AN INTRODUCTION TO QUEUING THEORY CONCEPTS

by 1264

JEANNE L. SEBAUGH

B.A., University of Kansas, 1962

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Mathematics

KANSAS STATE UNIVERSITY
Manhattan, Kansas
1970

Approved by:

S. Thomas Parker
Major Professor

TABLE OF CONTENTS

INTRODUCTION

Queuing theory involves the formal study of units arriving at some facility which services these units. The origin of queuing, or waiting-line, theory dates back to the early 1900's when A. K. Erlang developed telephone-traffic theory while working for the Copenhagen Telephone Company. Since that time, queuing theory has been applied to a wide variety of problems such as the landing of aircraft, the loading and unloading of ships, the timing of traffic lights, the design of automobile parking areas, the design of taxi waiting areas, the scheduling of patients in clinics, the processing of x-ray films, the passage of travelers through customs, the servicing of machines, and various aspects of the computer and teleprocessing fields.

Queuing theory attempts to solve problems that occur because of a fluctuating demand for a service to be performed over a period of time. Problems may arise because of too much or too little demand on the facilities or because of too few or too many facilities. In the case of too much demand on the facilities or too few facilities to meet the demand, a waiting-line forms. Too little demand or too many facilities results in idle facility time.

There are costs associated with both waiting time and idle time. Sometimes the actual costs associated with waiting time are difficult to calculate, since the monetary cost of inconvenience and congestion is intangible. But often an estimate of direct costs can be made. A restaurant can estimate

how many customers it loses because customers refuse to wait in a long line, and multiply this number by the average customer bill. The costs associated with idle facilities in terms of personnel and equipment are obvious.

One would like to obtain some optimum balance between the costs associated with waiting time and the costs associated with idle facility time. If the service facilities are fixed, some scheduling of the flow of units may be possible. An example of this situation is found in a study, made by the New York Port Authority, of automobile traffic through tunnels around Manhattan (11). The objective was to find some method of increasing the number of cars going through a tunnel in a given interval of time and, particularly, to try to eliminate the back-up of cars found in heavy traffic at the junction of the level part of the tunnel with the upgrade at either end.

Obviously, the facilities were fairly well fixed since the problem would have to become acute before building another tunnel would be justified. Therefore, a scheduling of the flow of units was tested. A computer program was written which accurately simulated the tunnel traffic, and then the effects of various speed limits and limits on inter-car spacing were tested. Indeed, it was found that it was possible to obtain more throughput if, instead of allowing the cars to enter the tunnel as fast as they arrived, the flow of cars were interrupted every so many cars and traffic were sent through in "platoons". Of course the optimum platoon size and inter-platoon distance had to be chosen correctly.

On the other hand, if the flow of units is not subject to control, then one tries to find the proper combination of personnel and equipment. An example of this technique is found in another study conducted by the New York Port Authority (5) concerning delays at toll booths at Port Authority tunnels and bridges. The result of this study was a recommendation of the optimum number and schedule for the toll collectors and the number of toll booths needed open at any time of day.

It may be possible to exert some control over both the flow of units and the available facilities. In this case, one seeks to schedule the flow of units and to provide the proper combination of facilities in order to minimize the over-all cost.

Whether one is studying the flow of automobiles in a tunnel, of customers in a supermarket, or of messages in a telecommunications system, common terminology and elements of study exist. One purpose of this paper is to acquaint the reader with queuing theory terminology, the four basic elements of any queuing situation, and the measures of congestion which may be examined.

Ideally, given these elements and their interactions, any of these systems could be represented in mathematical terms and the appropriate analyses applied to determine the expected effects of various modes of operation. In reality, however, it is often impossible to carry out such an extensive analysis. This is due in large part to the fact that the theory has not progressed beyond the point where easily manageable solutions are available for more than a few idealized systems.

The theoretical background and analytical solution through the use of differential/difference equations are presented for one of these idealized cases, the single-queue case with random arrivals and random service. The formulas for certain measures of congestion are presented for some of the other solvable queuing situations.

A study of these idealized problems should help one develop an insight into the character of queuing action. In some cases, it will provide upper bound and reasonableness checks for more accurate computer simulation, another topic of discussion in this paper.

THE ELEMENTS OF A QUEUING PROBLEM

A waiting line or queue occurs when a unit arrives at some busy service facility. In order to study this queuing situation one must first of all state the problem in a logical fashion. This involves a formal description of the situation and an enumeration of the results desired.

Before proceeding further some terms will be explicitly defined, since many definitions are not uniform in the literature. For example, some authors use the terms "queue" and "waiting line" to represent different phenomena while others use them interchangeably. In this paper the term waiting line will be used to denote those items actually waiting in line and will not include any items being serviced. A waiting line will be considered to be only part of a queue which will refer to all items in the system.

A service facility is alternatively referred to as a server or as a channel. Each of these terms refers to the person(s) and/or piece(s) of equipment providing the service.

In developing a model of the queuing situation certain assumptions must be made that specify it completely. The four basic areas of concern or elements of any queuing problem are the system's input, the waiting lines, the service channels, and the system's output.

The System's Input

The system's input is the area which is concerned with how units arrive and become part of the system. Once the unit becomes a part of the system it is generally called a "customer". The customers come from a population of potential customers. Usually each member of the population is considered to be functionally independent of all others.

One important characteristic of an input source is the population size, that is whether it is considered to be finite or infinite. In general the smaller the source or population, the greater is its effect on the arrival rate. If the population is assumed to be infinite, it is obvious that the supply of potential customers could never be depleted. Although a population is known to be finite, sometimes it may be assumed to be infinite if its behavior more closely resembles that of an infinite population.

The second important characteristic of an input source is the arrival rate. In order to describe the customer arrivals at a queue, some pattern must exist. It is possible that customers arrive at constant intervals, say one every twenty seconds. This would be the simplest pattern but is one which occurs infrequently. Usually the arrivals are somewhat irregular or variable. Because of this irregularity, arrival patterns are usually described in probabilistic terms in either of two basic ways. One way of doing this is to state the probability of a specific number of arrivals per time unit. The other is to state the probability that the time between arrivals (the interarrival

time) is less than a particular time t in the form of a probability distribution.

The arrival pattern most commonly used in queuing applications is that of completely random arrivals. It is also the simplest to deal with mathematically. Let a denote the mean arrival rate or the average number of customers arriving in each unit of the time interval T. That is, if T is measured in seconds, a is expressed as customers per second. The mean arrival rate, a, is sometimes referred to as the traffic rate and in some of the literature is denoted by λ instead of a. Let Δt denote a small increment in time. Then the following assumptions must hold if the input traffic is to be considered completely random:

1. The numbers of arrivals during non-overlapping time intervals are always independent. Thus neither is the arrival of a customer influenced by past arrivals nor does it affect future arrivals.
2. The probability that more than one arrival occurs in a given time Δt is negligible when compared with the probability of a single arrival, for small Δt. This probability is denoted by $o(\Delta t)$ where $o(\Delta t)$ denotes an infinitesimal of order higher than that of Δt.
3. The probability that an arrival occurs in Δt is approximately proportional to Δt when Δt is small. $P(\text{one customer arrives in any } \Delta t) = a\Delta t + o(\Delta t)$. This implies that the traffic load must be independent

of time and that peaking conditions must not be predictable within the time span being studied. If it is known when peak-period traffic occurs, that particular span of time may be studied independently.

Now divide t, a time period of fixed length, into m intervals of length Δt and let $P_k(t)$ denote the probability of exactly k arrivals in t, for all values of k = 0, 1, 2, ... Cox (6) outlines the following method of deriving $P_k(t)$.

$$P_1(\Delta t) = a\Delta t + o(\Delta t)$$

$$P_o(\Delta t) = 1 - a\Delta t + o(\Delta t)$$

Therefore, by the binomial probability law,

$$P_k(t) = \lim_{\Delta t \to 0} \frac{m!}{k!(m-k)!} (a\Delta t+o(\Delta t))^k (1-a\Delta t+o(\Delta t))^{m-k}$$

Since $t=m\Delta t$, $\Delta t=t/m$ and $m=t/\Delta t$; therefore as $\Delta t \to 0$, $m \to \infty$. Thus,

$$P_k(t) = \lim_{m\to\infty} \frac{m!}{m^k(m-k)!} a^k \frac{t^k}{m^k} \left(1-\frac{at}{m}\right)^{m-k}$$

$$= \frac{(at)^k}{k!} \lim_{m\to\infty} \frac{m!}{m^k(m-k)!} \lim_{m\to\infty} \left(1-\frac{at}{m}\right)^{m-k}$$

Now, for fixed k

$$\lim_{m\to\infty} \frac{m!}{m^k(m-k)!} = \lim_{m\to\infty} \frac{m.(m-1)\cdot\ldots\cdot(m-(k-1))(m-k)(m-(k+1))\cdot\ldots\cdot 1}{m\cdot m^{k-1}\cdot(m-k)(m-(k+1))\cdot\ldots\cdot 1}$$

$$= \lim_{m\to\infty} \left(\frac{m-1}{m}\right)\left(\frac{m-2}{m}\right)\ldots\left(\frac{m-(k-1)}{m}\right)$$

$$= \lim_{m\to\infty} \left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\ldots\left(1-\frac{k-1}{m}\right)$$

$$= 1$$

Also,

$$\lim_{m\to\infty}\left(1-\frac{at}{m}\right)^{m-k} = e^{-at}$$

Hence,

$$P_k(t) = \frac{(at)^k}{k!}\, e^{-at} \quad , \; a>0, \; k=0, 1, 2, \ldots$$

This is the Poisson distribution with both mean and variance of at.

