A GRAPHIC INTRODUCTION TO PROBLEMS IN QUEUEING THEORY FOR ARCHITECTS AND ENGINEERS

By

ARTHUR Y. KISHIYAMA

Bachelor of Science

California State Polytechnic College

San Luis Obispo, California

1963

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
MASTER OF ARCHITECTURAL ENGINEERING
May, 1968

A GRAPHIC INTRODUCTION TO PROBLEMS IN QUEUEING THEORY FOR ARCHITECTS AND ENGINEERS

Thesis Approved:

Thomas Scott Dean
Thesis Adviser

M. Palmer Terrell

N Durham
Dean of the Graduate College

PREFACE

One of the initial courses in my graduate program of study introduced me to a computer language known as GPSS, or General Purpose Systems Simulator, III. GPSS is a problem-oriented language, generally used in the electronic simulation of traffic systems. Using the extraordinary speed of the computer to its utmost advantage, the language simulates operational systems by creating electronic impulses representative of individual traffic units. Within the computer, the electronic units flow through a pattern of logic which has been defined by the analyst.

In structuring the pattern of logic, the model of a real-world system has been constructed. Real-world traffic systems are characterized by provisions for service, delays, points of decision, and transfers. The computer automatically maintains a complete record for each traffic unit as it passes through the system, including its time of origination, arrival, and departure to and from various points in the model. Upon the conclusion of simulation, data from these records provides a complete description of system behavior and performance.

Given the freedom to pursue my interests, I investigated various facets of GPSS by structuring many example problems. I was most impressed by the scope, versatility, and applicability of the language to the variety of traffic systems commonly occurring in architecture. In improvising hypothetical problems, I began to encounter information which would be extremely useful in the architectural design of buildings. From my own experiences and discussions with members of the faculty, many of whom were practicing architects, the conclusion was made that the investigation of architectural traffic systems has long been a neglected area of applications.

In continuing my studies with GPSS, I began to realize its limitations. GPSS is complex, demanding a great deal of programming experience, both in using the language and evaluating its results. Few practicing architects have access to the GPSS language or the equipment on which it is implemented. Finally, resultant inference of probable system behavior applies only to the model simulated so that no general solutions are available. Hence, a means of system analysis independent of the computer and surmounting these difficulties was necessary if investigations of traffic systems were to be practical.

In expanding my research, I found that the theories of probability have been extensively applied to traffic systems in the area broadly referred to as queueing theory. While GPSS gives a description of system performance by physical simulation, queueing theory accomplishes the same task mathematically. The great advantage of mathematical analysis was its use to achieve solutions which were general in nature.

In some respects, the mathematical approach to traffic analyses is even more difficult than simulation. The derivation of queueing parameters requires concepts well beyond the capabilities of most architects. Thus, the ultimate goal of this study consisted of an effort to present queueing theory in terms of easily understood variables and in a manner that would facilitate their application. In order to bypass the difficulties of mathematics and to provide a rapid means of investigation, a family of relationships descriptive of system performance have been presented in graphic form.

To accomplish these goals, over 50 separate computer programs were written for the IBM 7040 and 1620 computers. All of the illustrative material, some of which required several thousand calculations, were executed on plotting equipment attached to the 1620. Direct cross referencing of the subject material has been limited since only the most fundamental concepts of queueing theory have been discussed. Most texts dealing with basic operations research or queueing theory, including all of those in the bibliography, contain a detailed presentation of these concepts. Two of the included references, Queues, Inventories, and Maintenance by Philip M. Morse and Waiting-Line Models by Ernesto Ruiz-Pala, Carlos Avila-Beloso, and William W. Hines, were used extensively.

I would like to take this opportunity to acknowledge the efforts of Dr. Thomas S. Dean of the School of Architecture and Dr. Palmer M. Terrell of the Industrial Engineering Department. They provided me with encouragement, constructive criticism, and the inter-disciplinary guidance required for the completion of this study. In addition, the efforts of Mrs. Nancy Wolfe should be recognized for her excellent preparation of the manuscript.

As an active member of the United States Air Force, special acknowledgement is reserved for the Air Force Institute of Technology, to which I have been attached during my course of study. I hope that upon return to duty, I am able to fulfill the objectives of my assignment with a contribution of professionalism to military engineering.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Architectural Traffic Systems

A "system" may be defined as "an assemblage of objects united by some form of regular interaction or interdependence; an organic organized whole."[1] The complex buildings of today's technology typify Webster's definition of a system. In the modern passenger terminal, the assemblage of objects may be thought of as a collection of sub-systems. These sub-systems may consist of arterial roads, parking lots, ticket counters, restaurants, lobbies, baggage stands, and loading aprons. Interaction and interdependence within the system are characterized by traffic flows within and between sub-systems. Traffic flows may be composed of vehicles, people, baggage, communications, products, or innumerable like units.

Two distinct phases are common to traffic within any architectural system. First, the "arrival" of units to the system for "service"; and second, the actual performance of service, upon completion of which the unit is discharged. The arriving unit may receive immediate service if it is available. Otherwise, the unit must join a queue or waiting line and be delayed in its passage through the system.

[1] Webster's New Collegiate Dictionary, Second Edition (Springfield, Mass., 1960).

In the design of systems, the architect must insure that service capacities are capable of handling arrival traffic so that unreasonable waiting periods do not occur. If the system is not designed so that the service capacity is at least as large as arrival traffic, a waiting line will build until traffic is reduced or capacity is increased. Even with average capacities sufficient to handle average arrivals, temporary or even permanent congestion may occur because of fluctuation in the actual rates of service and arrivals. For example, the interval between arrivals and time required for service may be expected to vary from unit to unit. A series of short arrival intervals coupled with requirements for lengthy service will generally result in waiting. The variation of arrival and service rates is a measurable numeric quantity and must be recognized as inherent to any system of traffic.

In very simple terms, a description has been given of queueing, congestion, or waiting line problems, Queueing problems abound in architectural design. Buildings are not inanimate objects, but dynamic systems of traffic flow in which queueing situations are the rule rather than the exception. As the complexity of architectural structures increase, there is a resultant increase in the number of causes for waiting. As waiting increases, the necessity for the architect to satisfy service demands also increases. Traffic flows must be handled efficiently. Ticket and baggage counters must have sufficient "service channels" to avoid unreasonable queue lengths and waiting periods. Restaurants and parking lots must provide enough space so that service is available immediately or with very short delays.

It is the responsibility of the architect to evaluate the demand, establish the appropriate level of service, estimate the various costs associated with the satisfaction of demand, and determine the optimum level for system capacity. Unfortunately, he is not well-equipped to handle detailed analyses of these situations. His primary tools are past experience, judgement, and many questionable rules of thumb.

The architect would be better prepared in the analyses of traffic systems if he could answer any one of the following questions. What is the expected number of units in the system and in the queue at any given time? How many service channels are required so that waiting does not exceed a predetermined amount of time? How long will an arrival unit have to wait before service is performed and completed? What proportion of time will instantaneous service be available? What proportion of time will more than a given number of units be in the system? And, what is the "efficiency" of the system? Queueing theory, a recent development associated with the telephone industry, offers an approach to the evaluation of these questions.

Applications of Queueing Theory

In the design of automatic telephone exchanges, necessary information includes the effect of service demand fluctuations as varying numbers of customers dial different numbers. Most of the pioneering work in queueing theory is attributed to A. K. Erlang, a European electrical engineer. Beginning about 1905 and up to 15 years ago, most study on the theory of queues was accomplished by Erlang and others in connection with telephone problems. With the advent of post-World War II "operations research," this theory has been extended to other fields involving operational problems. Industrial applications of queueing theory in production and maintenance are now widely accepted. Particularly in highway design and airline operations, progress in queueing theory has been significant in many areas of transportation.

Arrival intervals and service times characterize the queueing system, each following a measurable probability distribution. A probabilistic "model" of the system may be mathematically constructed. The system may exist in any number of possible states as specified by the number of units in the system - those waiting for service, and if any, those that are in service. Based upon the laws of probability, the probability that the system is in each of its possible states may be computed. From these "state probabilities," numeric relationships involving system parameters may be derived.

System parameters may be considered as "measures of effectiveness" relating in numeric terms the long-run behavior of the system over time. Measures of effectiveness include any relationship which may be objectively or subjectively evaluated in determining the adequacy of system performance. As most queueing systems involving the architect will deal with the flow of people, the ultimate evaluation will generally be subjective.

Most published material on queueing theory adheres to a rigid mathematical discipline to the extent that several models have been studied more for their mathematical interest than potential applications. These publications require that the reader have at least a fundamental knowledge of statistics, probability theory, and advanced mathematics. The derivations of state probabilities are complex, tedious, and difficult for an individual unfamiliar with principles in these areas. Unfortunately, these basic prerequisites generally open a wide gap between the architect and the analysis of queueing problems.

Proposed Method of Study

The purpose of this paper is to bridge the gap between operations research and the architect. The primary goal is to provide the architect with a usable set of relationships, graphically presented, with which to gain an intuitive insight to the workings of elementary queueing systems. Instead of deriving complex state probabilities, a qualitative approach to queueing theory is taken. The presentation of measures of effectiveness is limited to a brief introduction followed by the relationship itself. In most cases, the derivations for these measures require their entire presentation for clarity. Any attempt to present a brief or abridged derivation would lead to confusion and merely clutter the objective of the text material. Almost all relationships have been graphically illustrated so that the behavior of the system under several conditions may be evaluated without repeating long, mathematical computations.

The first four chapters introduce the statistical, probabilistic, and structural concepts of queueing theory. The remaining chapters are devoted to the application of these concepts in the evaluation of system performance. Chapter II, an introduction to the fundamentals of statistics and probability, is included to provide the reader with a foundation on which to evaluate the cause and effects of random variation. These fundamentals are applied in Chapter III to the development of probability distributions for service and arrival times. The basic elements of queueing models are structured and defined in the form of an organization chart in Chapter IV.

The total number of possible queueing situations, and therefore models, approaches infinity in the real-world. The organization chart of Chapter IV defines system elements which have received mathematical attention and allows the construction of over 22,000 different models. Obviously, this study will consider a small number of these models. The few models selected have been chosen for their potential architectural applications and should adequately represent a great many queueing problems encountered in the design of real-world systems. Applicable assumptions to these models are clearly stated with particular attention to their proper use.

The material presented in the first four chapters will suggest a necessity for the collection, enumeration, and analysis of large amounts of data to determine the mean and distribution of arrival and service times. Because of cost, time, and insufficient personnel or capability, it is anticipated that many architects could not follow the procedures outlined. This does not negate the potential usefulness of this study, as most often rather simple assumptions and estimates may be made. In most cases, the architect will be able to reasonably estimate the average arrival and service rate. By making the additional assumption that arrivals and service follow specific distributions, many of which are quite valid for a large majority of architectural problems, the most important structural elements of a queueing system are defined.

"Traffic intensity" or the ratio of arrival to service rates is the primary variable in almost all measures of effectiveness. In many instances, it will be useful to make an optimistic and pessimistic estimate of traffic intensity. The model investigated may then be studied in terms of numeric intervals in which actual measures of effectiveness are likely to occur. Models developed in this study are presented in

Chapters V and VI. The objective of these chapters is to apply queueing theory to real-world problems as demonstrated through the use of several example problems.

CHAPTER II

FUNDAMENTALS OF STATISTICS AND PROBABILITY

It has been indicated that there are two distinct phases common to any queueing system. First, the arrival of units to the system; and second, the performance of service. Both the arrival and service rates may be expressed as numeric values in units of time. Estimates for these values are usually based upon data collected through observation.

Differences between one queueing system or another, ignoring their physical or theoretical structure, are most obvious in their differences in arrival or service rates. A system servicing 20 units per hour with 15 arrivals per hour will behave quite differently from a system servicing 20 units per hour but with 20 arrivals per hour. They are two distinctly different systems.

In their raw form, the collected data from which rates are derived usually communicate very little information. The body of analytical techniques directed to the description of collected data is called "descriptive statistics." It is the purpose of descriptive statistics to place raw data into a usable, compact form. The most commonly employed methods to accomplish this task are calculations of measures for "central tendency," measures for "dispersion," and the development of a "frequency distribution."

Before discussing the concepts of descriptive statistics, a particular characteristic of queueing systems should be understood. Most queueing systems involve the random occurrences of chance events. That is, in any interval of time, an arrival, service completion, or change in queue length may or may not have occurred as the result of chance. At any instantaneous point in time, these same events may or may not be about to occur as the result of chance. Because of random chance, variation is an inherent characteristic of most queueing systems. The concepts of random chance will be further discussed under the topic of probability. For the present, consider only that variation is also a measurable numeric quantity and in descriptive statistics is analyzed as a measure of dispersion.

Attempts to deal with queueing problems without acknowledgement of variation are usually made by providing for at least as many service completions as arrivals per unit time. For example, if there are always 20 arrivals per hour, at least 20 service completions per hour must always occur to prevent permanent congestion of the system. However, the inherent variation in arrival or service rates may cause temporary or even permanent congestion since in any particular hour, there could be 30 arrivals and only 10 service completions.

In the long run covering several one hour intervals, 20 arrivals or service completions are expected. In any particular one hour interval, the actual number of arrivals or service completions are subject to fluctuation. In this case, an engineer might attempt to increase the average service rate to 25 or 30 service completions per hour. As justification for this procedure, he might state that he is "allowing for a margin of error." In reality, he is not allowing for error but for natural variation due to fluctuation in the arrival and service rates of the system.

It is the purpose of queueing theory, based upon concepts of statistics and probability, to account for and recognize variation as an inherent characteristic. Once recognized, inference towards the predicted behavior of the system may be made with a greater degree of assurance.

Measures for Central Tendency and Dispersion

Several methods are available to describe the central tendency of a collection of data. The "median" is defined as that value lying in the middle of an ordered set of data. An ordered set is one which has been ranked from the smallest to largest value or largest to smallest value. The "mode" is defined as that value which occurs most frequently in a set. The most common and widely used measure for central tendency is the "arithmetic mean" or "average." The mean is that point about which all values of a set of numeric data tend to cluster. It may be expressed as:

$$\bar{\bar{X}} = \sum_{i=1}^{n} X_i/n = (X_1 + X_2 + X_3 + . . . + X_n)/n \qquad (2.1)$$

By analogy, the mean is identical in concept to the centroid or center of gravity in structural mechanics.

Consider the three sets of numeric data shown in Table I. Using the mean as an analytical tool, note that the summation of individual elements is 100 in each set, making the means equal to 100/5 or 20. By inspection, it is evident that the three sets of data are different so that the mean alone does not provide a unique description. In addition to a measure for central tendency, another method must be employed to make a numeric differentiation between sets.

TABLE I

THREE SETS OF NUMERIC DATA

X_i	SET A	SET B	SET C
$i = 1$	20	10	10
2	20	15	10
3	20	20	20
4	20	25	30
5	20	30	30
$\sum_{i=1}^{n} X_i$	100	100	100
$\bar{X} = \sum_{i=1}^{n} X_i/n$	20	20	20
$\sigma^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$	0	50	80

For this purpose, two measures of dispersion may be employed. The "range" is the least complex and is obtained by calculating the arithmetic difference between the largest and smallest value of the set. However, note that both Sets B and C of Table I have identical ranges of (30 - 10) or 20. While the ease of computing the range is a great advantage, a unique description of the set is not available since only two values of the set are utilized.

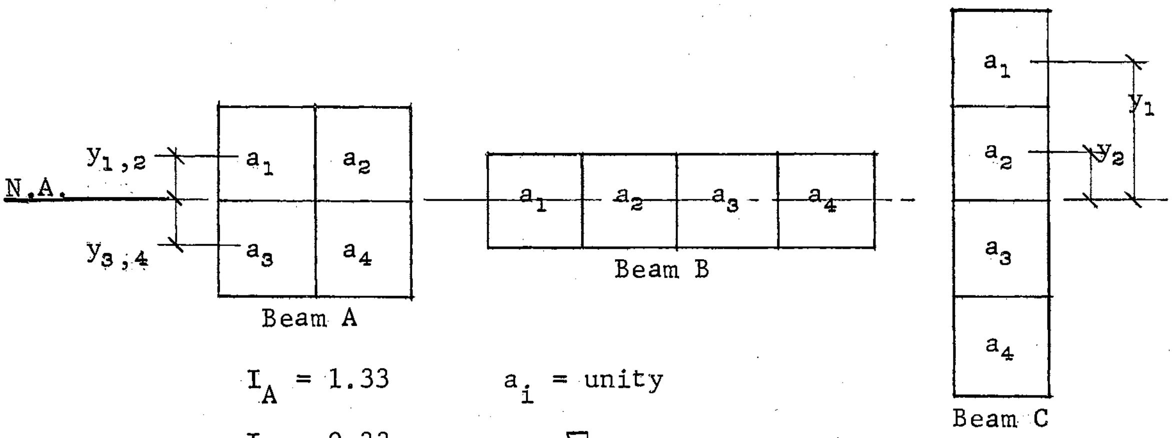
The "variance" is a much better measure of dispersion as it utilizes each element of the set. The variance may be expressed as:

$$\sigma^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n \qquad (2.2)$$

$$\sigma = \sqrt{\sigma^2} = \text{standard deviation}$$

The square-root of the variance is defined as the "standard deviation" or "root-mean-square-deviation." The variance for each of the three sets of data are shown in Table I. Note that the greater the variance or dispersion of individual elements about the mean, the greater the numeric value of the variance or standard deviation.

Engineers will recognize that the standard deviation is identical in concept to that of radius of gyration (r). In Figure 1, three timber beams of equal area (A) are shown. It should be obvious that the load capacity of the three beams are not equal. The load capacity is a



$I_A = 1.33$

$I_B = 0.33$

$I_C = 5.33$

$a_i = \text{unity}$

$$A = \sum a_i = n$$

$$I = \sum_{i=1}^{n} a_i y_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$r = \sqrt{(I/A)} = \sigma$$

Figure 1. Three Timber Beams

function of its moment of inertia (I) which in turn is a function of to what extent the area of the beam is distributed about its centroidal axis. The approximate moment of inertia is the summation of elemental areas (a_i), each multiplied by the square of the distance from the reference axis to its centroid (y_i^2). Radius of gyration is defined as $\sqrt{(I/A)}$. Considering each elemental area (a_i) as unity, y_i^2 is equivalent to $(X_i - \bar{X})$ and A is equivalent to n. The analogy should be clear. The greater the distribution of elemental areas about a neutral axis, the greater the numeric value of moment of inertia and radius of gyration. The greater the distribution of individual values in a set of numeric data, the greater the numeric value of the variance and standard deviation. As the description of a beam is not unique by its total area and location of its neutral axis, the description of a set of numeric data by average alone is insufficient. Some measure of dispersion, the extent to which individual values are distributed about the mean, is always required.

Frequency Distributions

When summarizing large masses of raw data, it is often useful to distribute the data into classes or categories and to determine the number of individuals belonging to each. A tabular arrangement of data by classes together with the corresponding class frequencies is called a "frequency distribution" or "frequency table."

Table II represents 30 items of raw data. It may be assumed that they represent the time required for service, in minutes, on randomly selected arriving units. Calculations are shown for the mean and standard deviation. It its present form, the data in Table II conveys

TABLE II

SERVICE TIMES FOR 30 RANDOM UNITS

Unit	Service Time	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	31	6	36
2	18	- 7	49
3	30	5	25
4	25	0	0
5	42	17	289
6	8	-17	289
7	32	7	49
8	25	0	0
9	27	2	4
10	14	-11	121
11	12	-13	169
12	44	19	361
13	34	9	81
14	31	6	36
15	52	27	729
16	28	3	9
17	21	- 4	16
18	31	6	36
19	9	-16	256
20	24	1	1
21	13	-12	144
22	11	-14	196
23	16	- 9	81
24	25	0	0
25	21	- 4	16
26	23	- 2	4
27	40	15	225
28	17	- 8	64
29	12	-13	169
30	34	9	81
	750		3536

$$\bar{X} = \sum_{i=1}^{n} X_i/n = 750/n = 25.0$$

$$\sigma^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n = 3536/30 - 117.8667$$

$$\sigma = \sqrt{117.8667} = 10.85$$

little information. A grouped frequency table is developed in Table III to express the same data in a more compact form. Six classes have been arbitrarily defined, each of 10 minute intervals. The number of times an observed service time falls within each class interval is tabulated and the tabulated frequencies are plotted in the form of a frequency histogram as shown in Figure 2.

