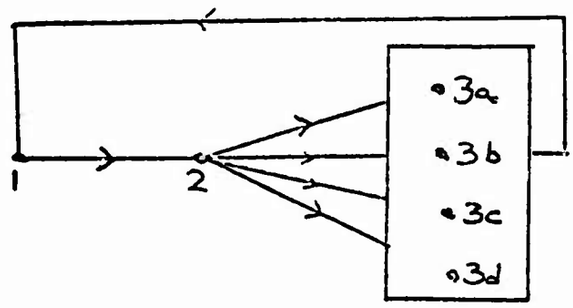


Figure 5

Present System - PCA Airlines

It consists of a 3 station cyclic queue, with 4 parallel channels at the third (or operating) station, and 5 planes in the system. Assuming that the distributions are exponential with mean service rates,

$$\mu_1 = 1/2$$
$$\mu_2 = 1/10$$
$$\mu_3 = 1/150$$

We can use the results of previous chapters to compute average lost revenue. Revenue is lost whenever there are fewer than 4 planes at center three. Therefore, (from 2.1 and summing to obtain the marginal distributions)

$$P(n_3 = 0) \approx 0$$
$$P(n_3 = 1) = .0017$$
$$P(n_3 = 2) = .0125$$
$$P(n_3 = 3) = .0623$$
$$P(n_3 \leq 3) = .0768$$

Hence 7.68 percent of the time at least one of the operating channels is idle.

The revenue statistics per plane are as follows:

Capacity	-	100 passengers
Avg. Load Factor	-	.90
Avg. Length of Flight	-	1 hour
Avg. Fare	-	\$12.00 per flight
Annual Flying Hours per Year	-	4,500

Avg. Annual revenue per plane = 100 (.90) \$12 x 4,500

= \$4.85 million

Avg. revenue lost per year = $\$4.85 \times 10^6 \, [1(.0623) + 2(.0125) + 3(.0017) + 4(0)]$ = \$447.500

Total avg. revenue per year = 4 ($\$4.85 \times 10^6$) - 447,500

= \$18,952,100

The lost revenue represents 2.3 percent of total revenue.

PCA is considering installing a new maintenance facility that will improve the breakdown rate of the planes but will require the same average service times. The new facility will increase the mean time between failures to 200 hours. Recalculating the measures of interest we get

$$P\ (n_3 \leq 3) = .0408$$

and

Avg. revenue lost per year = \$202,000

Hence, if the annual cost of the new facility is less than \$245,500, it pays to install it. We note here, that in these optimization problems <u>we are not directly concerned with flow rate but we are concerned with reducing the probability of not having enough planes to fly</u>. There is also the possibility that with the new system the servicing technology can change. With a little more time at the first station there is a 0.4

probability that a plane can go directly to the flight line and bypass the second stage. This structure is illustrated in Figure 6.

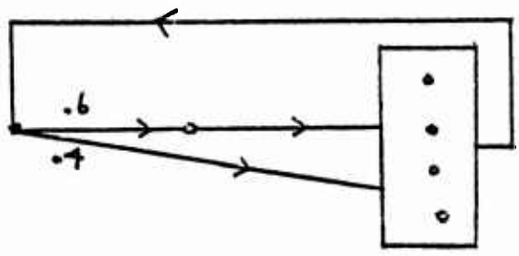


Figure 6

Modified System - PCA Airlines

Assume that the new service rates are

$\mu_1 = 1/4$ $\gamma_1 = 1/4$

$\mu_2 = 1/10$ $\gamma_2 = 1/6$

$\mu_3 = 1/200$ $\gamma_3 = 1/200$

Recalculating the results, we get

$$P\ (n_3 \leq 3) = .0273$$

Avg. revenue lost per year = $149,000

Hence, if the new facility and new technology costs less than $298,500, it pays to install it. One could go on and compute the effects of many other proposals; as long as we have closed form expressions the process could be computerized for larger problems.

It might be of interest to note the effects of having 5 planes in the system rather than 4. Under the last proposal, there is a probability of .81 of having all 5 planes available. This means that 81 percent of the time a plane is idle and not generating any revenue. Consider the same system with just 4 planes.

$$P\ (n_3 \leq 3) = .1825$$

and

Avg. revenue lost per year = $990,000.

The lost revenue is about twice the annual capital cost of one of these planes and therefore the extra plane is economically justified; note, that it may also be justified purely on a service level basis because an 18 percent chance of a cancelled flight may not be tolerable to the public.

REFERENCES

[1] Barlow, R. and Proschan, F., Mathematical Theory of Reliability, p. 33, John Wiley & Sons, Inc., 1965.

[2] Billingsley, S., "Statistical Inference for Markov Processes," Chicago University Press.

[3] Gordon, W. J., Jr., "Interaction in Closed Queueing Systems," Div. of Applied Math., Brown University, March 1965.

[4] Jackson, J. R., "Networks of Waiting Lines," Operations Research, pp. 518-521, 1957.

[5] Jackson, J. R., "Jobshop-Like Queueing Systems," Management Science, October 1963.

[6] Koenigsberg, E., "Cyclic Queues," Operational Research Quarterly, V9, No. 1, 1958.

[7] Koenigsberg, E., "Finite Queues and Cyclic Queues," Operations Research, pp. 246-259, 1960.

[8] Morse, P. M., Queues Inventories and Maintenance, pp. 39-58, John Wiley & Sons, Inc., 1958.

[9] Swersey, R. J., "Some Extensions of Cyclic Queue Theory," Operations Research Center, University of California, Berkeley, ORC 65-14 (June 1965).

[10] Wolff, R. W., "Problems of Statistical Inference for Birth and Death Queueing Models," Operations Research Center, University of California, Berkeley, ORC 63-3(RR), (March 1963).

[11] Wolff, R. W., private communication, September 1965.