As stated before, an alternate way to describe the arrival pattern is to give a probability distribution F(t) that the time between arrivals is less than a particular time t. This is known as the interarrival distribution. If a is the mean arrival rate, then 1/a is the mean interarrival rate or the mean arrival time, T_a.

If the input process is Poissonian, then $P_o(t)=e^{-at}$. The probability of no arrivals by time t is equivalent to the probability that periods of time between events will exceed t. Therefore,

$$P(\text{interarrival time } <t) = F(t) = 1-e^{-at}$$

This is referred to as the exponential distribution. The density function $F'(t)$ equals ae^{-at}. Note that this is a continuous distribution in time whereas the Poisson is discrete.

The Poisson or exponential distributions of arrivals are likely to be particularly good approximations when the customers are drawn from a large population all of whose members behave independently of one another. Flagle (10) states that it is demonstrable

> ...that as the divergence of events from scheduled times becomes great relative to the interval between scheduled events, the statistical properties of the phenomenon rapidly converge upon those of the Poisson process.

Indeed, the Poisson process may be looked on as the disintegration of a scheduled process.

If the pattern of arrivals is considered to be regular, it is usually assumed that customers arrive singly at equally spaced intervals 1/a. The rate of arrival of customers is a per unit time.

Of course, many other types of arrival patterns may occur. Those discussed by Cox (6) include general independent arrivals, regular arrivals with unpunctuality, aggregated arrivals, complex deterministic arrivals, discrete-time arrivals, non-stationary arrival patterns, arrivals correlated with other aspects of the

system, and arrivals in a continuous flow. But, as he states (6), p.18,

> ...The completely random and regular patterns are the most commonly used in applied mathematical work and it is only for these that mathematical solutions of any generality can be obtained; other arrival patterns usually require special investigation.

A check to roughly determine whether test data has a Poisson character is given in (1). It uses the squared coefficient of variation, C^2=variance/mean^2. The following criteria are suggested:

$0.0<C^2<0.7$: arrivals tend to be evenly spaced;

$0.7<C^2<1.3$: arrivals approximately Poisson;

$1.3<C^2$: arrivals tend to cluster.

The mean for the exponential distribution is 1/a and the variance is $1/a^2$. So it can be seen that in this case $C^2=1$.

Furthermore, (1) states that congestion becomes more acute as C^2 increases. Actually, an assumption of random or Poisson arrivals gives "worst-case" estimates for the behavior of queues with more regular arrivals. On the other hand, it will underestimate queues with clustering arrivals. This assumes the same mean arrival rate in each case.

The distribution of the system's input may be complicated by many factors which should be taken into consideration. Some control may be exerted on customers before their arrival. For example, an airport might notify incoming airplanes of inclement weather and ask them to take various courses of action. A unit might balk (not join the queue) if it sees that the line is too long. If there is more than one waiting line, the method of

choosing which one to join is important. Maybe the shortest line is joined or perhaps one is chosen at random, regardless of size. One might find collusion of customers over whether to join a line or which line to join. If a unit does not join a waiting line, it would be important to know whether it is entirely lost to the system or whether its entry is delayed for a certain period of time.

The Waiting Lines

Once a waiting line has formed, there are further properties of the line and customers to consider. Even after a customer joins a line it might be possible for him to renege or leave the system if he becomes impatient. Many times the waiting line has an upper bound on its length. In analysis, however, an infinite waiting area is often assumed and then probabilities for exceeding various finite levels are estimated.

Another characteristic of the waiting line to consider is referred to as the queue discipline. This is the rule by which the next customer is chosen from the waiting line to be serviced. The most common discipline is service in order of arrival or first-in/first-out. Other common disciplines include last-in/first-out and random selection. Customers may have assigned priorities, priorities which change, or preemptive priorities. Other less common disciplines are also mentioned in the literature.

The Service Channels

The service channels or servers also need to be described. Of course, one needs to know the number of channels available at any given time. If more than one channel is present, the arrangement of channels must be known. The channels may be in parallel, in series, or in any combination of these.

Service may be intermittent because of the server sometimes being unavailable. However, in this paper the assumption is made that a server is always present. A server is said to be idle if and only if there are no customers to be served. If the server is idle and a customer arrives, service begins immediately. Upon completion of the service, the next service begins immediately if a customer is waiting. Otherwise the server is again idle.

The most important characteristic of a channel is the amount of time which the customer spends in the channel. This is called the holding time or service time. This service time may vary from customer to customer and, as in the case of the arrival pattern, is described by a probability distribution, H(t). In the great majority of cases it is assumed that the holding times of different customers are mutually independent, identically distributed random variables. H(t) denotes the probability that the service time is less than a particular time t. The average number of customers serviced per unit time is s. The mean service time μ is the average time required to service a customer, $1/s=\mu$.

As with arrival rates, a great mathematical simplification can be made if the individual service times can be assumed to be distributed completely randomly. Under such a constraint the

three assumptions listed previously for random arrivals would hold for individual service times. In most practical applications, the service distribution is not exponential. However, an assumption of exponentially distributed service times, $H(t)=1-e^{-st}$, usually represents a worst-case assumption (1). Therefore, it is still often used because of its great simplification in the form of solution.

In (1) there is an explanation of how the assumption of exponentially distributed service times results in simplification of analysis. Let

$$H(t) = P(\text{service time}<t) = 1-e^{-st}$$

Then

$$P(\text{service time}>t) = e^{-st}\text{ , with mean equal to } 1/s.$$

Now assume that a service has been given for t_o units of time and let r denote the remaining time service is to be given. Let S equal the total service time so that $S=t_o+r$. One wants to find the mean of the remaining service time. The distribution of r is given by the conditional probability

$$\begin{aligned} P(r>t' \mid S>t_o) &= P(r+t_o>t'+t_o \mid S>t_o) \\ &= \frac{P(S>t'+t_o \text{ and } S>t_o)}{P(S>t_o)} \\ &= \frac{P(S>t'+t_o)}{P(S>t_o)} \\ &= \frac{e^{-s(t'+t_o)}}{e^{-st_o}} \\ &= e^{-st'} \end{aligned}$$

It is seen that the distribution of the remaining service time is the same as the original distribution, and the mean values are the same. This means that given a mean service time of 1/s, no matter how long a particular service has been in progress, the mean remaining service time is still 1/s. The simplification in analysis comes from that fact, namely, that a process governed by the exponential distribution is not affected by what has occured in the past and its future is affected only by its present state.

Although arrivals can often be approximated by a Poisson distribution, service times can rarely be realistically approximated by an exponential distribution. If the queue is simple (one server) and has a generalized service distribution, some mean queue statistics may be derived which depend only on the first two statistical moments of the service distribution. This type of analysis was first done by Pollaczek and was subsequently simplified by Khintchine. Therefore, the formulas are usually known as the Pollaczek-Khintchine formulas.

The squared coefficient of variation, C^2=*variance/mean*^{2}, can also be used with service distributions to determine what type of analysis is applicable. If service is constant, the variance is zero and, therefore, C^2 equals zero. If service is exponential, as before, C^2 equals one. If C^2 is greater than one, the general or arbitrary service theory should be used.

If C^2 lies between zero and one, there is a special service distribution which may be used (1), the Erlang-m distribution. It is actually a family of functions known as Erlang

distributions with parameter m. It is so named because Erlang used the functions in his studies of telephone traffic. The general form is

$$P(\text{service time}<t) = F_m(t)$$

$$= 1-e^{-D} \sum_{k=o}^{m-1} \frac{D^k}{k!} \text{ , } D=smt \text{ and } m=1,2,\ldots$$

The exponential distribution results if m=1.

$$D = s\cdot 1\cdot t = st \quad \text{and} \quad \sum_{k=0}^{0} \frac{D^k}{k!} = 1$$

Hence

$$F_1(t) = 1-e^{-st}$$

A constant results if m is infinite. As m increases from the value of one, more and more of the probability masses about the mean value 1/s.