TABLE III

FREQUENCY TABLE

Class	Interval	Freq	Fraction	Rel Freq	Cumul Freq
1	0-10	2	2/30	.0667	.0667
2	11-20	8	8/30	.2667	.3334
3	21-30	10	10/30	.3333	.6667
4	31-40	7	7/30	.2333	.9000
5	41-50	2	2/30	.0667	.9667
6	51-60	1	1/30	.0333	1.0000
		30	30/30	1.0000	

The relative frequency of a set of data is simply the frequency divided by the total number of observations. In Figure 3, the ordinate of Figure 2 has been changed from absolute to relative frequency by dividing by 30 so that the area under the relative frequency distribution is equal to unity. Relative frequency may be accumulated and plotted as a cumulative distribution. Each cell interval then represents the cumulative relative frequency up to and including that interval. The

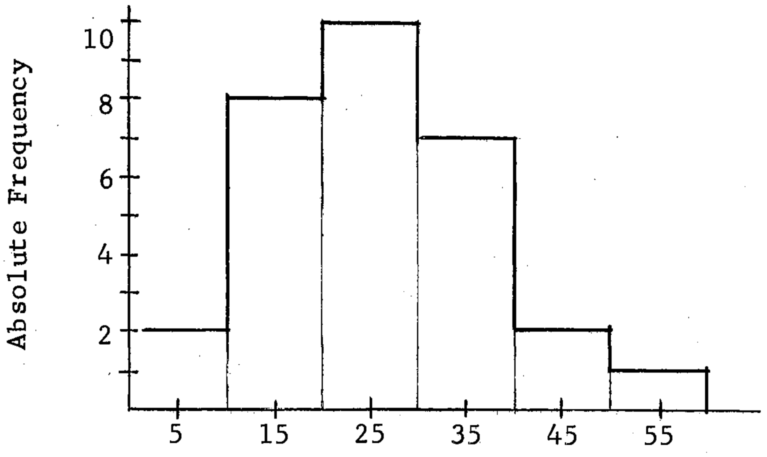


Figure 2. Frequency Diagram

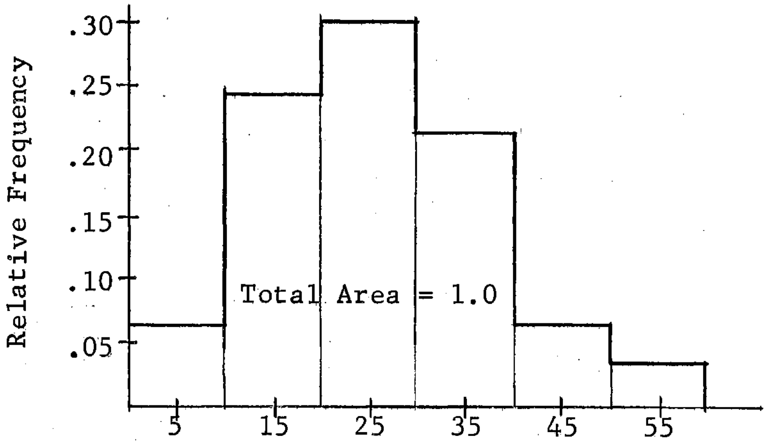


Figure 3. Relative Frequency

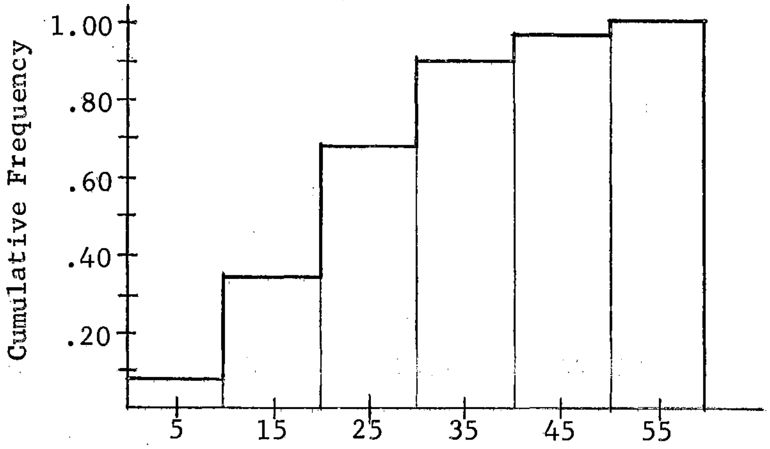


Figure 4. Cumulative Frequency

intervals will then have the appearance of a series of uneven ascending steps terminating at unity as shown in Figure 4.

Probability and Probability Distributions

Probability may be defined as either a measure of certainty or uncertainty. It provides a means for mathematically expressing a degree of assurance or doubt. As a concept, probability may be used to describe the outcome of a random event. An event with a probability of unity is certain to occur while an event with a probability of zero is certain not to occur.

There are two traditional definitions of probability. The "classical definition" has a mathematical origin and may be expressed as follows. If an event may happen in A ways and fail to happen in B ways, and all of these ways are mutually exclusive[1] and equally likely to occur, the probability of the event happening is A/(A+B), the number of ways favorable to the event divided by the total number of possible ways.

Suppose that the physical limitations to a space are limited so that a queue may never contain more than three persons. The possible lengths of the queue are therefore 0, 1, 2, or 3 persons. Furthermore, assume that these queue lengths are mutually exclusive and equally likely to occur. The probability of exactly 0, 1, 2, or 3 persons in the queue are 1/(1+3) or 1/4. The probability of 1 or less persons in the queue is 2/(2+2) or 2/4; of 2 or less 3/4; of 3 or less 4/4.

[1] Events are mutually exclusive if the occurrence of any one of them makes impossible the simultaneous occurrence of any of the others.

Conversely, the probability of 0 or more, 1 or more, 2 or more, and 3 or more persons in the queue are 4/4, 3/4, 2/4, and 1/4 respectively. Note that in the cases of 3 or less or 0 or more, the probability of occurance is equal to unity as the possibility of all events are included.

The classical definition above is necessary for mathematical manipulations involving probability statements. However, from the standpoint of useful applications, probability may be thought of as relative frequency in the long run. This is the "empirical or relative frequency" definition which may be stated as follows. If a large number of trials are made under the same conditions, the number of trials in which a certain event happens divided by the total number of trials will approach a limit as the total number of trials is increased indefinitely. This limit is called the probability that the event will happen under the same conditions.

The word "limit" in the frequency definition of probability is not used in its conventional mathematical sense. That is, a function does not asymptotically approach a limit as some variable increases or decreases. It is rather a "statistical or stochastic limit" which is continually fluctuating as additional trials are made. As the number of trials are increased indefinitely, the degree of fluctuation decreases to the extent that a relatively constant value is approached.

Table II represents a mass of service time data for a hypothetical facility. Table III and Figure 3, the relative frequency table and relative frequency diagram were developed from this data. Assuming that the service times are valid in their representation, Figure 3 may be considered a probability distribution from which future service times may be predicted. The probability of a service time falling in the interval 21-30 minutes is .3333. Similarly, from Figure 4, the probability of service time being less than 30 minutes is .6667.

Probability distributions should be differentiated from frequency distributions. The frequency distribution describes what has occured in the past while the probability distribution predicts what is expected to occur in the future. Probability distributions provide a means for assigning the likelihood of occurrence to all possible events. Variables described in terms of probability distributions are called "random variables." The specific value of a random variable is determined by the distribution and the occurrence of that value is governed by the associated probability.

Probability distributions may be either discrete or continuous, depending upon the nature of the event they are used to predict. If used to predict the number of persons in a queue, the distribution would be discrete. If used to predict service times, the distributions would be discrete over the interval of times selected. However, as the intervals are made small, the distribution will approach a continuous function. Continuous functions are often used to approximate discrete functions so that integration can be applied. Likewise, discrete functions are used to approximate continuous functions, particularly in applications using a digital computer, so that a summation process will perform the required integration.

Many mathematically derived probability distributions have been developed which closely approximate the occurrence of random events in real-life situations. The most important of all distributions is the normal or Gaussian probability distribution. It is defined as:

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \quad (-\infty \leq x \leq \infty) \qquad (2.3)$$

The mean and variance are μ and σ^2 respectively. The normal distribution is symmetric about the mean as a bell-shaped curve. The amount of humping, or steepness of the curve, about the mean is a function of its variance. The smaller the variance, the steeper the curve.

The normal distribution possesses several useful properties with regard to its shape. Where distances from the mean are expressed in terms of standard deviation, σ, the relative area defined between two such distances will be constant from one distribution to another. All normal distributions, when defined in terms of a common μ and σ, will be identical in form and corresponding probabilities may be tabulated. As the normal distribution describes every possible event, the total area under the curve is equal to unity. The cumulative probabilities from $-\infty$ to any value expressed in standard deviation units are given in Table IV.[a] The table gives the probability of a value falling within the range $-\infty$ to Z, where Z is a standard normal variate defined as:

$$Z = (x - \mu)/\sigma \qquad (2.4)$$

If the service time data of Table II were assumed normally distributed, the probability that service on a randomly selected unit would be less than 10.0 minutes would be computed as follows.

$$Z = (10.0 - 25.0)/10.85 = -1.382\sigma$$

[a] W. J. Fabrycky and Paul E. Torgersen, Operations Economy, (New Jersey, 1966), pp. 452-453.

From Table IV,

$$P(-\infty \text{ to } -1.382) = .0835$$

Similarly, the probability that service on a randomly selected unit would be less than 40.0 minutes would be computed as follows.

$$Z = 40.0 - 25.0/10.85 = +1.382\sigma$$

From Table IV,

$$P(-\infty \text{ to } +1.382) = .9165$$

The probability that service time on a randomly selected unit fell within the interval 10.0 to 40.0 minutes would be .9165 - .0835 = .8330. These calculations have been graphically portrayed in Figure 5.

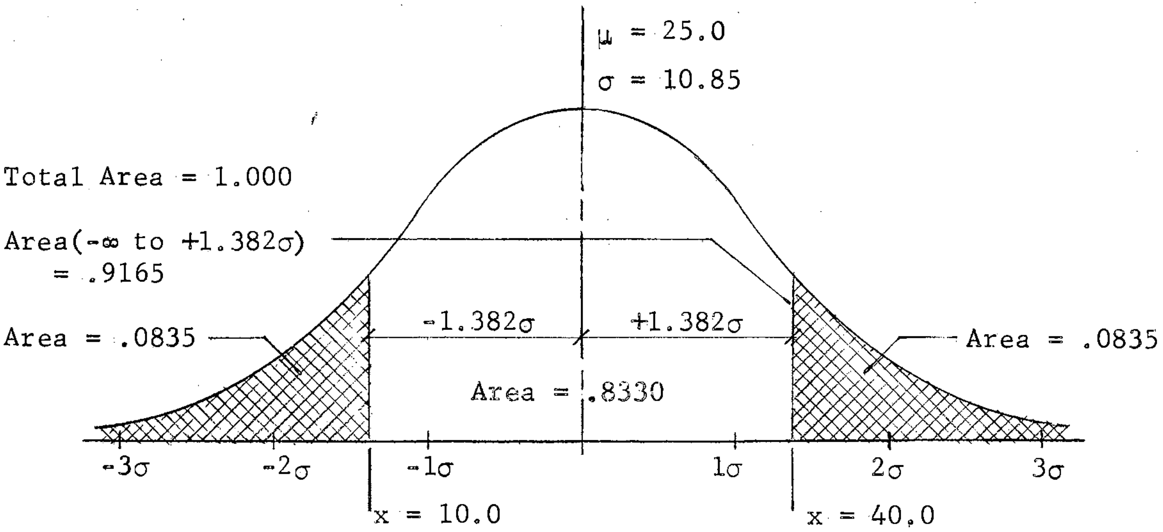


Figure 5. Service Times as a Normal Distribution

TABLE IV

CUMULATIVE NORMAL PROBABILITIES

Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.5	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023	+0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
−3.4	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00033	0.00034	+0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
−3.3	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048	+0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
−3.2	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069	+0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
−3.1	0.00071	0.00074	0.00076	0.00079	0.00082	0.00085	0.00087	0.00090	0.00094	0.00097	+0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
−3.0	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135	+0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019	+0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	+0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	+0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8079	0.8106	0.8133
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	+0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	+1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
−2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	+1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	+1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	+1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	+1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	+1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	+1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	+1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	+1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	+1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	+2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
−1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	+2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	+2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
−1.2	0.0985	0.1003	0.1020	0.1038	0.1057	0.1075	0.1093	0.1112	0.1131	0.1151	+2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	+2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	+2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	+2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	+2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
−0.7	0.2148	0.2177	0.2207	0.2236	0.2266	0.2297	0.2327	0.2358	0.2389	0.2420	+2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
−0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	+2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	+3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	+3.1	0.99903	0.99906	0.99910	0.99913	0.99915	0.99918	0.99921	0.99924	0.99926	0.99929
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	+3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	+3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	+3.4	0.99966	0.99967	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	+3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

Two particular probability distributions of great importance to queueing theory are the exponential and Poisson distribution. The number of arrivals and service completions per unit time for most models illustrated in this study are assumed to follow a Poisson distribution. Most of the service time and all of the arrival interval distributions are assumed to be exponential. Because of the significance of these assumptions, the exponential and Poisson distributions are introduced in Chapter III.

Inferential Statistics and Sampling

Thus far, masses of raw data have been described by calculating measures for central tendency and dispersion. It was stated that the data was collected through observation and was representative of a characteristic of interest. In Table II, 30 items of data were introduced as individual measurements of service time for a hypothetical facility. From this data, a frequency distribution was developed from which inference was made towards the future. An alternative inference was made by assuming that the service times measured followed a normal distribution. To the extent that the data in Table II is representative of the true behavior of the facility over time, these inferences may be correct. The body of statistics that deals with the formulation of inferences or conclusions from raw data is called "inferential statistics."

A "population" consists of all possible objects, states, or events within an arbitrarily defined boundary and may be either finite or infinite. The population of service times for the hypothetical facility above consists of the time required to service every arriving unit throughout the life-time of the facility. The total number of service completions accomplished by the facility is therefore a very large, finite population. It is usual to consider very large populations as infinite in size for computational purposes.

Large or infinite populations are a characteristic in most architectural queueing problems. For example, the population of arriving automobiles to a shopping center parking lot consists of every automobile in the surrounding state, county, city, or neighborhood, depending on where the arbitrary boundary is defined. Similarly, customers at a supermarket, passengers arriving at a terminal, or a life-time of service times for a facility comprise extremely large populations. From these large populations, a means by which arrival and service rates may be obtained with a degree of assurance is necessary.

Complete enumeration of each element from a very large population is generally impractical or uneconomical. In many cases, it may be inaccessible as a whole. For these reasons, a "sample" is drawn from the population. A sample is simply a part or portion of the total population and is usually assumed as typical of the population, at least in regard to the parameter under consideration. Samples taken must be selected at "random" where randomness implies that each element of the population has an equal chance for selection.

Properties of a population as the mean or variance are termed "parameters" while properties of samples are termed "statistics." The distinction is important as a population has only one mean or variance. Samples may have different means and variances as each sample is composed of different, randomly selected elements of the population. Through sampling, the limited observation of a population, inferential statistics attempts to estimate population parameters.

The means of samples taken from a population in turn from another sub-population of sample means. The mean and variance of sample means may be calculated as any other set of numeric data. The mean of sample means is considered the best estimate of the population mean. The true population mean will be approached as a statistical limit as samples are taken indefinitely. The variance of sample means is related to the variance of the populations by the relationship:

$$\sigma^2_{\bar{x}} = \sigma^2/n \qquad (2.5)$$

where n is the sample size. Equation (2.5) holds true for any population, regardless of its distribution. Recall that if the population is normal and its mean and variance are known, the population is fully described. However, populations in real-life situations will very often follow a distribution other than normal.

The "central limit theorem" is a mathematical proof which states that the population of sample means will approach normality as the size of the sample and number of samples taken tends towards infinity. It has been demonstrated that sample means taken from any population, regardless of its distribution, will approach normality even with sample sizes as small as four.

Once again, consider the data in Table II. Assume that each of the values represent the mean of a sample size four so that 120 observations have been made. The population mean is therefore estimated as the mean of sample means or 25.0. The standard deviation of sample means, $\sigma_{\bar{x}}$, is 10.85. From this information, it is desired to calculate the .95 "confidence interval." The confidence interval is defined by limits between which the stated proportion of observations will be expected to fall.

Table IV may be used since, according to the central limit theorem, the sample means are expected to approach a normal distribution. Since the normal distribution is symmetric, enter the body of Table IV with the probabilities .025 and .975 which yield Z values of ± 1.96. From Equation (2.4), $X = \mu \pm 1.96\sigma_{\bar{X}} = 25.0 \pm (1.96)(10.85) = 3.73$ and 46.27. The preceding calculations allow the following statement. If sampling is continued from the same population, in the long run 95 out of 100 sample means may be expected to fall between 3.73 and 46.27.

From Equation (2.5), the variance of the population is equal to the variance of sample means multiplied by the sample size. Taking the square-root of both sides of the equation, the standard deviation of the population, σ, is seen to equal twice the standard deviation of sample means, $2(\sigma_{\bar{X}})$, with a sample size four. Table IV may be used to infer any probability of interest concerning the population only if the population is assumed normally distributed. It should be intuitively clear that the smaller the variance of the population, the smaller the interval for a stated percentage of confidence. The calculations above are illustrated in Figure 6.

Summary of Statistics and Probability

In the analysis of any queueing system, the arrival and service rates must be known or estimated. These rates are generally derived from very large populations making the techniques of random sampling a necessary procedure. The result of observing every possible element of a population would be a specific distribution of numeric values. A complete description of a population is available when the distribution of individual elements is known. The distribution of elements within a

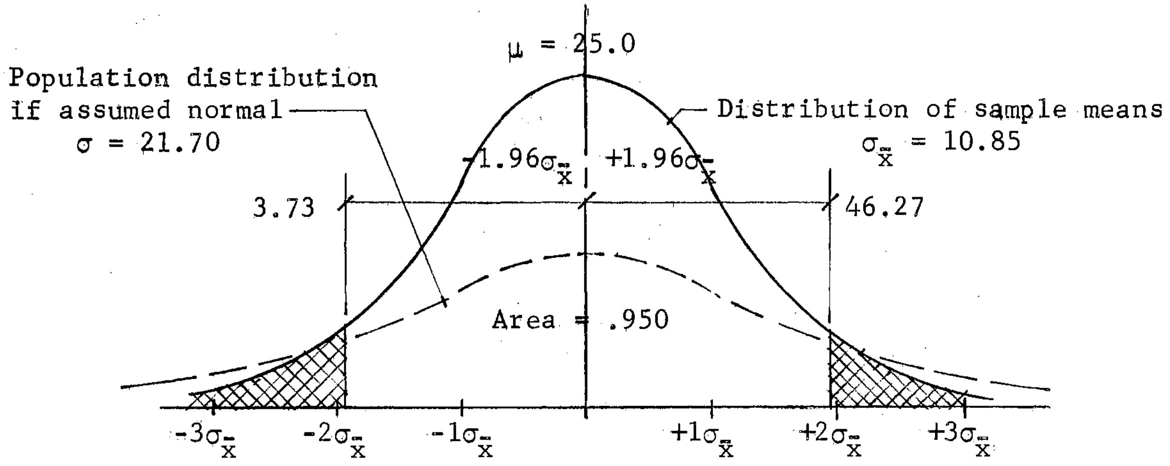


Figure 6. A Confidence Interval for Sample Means

population may be numerically described by calculating the mean and variance or standard deviation. Without recognition of variation within the arrival and service rate populations, proper inference about the predicted behavior of a queueing system cannot be made with assurance. The degree of assurance associated with any inferential statement is expressed in terms of a probability. Probability statements give the proportion of times a particular event may be expected to occur or not occur as repeated opportunity for occurrence is extended indefinitely.

Even a superficial study of statistics and probability would entail a complete volume of work. Hundreds of books and a countless number of technical papers and articles have been published in these subject areas. The purpose of this chapter has been limited to the introduction of the language of statistics and probability as they broadly apply to queueing theory. A rudimentary knowledge of the concepts involved will facilitate understanding of topics introduced in subsequent chapters.

CHAPTER III

PROBABILITY DISTRIBUTIONS OF SERVICE AND ARRIVAL TIMES

The rate at which units arrive and are serviced in the queueing model has been only briefly discussed. Further study requires a more definitive approach. It should be understood, however, that units arrive for service in a more or less irregular pattern with service performance subject to random variability. In subsequent chapters, arrival and service times will be assumed as independent random variables with probability distributions having known form and parameters. Upon this assumption, the probability distribution defines a population of arrival and service times for all units consecutively entering the system.

Distribution of Service Times

Service time is simply the amount of time that has passed from the beginning of service to its completion. In the case of a sales counter, service begins when the "customer" arrives at the head of the line, if one exists, and received attention from the clerk. Service ends when all transactions have been completed and the customer moves away from the counter. The mean service time will be represented by T_s while its reciprocal, $1/T_s$ will be represented by μ and defined as the service rate. The service rate is seen to represent the mean number of service completions per unit time.

Once the service times of a considerable number of customers are obtained, preferably 200 or more, they should be plotted on a time base to establish their stability. Tests for stability indicate whether observed service times have been taken from the same population. They are particularly important in architectural applications as rush-hour peaks and off-hour lulls are typical in many building types.

The use of "control chart" techniques[1] as applied to industrial quality control are based upon the central limit theorem and the normal distribution introduced in Chapter II. Recall that sample means taken from a population of any distribution may be expected to approach a normal distribution. The mean and standard deviation of sample means therfore define a particular normal distribution or population. From Table IV, it is seen that the interval ± $3\sigma_{\bar{x}}$ define boundaries within 99.7% of the sample means may be expected to fall. The inclusive range of this interval leads to the conclusion that sample means falling outside of the boundaries have been drawn from a different population.

Experience has shown that population parameters may be reasonable estimated after 20 samples have been drawn. Once established, these parameters may be used to graphically construct "upper and lower control chart limits" corresponding to the ± $3\sigma_{\bar{x}}$ limits of the normal curve. A central line, midway between the two, corresponds to the mean of sample means. Values of the sample means may then be plotted as a function of time by maintaining the order of sampling. As long as the plotted values remain within the control chart limits, there is reasonable

[1] Eugene L. Grant, Statistical Quality Control (New York, 1964), pp. 65-90.

assurance that the sampling process is continuing from the same population. Prolonged runs above or below the central line indicate that a shift in population mean has occurred. Points falling outside both control limits indicate a change in the population dispersion about the mean. In either case, a new population of service times exist which differs from the original population whose parameters were originally estimated.

Tests for independence are made to insure that the service time for any particular unit does not depend in any way upon the service time of the preceding unit or units. Several methods are available for such tests but require statistical concepts beyond those introduced in this study. Tests for independence make up a significant portion of many texts on statistics, some of which are listed in the bibliography. It is a reasonable assumption, however, that the types of service times encountered in architectural queueing models are independent. Service time populations are generally very large and center about the activities of people. The service time required of one arriving individual will usually not depend upon the service time required of preceding individuals. The preceding comments pertaining to stability and independence also apply to arrival intervals.

Any logical procedure may be used to construct a frequency diagram so that the sequence of service times in Table II may be graphed in order of decreasing length. The resultant plot is the cumulative number of service operations that take longer than a given time (t) as illustrated in Figure 7. By dividing the ordinate by the total number of service times, a scale that represents an estimate of the probability, $S_o(t)$, that a service operation will take longer than time t will be

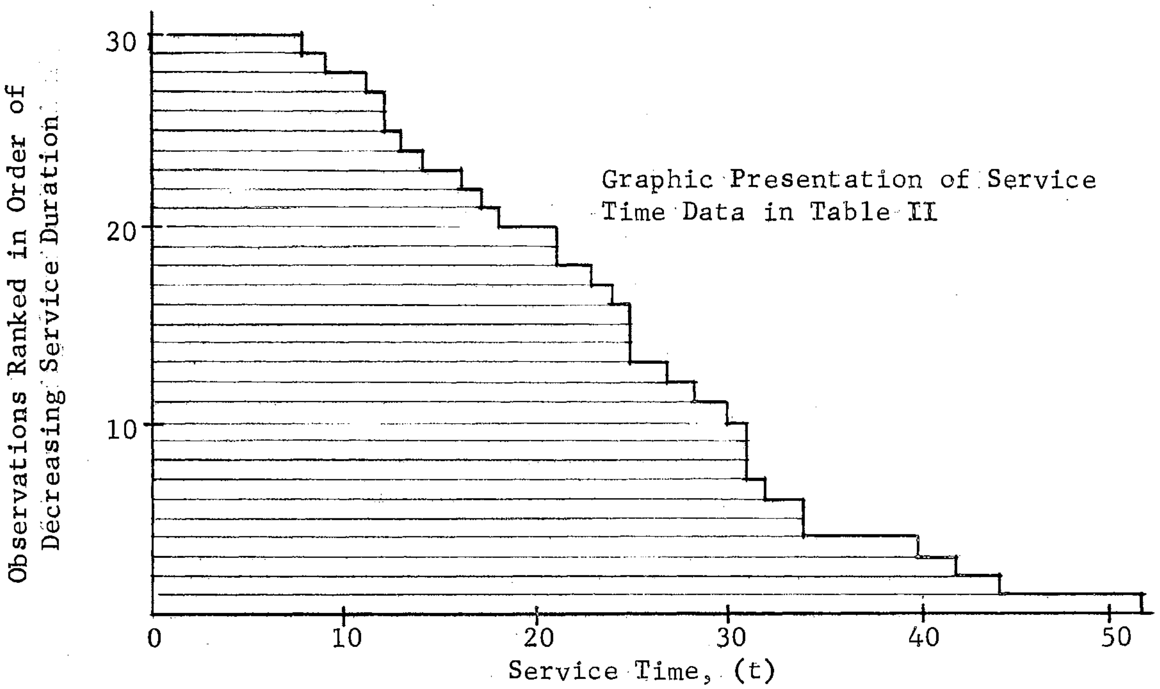


Figure 7. Cumulative Service Time Data From Table II

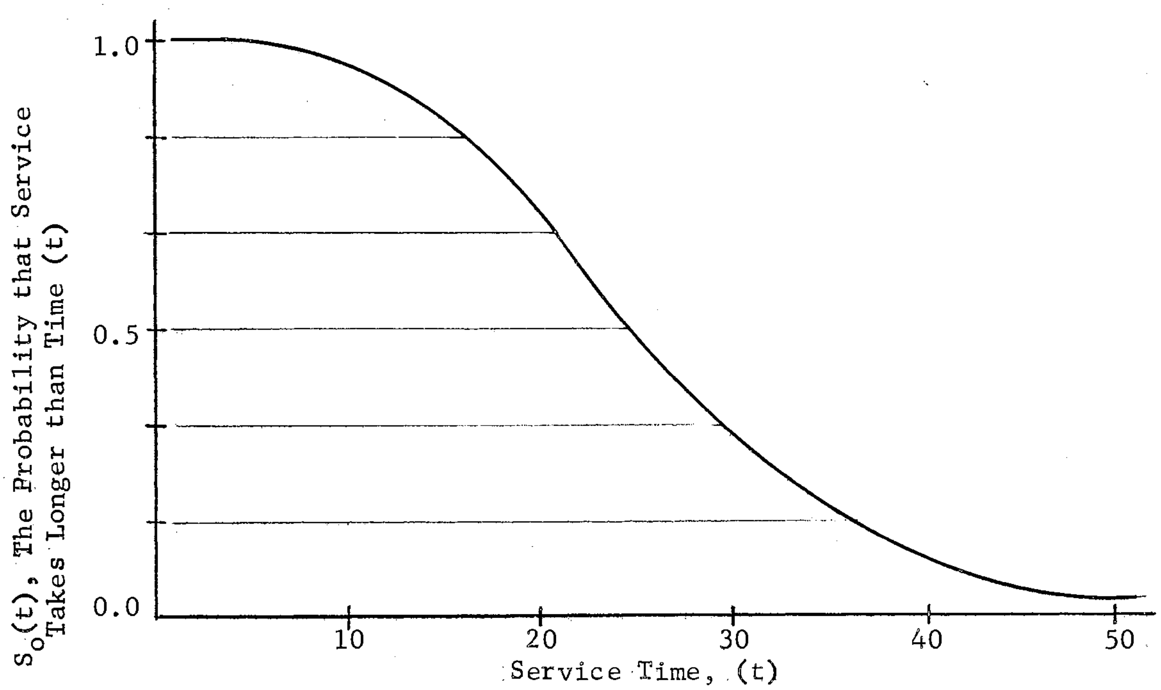


Figure 8. $S_o(t)$, Service Time Distribution Function

obtained. As sampling is continued indefinitely, the curve will approach a continuous function as illustrated in Figure 8. If the situation remains the same, another sample of measured times will yield another empirically determined probability curve which will be roughly equal to the first.

This probability function or "service-time-distribution," $S_o(t)$, is all that is required to represent the service facility since it defines the service time for all units consecutively entering the system. All curves of $S_o(t)$ will start with a probability of unity at $t = 0$ since it is certain that a service operation will take longer than zero time. A special case of $S_o(t)$ occurs when all operations take the same amount of time. This is the case of constant service which is illustrated in Figure 9 by the dashed line. The service time for every arriving unit

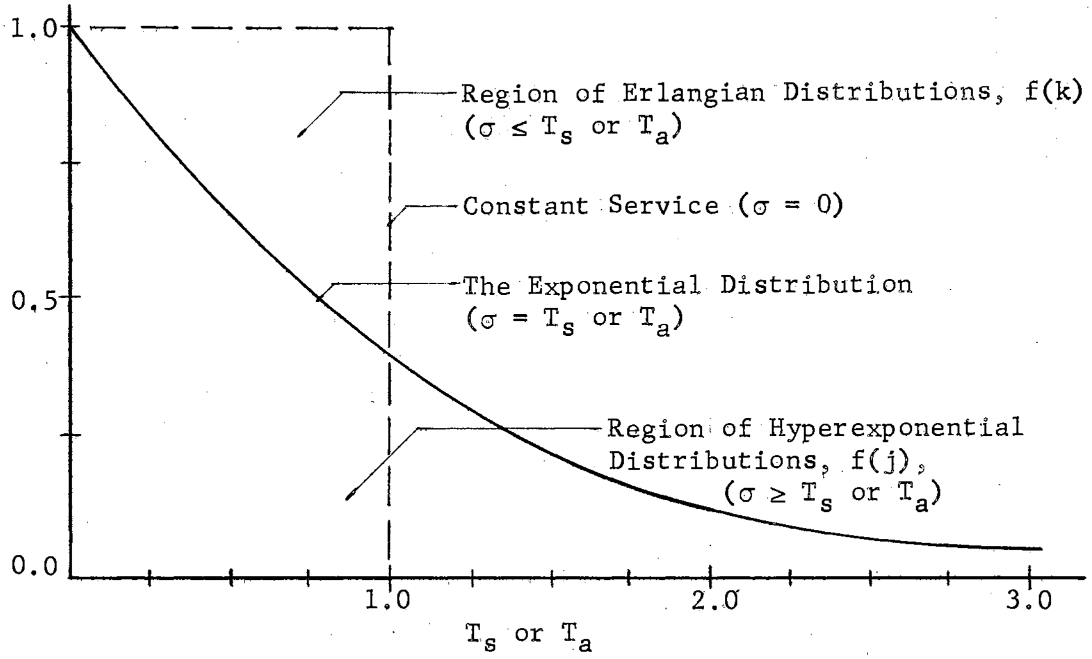


Figure 9. $S_o(t)$ or $A_o(t)$ Curves

takes exactly time T_s so that it is certain [$S_o(t) = 1$] that every service takes longer than (t) if (t) is less than T_s and certain that no service [$S_o(t) = 0$] is longer than (t) if (t) is greater than T_s. Constant services are very usual for architectural queueing systems and most $S_o(t)$ curves will tend monotonocally toward zero as (t) approaches infinity.

A large number of service operations will exhibit $S_o(t)$ distributions which are closely approximated by the exponential curve. The exponential curve is illustrated as the solid line in Figure 9 and represents the case where the probability of prolongation of service is independent of how long ago the service started. The probability function $S_o(t)$ for the exponential case is expressed as:

$$S_o(t) = e^{-\mu t} \tag{2.1}$$

where e = 2.718182..., the base of the natural logarithms and $\mu = 1/T_s$, the mean number of expected service completions per unit time. The exponential distribution is peculiar since its mean, T_s, and standard deviation, σ, are exactly equal. Exponential distributions are extremely important because of their wide scope of applicability and will be used extensively within this study. They have a further theoretical importance because they enable the mathematical solution of queueing problems by the use of linear equations.

The curves in Figure 9 have been plotted as a function of time in units of T_s as the term μt is equivalent to t/T_s. It is interesting to note that when the multiple of T_s is unity, $S_o(t) = e^{-1} = 1/2.718182 = .368$ which means that 36.8% of the time the expected service duration will be greater than T_s. The complementary statement would be that

63.2% of the time the expected service duration will be less than T_s. The significance of these statements is that where exponential distributions apply, service times less than T_s are more frequent than service times greater than T_s.

While exponential distributions are very common, it will be found that in many cases the service time distribution departs significantly from the $\sigma = T_s$ or exponential distribution. In other words, the numeric value of the disperson about the mean, expressed in terms of standard deviation, will be greater or less than the mean. Those distributions with less variation ($\sigma < T_s$) than the exponential case are known as Erlagian distributions while those distributions with greater variation ($\sigma > T_s$) are known as hyperexponential distributions.

The probability function, $S_o(t)$, for Erlagian service distributions may be expressed as:

$$S_o(t) = e^{-k\mu t} \sum_{n=0}^{k-1} (k\mu t)^n/n! \qquad (2.2)$$

where k may be considered an integer constant indicating the degree of departure from the exponential case. When $k = 1$, Equation (2.2) reduces to Equation (2.1), the exponential distribution while when k approaches infinity, $S_o(t)$ approaches the special case of constant service times.

The standard deviation of an Erlagian distribution may be expressed as:

$$\sigma = T_s/\sqrt{k} \qquad (2.3)$$

from which the distribution plotted in Figure 8 is the case of k approximately equal to five. Erlagian distributions are plotted for $k = 1$, 2, 4, 10, 50, and infinity in Figure 10. These types of distribution may be expected where service performance is relatively the same for a majority of arrivals. A typical example might be an airline ticket counter handling only those passengers traveling short distances.

The probability function, $S_o(t)$ for hyperexponential service distributions may be expressed as:

$$S_o(t) = \pi e^{-2\pi\mu t} + (1 - \pi)e^{-2\mu t(1 - \pi)} \qquad (2.4)$$

where π is defined in terms of an integer j by the relationship:

$$(j) = \left[1 + \frac{(1 - 2\pi)^2}{2\pi(1 - \pi)}\right]. \qquad (2.5)$$

(j) may be considered an integer constant indicating the degree of departure from the exponential case. When $j = 1$, $\pi = .500$ so that Equation (2.4) reduces to Equation (2.1), the exponential distribution. As j increases, the degree of variation increases. From Equation (2.5), when $\pi = .2113$, $j = 2$; when $\pi = .1127$, $j = 4$; when $\pi = .0478$, $j = 10$; and when $\pi = .0244$, $j = 20$.

The standard deviation of a hyperexponential distribution may be expressed as:

$$\sigma = T_s\sqrt{j} \qquad (2.6)$$

Hyperexponential distributions for $j = 1, 2, 4, 10$, and 20 are plotted in Figure 10. These types of distributions may be expected where service performance is characterized by very long and very short durations. In the ticket counter example, suppose that transcontinental as well as short distance tickets were handled. Assuming that arriving passengers

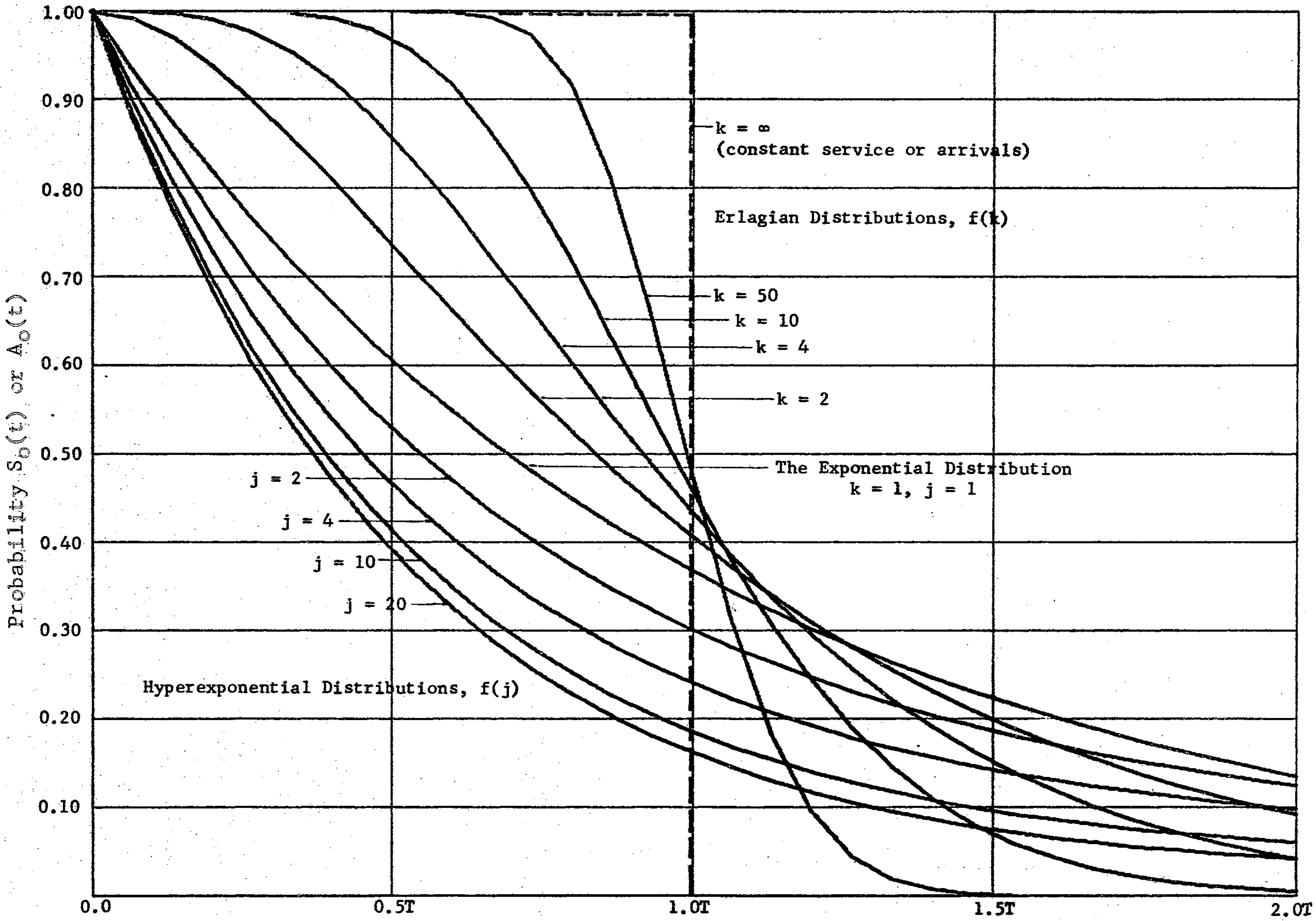


Figure 10. Probability Distributions for Mean Service Time, T_s, or Mean Interval Between Arrivals, T_a

traveling transcontinental distances require substantially longer service performance, long and short durations would predominate yielding a large variation about the mean. In both Erlagian and hyperexponential distributions, it is not the magnitude of service times that govern but the extent that individual service times are distributed about their mean.