The density function is then given by

$$\frac{dF_m(t)}{dt} = f_m(t) = \frac{t^{m-1}(sm)^m e^{-smt}}{(m-1)!} \text{ .}$$

The first and second moments may be obtained by a progressive integration by parts. The first moment is 1/s and the second moment is $(m+1)/s^2m$ giving a variance of $1/s^2m$.

The squared coefficient of service variation for the Erlang-m distribution is

$$C^2 = \frac{\text{variance}}{\text{mean}^2} = \frac{s^2}{s^2 m} = \frac{1}{m} \text{ .}$$

Therefore, given the mean and variance of the service times, one can determine the order m of the Erlang distribution that approximates most closely a given distribution (1). It should be remembered that the Erlang distribution is not valid if C^2 is greater than one.

There are several other more complicated ways of stating the service time or time spent in the channels. Some of these are special Erlangian and gamma type distributions. Others take into account customers of several different types.

The System's Output

The system's output is a result of all the previously discussed elements plus any additional factors inherent in the output process itself, if such exist. In the simplest case, with a single server and first-in/first-out queue discipline, the system's output can be determined from the arrival and service distributions. As the system becomes more complicated, the system's output will be influenced by the growing number of factors.

The process of a unit leaving the system may in itself become more complicated. For example, the customers leaving the system may back-up and disrupt the servicing process. There may be some cycling of customers with some of the system's output again becoming part of the system's input.

MEASURES OF CONGESTION

Certain measures of congestion drawn from statistics are required to describe the effects of a queuing system under a fluctuating traffic load. The simplest measure of congestion is the concept of traffic intensity. It is defined as

$$\frac{\text{mean arrival rate}}{\text{mean service rate}} = \frac{\text{mean service time for customers}}{\text{mean arrival time for customers}}$$

$$= \frac{a}{s}$$

This dimensionless ratio, called an "erlang" in honor of A. K. Erlang, is equal to the total expected service time per unit time.

As explained by (1), aT customers are expected to arrive in a long time interval T. Each of these customers requires a service of mean time $1/s=T_s$. The total expected time to sequentially service all the expected customers is aTT_s. The ratio of the total expected service time per unit time may be obtained by dividing by the length of the period T, resulting in aT_s or a/s.

If this ratio is greater than one, customers are arriving faster than one server can serve them. It can be seen that the traffic intensity of any queuing system identifies the minimum number of servers required to handle a given traffic system with no loss of customers.

In order to generalize the above discussion of traffic intensity to compensate for systems with loss of customers or

with multiserver facilities in which traffic is evenly divided among the servers, another closely related ratio may be introduced. It is the server utilization, denoted by p, and measures the fraction of time that a single server is busy.

$$p = a'T_s = a'/s$$

The variable a' is introduced which represents the traffic rate of customers actually served by the server. It will be less than or equal to a, the total arrival rate of customers to the queue. In a single-server system with no loss of customers, a' equals a. Indeed, much of the literature defines p=a/s. In a system in which traffic is evenly distributed among c servers, a'=a/c.

Empirically, p is restricted to a value less than or equal to one, since it is physically impossible for a server to be more than 100% busy (p=1). As the rate of traffic passing through the system, a', increases, so does the server utilization and customer congestion. That is, the waiting line and, consequently, customer waiting time, becomes longer.

The theoretical maximum input to a single-server, no-loss queue would be $1/T_s=s$, or the mean service rate. However, waiting lines become quite large near system saturation and grow without bound when p=1. As (1) points out, practicality usually limits the input for a single server to 70-90% of the theoretical maximum.

These two ratios, traffic intensity and server utilization, provide rough measures of the reasonableness of a system's capacity for handling given traffic loads. They provide a limit beyond

which the system cannot operate unless corrective action is taken. However, they do not provide a direct measure of how well a system is operating within rated limits.

Saaty (18) gives an extensive list of measurements that could be important in studying a queuing system. It includes, in addition to those mentioned previously:

(1) the probability that the number of customers in the waiting line or in the system equals n,

(2) the mean and variance of the number of customers in the waiting line or in the system,

(3) the distribution of the time spent in the waiting line or in the system,

(4) the mean and variance of the time spent in the waiting line or in the system,

(5) the probability that the waiting time is longer or is not longer than a given period of time,

(6) the probability that there is someone waiting,

(7) the mean length of a busy period,

(8) the ratio of the mean waiting time to the mean service time,

(9) the probability that not more than a certain number of channels is occupied,

(10) the mean number of idle channels,

(11) the probability of a lost call.

It is not within the scope of this paper to study all of these measures of congestion. The two areas of greatest practical interest are the number of customers in the queue and the amount

of time a customer spends in the system. In particular, this paper examines the probability of any specified number of customers in the queue or in the waiting line alone, and the mean and variance of the number of customers in the queue or in the waiting line. It examines the mean and variance of the queuing time and waiting time and the probability that the queuing time and waiting time are less than some specified amount.

NOTATION

It seems appropriate at this time to summarize previous notation and introduce notation for the measures of congestion to be studied throughout the rest of this report.

T = time interval under study

t = an instant in time

Δt = a small increment in time

a = mean arrival rate (if T is in seconds, the average number of customers arriving per second)

$1/a = T_a$ = mean interarrival time (if T is in seconds, the average time between customer arrivals in seconds)

s = mean service rate (if T is in seconds, the average number of customers serviced per second)

$1/s = T_s$ = mean service time (if T is in seconds, the average time required to service a customer in seconds)

$p = a'T_s = a'/s$ = server utilization

P_n = Probability (n customers in the system)

MCq = mean number of customers in the system (queue)

VCq = variance of the number of customers in the system

MCw = mean number of customers in the waiting line alone

VCw = variance of the number of customers in the waiting

line

Q(t) = Probability (queuing time, including service <some value t)

MTq = mean queuing time

VTq = variance of the queuing time

W(t) = Probability (waiting time <some value t)

MTw = mean waiting time for service

VTw = variance of waiting time for service

c = number of channels

STATIONARITY

There are two basic types of servicing operations, determinate and indeterminate. In a determinate type of operation, all aspects of the system, such as the arrival and servicing rates, are precisely known as functions of time. Therefore, the state of the system can be precisely predicted at any time. However, if any of the aspects of the system vary in an unknown, or random, fashion, then an indeterminate situation exists.

Thus in an indeterminate mode of operation, the measurable quantities associated with the operation will be stochastic variables, which, over a span of time, fluctuate about some average values or rates. The system can then be defined as being in a number of possible states, specified by such things as the number of customers in a queue, waiting for service, or in service, the particular phase of each customer in its service channel, etc.

Now instead of trying to predict precisely how the state under study changes with time, the probabilities that the system is in each of its possible states, the state probabilities, may be calculated. From these probabilities, one can calculate the means and variances of the various quantities of interest, such as the average length of the queues, and derived probabilities, such as the probability that the queue is not longer than a certain amount.

If the mean arrival and service rates are constant (that is, if the fluctuations in both are short term around constant

mean values), the state probabilities and the derived averages will be independent of time. That is, starting from an initial state, the process tends toward equilibrium irrespective of the initial state. In the state of equilibrium, the process shows only statistical fluctuation with no tendency toward any particular state. In all practical cases, this stationary condition will occur after a long enough time if p is less than one and if a and s remain constant.

ANALYTICAL SOLUTIONS TO SOME QUEUING PROBLEMS

In solving queuing problems analytically, one starts by surveying the structure of the system and rules which govern its behavior. Assumptions are made concerning the arrival and service probability distributions. From this foundation, one then tries to compute the state probabilities and, finally, to calculate the mean values of the various quantities of interest.

It becomes apparent that the number of combinations of arrival rates and service rates, queue disciplines and channel designs is quite large. In a paper of this type it is impossible to do more than outline some of the more important techniques used and to give results for only the few commonly occurring, idealized situations.

In this section, it is assumed that the customers arrive from an infinite source and are served on a first-in/first-out basis. The servicing rate is assumed to be independent of the number of customers in line. In the case of a single server, the server is assumed to be busy as long as there is a customer in the queue.

Single-Server Queue: Random Arrivals/Random Service

Queue Size. Let $P_n(t)$ equal the probability that n customers are in the queue at time t. Then by the assumption of stationarity P_n may be obtained and from it the means and variances can be calculated. The following derivation is taken from Churchman (5).