Distribution of Arrival Times

Irregular arrivals to the queueing system may be described in terms of probabilities quite analogous to service times. The arrival time is the amount of time that has elapsed between successive arrivals to the system. The mean arrival time or interval between arrivals shall be symbolically represented by T_a. The reciprocal of the mean arrival time, $1/T_a$ shall be represented by λ and defined as the arrival rate or expected mean number of arrivals per unit time.

The interval between arrivals may be collected as data and plotted identically as the service time of Figure 7. By dividing by the total number of arrivals, the arrival distribution function $A_o(t)$ similar to the curve in Figure 8 may be derived. The curve $A_o(t)$ yields the probability that the next arrival comes later than time t after the previous arrival, or than no arrival occurs in time t after the previous arrival.

The exponential curve closely approximates the real-world condition where the probability of occurrence for the next arrival is independent of the elapsed time since the last arrival. The distribution function for exponential arrival times may be expressed as:

$$A_o(t) = e^{-\lambda t} \tag{2.7}$$

In most architectural systems, there are a greater number of chance factors influencing arrival times than service times. It is therefore reasonable to assume that arrival times come closer to being random in the sense that they are exponentially distributed. While the service times of a great many facilities may tend towards regularity, the arrival times are in most cases completely unpredictable. For example, it is all but impossible to predict the moment a customer will arrive at a sales counter, or passenger to a ticket counter, or automobile to a parking lot.

It may be shown that when the distribution function $A_o(t)$ is exponential, the distribution $A_n(t)$ follows the well known and widely applied Poisson distribution. $A_n(t)$ gives the probability of exactly n arrivals within an interval of time t and may be expressed as:

$$A_n(t) = (\lambda t)^n e^{-\lambda t}/n! \tag{2.8}$$

Situations to which the Poisson distribution has been shown to be applicable are so numerous and so diversified that it has sometimes been call called the law of small numbers. For example, many architects will be familiar with rainfall intensity charts which give the number of yearly periods in which a specified amount of rainfall accumulated within a specified amount of time. These particular charts follow the Poisson distribution.

A special distinction is characteristic of the Poisson distribution which may be stated as follows. The area for opportunity of occurrence for an event is extremely large relative to the chance that the event will occur at any given opportunity. There are many opportunities for a rainstorm to occur. However the chance that a rainfall of one-half inch or more will occur in any particular 10 minute time period is extremely small. Similarly, there are many opportunities for an automobile to arrive at a large shopping center parking lot or passengers to arrive at a baggage check-out counter. However, in both cases, the chance that an arrival occurs within any particular short interval is small in relation to its opportunity for occurrence.

Because of these distinctive characteristics and its wide scope of applicability, the Poisson distribution is particularly useful to queueing theory. In addition, with the relative simplicity of the exponential distribution function, there is much less difficulty encountered in the mathematical derivations of queueing model parameters. Most queueing theory texts devote extensive coverage to models incorporating this particular distribution function. For these reasons, all models in this study consider only those systems with Poisson arrivals.

Of course, it must be acknowledged that other arrival distributions exist. Under these conditions, the $A_o(t)$ curves may be described in terms of Erlagian or hyperexponential classifications. The relationships for Erlagian or hyperexponential arrival distributions are identical to those of Equations (2.2) and (2.4) replacing the value μ by λ. The mathematical derivation of parameters for these conditions becomes so complex that they are impractical for inclusion to this study. In many cases, other methods such as "Monte Carlo analysis" or "computer simulation" are far superior. A family of $A_o(t)$ curves is shown corresponding to $S_o(t)$ curves in Figure 10.

CHAPTER IV

THE STRUCTURAL ASPECTS OF QUEUEING SYSTEMS

It has been established previously that the arrival and service distributions are essential elements to any queueing system. However, a complete description of queueing systems requires additional statements concerning their physical and theoretical structure. Statements involving physical structure relate to aspects as the total number of service facilities, how service facilities are physically arranged, or possible limitations to the maximum length of queue. Theoretical considerations include assumptions concerning the logical order in which units are serviced or the manner in which arriving units enter the system.

Several methods of classification have been developed to describe queueing systems. The objective of these systems is mainly to provide a concise notation for those persons dealing in mathematical applications of queueing theory. As this is an introductory study of real-world applications, a more qualitative approach is desirable. James M. Moore[1] has provided this approach by summarizing the interrelationships of elements within a queueing system in terms of an organizational chart. Moore's chart, which describes models for which a relatively large

[1] James M. Moore, "To Queue or Not to Queue," Journal of Industrial Engineering, Vol. XII, No. 2, March-April, 1961, pp. 119-121.

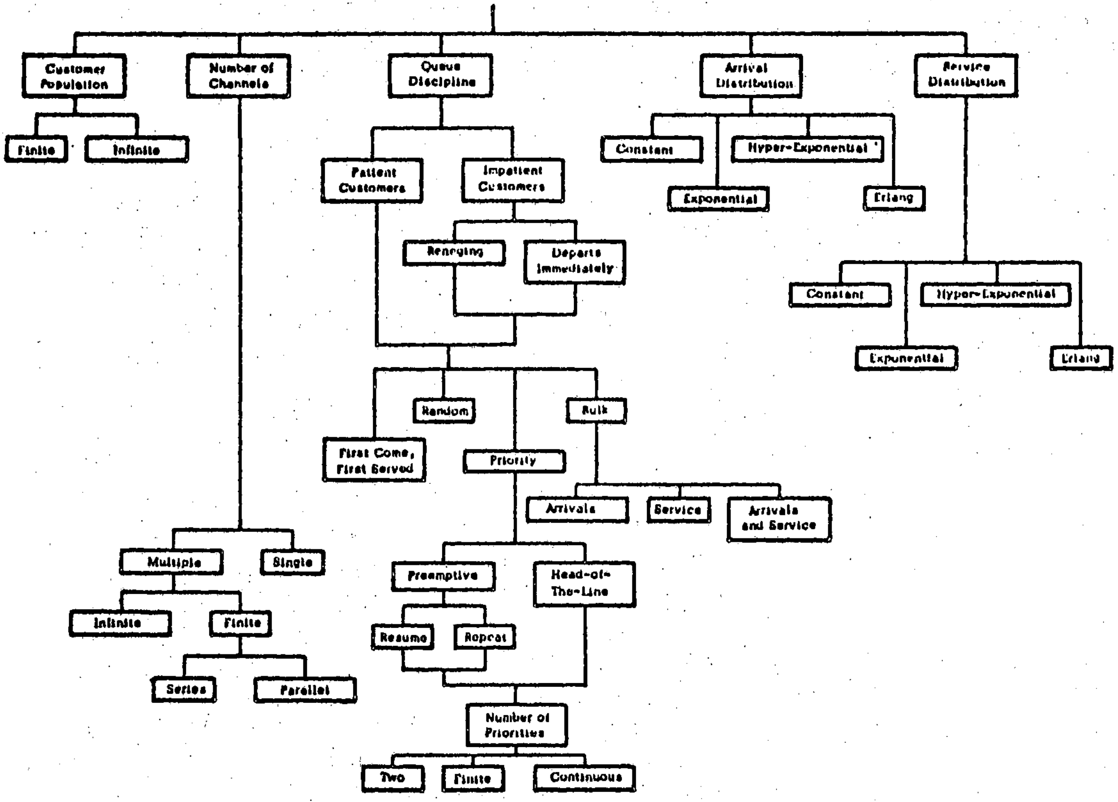


Figure 11. Moore's Organization Chart

amount of published material is available, is shown in Figure 11.

In order to define a given queueing model, information must be obtained for each of the five blocks at the top of the chart. These five basic structural elements include the customer population, number of channels, queue discipline, arrival distribution, and service distribution. Each particular element is fully described when a dead-end branch is reached. Where the appropriate information is not available, reasonable assumptions must be made.

The Customer Population

The customer population refers to the population from which arrivals enter the system. When the number of customers is very large and their demands for service are correspondingly small, it is convenient to assume that the customer population is "infinite." This assumption simplifies the computational effort in deriving system parameters and is the more usual case in architectural situations. Where this assumption cannot be made, the customer population must be considered "finite." The importance of this differentiation may be illustrated in a probabilistic example.

Consider two hypothetical restaurants, one open to the public and the other private, limited to a membership of 200 persons. The public restaurant has a customer population which includes all persons living within reasonable distance and may be considered very large. The demands for service placed upon the public restaurant are relatively small when compared to the population size. Because the population is large, the demand and therefore the probability of an arrival is relatively constant regardless of the number of arrivals which have previously occurred. However, suppose that 100 customers have arrived at the private restaurant. This small population has been reduced by one-half so that a substantial decrease in demand may be expected. The arrival distribution for the public restaurant is independent of the number of previous arrivals while the private restaurant is highly dependent upon the number of previous arrivals.

The Number of Channels

The number of channels refers to the number of facilities available for service. A theater with one ticket booth is a "single" channel facility while a bank with several tellers is a "multiple" channel facility. When a customer has the option of service from any one of several channels, the channels are said to be in "parallel." The vast majority of queueing problems found in architecture involve parallel, multi-channel facilities. Service facilities are in "series" when service is rendered consecutively by more than one channel. For example, departure from an inter-continental airline terminal might require the consecutive service of a ticket counter, baggage counter, customs inspection, and passport inspection in that specific order. "Infinite," multi-channel facilities rarely, if ever, exist in practical applications. They are primarily of mathematical interest because of the ease in computing their system parameters.

The Queue Discipline

The queue discipline is the logical procedure by which customers receive service. In some respects, the queue discipline also refers to the manner in which customers arrive. Figure 11 indicates many variations in queue discipline as several either/or decisions must be made. For example, customers may be "patient" or "impatient." Patient customers refer to the possibility of an unlimited or infinite queue. All arrivals enter the system and remain for service regardless of the system's condition. This situation is particularly applicable where no other alternatives are available as in a single exit from a parking lot. However, the patient customers or unlimited queue condition may also be applied in less rigid situations as, for example, a theater ticket line. The theater queue will certainly not reach infinity, but all customers will remain in the system as long as they are reasonably assured of entry. On the other hand, impatient customers will either "depart immediately" or "renege" if the system is in any condition which discourages their entry. The latter condition assumes some measurable amount of time was spent in the system before departure.

Once the customer has decided to remain in the system, he may receive service in several ways. The most common, particularly in architectural situations, is the "first come, first serve" discipline which requires no explanation. In other instances, customers may receive service at "random," for example as in a crowded bar or concession stand where no defined queue has formed.

The "priority" discipline exists where one type of customer has preference over another. The "head-of-line" priority is the condition where customers with higher priorities move directly to the front of the waiting line. The "preemptive" priority occurs when an arriving customer with higher priority bumps out of the facility the customer receiving service. This is the hospital case where the routine treatment of a patient is interrupted to provide emergency treatment for an accident victim. The bumped patient may "resume" service at the point of interruption or be required to "repeat" the service cycle from the beginning. The number of priorities by which the system is governed may vary from the simple case of "two" to any "finite" number if three or more levels of customers exist. When many finite levels of customers are considered, their priorities are often treated as "continuous" functions to simplify their mathematical computations.

The "bulk" discipline refers to conditions in which customers arrive or are serviced as a group. For example, passengers disembarking from an airplane arrive as a group while the passengers in an elevator arrive individually but are serviced as a group. Sometimes both the arrivals and service use the bulk queue discipline. Under these conditions, it is much simpler to consider the entire group as one arrival and the service time as the amount of time required to service the entire group.

The Arrival and Service Distributions

Arrival and service distributions were discussed in Chapter III in terms of probability curves. These curves were classified into families described as constant, Erlagian, exponential, or hyperexponential distributions, corresponding with Moore's chart. When the probability curve or distribution is known, together with the mean interval between arrivals, T_a, or the mean service time, T_s, the arrival or service distributions are completely defined.

Traffic Intensity

The arrival and service times may be expressed in terms of a single variable, ρ, the "traffic intensity." Traffic intensity is the ratio of the mean service time and the mean interval between arrivals. Written in terms of their rates, ρ is the ratio of the mean arrival and service rates.

$$\rho = T_s/T_a = (1/\mu)/(1/\lambda) = \lambda/\mu \tag{4.1}$$

When $\rho < 1$, on the average, more service completions than arrivals occur per unit time. As λ increases or μ decreases, traffic intensity increases and it becomes more likely that a customer will have to wait.

When $\rho \geq 1$, the arrival rate is greater than the service rate and a greater number of customers arrive than are serviced per unit time.

Two conditions will result. First, the queue will grow without bound; and second, a steady state condition will not be reached. A steady state occurs when "state probabilities" reach a point of equilibrium and become independent of time. Independence of time means that a specific state probability is constant for a particular system at any point in time during its operation.

Since all measures of effectiveness developed in this study are independent of time, investigation will be limited to the condition of $\rho < 1$. This is a logical approach since unbounded systems are unrealistic in most architectural problems. However, it should be recognized that the condition $\rho \geq 1$ may exist irregularly for short periods of time in real-world systems. One of the primary advantages of queueing theory is that it recognizes the occurrence of these irregularities in its measures of effectiveness on a probabilistic basis.

CHAPTER V

SINGLE CHANNEL MODELS

As discussed in Chapter IV, all models considered in this study shall have an infinite customer population, a Poisson's arrival distribution, and shall be serviced in a first come, first serve discipline. Except as discussed in the next section, the service time distribution shall be assumed exponential. In this particular chapter, models consisting of a single service facility will be considered under conditions of either patient or impatient customers. It is assumed that impatient customers immediately depart the system. Figure 11 shows that a queueing model is completely described by the assumptions above as dead-end branches are reached under each of the five major blocks.

Patient Customers

Models with patient customers assume that all arrivals enter the system to remain for service regardless of the system condition. Graphically represented in Figure 12, the open-ended queue indicates that this model is the case for which an infinite queue is allowed. The symbol n represents the number of units in the system at any point in time. The simplicity of this model allows consideration of a few measures of effectiveness for non-exponential service distributions. These measures include the expected mean number of units in the system and in the queue, L and L_q, and the expected mean waiting time of units in the

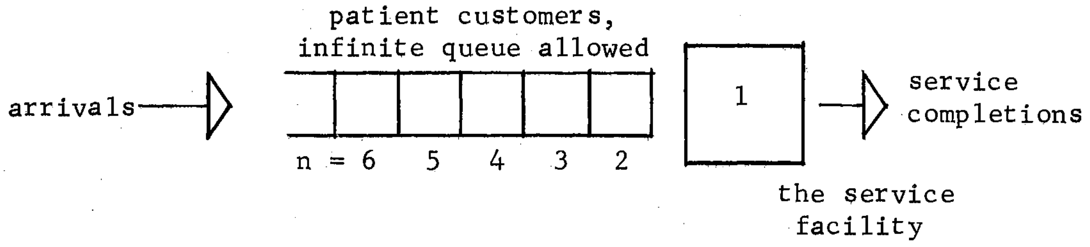


Figure 12. Single Channel Service, Infinite Queue Allowed

system and in the queue, W and W_q.

Exponential Services: L, L_q, W, and W_q

The general state probability, P_n, is the probability that there are exactly n units in the system at any point in time. It may be shown that when the service distribution is exponential, P_n is expressed as:

$$P_n = (1 - \rho)\rho^n. \tag{5.1}$$

In the case of n = 0, the probability of zero units in the system, Equation (5.1) reduces to $P_0 = (1 - \rho)$. P_0 represents the proportion of time that the facility is completely idle. Hence, the expression $(1 - P_0)$ always represents the proportion of time that the facility is occupied. It is a direct measure of facility efficiency and shall be defined as the "utilization factor."

For single channel systems with infinite queues allowed, P_0 is independent of the service distribution. Whether the service distribution is exponential, Erlagian, hyperexponential, or constant, P_0 is always $(1 - \rho)$. For these particular models, the traffic intensity,

$\rho = (1 - P_0)$ is the utilization factor, making it a useful measure of effectiveness.

The summation of probabilities for all possible states of the system must equal unity as the probability of occurrence for every possible event is considered. The number of possible states is infinite as an unlimited queue is allowed when all customers are assumed patient. Since each value of P_n represents the proportion of time that the system contains exactly n units, the product $n(P_n)$ summed from 0 to ∞ yields the mean number of units in the system. The value of L may be expressed and evaluated as:

$$L = \sum_{n=0}^{\infty} n(P_n) = \rho/(1 - \rho). \tag{5.2}$$

Values for L are plotted as a function of ρ in Figure 13 as the curve designated for exponential services.

By similar reasoning, the expected mean number of units in the queue may be expressed and evaluated as:

$$L_q = \sum_{n=1}^{\infty} (n - 1)P_n = \rho^2/(1 - \rho) \tag{5.3a}$$

$$L_q = L - \rho. \tag{5.3b}$$

Since ρ represents the proportion of time that the facility is occupied, L_q expressed in terms of L as in Equation (5.3b) is valid regardless of the service distribution. Figure 13 may be used to determine L_q by first determining L and making the subtraction of ρ.

The expected mean time a unit spends in the system, W, may be

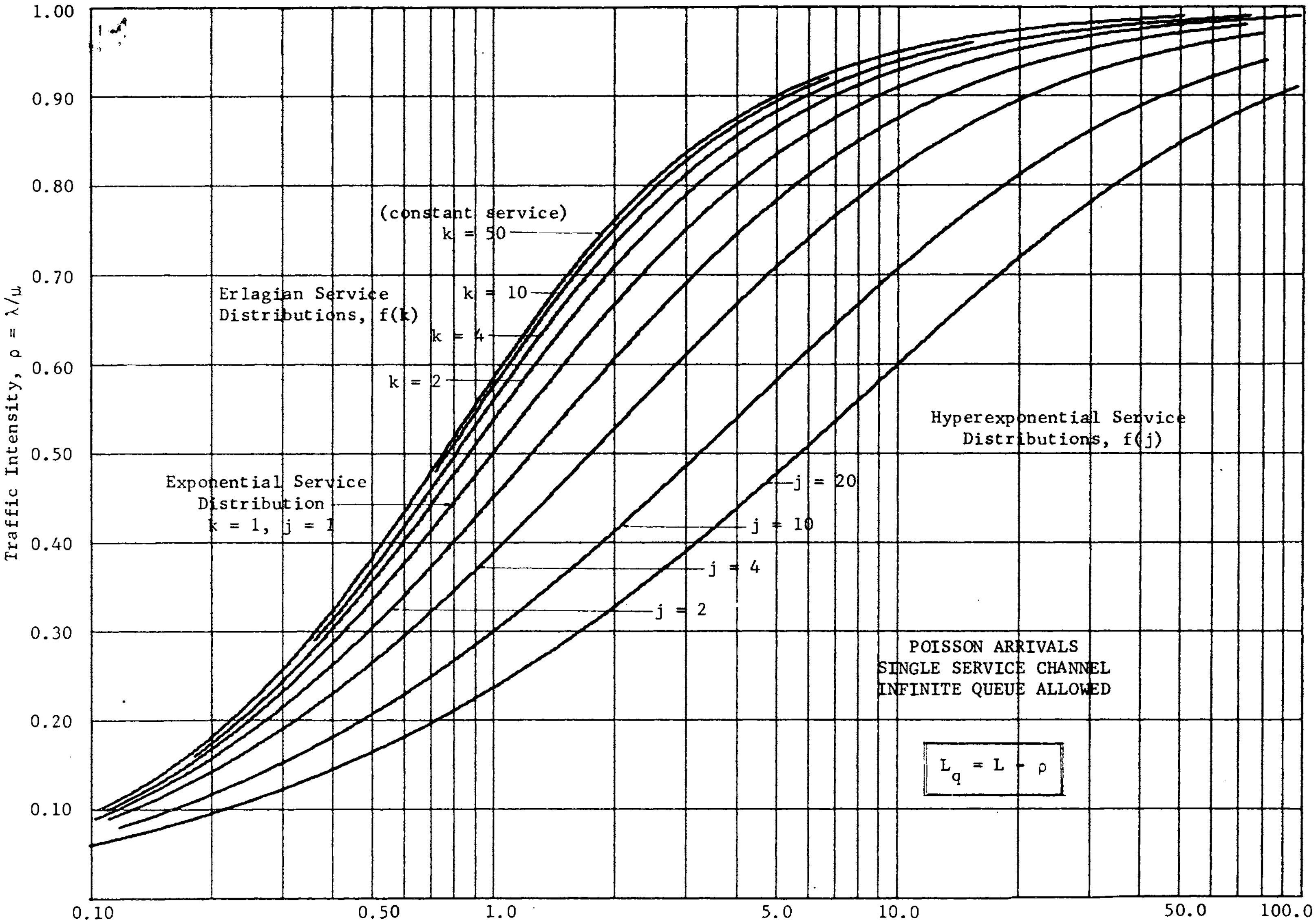


Figure 13. L, The Mean Number of Units in the System, Single Channel Systems, Infinite Queue Allowed

expressed as:

$$W = 1/(\mu - \lambda) \tag{5.4a}$$

$$W = L/\lambda \tag{5.4b}$$

and the expected mean time a unit spends in the queue, W_q, may be shown to be:

$$W_q = \lambda/\mu(\mu - \lambda) = L_q/\lambda \tag{5.5a}$$

$$W_q = (L - \rho)/\lambda \tag{5.5b}$$

Values for W or W_q are plotted as functions of L or L_q for incremental values of λ in Figure 14. When W and W_q are expressed in terms of L as in Equations (5.4b) and (5.5b), the relationships are valid, regardless of the service distribution. Thus, Equations (5.3b), (5.4b), and (5.5b) may be used whether the distribution is constant, exponential, Erlagian, or hyperexponential after the proper value of L has been determined.

Erlagian and Hyperexponential Services: L

Because of the relationships just discussed, measures of effectiveness for Erlagian and hyperexponential service distributions are limited to L alone. The derivation and expression of the general state probability, P_n, is well beyond the scope of this study. However, for Erlagian service distributions, L may be expressed as:

$$L = \frac{2k\rho - \rho^2(k - 1)}{2k(1 - \rho)} \tag{5.6}$$

where k may be considered an integer constant indicating the degree of

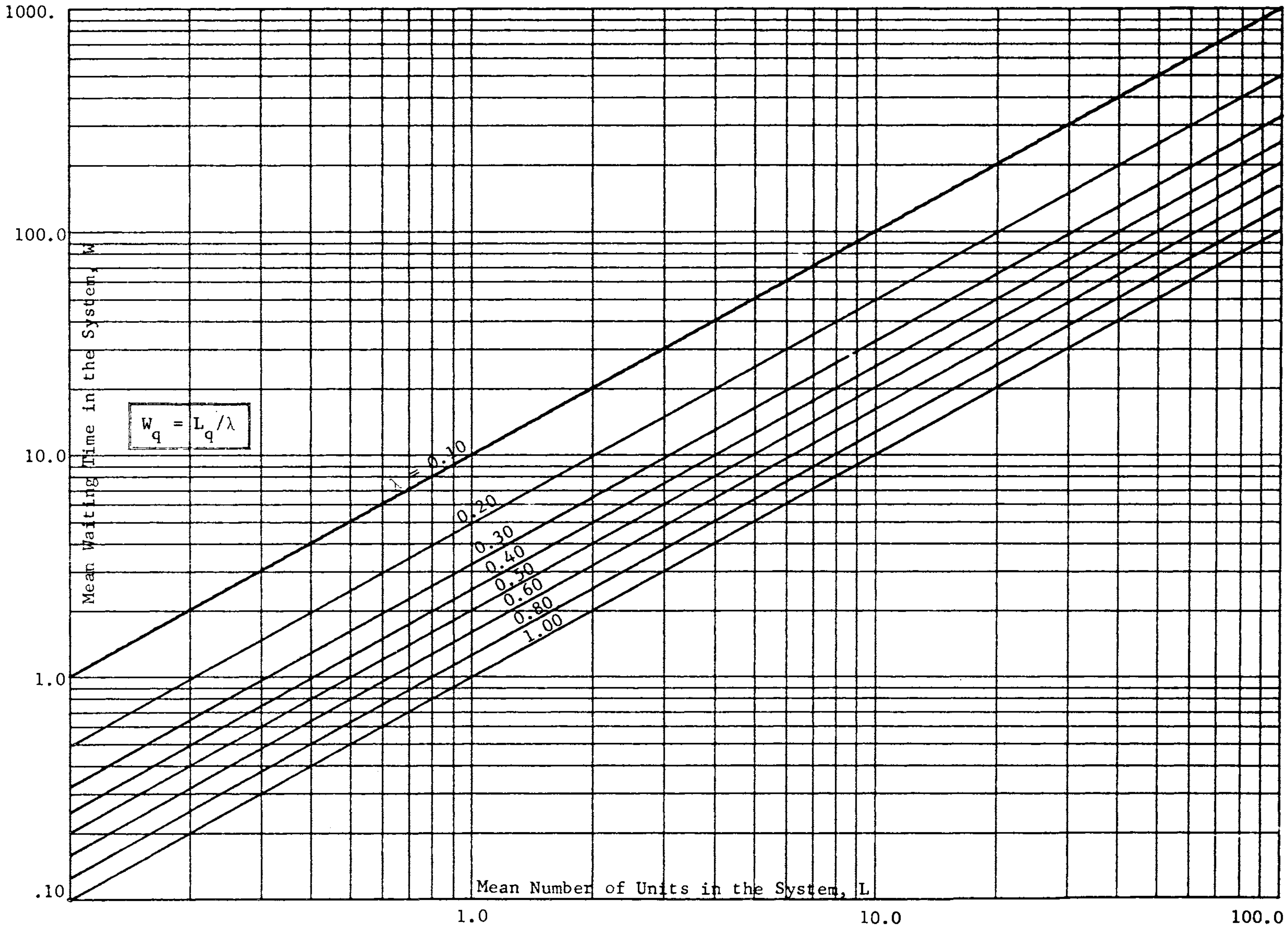


Figure 14. W, The Mean Waiting Time Spent in the System, All Systems

departure from the exponential case. When k = 1, the service distribution is exponential and Equation (5.6) reduces to Equation (5.2). As k increases service time variability decreases. Constant service times are approached as k approaches ∞. Values for L are plotted as functions of ρ for k = 1, 2, 4, 10, and 50 in Figure 13. When k = 50, k may be considered as infinite or representative of the constant service time condition.

Hyperexponential service distributions are characterized by very long and very short service times. The mean number of units in the system, L, may be expressed as:

$$L = \frac{\rho^2 + \rho(1 - \rho)4\pi(1 - \pi)}{4\pi(1 - \rho)(1 - \pi)} \qquad (5.7)$$

where π may be expressed in terms of an integer constant, j, [see Equation (2.5)] indicating the degree of departure from the exponential case. When π = .50, j = 1 and Equation (5.7) reduced to Equation (5.2), the exponential case. Values of L are plotted as functions of ρ for j = 1, 2, 4, 10, and 20 in Figure 13.

Example Problems

The use of Figures 13 and 14 to determine the preceding measures of effectiveness is best illustrated through the use of example problems. The presentation of these measures in graphic form allows rapid investigation of the system under several different conditions. In addition, many characteristics of the measures which are not readily apparent in their mathematical form become clearly evident when presented as shown.

Problem No. 1: An airline ticket counter is subject to Poisson arrivals, as all models of this study, with a mean interval between arrivals of 10 minutes. A mean service duration of 5 minutes per customer has been determined. As both local and intercontinental tickets are sold, very short and very long service durations predominate. A greater number of local tickets are sold so that there are more short service durations than long. A cumulative frequency diagram of service times is constructed, as in Figure 7, and compared to the family of curves in Figure 10. Assume that the service distribution is determined as hyperexponential with j = 10. What are the operating characteristics of the system expressed in numeric terms as measures of effectiveness?

Solution: $T_a = 10$ minutes; $\lambda = 1/T_a = 0.10$ arrivals/minute. $T_s =$ 5 minutes; $\mu = 1/T_s = 0.20$ service completions/minute. Traffic intensity, $\rho = \lambda/\mu = 0.10/0.20 = 0.50$.

Enter Figure 13 on the ordinate with $\rho = 0.50$ and move horizontally until intersecting the curve $j = 10$. Move vertically from the intersection to the abscissa where L = 3.2 customers. From Equation (5.3b), $L_q = L - \rho = 3.2 - 0.5 = 2.7$ customers.

Enter Figure 14 on the abscissa with L = 3.2 customers and move vertically until intersecting the curve $\lambda = .10$. Move horizontally from the intersection to the ordinate where W = 32 minutes. To determine W_q, enter Figure 14 with $L_q = 2.7$ for which $W_q = 27$ minutes. Alternatively, the relationships $W = L/\lambda$ or $W_q = L_q/\lambda$ may be used in lieu of Figure 14.

Comments: For any one hour interval or other short period of time, the mean values determined may vary significantly from the values above. However, in the long run operation of the system, the values above will be approached as statistical limits. The expression, $P_0 = (1 - \rho) =$ .50, indicates that 50% of the time an arriving customer will find the counter empty and enter for immediate service. The utilization factor, $(1 - P_o) = 0.50$, indicates the proportion of time that the counter is occupied. It seems incongruous that although the facility is idle 50% of the time, the mean number of units in the system is still 3.2 customers. From Figure 10, 82% of the customers will require services taking less than T_s or 5 minutes while 58% will take less than $.5T_s$ or 2.5 minutes. As the mean interval between arrivals is 10 minutes, the large idle period appears to be a reasonable result. Figure 10 also indicates that 7% of the customers will require more than $2T_s$ or 10 minutes for service. The arrival of these few customers, although infrequent in occurrence, blocks the service channel causing the build-up of units in the system.

It would seem intuitively correct for the mean time spend in the system to equal the mean time spent in the queue plus the mean duration of service. This may be expressed as $W = W_q + T_s$. Substituting values, $32 = 27 + 5 = 32$ minutes.

Problem No. 2: In the same airline ticket counter, assume that sales are now limited to intercontinental tickets so that the measured variation of service times has decreased. Furthermore, assume that the service distribution is now exponential with the same arrival and service rates as in the previous problem.

Solution: The traffic intensity, $\rho = 0.50$ remains the same. Entering Figure 13 with $\rho = 0.50$ and using the exponential service curve, $L = 1.0$ and $L_q = L - \rho = 0.50$. From Figure 14, $W = 10.0$ and $W_q = 5.0$ minutes.

Comments: The effect of service time variability is demonstrated by comparing the measures of effectiveness for these problems. Where the service distribution was hyperexponential with $j = 10$, the standard deviation, $\sigma = T_s\sqrt{j} = 15.81$. Where the distribution was exponential, $\sigma = T_s = 5.0$. The corresponding change in L was from 3.2 to 1.0 customers. It may be concluded that as the variation within the service distribution increases or decreases, there is a corresponding increase or decrease in L, L_q, W, and W_q. This is clearly illustrated in Figure 13 as the curves are plotted from left to right in terms of increasing service time variability.

The corresponding increase in L when the service distribution is changed from the constant curve, $k = 50$, to the exponential curve, $k = 1$ is relatively small. A careful examination of Figure 13 will show that the greatest possible change in magnitude is by a factor of two which occurs only as ρ approaches unity.

The service distributions for most architectural service mechanisms exhibit variability somewhere between constant and exponential distributions, that is, they have standard deviations between zero and T_s. If the exponential case is assumed where the distribution is less variable, resültant inferences will be conservative and never more incorrect than by a factor of two. It is for this reason that all further models will be limited to exponential service distributions.

Problem No. 3: Assume that with the addition of computerized ticket handling aids, the mean service time of the counter is reduced to 2.5 minutes. The service distribution remains exponential and the arrival rate remains as before. What are the effects upon the measures of effectiveness?

Solution: $\mu = 1/2.5 = 0.40$; $\rho = .10/.40 = 0.25$. From Figure 13, using the exponential service curve and entering with $\rho = 0.25$, $L = 0.33$

customers, $L_q = 0.33 - 0.25 = 0.08$ customers. From Figure 14, W and W_q are 3.3 and 0.8 minutes respectively.

Comments: A comparison of these results with the previous problem will indicate that as ρ decreases, a corresponding decrease in the measures of effectiveness will result. However, since $P_0 = (1 - \rho)$ indicates the proportion of idle periods, a decrease in ρ also indicates a decrease in system "efficiency." In this case, customer service has improved by reducing W from 10.0 to 3.3 minutes but at a "cost" of doubling the speed of service which increases the proportion of idle time from 50% to 75%.

The effect of facility utilization, $\rho=(1 - P_0)$, upon L are clearly evident in Figure 13. It should now be apparent that the magnitude of L depends upon both the speed and the variability of service times.

Problem No. 4: It has been determined that the mean wait time in the system, W, should not exceed 20 minutes. The arrival rate remains $\lambda = .10$ arrivals/minute. Assuming that the service distribution is exponential, at what rate must the service facility operate to satisfy the condition above?

Solution: Enter Figure 14 on the ordinate with W = 20 minutes and move horizontally until intersecting the $\lambda = .10$ curve. Move vertically from the intersection to the abscissa where L = 2.0 customers.

Enter Figure 13 on the abscissa with L = 2.0 and move vertically until intersecting the exponential service curve. Move horizontally from the intersection to the ordinate where $\rho = .67$.

Since $\rho = \lambda/\mu$, $\mu = \lambda/\rho = .10/.67 = .15$ service completions/minute. $T_s = 1/\mu = 1/.15 = 6.68$ minutes/customer. $L_q = L - \rho = 2.0 - .67 = 1.33$ customers. $W_q = 1.33/.10 = 13.3$ minutes.

Comments: This problem demonstrates than any one of the measures of effectiveness may be predetermined, after which, system behavior may be investigated. In most cases, an architect has no control over the arrival rate but to some degree is able to control the service rate.

In some conditions, it may be desirable to solve this type of problem in reverse. For example, assume that the service rate is fixed. The parameter of interest would now be the maximum arrival rate that the system is capable of handling. The versatility of queueing theory and the ease of graphic computations will become more apparent as further graphs and measures of effectiveness are introduced.

Exponential Service: Q_N, Q_{Nq}, $G(T_s)$, $G_q(T_s)$

Discussion to this point has been limited to mean or average values for measures of effectiveness. While averages are very useful, particularly in an economic analysis, most architectural problems involve traffic units composed of individual persons. Therefore, it is far more critical to investigate system behavior as it affects individuals rather than grouped units. For this purpose, two additional measures of effectiveness are introduced which apply only to exponential service distributions. The first, Q_N and Q_{Nq}, are the probabilities of N or more units in the system and in the queue. The second, $G(T_s)$ and $G_q(T_s)$, the latter being read, "G sub q, a function of T_s", are the probabilities that time spent in the system and queue exceeds a multiple of T_s. In both cases, probability represents the relative proportion of time that the stated event may be expected to occur.

Recall that the general, state probability of an exponential service system was expressed as $P_n = (1 - \rho)\rho^n$, where $\sum_{n=0}^{\infty} P_n = 1.0$. Therefore, Q_N may be expressed as:

$$P_{N \text{ or more}} = Q_N = \sum_{n=N}^{\infty} P_n = \sum_{n=N}^{\infty} (1 - \rho)\rho^n$$

$$Q_N = \rho^N \qquad (5.8)$$

A family of curves for Q_N has been plotted for several values of N in Figure 15. The complimentary statement, the probability of N <u>or less</u> units in the system can therefore be written in terms of Q_N as:

$$P_{N \text{ of less}} = \sum_{n=0}^{N} P_n = \sum_{n=0}^{N} (1 - \rho)\rho^n = 1.0 - Q_{N+1} \qquad (5.9)$$

Equations (5.8) and (5.9) are cumulative probabilities comprised of the summation of individual state probabilities. Figure 15 may be adapted to compute a state probability by rewriting P_n in terms of Q_N.

$$P_{n=N} = (1 - \rho)\rho^N = \rho^N - \rho^{N+1} = Q_N - Q_{N+1} \qquad (5.10)$$

<u>Problem No. 5</u>: In Problem No. 2, traffic intensity was $\rho = 0.50$ and L = 1.0 customers. Determine Q_N, $P_{N \text{ or less}}$, and P_n for n or N = 0 to 6 customers in the system. State any conclusions that may be drawn from the computation of these probabilities.

<u>Solution</u>: The results are shown in Table V. For demonstration purposes, six decimal places have been carried. The values of Q_N may be checked by using Figure 15.

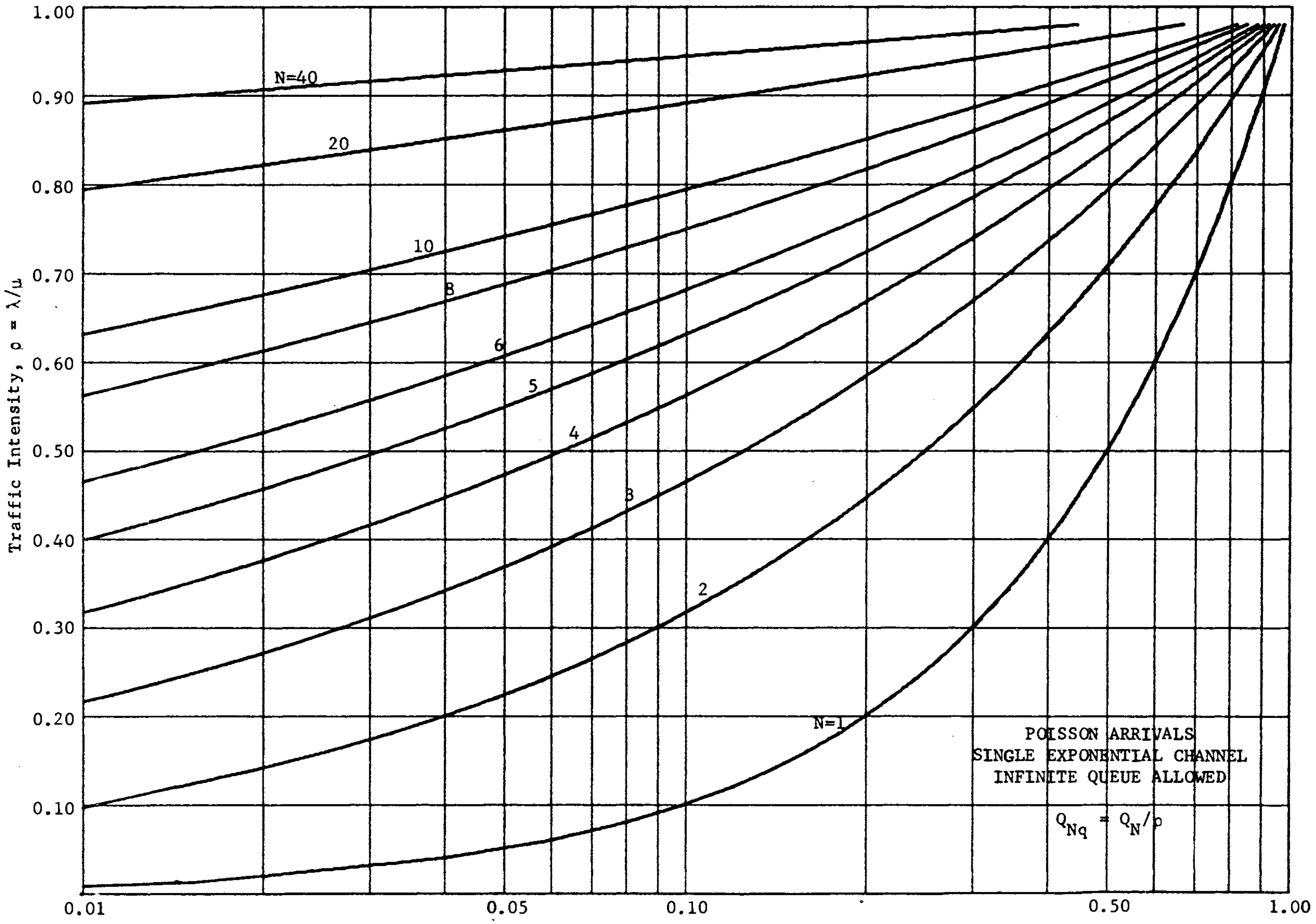


Figure 15. Q_N, The Probability of N or More Units in the System, Single Exponential Channel, Infinite Queue Allowed

TABLE V

SOLUTIONS TO PROBLEM NO. 5

N	$\rho^N = Q_N$	$1.0 - Q_{N+1} = P_{N \text{ or less}}$	$Q_N - Q_{N+1} = P_{n=N}$
0	$.50^0$=1.000000	1.0-.500000=.500000	1.000000-.500000=.500000
1	$.50^1$=0.500000	1.0-.250000=.750000	0.500000-.240000=.250000
2	$.50^2$=0.250000	1.0-.125000=.875000	0.250000-.125000=.125000
3	$.50^3$=0.125000	1.0-.062500=.937500	0.125000-.062500=.062500
4	$.50^4$=0.062500	1.0-.031250=.968750	0.062500-.031250=.031250
5	$.50^5$=0.031250	1.0-.015625=.984375	0.031250-.015625=.015625
6	$.50^6$=0.015626	1.0-.007812=.992817	0.015625-.007812=.007812
7	$.50^7$=0.007812	.	.
.	.	.	.
.	.	.	.
.	.	.	.
∞	$.50^\infty$=0.000000	1.0-.000000=1.00000	0.000000-.000000=.000000
			$\sum_{n=0}^{\infty} P_n = 1.00000$

Comments: Observe that the mean number of units in the system, L = 1.0 customer or P_1 is expected to occur 25% of the time. The mean number or less, $P_{1 \text{ or less}}$, may be expected 75% of the time while the mean or greater, Q_1, may be expected 50% of the time. P_6 = .0078 indicates that exactly six customers in the system will occur .78% of the time. Q_6 = .0156 indicates that six or more customers in the system may be expected 1.56% of the time. Thus, the practical limit to the number of customers in the system is about six.