First of all, one develops a set of differential equations from which to obtain $P_n(t)$. Note that $a\Delta t$ equals the probability of a new customer entering the line in the time interval t to $t+\Delta t$. $s\Delta t$ equals the probability that the servicing of a customer is completed in the time interval t to $t+\Delta t$. Therefore, the probability that there will be n customers ($n>0$) in the queue at time $(t+\Delta t)$ is expressed as the sum of the following four independent compound probabilities:

1. The product of the probabilities that
   a. There are n customers in line at time t $(P_n(t))$
   b. There are no arrivals during the Δt interval $(1-a\Delta t)$
   c. There are no customers serviced during the Δt interval $(1-s\Delta t)$
2. The product of the probabilities that
   a. There are (n+1) customers in line at time t $(P_{n+1}(t))$
   b. There are no arrivals during the Δt interval $(1-a\Delta t)$
   c. There is one customer serviced during the Δt interval $(s\Delta t)$
3. The product of the probabilities that
   a. There are (n-1) customers in line at time t $(P_{n-1}(t))$
   b. There is one arrival during the Δt interval $(a\Delta t)$

c. There are no customers serviced during the Δt interval $(1-s\Delta t)$

4. The product of the probabilities that

a. There are n customers in line at time t $(P_n(t))$

b. There is one arrival during the Δt interval $(a\Delta t)$

c. There is one customer serviced during the Δt interval $(s\Delta t)$

The probabilities that more than one customer arrives or that more than one customer is serviced during the Δt interval may be assumed to be negligible since the arrival and service rates are random.

The above four probabilities may be transformed as follows:

1. $P_n(t)(1-a\Delta t)(1-s\Delta t) = P_n(t)(1-a\Delta t-s\Delta t) + o_1(\Delta t)$.
2. $P_{n+1}(t)(1-a\Delta t)(s\Delta t) = P_{n+1}(t)s\Delta t + o_2(\Delta t)$.
3. $P_{n-1}(t)(a\Delta t)(1-s\Delta t) = P_{n-1}(t)a\Delta t + o_3(\Delta t)$.
4. $P_n(t)(a\Delta t)(s\Delta t) = o_4(\Delta t)$.

The $o_i(\Delta t)$ are higher order terms of Δt that are assumed to be negligible when compared to those in Δt.

By adding these probabilities, one obtains for the probability of n customers in line at time $(t+\Delta t)$,

$$(1)\quad P_n(t+\Delta t) = P_n(t)(1-a\Delta t-s\Delta t) + P_{n+1}(t)s\Delta t + P_{n-1}(t)a\Delta t + o_1(\Delta t) + o_2(\Delta t) + o_3(\Delta t) + o_4(\Delta t).$$

This equation may be rewritten as follows

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = aP_{n-1}(t) + sP_{n+1}(t) - (a+s)P_n(t) +$$

$$\frac{1}{\Delta t}(o_1(\Delta t) + o_2(\Delta t) + o_3(\Delta t) + o_4(\Delta t)).$$

By letting Δt approach 0, one obtains the differential equation

$$(2) \quad \frac{dP_n(t)}{dt} = aP_{n-1}(t) + sP_{n+1}(t) - (a+s)P_n(t) \quad , n>0.$$

The assumption has been made so far that n>0. Now consider the case when n=0. In this situation, the probability that there will be 0 customers in the queue at time $(t+\Delta t)$ is the sum of the two independent probabilities:

1. The product of the probabilities that
   a. There are 0 customers in line at time t $(P_o(t))$
   b. There are no arrivals during the Δt interval $(1-a\Delta t)$
2. The product of the probabilities that
   a. There is one customer in line at time t $(P_1(t))$
   b. There is one customer serviced during the Δt interval $(s\Delta t)$
   c. There are no arrivals during the Δt interval $(1-a\Delta t)$

By adding these two probabilities, one obtains for the probability of a queue of length 0 at time $(t+\Delta t)$

$$(3)\quad P_o(t+\Delta t) = P_o(t)(1-a\Delta t) + P_1(t)s\Delta t - asP_1(t)(\Delta t)^2,$$

from which it follows that

$$\frac{P_o(t+\Delta t) - P_o(t)}{\Delta t} = -aP_o(t) + sP_1(t) - asP_1(t)(\Delta t)$$

and

$$(4)\quad \frac{dP_o(t)}{dt} = -aP_o(t) + sP_1(t) \quad , n=0.$$

The differential equations 2 and 4 express implicitly the relationships between waiting time and servicing time and thus furnish the basis for solutions to many queuing problems. Solutions are usually difficult to obtain, depending upon the complexity of $P_n(t)$.

However, one readily obtains a solution in the case in which it is assumed that stationarity is obtained and that $P_n(t)$ is independent of t and, in fact, equals P_n. Then, since this probability does not change with time, its rate of change is equal to zero:

$$\frac{dP_n}{dt} = 0 \quad , n=0,1,2,\ldots$$

Equations 2 and 4 then become

$$(2a) \quad 0 = aP_{n-1} + sP_{n+1} - (a+s)P_n \quad (n>0),$$

$$(4a) \quad 0 = -aP_o + sP_1 \quad (n=0).$$

Equations 2a and 4a are difference rather than differential equations and may be solved for P_o, P_1, ..., P_n, ... by successive substitution and utilization of the fact that

$$\sum_{i=1}^{\infty} P_i = 1.$$

The procedure is as follows:

$P_o = P_o$,

$P_1 = pP_o$, (From eq. 4a and setting p=a/s)

$P_2 = p^2P_o$, (From setting n=1 in eq. 2a and substituting for P_1)

$P_3 = p^3P_o$, (From setting n=2 in eq. 2a and substituting for $P_2$0,

.
.
.

$P_n = p^nP_o$.

Summing corresponding members of these equations, one obtains

$$(5) \quad \sum_{i=0}^{\infty} P_i = P_0 \sum_{n=0}^{\infty} p^n .$$

The assumption is made that $p<1$, a condition that must hold to prevent queue length without bound. Since

$$\sum_{i=0}^{\infty} P_i = 1 \qquad \text{and} \qquad \sum_{n=0}^{\infty} p^n = \frac{1}{1-p},$$

then by the equation for the sum of an infinite geometric series, one has

$$\frac{1}{1-p} P_0 = 1.$$

Hence,

$$(6) \quad P_0 = 1-p.$$

By substituting this value of P_0 in the foregoing expression for P_n, it follows that the probability of a waiting line of length n is given by

$$(7) \quad P_n = p^n(1-p) \quad , \; p<1.$$

Now one may find MCq, the mean length of the queue. By definition, since $\Sigma P_n = 1$,

$$(8)\quad MCq = \sum_{n=0}^{\infty} nP_n.$$

Substituting in equation (8) the value of P_n given in equation (7), equation (8) becomes

$$MCq = \sum_{n=0}^{\infty} np^n(1-p)$$

$$= (1-p) \sum_{n=0}^{\infty} np^n$$

$$= (1-p)(p+2p^2+3p^3+\ldots)$$

$$(9)\qquad = p(1-p)(1+2p+3p^2+\ldots).$$

To evaluate this expression one first obtains the sum of the series within the last set of parenthesis by the use of integration and differentiation. Call the series S(q) and integrate it term by term to obtain

$$\int_0^p S(q)dq = p+p^2+p^3+\ldots$$

which is a geometric series having the sum $p/(1-p)$. Now differentiate this sum with respect to p and obtain $1/(1-p)^2$. This means that

$$(10)\quad S(p) = \frac{1}{(1-p)^2}.$$

Hence, substituting this value in equation (9), one obtains

$$(11)\quad p(1-p)\frac{1}{(1-p)^2} = \frac{p}{1-p},$$

so that, for the given conditions, the mean length of the queue is given by

$$(12)\quad MCq = \frac{p}{1-p},\ p<1.$$

The variance of the queue size can be calculated by using the procedure above to find the second moment, MC^2q. Then

$$VCq = MC^2q - (MCq)^2$$

$$= \frac{p+p^2}{(1-p)^2} - \frac{p^2}{(1-p)^2}$$

$$(13)\qquad = \frac{p}{(1-p)^2}.$$

It can be seen that as p approaches 1, the queue size grows without bound, giving rise to an infinite queue. Similarly, the variance increases sharply with p. This explains the great instability of highly utilized queues.

In the single-server case, the mean size of the waiting line is

$$MCw = \sum_{n=1}^{\infty} (n-1)P_n$$

$$= \sum_{n=1}^{\infty} (n-1)p^n(1-p)$$

$$= (1-p) \sum_{n=1}^{\infty} (n-1)p^n$$

$$= (1-p)(p^2+2p^3+3p^4+\ldots)$$

$$= p^2(1-p)(1+2p+3p^2+\ldots)$$

$$= p^2(1-p)\frac{1}{(1-p)^2}$$

$$= \frac{p^2}{1-p} . \tag{14}$$

Similarly,

$$VCw = MC^2w - (MCw)^2$$

$$= \frac{p^2(1+p-p^2)}{(1-p)^2} . \tag{15}$$

Note that MCw is shorter than MCq by the quantity p, MCq=MCw+p. The difference is, on the average then, less than one customer.

Queuing Time. The derivation of the probability distributions for queuing time and waiting time is quite lengthy, involving Laplace transforms, and only the results will be given here (1).

The distribution for total time spent in the system is

$$Q(t) = \text{Prob(queuing time<t)}$$

(16) $= 1-e^{-(1-p)t/T_s}$

This distribution is of the exponential form. As p approaches 1, the (1-p) factor causes the distribution to stretch out.

The mean queuing time of customers at a single channel with random input can be formulated as follows:

(17) $MCq = aMTq$

Therefore,

$$MTq = \frac{MCq}{a}$$

$$= \frac{p}{a(1-p)}$$

$$= \frac{a/s}{a(1-p)}$$

$$= \frac{1}{s(1-p)}$$

(18) $$= \frac{T_s}{1-p}$$

Of course, MTq could also be obtained as follows:

$$MTq = \int_0^\infty t\,dQ(t)$$

$$= \int_0^\infty \frac{t(1-p)}{T_s} e^{-(1-p)t/T_s} dt.$$

After integrating by parts,

$$MTq = -\frac{T_s}{1-p} e^{-(1-p)t/T_s} \Big]_0^\infty$$

$$= \frac{T_s}{1-p} , \tag{18a}$$

which agrees with equation 18.

The variance of the queuing time is

$$VTq = \int_0^\infty (t-MTq)^2 dQ(t)$$

$$= \frac{T_s^2}{(1-p)^2} . \tag{19}$$

The distribution for the waiting time, not including service, is

$$W(t) = \text{Prob(waiting time<t)}$$

$$= 1-pe^{-(1-p)t/T_s}. \tag{20}$$

It should be noted that W(0)=1-p. This represents the fraction of arrivals which find the server free and therefore have zero waiting time. The waiting time distribution has a modified exponential form.

The mean waiting time could be derived in the following manner. Since queuing time is the sum of the waiting time and the service time, the mean values are additive.

$$MCq = aMTq \qquad \text{(Eq. 17)}$$

$$= a(MTw+T_s). \qquad \text{(Additive property)} \tag{21}$$

Therefore,

$$MTw = \frac{MCq}{a} - T_s$$

$$= \frac{p}{a(1-p)} - \frac{1}{s} \qquad \text{(Substituting for MCq and } T_s\text{)}$$

$$= \frac{sp - a(1-p)}{as(1-p)}$$

$$= \frac{s(a/s) - a(1-a/s)}{as(1-p)} \qquad \text{(Substituting for p)}$$

$$= \frac{a - a + a^2/s}{as(1-p)}$$

$$= \frac{a/s^2}{(1-p)}$$

$$= \frac{pT_s}{1-p} \,. \tag{22}$$

The mean waiting time could be obtained in the normal fashion:

$$MTw = \int_0^\infty tdW(t)$$

$$= \frac{pT_s}{1-p} \,. \tag{22a}$$

The variance of waiting time for service is

$$VTw = \int_0^\infty (t-MTw)^2 dW(t)$$

$$= \frac{(2-p)pT_s^2}{(1-p)^2} \qquad (23)$$

Example [(1),p.10.]. Messages arrive at a telecommunications switching center for a particular outgoing communication line in a Poisson manner with a mean arrival rate of 180 messages per hour. The outgoing transmission times are proportional to the message lengths, which are distributed approximately exponentially, with mean length of 144 characters. Line speed is 12 characters per second.

The mean arrival rate, or average number of messages arriving per second is

$$a = 180 \text{ mgs/hr} = 0.05 \text{ mgs/sec.}$$

The reciprocal, or average time between message arrivals is

$$1/a = T_a = 20 \text{ sec/mg.}$$

The average time required to service a message is

$$1/s = T_s = 144 \text{ char}/(12 \text{ char/sec}) = 12 \text{ sec/mg.}$$

s, the average number of messages per second is then

$$s = 1/T_s = 0.08 \text{ mgs/sec.}$$

The server in this example is the outgoing transmission line. Its utilization is given by

$$p = a/s = aT_s = (0.05 \text{ mgs/sec}) \times (12.0 \text{ sec/mg}) = 0.6$$

Knowing the value of p and using equations 12-15, values for the following table may be calculated.

	Numbers of messages	
	In the system (queue)	In the waiting line
Mean	1.5	0.9
Variance	3.75	2.79

Then using equations 18, 19, 23, and the values of p and T_s, the following values regarding queuing time and waiting time can be evaluated.

	Time in seconds	
	Queuing time	Waiting time
Mean	30	18
Variance	900	756

Single-Server Queue: Random Arrivals/General Service

As has been stated previously, the assumption of an exponential (random) service time distribution is usually unrealistic. An approximation technique using the Pollaczek-Khintchine formulas may be used for a simple queue with generalized service time distribution. As it turns out, the mean queue statistics depend only on the first two statistical moments of the service distribution. The following discussion is taken in large part from (1).

Arrivals and Service. Let H(t) denote the generalized service time distribution of a simple queue, with moments b_n:

$$H(t) = \text{Prob(service time<t)}$$

$$(24) \quad b_n = \int_0^\infty t^n dH(t).$$

The mean service time, T_s, and the variance of the service time, VT_s, are then given by

$$(25) \quad T_s = b_1,$$

$$(26) \quad VT_s = b_2 - b_1^2.$$

Assuming a Poisson input with mean arrival rate a, the server utilization is then given as

(27) $p = ab_1$.

Unlike the exponential service case, explicit expressions for the queue size distribution and the waiting time distribution are not known in the general case. Instead, expressions are derived for the moments of these distributions using Laplace transform theory.

Queue Size. The mean and variance of the queue size in terms of the service moments, b_n, are (see (1))

$$(28)\quad MCq = \frac{a^2 b_2}{2(1-p)} + p$$

$$(29)\quad VCq = \frac{a^3 b_3}{3(1-p)} + \frac{a^4 {b_2}^2}{4(1-p)^2} + \frac{a^2(3-2p)b_2}{2(1-p)} + p(1-p)$$

If the service time distribution is indeed exponential, the service moments, b_n, are found to be

$$b_n = \int_0^\infty t^n dH(t)$$

$$= \int_0^\infty t^n d(1-e^{-t/T_s})$$

$$= \int_0^\infty t^n ((1/T_s)\, e^{-t/T_s})dt$$

$$= n!\, {T_s}^n ,$$

so that $b_1 = T_s$, $b_2 = 2{T_s}^2$, and $b_3 = 6{T_s}^3$.

Substituting these values into equations 28 and 29, the queue size formulas for exponential service are obtained as given in equations 12 and 13. For example, substituting into equation 28,

$$MCq = \frac{a^2 b_2}{2(1-p)} + p$$

$$= \frac{a^2(2T_s^2)}{2(1-aT_s)} + aT_s$$

$$= \frac{2a^2T_s^2+2aT_s-2a^2T_s^2}{2(1-aT_s)}$$

$$= \frac{aT_s}{1-aT_s}$$

$$= \frac{p}{1-p} . \qquad \text{(Same as eq. 12)}$$

Since MCq = MCw + p,

it follows that

$$MCw = MCq-p$$

$$= \frac{p}{1-p} - p$$

$$= \frac{p-p+p^2}{1-p}$$

$$= \frac{p^2}{1-p} . \qquad \text{(Same as eq. 14)}$$

Queuing Time. The first three moments of the waiting time distribution, W(t), are as follows:

$$(30)\quad W_1 = MTw = \frac{ab_2}{2(1-p)} ,$$

$$(31)\quad W_2 = \frac{ab_3}{3(1-p)} + \frac{a^2b_2^2}{2(1-p)^2}$$

$$(32)\quad W_3 = \frac{ab_4}{4(1-p)} + \frac{a^2b_2b_3}{(1-p)^2} + \frac{3a^3b_2^3}{4(1-p)^3} .$$

Likewise, the first three moments of the queuing time distribution, Q(t), are

$$(33)\quad q_1 = MTq = \frac{ab_2}{2(1-p)} + b_1 ,$$

$$(34)\quad q_2 = \frac{ab_3+3b_2}{3(1-p)} + \frac{a^2b_2^2}{2(1-p)^2} ,$$

$$(35)\quad q_3 = \frac{ab_4+4b_3}{4(1-p)} + \frac{2a^2b_2b_3+3ab_2^2}{2(1-p)^2} + \frac{3a^3b_2^3}{4(1-p)^3} .$$

In the special case in which the service is a constant, T_s, the service time distribution, H(t), is defined as

$$H(t) = 0, \text{ when } 0<t<T_s$$

$$1, \text{ when } t>T_s$$

The moments, b_n, of H(t) are then

$$b_n = \int_0^\infty t^n dH(t)$$

$$= \int_0^\infty t^n (t-T_s)dt$$

$$= T_s{}^n ,$$

so that $b_1 = T_s$, $b_2 = T_s{}^2$, and $b_3 = T_s{}^3$.