Problem No. 7: As in all previous problems, assume that the arrival rate has been determined as λ = 0.10. Suppose that no more than three customers in the system are desired five or less percent of the time. What is the required service rate to satisfy the condition above?

Solution: Enter Figure 15 on the abscissa with $Q_N = .05$ and move vertically until intersecting the N = 3 curve. Move horizontally from the intersection to the ordinate where $\rho = 0.37$. The required service rate is therefore $\mu = \lambda/\rho = .10/.37 = .27$ customers/minute. $T_s = 1/\mu = 1/.27 = 3.7$ minutes/customer. From Figure 13, with $\rho = .37$, L = .59 customers. From Figure 14, W = 5.9 minutes.

Comments: This problem again demonstrates that any measure of effectiveness may be established as a governing relationship. Once established, all other measures of effectiveness may be determined. The conditions of this problem are more demanding than those in Problem No. 6 so that an increase in required service speed results. The proportion of time that the facility is occupied decreases, increasing the idle periods and decreasing all other measures of effectiveness.

The probability of N or more units in the queue is the probability of N + 1 or more units in the system and may be expressed as:

$$Q_{Nq} = \sum_{n=N+1}^{\infty} P_n = \rho^{n+1} = Q_{N+1} \qquad (5.11)$$

Similar types of probabilistic statements may be made concerning the time spent in the system by individual customers. $G(T_s)$, the probability that time spent in the system exceeds a multiple of T_s may be expressed as:

$$G(T_s) = e^{-(1-\rho)cT_s} \qquad (5.12)$$

where c is a constant indicating a multiple of T_s and e = 2.718182..., the base of natural logarithms. The probability that time spent in the queue will exceed a multiple of T_s may be expressed as:

$$G_q(T_s) = \rho G(T_s). \tag{5.13}$$

A family of curves for $G(T_s)$ has been plotted in Figure 16 for several multiples of T_s.

Problem No. 8: In Problem No. 2, $\rho = .50$, $T_s = 5$ minutes/customer, and W = 10 minutes. What proportion of time will time spent in the system exceed 5, 10, 15, 20, 25, 30, and 40 minutes? If the arrival rate is $\lambda = .10$ arrivals/minute, what service rate is necessary to insure that 10 percent or less of the time, customers spend more than 20 minutes in the system?

Solution: Enter Figure 16 on the ordinate with $\rho = .50$ and move horizontally until intersecting the curve $1.0T_s$. Move vertically from the intersection to the abscissa where G(5) = .61 or 61% of the time, the total time spent in the system will exceed 5 minutes. Similarly, G(10) = .37; G(15) = .23; G(20) = .14; G(25) = .082; G(30) = .050; and G(40) = .018. For the second part of the problem, enter Figure 16 on the abscissa with $G(T_s) = .10$ and move vertically until intersecting the $4.0T_s$ or 20 minute curve. Move horizontally from the intersection to the ordinate where $\rho = .425$. The required service rate, $\mu = \lambda/\rho = .10/.425 = .235$ service completions/minute. The required mean service duration, $T_s = 1/.235 = 4.26$ minutes/customer.

Comments: This problem demonstrates the procedure to rapidly determine the proportion of time total time spent in the system exceeds a given multiple of T_s. Stated in other words, it determines the

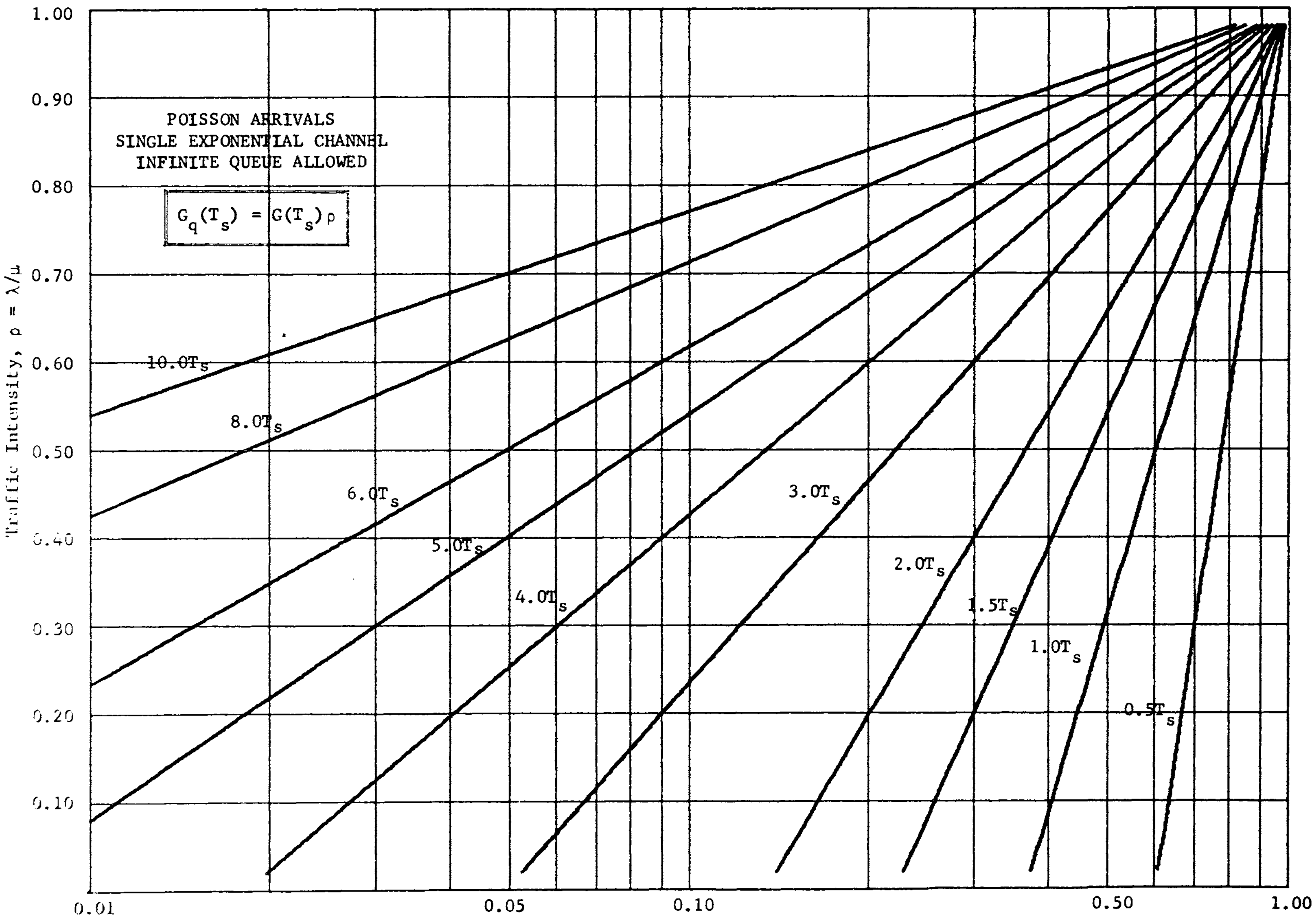


Figure 16. $G(T_s)$: The Probability that Time Spent in the System Exceeds a Multiple of T_s, Single Exponential Channel, Infinite Queue Allowed

proportion of customers for which time spent in the system will exceed a given value. This problem also demonstrates how $G(T_s)$ may be used as a governing measure of effectiveness.

Impatient Customers

In this study, models considered shall assume that impatient customers depart the system immediately. A customer is defined as impatient whenever there are exactly N' units in the system or (N' - 1) units in the queue. For this reason, systems of this kind are often called "limited queue systems." For example, where N' = 3, the queue length is limited to two customers. When the queue is full, an arriving customer is refused entry to the system and departs immediately. The departing customer is lost to the system, becoming anonymous by rejoining the customer population. This means that the lost customer is treated as any other eligible unit within the infinitely large population, receiving no special consideration upon attempting to re-enter the system. The limited queue model is schematically illustrated in Figure 17.

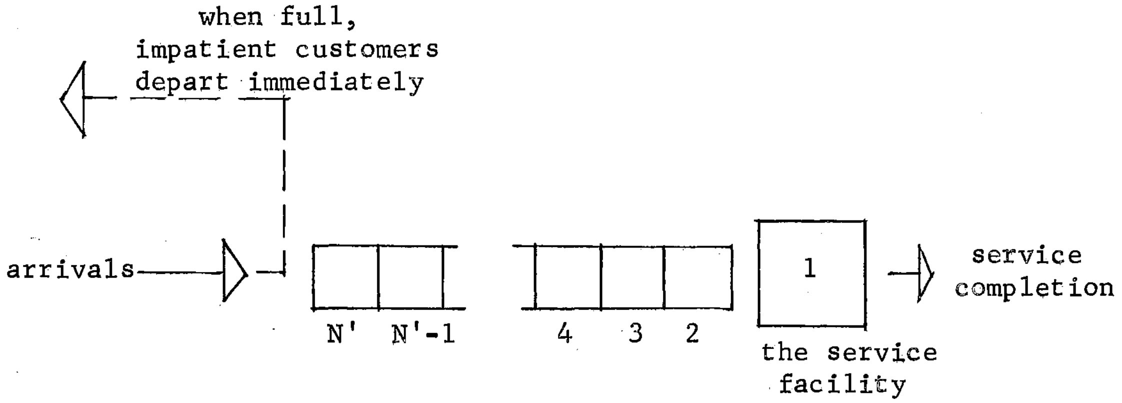


Figure 17. Single Channel Service, Limited Queue

This model is applicable wherever the value of N' may be objectively or subjectively determined. In many cases, physical limitations to queue length may govern. For example, the available space in front of a counter may limit queue length to four persons. All arrivals occurring while the queue length is four are lost to the system. Thus, the probability of a full system must be made small if it is desired that few or no customers are to be lost.

The State Probability

The state probability of exactly n units in the system may be expressed as:

$$P_n = \left[\frac{1 - \rho}{1 - \rho^{N'+1}}\right] \rho^n . \tag{5.14}$$

In this case, the summation of state probabilities from n = 0 to N' must equal unity. The probability of no units in the system, P_0, describes the proportion of time that the facility is completely idle.

$$P_0 = \frac{1 - \rho}{1 - \rho^{N'+1}} \tag{5.15}$$

The probability that the facility is busy is the complement of Equation (5.15) or $(1 - P_0)$ and is the measure of facility utilization. Since ρ is always less than unity, the term $\rho^{N'+1}$ in Equations (5.14) and (5.15) approaches zero as N' increases. Except for values of $\rho > 0.90$, this term may be considered as zero for $N' > 20$ without significant loss of accuracy. Substituting into Equations (5.14) and (5.15) with $N' > 20$, they become $P_n = (1 - \rho)\rho^n$ and $P_0 = (1 - \rho)$ respectively. Since these are the state probabilities for the infinite queue or patient customer condition, the infinite queue model may be used where N' is greater than 20. A review of Figure 15 will indicate the validity of this conclusion. The probability of more than 20 units in the system is insignificant so that proportion of lost customers is small. If no customers are lost, all customers must be patient so that the necessity to use a limited queue model does not exist.

A family of curves for P_0 have been plotted for different values of N' in Figure 18. Equally important to the investigator is the probability that the system is full, $P_{N'}$, which indicates the proportion of lost customers. Substituting N' for n in Equation (5.14),

$$P_{N'} = P_{\text{full}} = \left[\frac{1 - \rho}{1 - \rho^{N'+1}}\right] \rho^{N'} \qquad (5.16)$$

A family of curves for $P_{N'}$ have been plotted for different values of N' in Figure 19.

$Q_{N,N'}$ shall be defined as the probability of N or more units in a system limited to a maximum of N' units. It may be expressed as:

$$Q_{N,N'} = \sum_{n=N}^{N'} P_n = \sum_{n=N}^{N'} \left[\frac{1 - \rho}{1 - \rho^{N'+1}}\right] \rho^n = \frac{\rho^N - \rho^{N'}}{1 - \rho^{N'+1}} \qquad (5.17)$$

A complete set of graphs for the relationship above would require a separate figure for each value of N'. For expediency, Figures 20, 21, and 22 illustrate several combinations of N and N' to demonstrate the behavior of systems under these conditions. $Q_{2,3}$ is found by entering Figure 20 on the ordinate with a predetermined value of ρ and moving horizontally until intersecting the N,N' = 2,3 curve. Moving vertically

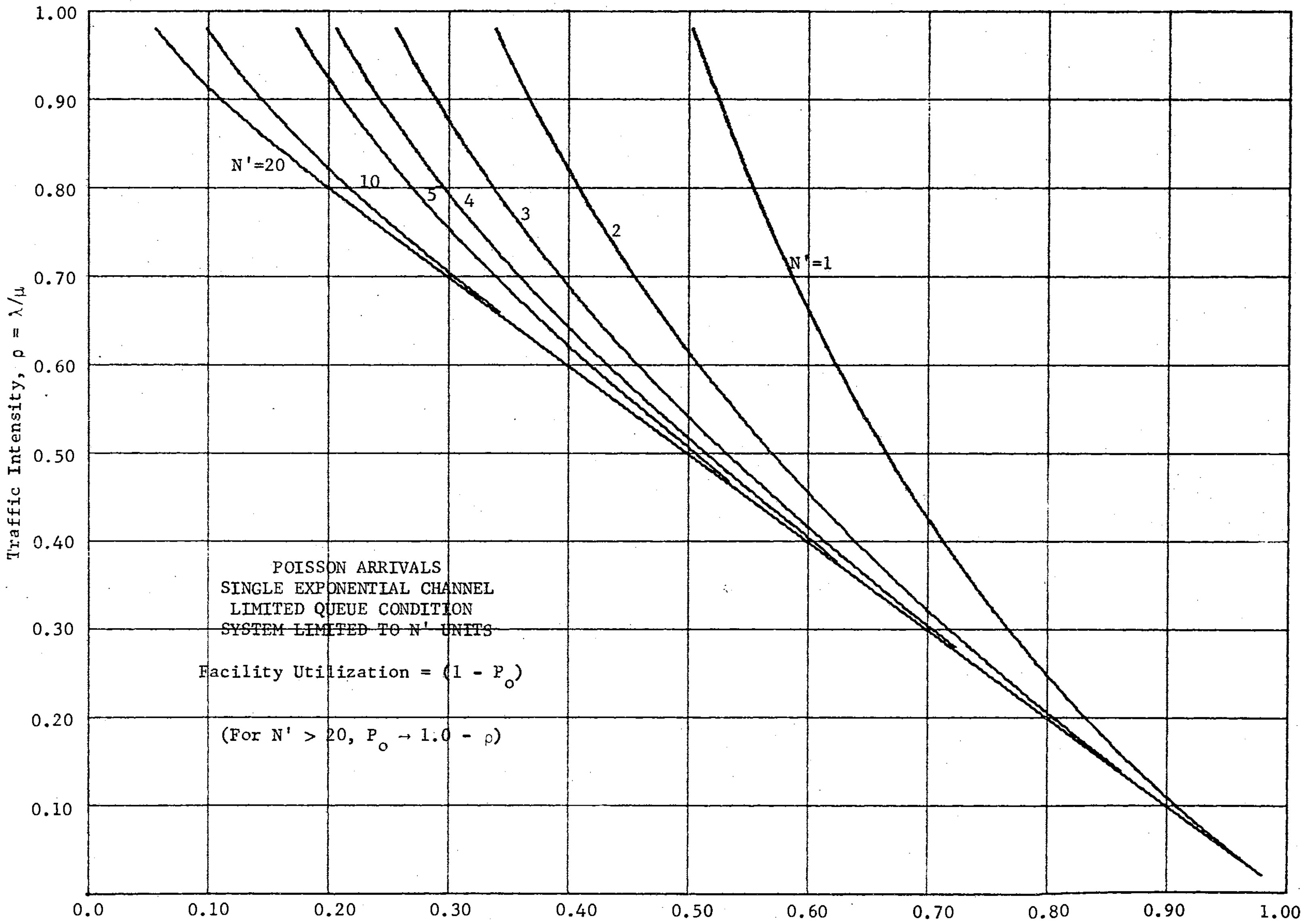


Figure 18. P_0, The Probability of No Units in the System, Single Exponential Channel, Limited Queue Allowed

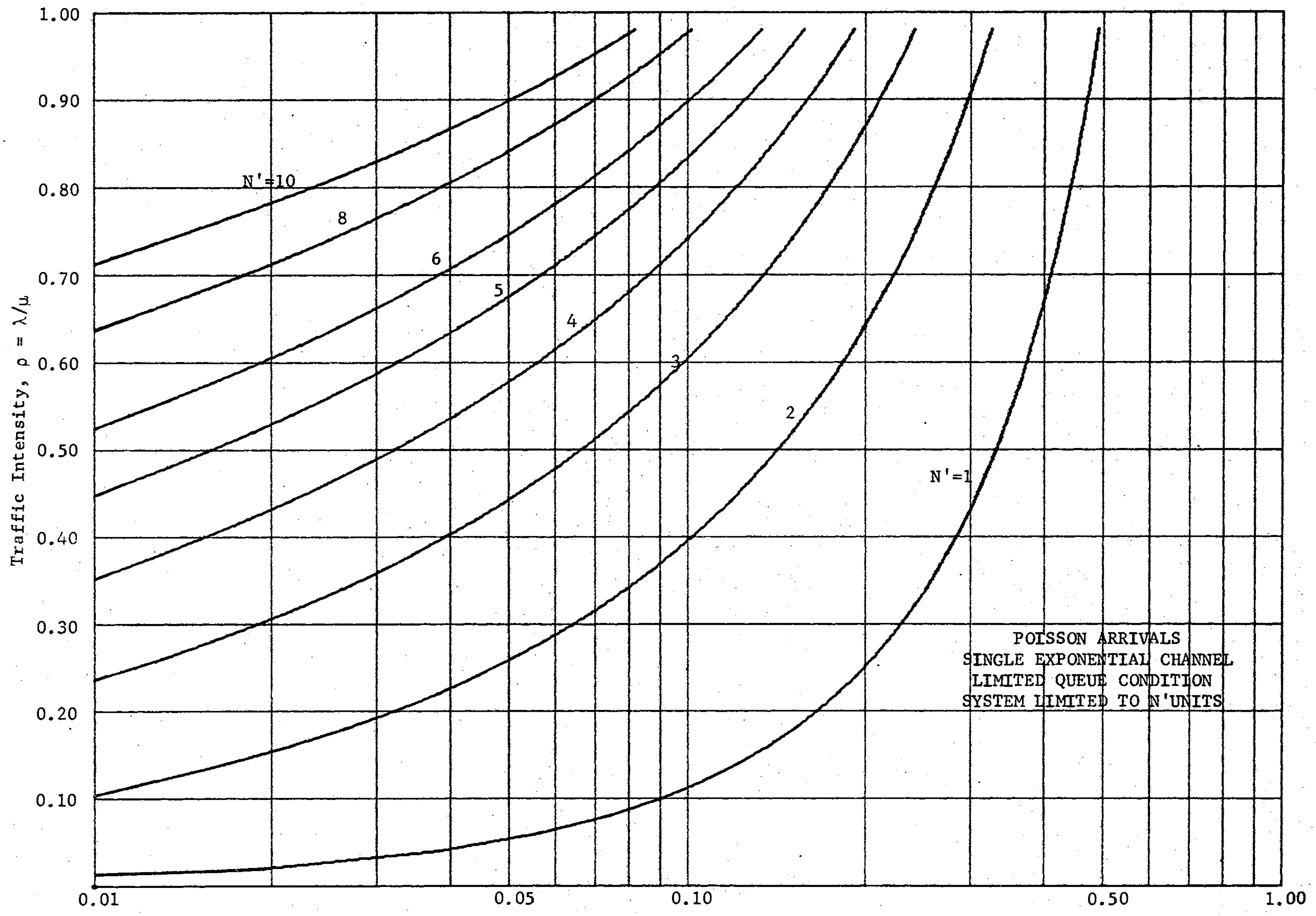


Figure 19. $P_{N'}$, The Probability that the System is Full, Single Exponential Channel, Limited Queue Allowed