Substituting the values for these moments into equation 30, the mean waiting time for constant service is found to be

$$MTw = \frac{ab_2}{2(1-p)}$$

$$= \frac{aT_s{}^2}{2(1-p)}$$

$$= \frac{pT_s}{2(1-p)} . \qquad (36)$$

By comparing this with the waiting time found using the exponentially distributed service time given in equation 22, one can see that the wait for a constant server is, on the average, one-half the wait for a random server.

Equation 33 may be rewritten for mean queuing time in terms of the mean service, T_s, and the service variance, VT_s. Using the relationships of equations 25 and 26 in equation 33, one obtains

$$MTq = \frac{ab_2}{2(1-p)} + b_1$$

$$= \frac{1}{2(1-p)} (ab_2+2b_1-2b_1p)$$

$$= \frac{1}{2(1-p)} (2b_1-2ab_1^2+ab_2)$$

$$= \frac{1}{2(1-p)} \left(\frac{(2b_1^2-2ab_1^3+ab_1b_2)}{b_1}\right)$$

$$= \frac{b_1}{1-p} \left(\frac{(2b_1^2-2ab_1^3+ab_1b_2)}{2b_1^2}\right)$$

$$= \frac{T_s}{1-p} \left(1- \frac{2ab_1^3-ab_1b_2}{2b_1^2}\right)$$

$$= \frac{T_s}{1-p} \left(1- \frac{ab_1(2b_1^2-b_2))}{2\left(b_1^2\right)}\right)$$

$$= \frac{T_s}{1-p} \left(1- \frac{p}{2}\left(\frac{(b_1^2-b_2+b_1^2))}{b_1^2}\right)\right)$$

$$= \frac{T_s}{1-p} \left(1- \frac{p}{2}\left(\frac{(1-b_2-b_1^2))}{b_1^2}\right)\right)$$

(37) $$= \frac{T_s}{1-p} \left(1- \frac{p}{2}\left(\frac{(1- VT_s))}{T_s^2}\right)\right)$$

It should be noted that the last term in equation 37 is the squared coefficient of variation of service time. As has been mentioned, for most useful service distributions, C^2 lies between 0 (constant service) and 1 (exponential service). Thus mean queuing time is usually found within limits given by the following:

$$(38)\quad \frac{T_s}{1-p}\,(1-p/2) \le MTq \le \frac{T_s}{1-p}\,.$$

These bounds are close for small server utilization and diverge for increasing utilization. This indicates that service distribution is relatively unimportant in low-use systems but becomes more important with heavy traffic.

Single-Server Queue: Random Arrivals/Erlang Service

In order to classify those service distributions with squared coefficient of service variation between 0 and 1, the Erlang-m distribution is used. The general form with mean T_s is

$$(39)\quad F_m(t) = 1 - e^{-D} \sum_{k=0}^{m-1} \frac{D^k}{k!}, \quad D = \frac{mt}{T_s} = smt.$$

As was pointed out previously, the value of m can be obtained by knowing the mean and variance of the service distribution since

$$m = \frac{1}{C^2} = \frac{\text{Mean}^2}{\text{Variance}}.$$

The nearest integer value of m must be chosen.

Moments of the Erlang-m Distribution. Using the Erlang-m distribution for service time, one can derive formulas containing only the mean and the parameter m by substituting the Erlang-m moments into the queuing formulas for random arrivals and general service. The moments for the Erlang distribution are

$$(40)\quad b_n = \frac{(n+m-1)!}{(m-1)!} \frac{(T_s)^n}{(m)^n}.$$

In particular

$$b_1 = \frac{(1+m-1)!}{(m-1)!} \frac{(T_s)}{(m)}$$

$$= \frac{m!}{(m-1)!} \frac{(T_s)}{(m)}$$

$$= m \frac{(T_s)}{(\overline{m})}$$

$$= T_s \quad , \tag{41}$$

$$b_2 = \frac{(2+m-1)!}{(m-1)!} \frac{(T_s)^2}{(\overline{m})}$$

$$= \frac{(m+1)!}{(m-1)!} \frac{(T_s)^2}{(\overline{m})}$$

$$= m(m+1) \frac{(T_s)^2}{(\overline{m})}$$

$$= \frac{(m+1)T_s^{\,2}}{m} \quad . \tag{42}$$

Queue Size. Substituting the first and second moments of the Erlang-m distribution into equation 28 and remembering that $p=ab_1=aT_s$, one obtains for the mean queue size

$$MCq = \frac{a^2 b_2}{2(1-p)} + p$$

$$= \frac{a^2(m+1)T_s^{\,2}}{2m(1-p)} + p$$

$$= \frac{p^2(m+1)}{2m(1-p)} + p$$

$$= \frac{1}{2m(1-p)} (mp^2+p^2+2mp-2mp^2)$$

$$= \frac{1}{2m(1-p)} (2mp-p^2m+p^2)$$

$$= \frac{p}{1-p}\,\frac{(2m-pm+p)}{2m}$$

$$= \frac{p}{1-p}\left(1-\frac{pm-p}{2m}\right)$$

$$= \frac{p}{1-p}\left[1-\frac{p}{2}\left(\frac{m-1}{m}\right)\right]$$

$$(43) \qquad = \frac{p}{1-p}\left[1-\frac{p}{2}\left(1-\frac{1}{m}\right)\right].$$

Using the same method, one obtains the variance of the queue size

$$(44) \quad VCq = \frac{p}{(1-p)^2}\left[1-\frac{p}{2}\left(3-\frac{p(10-p)}{6}-\frac{3-3p+p^2}{m}-\frac{p(8-5p)}{6m^2}\right)\right].$$

Queuing Time. These formulas are obtained in the same fashion as above. The results are

$$(45) \quad MTw = \frac{pT_s}{2(1-p)}\left(1+\frac{1}{m}\right),$$

$$(46) \quad MTq = \frac{T_s}{1-p}\left[1-\frac{p}{2}\left(1-\frac{1}{m}\right)\right],$$

$$(47) \quad VTq = \left(\frac{T_s}{1-p}\right)^2\left\{\left[1-\frac{p(4-p)}{6}\left(1-\frac{1}{m}\right)\right]\left(1+\frac{1}{m}\right)-\left[1-\frac{p}{2}\left(1-\frac{1}{m}\right)\right]^2\right\}.$$

Since these formulas are somewhat complicated, it is usually easier to graph them as functions of p and m.

Example. Messages arrive at a telecommunications switching center for a particular outgoing line in a Poisson manner with a mean arrival rate of 0.20 messages per second. The messages are of fixed length and the service time is a constant of 4.0 seconds. The variance of the service time is thus zero, causing m to be infinite since $m=\text{mean}^2/\text{variance}$. If m is infinite, then 1/m is zero.

The average number of messages arriving per second is

$$a = 0.20 \text{ mgs/sec.}$$

The average time required to service a message is the constant time

$$T_s = 4.0 \text{ sec/mg.}$$

The utilization of the server, that is, the outgoing transmission line is

$$p = aT_s = 0.80$$

The means of the number of messages in the system (queue) and in the waiting line alone may then be calculated from equation (43) to be

MCq = 2.4 messages, and

MCw = MCq - p = 1.6 messages.

The means of the queuing and waiting times may be calculated from equation (46) to be

MTq = 12.0 seconds, and

MTw = MTq - T_s = 8.0 seconds.

Multiserver Queue: Infinitely Many Servers

In this situation, it is assumed that each of the infinite number of servers offers identical service to customers arriving at random, each finding a server available immediately no matter how many others are already being served. Hence, no waiting line can form and the number of customers in the queue consists only of those being served.

The problem in such a situation is to find the distribution, mean, and variance of the number of busy servers. Then a practical upper bound for the number of servers can be established. Random or Poisson input has been assumed with mean rate a. Each server is assumed to have an identical general service distribution with mean service time T_s. Upon letting P_n be the probability that an observer finds n customers in the system, causing n servers, or channels, to be busy, one obtains the relation

$$(48) \quad P_n = \frac{a(aT_s)^n}{n!} e^{-aT_s} ,$$

or the Poisson distribution with parameter aT_s.

The mean and variance of the number of customer holding servers is then

$$(49) \quad MCq = aT_s,$$

$$(50) \quad VCq = aT_s.$$

Note that the traffic intensity, aT_s, is in this case not limited to less than one. It indicates the mean number of busy servers.

The probability that the number of customers is less than some given value of m is

$$(51) \quad \text{Prob}(n \leq m) = e^{-aT_s} \sum_{n=1}^{m} \frac{(aT_s)^n}{n!} .$$

The values of this function can be found in a cumulative Poisson function table.

Multiserver Queue: Finite Number of Servers

In this model it is assumed that arriving customers join a common waiting line and are serviced by a finite number of identical servers in parallel. In actuality any rule could be used to assign a customer to several available servers. The rule which, on the average, evens out the traffic load among the servers is that allowing the customers to choose a free server at random.