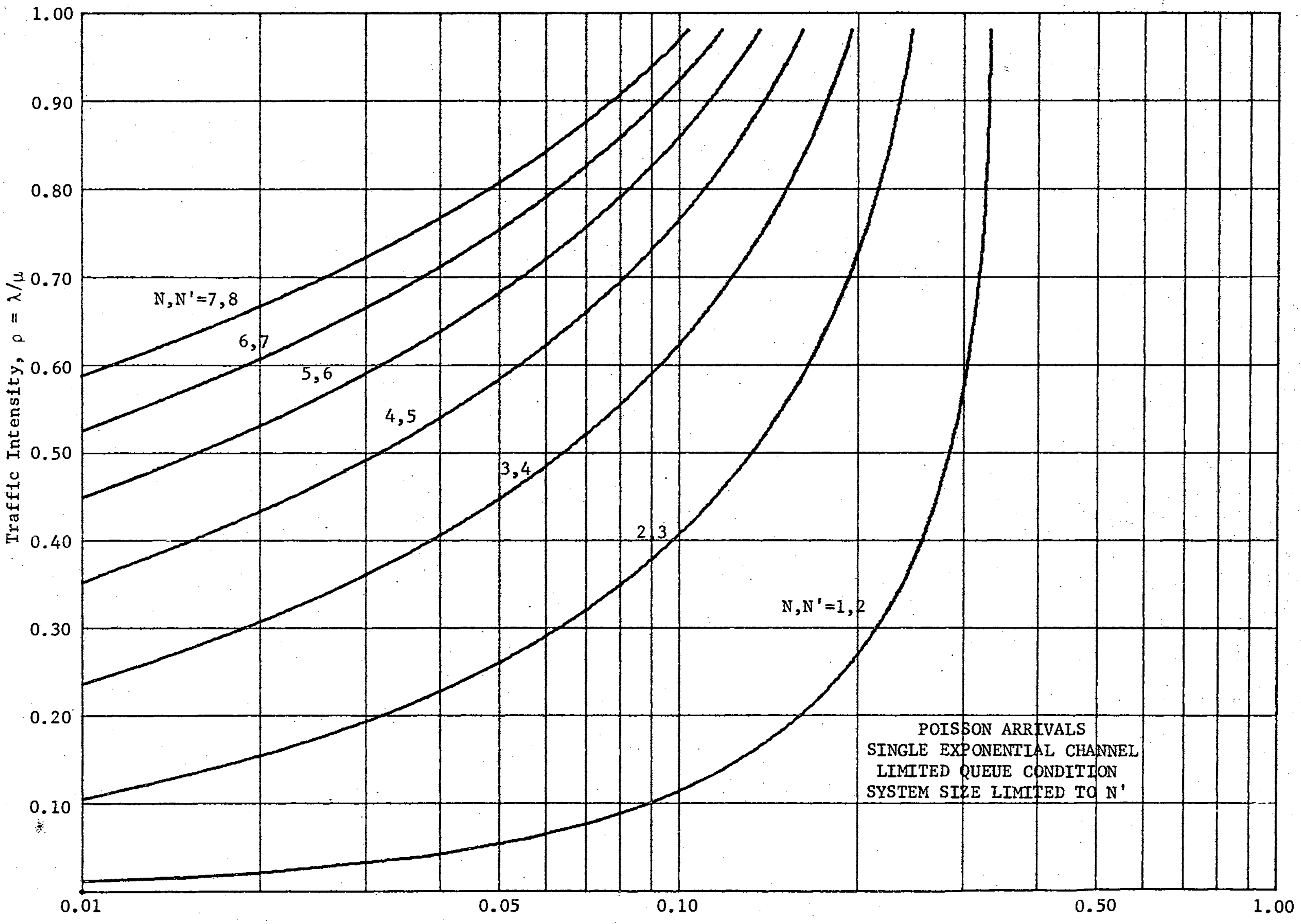


Figure 20. $Q_{N,N'}$, The Probability of N or More Units in the System, N = N' - 1
Single Exponential Channel, Limited Queue Allowed

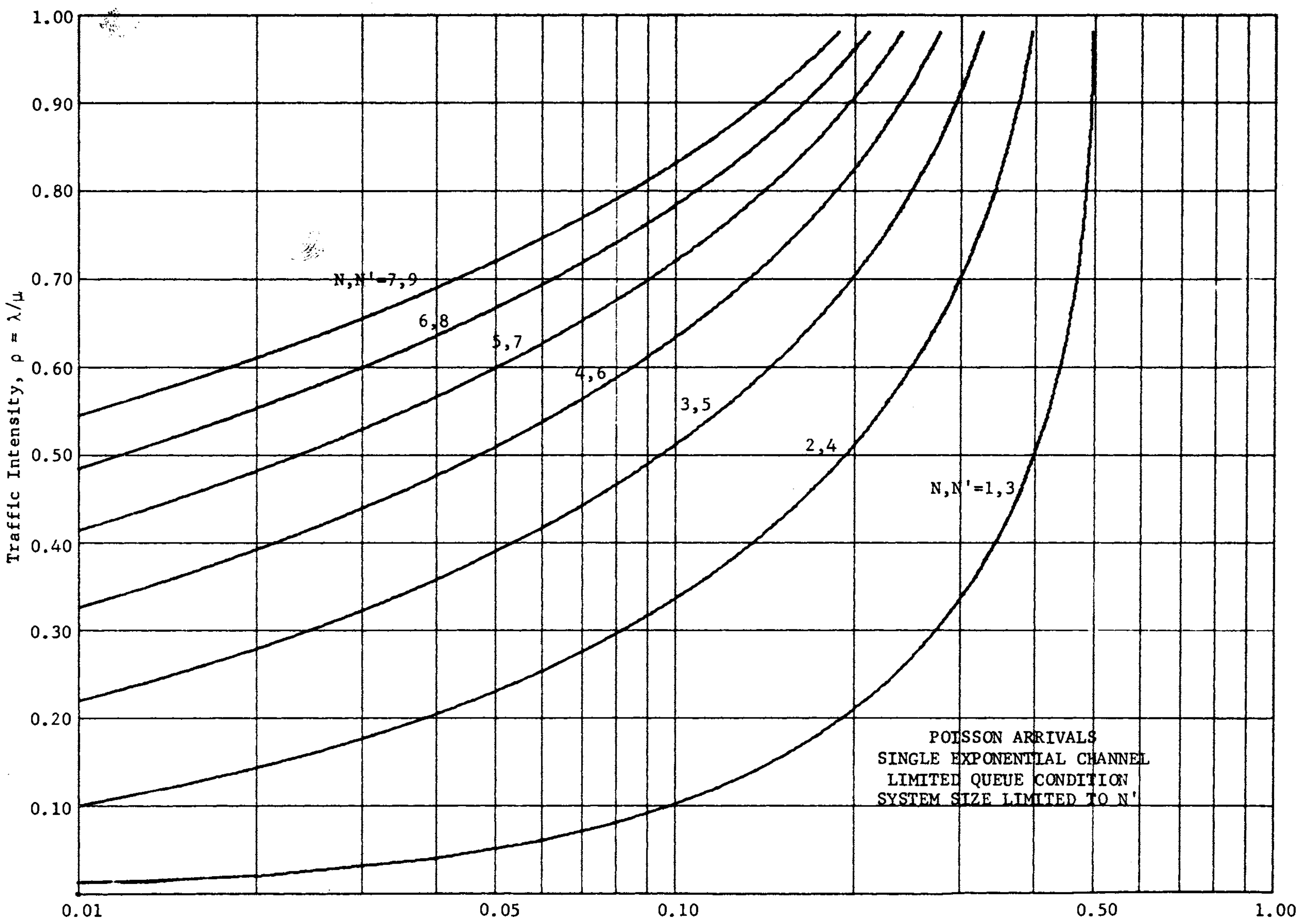


Figure 21. $Q_{N,N'}$, The Probability of N or More Units in the System, N = N' - 2
Single Exponential Channel, Limited Queue Allowed

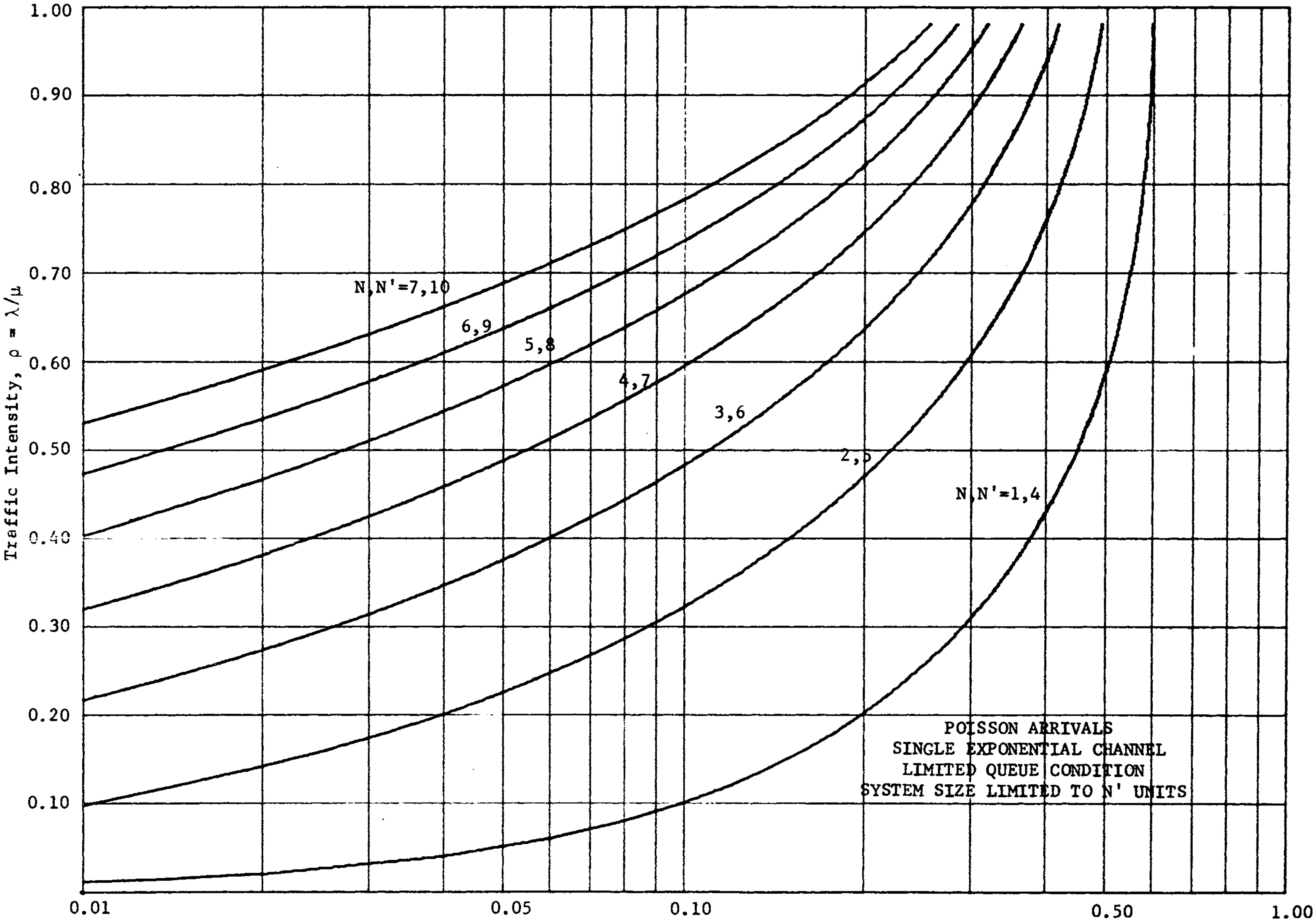


Figure 22. $Q_{N,N'}$, The Probability of N or More Units in the System, N = N' - 3
Single Exponential Channel, Limited Queue Allowed

from the intersection, $Q_{2,3}$ is on the abscissa and indicates the proportion of time that a system limited to size 3 contains either 2 or 3 customers. Stated in other words, $Q_{2,3}$ is the probability of this particular system being 2/3 or 67% or more full. The use of Figures 18 thru 22 as governing system relationships are identical to those presented in previous sections.

The Entry Rate: λ_e

Since all customers who arrive when the system is full are lost, the arrival rate to the system, λ, is not the same as the arrival rate to the service facility. Thus, λ_e is defined as the "entry rate" at the service facility in a limited queue system. If the system is never full, the arrival rate must equal the entry rate as no customers are lost. It may therefore be concluded that the entry rate is directly proportional to the amount of time that the system is not full, or:

$$\lambda_e = \lambda \sum_{n=0}^{N'-1} P_n = \lambda(1 - P_{N'}) \qquad (5.18)$$

where values of $P_{N'}$ may be found in Figure 18. In Equation (5.18), it is seen that λ_e is always less than λ. Thus, the consequence of a limited queue is to increase the mean interval between arrivals or decrease the total number of customers serviced per unit time. For this reason, all measures of effectiveness for limited queue conditions are less than the corresponding measures where infinite queues are allowed. As the length of queue allowed becomes more limited, N' decreases and the corresponding decrease in measures of effectiveness from the infinite queue condition become larger. This is clearly evident in the comparison of related figures for the limited and infinite queue conditions.

Measures of Effectiveness: L, L_q, W and W_q

The mean number of units in the system may be expressed as:

$$L = \sum_{n=0}^{N'} n(P_n) = \frac{1 - (N' + 1)\rho^{N'} + N'\rho^{N'+1}}{(1 - \rho)(1 - \rho^{N'+1})} \tag{5.19}$$

A family of curves for L are shown in Figure 23 for several values of N'. The arrival rate, λ, is used for determining ρ as Equation (5.19) accounts for lost customers. The use of Figure 23 is identical to the use of Figure 13.

The mean number of units in the queue, L_q, was developed as $L - \rho$ for infinite queue systems where ρ represented the proportion of time that the system was occupied with one unit in service. Thus, L_q is logically equal to L minus the proportion of time that the facility is busy multiplied by one, representing the expected number of units in service. The proportion of time that the facility is occupied for limited queue systems is $1 - P_0$ where P_0 is defined in Equation (5.15). Hence, L_q may be expressed as:

$$L_q = L - 1 + P_0 \tag{5.20a}$$

$$L_q = \rho^2 \left[\frac{1 - N'\rho^{N'-1} + (N' - 1)\rho^{N'}}{(1 - \rho)(1 - \rho^{N'})} \right] \tag{5.20b}$$

L_q may be determined from Equation (5.20a) by using Figures 18 and 23. The expression for Equation (5.20b), however, is shown as a family of

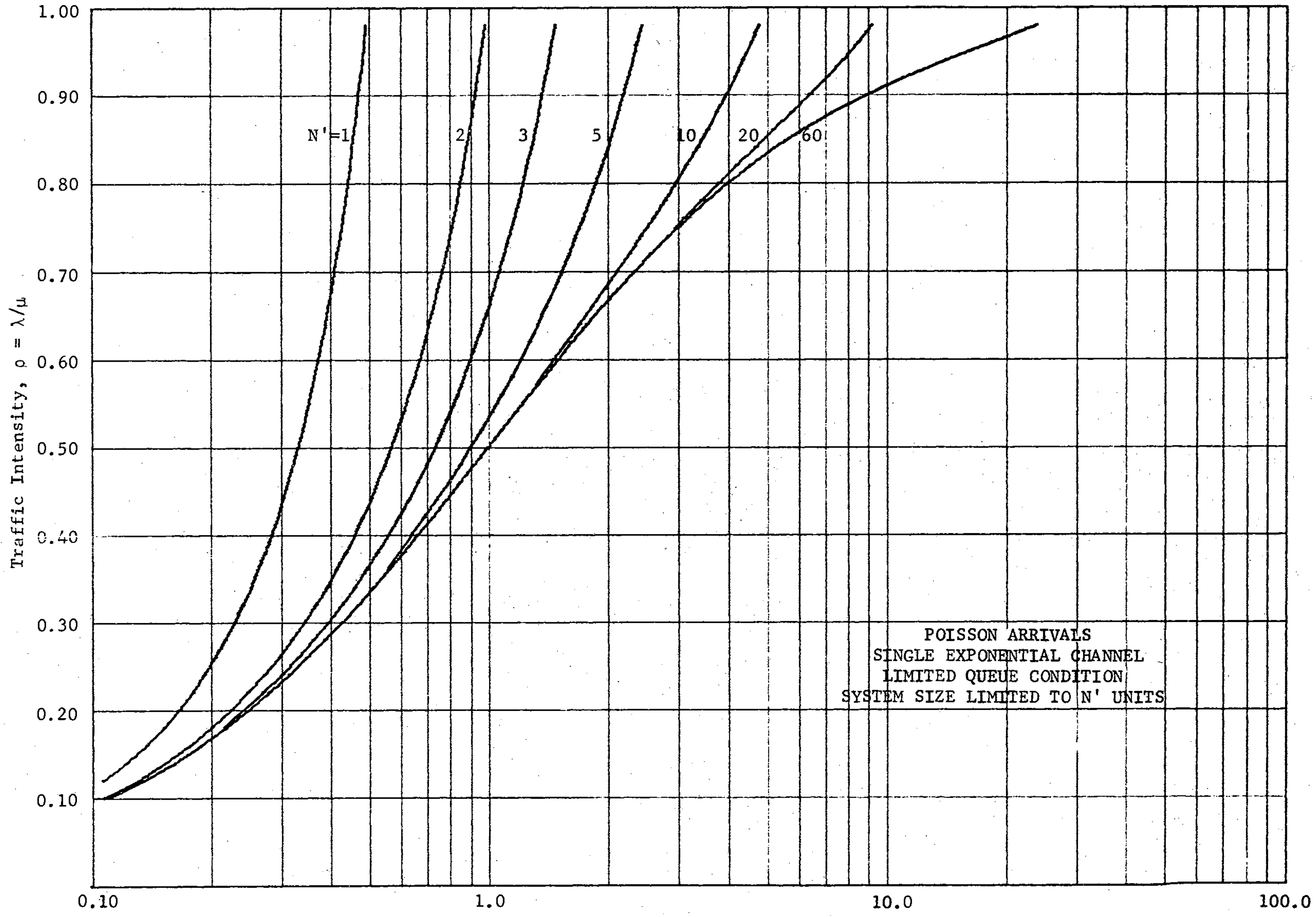


Figure 23. L, The Mean Number of Units in the System, Single Exponential Channel, Limited Queue Allowed