(1) states that useful measures of congestion in this case have been obtained only by using the following assumptions: Random input with mean arrival rate=a customers per unit time and identical exponential service time distribution with mean T_s for all servers.

The traffic intensity is again aT_s. But now, since there is more than one server, $a'=a/c$ and $p=aT_s/c$, which for stable operation must be less than one.

The probability of n customers in the system, P_n, can be found by using differential-difference equations as in the single-server case. It can be shown that

$$(52)\quad P_n = \begin{cases} \dfrac{(aT_s)^n}{n!} P_0 \text{ , if } n<c \text{ ,} \\ \dfrac{(aT_s)^n}{c!c^{n-c}} P_0 \text{ , if } n\geq c \text{ .} \end{cases}$$

As before, P_0 can be determined by using the law of total probability, which states that the sum of the probabilities P_n is equal to unity. The result is

$$(53)\quad P_0 = \frac{1}{\sum_{n=0}^{c-1} \frac{(aT_s)^n}{n!} + \frac{c}{c-aT_s}\frac{(aT_s)^c}{c!}} .$$

The mean number of busy servers is given by

$$\sum_{n=0}^{c-1} nP_n + \sum_{n=c}^{\infty} cP_n = aT_s ,$$

or the traffic intensity.

The probability that all c servers are busy is the same as the probability that there are c or more customers in the system.

$$(54)\quad P(n\geq c) = \sum_{n=c}^{\infty} P_n = \frac{1-r_c}{1-pr_c} , \quad r_c = 1 - \frac{e^{-aT_s}\frac{(aT_s)^c}{c!}}{e^{-aT_s}\frac{(aT_s)^n}{n!}} .$$

In this form r_c can be calculated from tables of the Poisson distribution.

Equation 54 can also be viewed as the probability that an arriving customer, finding all servers busy, must wait for service in the common waiting line.

Queue Size. The formula for the mean length of the waiting line is

$$(55)\quad MCw = \sum_{n=c}^{\infty} (n-c)P_n$$

$$= \frac{p}{1-p}\ \text{Prob}(n>c).$$

Queuing Time. The mean time waiting for service, MTw, is obtained from the relation MTw = MCw/a:

$$(56)\quad MTw = \frac{T_s\ \text{Prob}(n>c)}{c(1-p)}\ .$$

The waiting time distribution is given by

$$(57)\quad W(t) = 1 - \text{Prob}(n>c)\ e^{-(c-aT_s)t/T_s}.$$

An important use of the distribution W(t) is to determine the number of servers required to satisfy a given waiting time criterion.

THE USE OF SIMULATION IN SOLVING QUEUING PROBLEMS

In constructing a model of a queuing situation, it is desirable to be able to use variables whose values can be obtained analytically without too much difficulty. In order to do this, simplifying assumptions, such as random input, are often made. However, some of the expressions in the model which are constructed from even very simple variables may themselves become complex. This is particularly true when probability concepts are involved.

Monte Carlo Technique

In some instances, it is not even possible, let alone practical, to evaluate such a function by mathematical analysis. Fortunately, such expressions can often be evaluated by a Monte Carlo technique. In essence, a Monte Carlo technique consists of simulating an experiment to determine some probabilistic property of a population of objects or events by the use of random sampling applied to the components of the objects or events. Examples and descriptions of how a Monte Carlo technique can be applied to queuing problems are found in (5), (6), (12), and (19).

The Monte Carlo technique can be employed by the use of random number tables and manual calculations. However, many computerized routines have been developed to enable the simulation to be done on a digital computer.

Computer Simulation

Distinct programming languages exist by which computers may be made to show how a particular model of a queuing situation would perform. One such language, written by IBM, is called the General Purpose Systems Simulator, GPSS (13).

GPSS allows the user to study the logical structure of the system, to follow the flow of traffic through the system, and to measure the effects of blocking which might be caused by the limited capacity of parts of the system. Output of such a program gives such information as the amount of traffic that flows through a part of the system or the whole system, the average time to pass between selected points of the system or the whole system, the extent to which each part of the system is used, and the maximum and average queue lengths occurring in various parts of the system.

It does not seem worthwhile to go into any detailed explanation of how to use such a language. It should be pointed out, however, that the user need only know the language used to describe the model and does not need to be able to program for the computer on which the program operates.

An example of a problem which is quite simple to simulate on the computer, but which would prove more difficult to solve analytically, is the following supermarket problem.

Customers arrive every 30 ± 30 seconds at a supermarket with 100 pushcarts. If no carts are available, the customer goes away.

Twenty percent of the customers are express (8 items or less) and complete their shopping in 6±5 minutes. They are serviced by a single express checkout counter which takes 90±60 seconds per customer.

The remaining eighty percent of the customers take 50±40 minutes to do their shopping. They are serviced by 4 regular checkout counters which take 120±90 seconds per customer.

There is a separate line for each regular checkout counter. Customers pick a checkout counter at random without regard to length of line since they cannot observe all checkout counters simultaneously.

An advantage of computer simulation is that once the model of such a system has been set up, various of the parameters may be changed and the model run again with relative ease. In the example above, the arrival and service rates, and the numbers of pushcarts and counters could all be varied and the best arrangement found.

The biggest disadvantage of computer simulation is that it is expensive in terms of computer time. It is characteristic of many queuing systems, especially those with extremely variable service times or arrival patterns, or with high traffic intensity, that quite variable results are obtained. This means that the whole procedure should be repeated a number of times independently from the beginning and run for a sufficiently long time to avoid any difficulties associated with initial conditions.

This underlies the desirability of first trying to get at least some rough "ball-park" analytic solution if at all

possible in order to try to determine the feasibility of computer simulation. It is also desirable to obtain analytic solutions where possible in the model, even if they are applicable only to small parts of the problem.

CONCLUSION

This paper describes the basic elements in a queuing situation, presents the basic theory which describes the relationships among these elements, and reviews some basic techniques used to study queuing problems. Indeed, the purpose of queuing theory is to provide various techniques for obtaining such measures of congestion as the average queue length and average waiting time when the arrival and service rates are known. These quantities may be found by analytical techniques if possible, by computer simulation, or often by some combination of these two techniques. If costs can be assigned to waiting time and service time, the problem of establishing a proper balance between these costs can be determined.

Further areas of interest might involve the study of combinations of arrival and service distributions other than those presented, the study of queuing disciplines other than the first-in/first-out discipline assumed in this paper, the study of customer priorities, the study of systems in which there is customer loss, and the study of networks of queues. A brief description and the results of many of these variations can be found in reference (1).

There are many other, more advanced, mathematical techniques which are employed in the study of queues. Boudreau and others set up their analyses of a computer study (3) and a telecommunications study (4) using the theory of Markov chains. Schay (20) studied a multiserver problem using methods that were

based on an analogy to statistical mechanics. Leibowitz (15) suggested the investigation of the extensive theory of many-body systems in mathematical physics as a clue to techniques for approaching queuing structures.

The number of books and articles on queuing theory describing the types of problems that have been approached and the methods that have been used to solve these problems is quite large. An article by Doig (8) contains a bibliography of published work on queuing theory up to 1957. A more recent, excellent bibliography is found in Saaty's book (18). New British and American work appears mostly in the Bell System Technical Journal, in the Journal of the Royal Statistical Society, Series B, and in Operations Research.

REFERENCES

(1) Analysis of some queuing models in real-time systems. IBM data processing techniques publication, form F20-0007-0. n.d.

(2) Benes, Vaclav E.
General stochastic processes in the theory of queues. Reading, Mass.: Addison-Wesley, 1963.

(3) Boudreau, P.E. and M. Kac.
Analysis of a basic queuing problem arising in computer systems. IBM Jour. of Res. and Develop. 5(2): 132-140. April, 1961.

(4) --------, J.S. Griffin, Jr., and M. Kac.
A discrete queuing problem with variable service times. IBM Jour. of Res. and Develop. 6(4): 407-418. October, 1962.

(5) Churchman, C.W., R.L. Ackoff, and E. Leonard Arnoff
Introduction to operations research. New York: John Wiley and Sons, 1957.

(6) Cox, D.R., and Walter L. Smith.
Queues. New York: John Wiley and Sons, 1961.

(7) Cramer, Harold.
Mathematical methods of statistics. Princeton: Princeton University Press, 1946.

(8) Doig, A.
A bibliography on the theory of queues. Biometrika 44: 490-514, 1957.

(9) Feller, William.
An introduction to probability theory and its applications. Second edition. New York: John Wiley and Sons, 1957.

(10) Flagle, Charles D.
Queuing theory. Operations research and systems engineering, chapter 14. Baltimore: The Johns Hopkins Press, 1960.

(11) Freiberger, Walter F., and William Prager, editors.
Computers and operations research. Applications of digital computers, 1-10. Boston: Ginn and Co., 1963.

(12) Goode, H.H., and R.E. Machol.
High traffic-queuing theory. System engineering, chapter 23. New York: McGraw-Hill, 1957.

(13) Gordon, G.
A general purpose systems simulator. IBM Systems Journal. September, 1962.

(14) Khintchine, A.Y.
Mathematical methods in the theory of queuing. New York: Hafner, 1960.

(15) Leibowitz, M.A.
An approximate method for treating a class of multi-queue problems. IBM Jour. of Res. and Develop. 5(3): 204-209. July, 1961.

(16) McCloskey, J.F., and J.M. Coppinger, editors.
Queuing theory. Operations research for management, volume 1, 134-148. Baltimore: The Johns Hopkins Press, 1956.

(17) Morse, Philip M.
Queues, inventories and maintenance. New York: John Wiley and Sons, 1958.

(18) Saaty, Thomas L.
Elements of queuing theory. New York: McGraw-Hill, 1961.

(19) --------.
Mathematical methods of operations research. New York: McGraw-Hill, 1959.

(20) Schay, Jr., G.
Approximate methods for a multiqueue problem. IBM Jour. of Res. and Develop. 6(2): 246-249. April, 1962.

(21) Takacs, Lajos.
Introduction to the theory of queues. New York: Oxford University Press, 1962.

SEQUEL

Since some of the preceding portions of this report were written in 1966, it was felt that a 1969 sequel would be appropriate in order to review recent literature concerning the theory and application of queuing theory. It was soon evident that the flood of papers relating to queuing theory has not ceased.

Three important abstract sources for material relating to queuing theory are Mathematical Reviews, the International Journal of Abstracts: Statistical Theory and Method, and International Abstracts in Operations Research. To illustrate the abundance of literature being published in this area, it was discovered that for the period 1966-1969 the annual average of references in the International Abstracts in Operations Research (26) alone was over one hundred per year.

In 1965, Morse (29) presented an analysis of queuing theory literature. He stated:

> Publication increased exponentially until a few years ago. From 1910 to 1955 the number of papers on queuing published per year doubled every five years; in other words, in each five-year period the number of papers published equalled the total number published previously. This rapid growth slackened somewhat in the past eight years; it is still growing but the growth is slower. The rate is even now considerably faster than the doubling every fifteen years, which is characteristic of scientific publication as a whole.

Several authors now use the term 'stochastic service system' rather than 'queue'. Riordan (31) argues that the use of the term queue has many weaknesses and prefers not to

use it. For example, in some problems no physical queue forms. However, as Morse (29) points out, the "vocal simplicity" of the word 'queue' will probably result in its being used in somewhat inappropriate situations. Meanwhile, when researching the literature, one must examine both subject headings, queues and stochastic service systems.

Another problem confounding an evaluation of recent developments in queuing theory is the criticism that much of the work is receiving. In particular, it appears that more and more of the publications are oriented towards theory (29). This tendency towards more formal mathematics does not delight the applications oriented operations research people, since they feel that most of the results are of little practical use. Lee (27) is most critical of current work:

> The majority of these papers originate, it appears, in the twilight zone of academic graduate research. In them, the remoter mysteries of the simpler models of the more familiar queuing-processes are courageously explored; and the well-known properties of the classical models are repeatedly derived anew. Whilst the mathematical apparatus becomes even more elaborate, it remains difficult to find reports of experimental work with queueing-processes, or of empirical observation of queueing-processes, or even of applications of existing theory. There seems to be a great deal of what passes for research, and very little of what might pass for application.

Morse (29) relates his fear that "...the theory may lose its contact with applications and become ingrown". He suggests that a change in view may help keep interest in the field and, in fact, stimulate its growth. His main suggestion is to invert the study of a typical queuing problem and pose such questions as to whether the arrival distribution can be determined from

the queue-length and service-time distributions. Morse's hope is that "...future developments of queuing theory will not just be further embellishments of other special cases, but will also include consolidation of concepts."

The problem of application of queuing theory concepts is indeed a profound one. Alec M. Lee, Director of Operational Research, Air Canada, at the time of publication of his book (27), has perhaps written the best exposition on application of existing queuing theory tools. He believes that a great deal of service improvement can be made through use of existing tools. His argument is that any logical theory is better than none at all as long as one remembers its limitations. From his position as an applications specialist, however, Lee states:

> We need to know more about the real behavior of people in queues: until such time as we have that information, much of the theory that appears in the journals will remain no more than a collection of charming, mathematical acrostics. Mathematics, however ingenious, is not a proper substitute for knowledge.

The following may very well be an illustrative example of the situation about which Lee is speaking. The Association for Computing Machinery recently began publication of a new survey and tutorial journal. The editor saw fit to include in the second publication of this new journal an article (28) based upon the queuing problem of many remote computer users demanding simultaneous access over telephone lines to the central computer facilites. The editor certainly must have believed this subject to have great significance to include it so early in the life of a new journal. However, he himself states (25): "After reading this survey, one cannot fail to be struck by the primitiveness

of all the models which have been studied...". It appeared to the editor that assumptions regarding arrival and service distributions and queue discipline were made merely to simplify the mathematics without due regard to the real-life situation.

As a result of this recent review of the state-of-the-art in queuing theory, one must say that the field of queuing theory could certainly benefit from research resulting in Morse's "consolidation of concepts" and more true experimentation in the scientific sense. It would appear that much work needs to be done in the design and control of experiments (using either real or computer models) in order to test various queuing theory models. Often experimental results suggest theory where little existed before.

SEQUEL REFERENCES

(22) Bartlett, Maurice Stevenson.
An Introduction to Stochastic Processes. Second Edition. Cambridge: Cambridge U.P., 1966.

(23) Beckmann, Petr.
Introduction to Elementary Queuing Theory and Telephone Traffic. Boulder, Colorado: The Golem Press, 1966.

(24) Cox, D. R. and H. D. Miller.
The Theory of Stochastic Processes. New York: John Wiley and Sons, 1965.

(25) Dorn, William S.
Editor's Preview. Computing Surveys 1(2): 81-84. June, 1969.

(26) International Abstracts in Operations Research.
Volumes 6-9. Baltimore: Waverly Press, Inc., 1966-1969.

(27) Lee, Alec M.
Applied Queueing Theory. London: Macmillan, 1966.

(28) McKinney, J. M.
A Survey of Analytical Time-Sharing Models. Computing Surveys 1(2): 105-116. June, 1969.

(29) Morse, Philip.
The Application of Queuing Theory in Operations Research. Queuing Theory: Recent Developments and Applications, introduction. New York: American Elsevier Publ. Co., Inc., 1967.

(30) Prabhu, Narahari Umanath.
Queues and Inventories, a Study of their Stochastic Processes. New York: John Wiley and Sons, 1965.

(31) Riordan, John.
Stochastic Service Systems. New York: John Wiley and Sons, 1962.

(32) Ruiz-Pala, Ernesto, Carlos Ávila-Beloso, and William W. Hines.
Waiting-line Models: An Introduction to their Theory and Application. New York: Reinhold Publ. Corp., 1967.

ACKNOWLEDGEMENT

It is difficult for this writer to say whether the knowledge, the encouragement, the patience, or the sense of humor contributed by Professor S. T. Parker helped the most in the completion of this paper. All of these qualities were presented in the timely fashion of a great teacher. This student wishes to express her deepest thanks.

AN INTRODUCTION TO QUEUING THEORY CONCEPTS

by

JEANNE L. SEBAUGH

B.A., University of Kansas, 1962

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Mathematics

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1970

ABSTRACT

This report serves as an introduction to the concepts of queuing theory. Queuing theory is a study of units arriving at some facility which services these units. The terminology and elements of study in any queuing situation are defined and discussed.

The four basic elements of any queuing situation are the system's input, the waiting lines, the service channels, and the system's output. The system's input is generally described in probabilistic terms with assumptions made as to the size of the source population. The primary waiting-line consideration involves the order in which units are chosen from the line for service. The service channels or servers may vary in number and/or arrangement. The service time itself is also generally described in probabilistic terms.

The system's output may, in the simpler cases, be derived from an analysis of the previous three elements. Various measures of congestion concerning queuing time and queue lengths may be estimated if assumptions of stationarity may be made. In this paper, analytic derivations for measures of congestion are made for a single-server queue with random arrivals and random service. Results are given for the single-server case with random arrivals/general service and random arrivals/Erlang service. Results are also given for the multiserver case with random arrivals/general service and

both a finite and infinite number of servers.

The use of simulation techniques in solving queuing theory problems may be necessary for a complex model. A Monte Carlo approach and the use of a general purpose computer simulation language are discussed.

A sequel has been appended to the main body of the paper. It serves as a recent review of the state-of-the-art in queuing theory and applications.