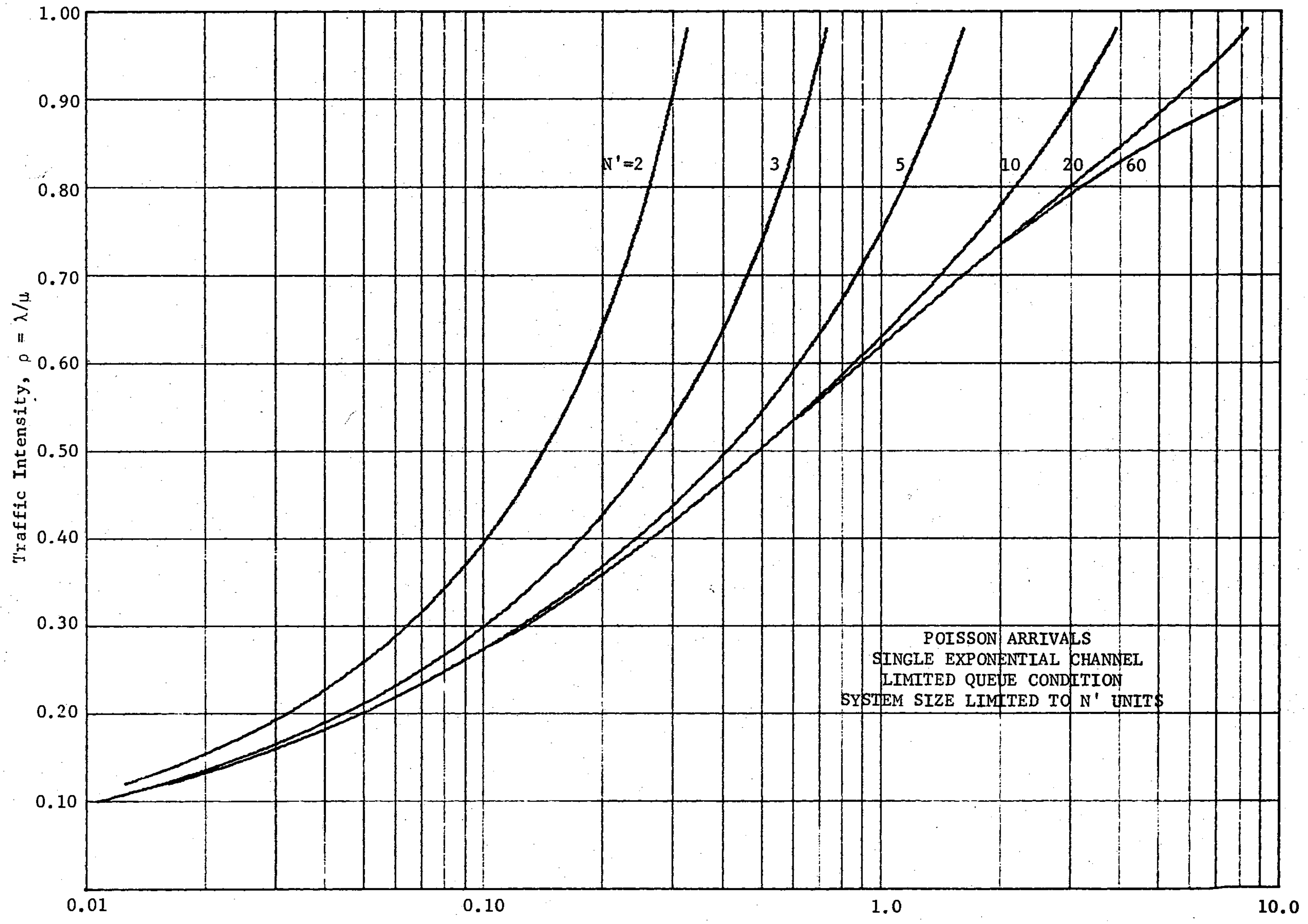


Figure 24. L_q, The Mean Number of Units in the Queue, Single Exponential Channel, Limited Queue Allowed

curves for several values of N' in Figure 24.

The mean wait time in the system and in the queue may be expressed as Equations (5.21a) and (5.21b) respectively.

$$W = L/\lambda_e \tag{5.21a}$$

$$W_q = L_q/\lambda_e \tag{5.21b}$$

Problem No. 9: Using the same ticket counter introduced in previous problems and making the assumption that all customers arriving when there are three persons in the system are lost, what general conclusions may be made of system behavior?

Solution: From previous problems, T_a = 10 minutes; T_s = 5 minutes; λ = .10; μ = .20; ρ = .50; and the arrival and service distributions are exponential. Where direct comparisons are applicable, solutions for the infinite queue condition are shown in brackets.

N' = 3. From Figure 18, P_0 = 0.53 which indicates that the facility will be completely idle 53% of the time. [$P_0 = 1 - \rho = .50$] From Figure 19, $P_{N'}$ = .067 which indicates that 6.7% of all arriving customers will be lost to the system. It also indicates that 1 - .067 or 93.3% of the time, there will be 3 or less units in the system. [$P_{3 \text{ or less}} = .9375$] From Figure 20, $Q_{2,3}$ = .133 which indicates that 13.3% of the time, there will be two or three units in the system. [Q_3 = .250, which includes two to an infinite number of units in the system] From Figure 23, L = .72 customers. [L = 1.0 customers] From Figure 24, L_q = .25 customers. Alternatively, $L_q = L - 1 + P_0 = .72 - 1.0 + .53 = .25$ customers. [$L_q = 1 - \rho = .50$] $\lambda_e = (1 - P_N') = .10(1.0 - .067) = .0933$. $W = L/\lambda_e = .72/.0933 = 7.72$ minutes.

[W = 10.0 minutes] $W_q = L_q/\lambda_e = .25/.0933 = 2.68$ minutes. [$W_q = 5.0$ minutes]

Comments: As expected, all measures of effectiveness in this example are less than those found with an infinite queue allowed. As N' is increased, the measures of effectiveness will approach the infinite queue case, in this case at about N' = 6. Any measure of effectiveness may be predetermined as a governing limit from which all other measures may be found.

CHAPTER VI

MULTIPLE EXPONENTIAL CHANNEL MODELS

The most common queueing systems found in architecture are those with multiple service channels in parallel. Banks with several tellers, supermarket checkout stands, entrances and exits of stadiums, parking lots, and convention halls, or an array of ticket boothes are all examples of multiple channel service facilities. When the arriving unit has the option of receiving service from any one of several channels, the channels are said to be in parallel.

This chapter will consider systems having M, parallel exponential channels, each with equal mean service rates μ. Arriving units are assumed to follow a Poisson distribution with mean arrival rate λ. If the operational system allows a queue, a single queue is formed when all M channels are occupied. Customers are served in a first come, first serve discipline. Upon reaching the front of the queue, the customer will depart the queue and enter the first unoccupied service channel.

As in single channel systems, the state of the system is characterized by n, the total number of units present. When n is less than M, there is no queue since all units present are occupying a service channel. When n is larger than M, there is a queue of length N_q equal to the quantity $(n - M)$. Note that the single channel systems of Chapter V are the special case of M equal to one.

Traffic Intensity for Multiple Channel Models

Traffic intensity for the entire system is defined as the ratio of the mean rate of arrivals and the maximum possible rate of service for all M channels combined. System traffic intensity may be expressed as:

$$\rho = \lambda/M\mu \qquad (6.1a)$$

The traffic intensity for a single service channel is λ/μ and is represented by the symbol φ. In terms of Equation (6.1a), φ may be expressed as:

$$\varphi = \lambda/\mu = M\rho \qquad (6.1b)$$

from which Equation (6.1a) may be rewritten,

$$\rho = \varphi/M \qquad (6.1c)$$

As discussed in Chapter V, a steady state solution is obtained only when ρ is less than one. All models considered in this study will satisfy this condition where the combined rate of all M channels is greater than the rate of arrivals.

The General State Probabilities

The general state probabilities for any multiple channel system with exponential services and Poisson arrivals may be expressed as:

$$P_n = \frac{(\varphi)^n}{n!} P_0 \qquad (0 \le n \le M) \qquad (6.2a)$$

$$P_n = \frac{M^M}{M!} \rho^n P_0 \qquad (M \le n \le N) \qquad (6.2b)$$

where P_0 is the probability of no units in the system and N is the maximum size of the system if it is limited.

The Functional Variables

To express subsequent measures of effectiveness compactly and to facilitate their computation, a set of functional variables, $E_m(x)$, $D_m(x)$, and $e_n(x)$ are defined as:

$$E_m(x) = e^{-x} \sum_{n=0}^{m} (x^n/n!) \qquad (6.3)$$

$$D_{m-1}(x) = E_m(x) - E_{m-1}(x)[x/m] \qquad (6.4)$$

$$e_n(x) = (x^n e^{-x}/n!) \qquad (6.5)$$

The definitions above are in general form. The function (x) is always φ or $M\rho$, the traffic intensity of one service facility. The numeric expansion of these functional variables are found in Table VI as functions of ρ and M. Linear interpolation may be accomplished with no significant loss of accuracy.

Given a system of four service channels ($M = 4$), each with a traffic intensity of 2.0 ($\varphi = 2.0$), system traffic intensity, ρ, would equal φ/M or $2.0/4.0 = 0.50$. From Table VI, $E_M(\varphi) = 0.9473$; $E_{M-1}(\varphi) = 0.8571$; $E_{M+1}(\varphi) = 0.9834$; $D_{M-1}(\varphi) = 0.5188$; and $e_M(\varphi) = 0.0902$. For large systems of $M \geq 100$ and $\rho < 0.80$, the functional variables approach the constants listed in the last line of Table VI.

The remainder of this chapter will be concerned with two extreme queueing conditions. In the first, system size will be limited to M or less units and no queue will be allowed to form. In the second, the formation of an infinite queue will be allowed.

TABLE VI

FUNCTIONAL VARIABLES

M	ρ	$E_M(\varphi)$	$E_{M-1}(\varphi)$	$E_{M+1}(\varphi)$	$D_{M-1}(\varphi)$	$e_M(\varphi)$
1	.10	.9953	.9048	.9998	.9048	.0905
1	.20	.9825	.8187	.9989	.8187	.1637
1	.30	.9631	.7408	.9964	.7408	.2222
1	.40	.9384	.6703	.9921	.6703	.2681
1	.50	.9098	.6065	.9856	.6065	.3033
1	.60	.8781	.5488	.9769	.5488	.3293
1	.70	.8442	.4966	.9659	.4966	.3476
1	.80	.8088	.4493	.9526	.4493	.3595
1	.90	.7725	.4066	.9371	.4066	.3659
1	1.00	.7358	.3679	.9197	.3679	.3679
2	.10	.9989	.9825	.9999	.9006	.0164
2	.20	.9921	.9384	.9992	.8044	.0536
2	.30	.9769	.8781	.9966	.7135	.0988
2	.40	.9526	.8088	.9909	.6291	.1438
2	.50	.9197	.7358	.9810	.5518	.1839
2	.60	.8795	.6626	.9662	.4819	.2169
2	.70	.8335	.5918	.9463	.4192	.2417
2	.80	.7834	.5249	.9212	.3634	.2584
2	.90	.7306	.4628	.8913	.3141	.2678
2	1.00	.6767	.4060	.8571	.2707	.2707
3	.10	.9997	.9964	1.0000	.9001	.0033
3	.20	.9966	.9769	.9996	.8013	.0198
3	.30	.9865	.9371	.9977	.7054	.0494
3	.40	.9662	.8795	.9923	.6144	.0867
3	.50	.9344	.8088	.9814	.5299	.1255
3	.60	.8913	.7306	.9636	.4529	.1607
3	.70	.8386	.6496	.9379	.3839	.1890
3	.80	.7787	.5697	.9041	.3230	.2090
3	.90	.7141	.4936	.8629	.2698	.2205
3	1.00	.6472	.4232	.8153	.2240	.2240
4	.10	.9999	.9992	1.0000	.9000	.0007
4	.20	.9986	.9909	.9998	.8004	.0077
4	.30	.9923	.9662	.9985	.7024	.0260
4	.40	.9763	.9212	.9940	.6078	.0551
4	.50	.9473	.8571	.9834	.5188	.0902
4	.60	.9041	.7787	.9643	.4369	.1254
4	.70	.8477	.6919	.9349	.3633	.1557
4	.80	.7806	.6025	.8946	.2986	.1781
4	.90	.7064	.5152	.8441	.2427	.1912
4	1.00	.6288	.4335	.7851	.1954	.1954
5	.10	1.0000	.9998	1.0000	.9000	.0002
5	.20	.9994	.9963	.9999	.8001	.0031
5	.30	.9955	.9814	.9991	.7011	.0141
5	.40	.9834	.9473	.9955	.6045	.0361
5	.50	.9580	.8912	.9858	.5124	.0668
5	.60	.9161	.8153	.9665	.4269	.1008
5	.70	.8576	.7254	.9347	.3498	.1322
5	.80	.7851	.6288	.8893	.2821	.1563
5	.90	.7029	.5321	.8311	.2240	.1708
5	1.00	.6160	.4405	.7622	.1755	.1755

TABLE VI - Continued

M	ρ	$E_M(\varphi)$	$E_{M-1}(\varphi)$	$E_{M+1}(\varphi)$	$D_{M-1}(\varphi)$	$e_M(\varphi)$
10	.10	1.0000	1.0000	1.0000	.9000	0.0000
10	.20	1.0000	1.0000	1.0000	.8000	0.0000
10	.30	.9997	.9989	.9999	.7000	.0008
10	.40	.9972	.9919	.9991	.6004	.0053
10	.50	.9863	.9682	.9945	.5022	.0181
10	.60	.9574	.9161	.9799	.4077	.0413
10	.70	.9015	.8305	.9467	.3201	.0710
10	.80	.8159	.7166	.8881	.2426	.0993
10	.90	.7060	.5874	.8030	.1773	.1186
10	1.00	.5830	.4579	.6968	.1251	.1251
20	.10	1.0000	1.0000	1.0000	.9000	0.0000
20	.20	1.0000	1.0000	1.0000	.8000	0.0000
20	.30	1.0000	1.0000	1.0000	.7000	0.0000
20	.40	1.0000	1.0000	1.0000	.6000	0.0000
20	.50	.9984	.9965	.9993	.5001	.0019
20	.60	.9884	.9787	.9939	.4012	.0097
20	.70	.9521	.9235	.9712	.3056	.0286
20	.80	.8682	.8122	.9108	.2184	.0559
20	.90	.7307	.6509	.7991	.1449	.0798
20	1.00	.5591	.4703	.6437	.0888	.0888
40	.10	1.0000	1.0000	1.0000	.9000	0.0000
40	.20	1.0000	1.0000	1.0000	.8000	0.0000
40	.30	1.0000	1.0000	1.0000	.7000	0.0000
40	.40	1.0000	1.0000	1.0000	.6000	0.0000
40	.50	1.0000	1.0000	1.0000	.5000	0.0000
40	.60	.9990	.9983	.9995	.4001	.0007
40	.70	.9875	.9810	.9920	.3008	.0065
40	.80	.9293	.9044	.9488	.2058	.0249
40	.90	.7771	.7263	.8217	.1234	.0508
40	1.00	.5419	.4790	.6033	.0629	.0629
60	.10	1.0000	1.0000	1.0000	.9000	0.0000
60	.20	1.0000	1.0000	1.0000	.8000	0.0000
60	.30	1.0000	1.0000	1.0000	.7000	0.0000
60	.40	1.0000	1.0000	1.0000	.6000	0.0000
60	.50	1.0000	1.0000	1.0000	.5000	0.0000
60	.60	1.0000	1.0000	1.0000	.4000	0.0000
60	.70	.9965	.9948	.9977	.3002	.0017
60	.80	.9605	.9477	.9706	.2024	.0128
60	.90	.8133	.7760	.8463	.1149	.0373
60	1.00	.5343	.4828	.5849	.0514	.0514
100	.80	.9869	.9829	.9900	.2005	.0039
100	.82	.9768	.9705	.9819	.1810	.0063
100	.84	.9611	.9516	.9690	.1618	.0095
100	.86	.9383	.9248	.9498	.1430	.0135
100	.88	.9066	.8884	.9225	.1248	.0182
100	.90	.8651	.8418	.8859	.1075	.0233
100	.92	.8134	.7849	.8393	.0912	.0284
100	.94	.7518	.7187	.7825	.0762	.0330
100	.96	.6818	.6451	.7167	.0625	.0367
100	.98	.6058	.5667	.6437	.0504	.0391
100	1.00	.5266	.4867	.5661	.0399	.0399

TABLE VI - Continued

M	ρ	$E_M(\varphi)$	$E_{M-1}(\varphi)$	$E_{M+1}(\varphi)$	$D_{M-1}(\varphi)$	$e_M(\varphi)$
200	.80	.9993	.9990	.9995	.2001	.0003
200	.82	.9974	.9967	.9980	.1801	.0007
200	.84	.9930	.9914	.9944	.1602	.0016
200	.86	.9837	.9804	.9864	.1405	.0032
200	.88	.9658	.9599	.9710	.1211	.0059
200	.90	.9351	.9254	.9437	.1022	.0097
200	.92	.8873	.8730	.9005	.0842	.0143
200	.94	.8199	.8005	.8380	.0674	.0194
200	.96	.7329	.7090	.7558	.0523	.0239
200	.98	.6303	.6032	.6567	.0392	.0271
200	1.00	.5190	.4908	.5471	.0282	.0282
300	.80	1.0000	1.0000	1.0000	.2000	0.0000
300	.82	1.0000	1.0000	1.0000	.1800	0.0000
300	.84	.9992	.9989	.9994	.1601	.0003
300	.86	.9958	.9950	.9966	.1402	.0009
300	.88	.9871	.9849	.9890	.1204	.0022
300	.90	.9673	.9627	.9715	.1009	.0046
300	.92	.9291	.9207	.9368	.0820	.0084
300	.94	.8650	.8519	.8773	.0642	.0131
300	.96	.7714	.7533	.7886	.0481	.0180
300	.98	.6513	.6297	.6725	.0343	.0217
300	1.00	.5158	.4928	.5388	.0231	.0231
400	.80	1.0000	1.0000	1.0000	.2000	0.0000
400	.82	1.0000	1.0000	1.0000	.1800	0.0000
400	.84	1.0000	1.0000	1.0000	.1600	0.0000
400	.86	1.0000	.9997	1.0002	.1402	.0003
400	.88	.9959	.9950	.9967	.1203	.0009
400	.90	.9839	.9815	.9860	.1005	.0023
400	.92	.9549	.9498	.9597	.0811	.0052
400	.94	.8974	.8880	.9063	.0627	.0094
400	.96	.8022	.7878	.8159	.0459	.0144
400	.98	.6698	.6514	.6878	.0315	.0184
400	1.00	.5142	.4942	.5341	.0200	.0200
500	.80	1.0000	1.0000	1.0000	.2000	0.0000
500	.82	1.0000	1.0000	1.0000	.1800	0.0000
500	.84	1.0000	1.0000	1.0000	.1600	0.0000
500	.86	1.0000	1.0000	1.0000	.1400	0.0000
500	.88	.9999	.9996	1.0002	.1203	.0004
500	.90	.9928	.9916	.9939	.1004	.0012
500	.92	.9716	.9683	.9747	.0808	.0033
500	.94	.9216	.9146	.9281	.0619	.0070
500	.96	.8278	.8159	.8391	.0445	.0119
500	.98	.6863	.6702	.7022	.0296	.0162
500	1.00	.5134	.4955	.5312	.0179	.0179
100 to 500	0.01 to 0.79	1.0000	1.0000	1.0000	1.0 - ρ	0.0000

The No Queue Condition

The particular condition of no queue allowed receives special attention because of its application to a common architectural system - the parking lot. To some extent, it also applies to certain restaurants. Where no queues are allowed, system size has been limited to the number of available service channels, M. It is assumed that all arrivals occuring when all M service channels are occupied immediately depart the system. In a sense, they are refused service and entry to the system and may be considered as lost customers. In parking lots and restaurants, the number of available channels corresponds to the number of parking spaces or boothes and tables, respectively.

It should be recognized that restaurant arrivals must be considered as bulk arrivals and not as individual units. The arriving unit is composed of any number of individuals that will occupy one service channel. The no queue condition is schematically illustrated in Figure 25.

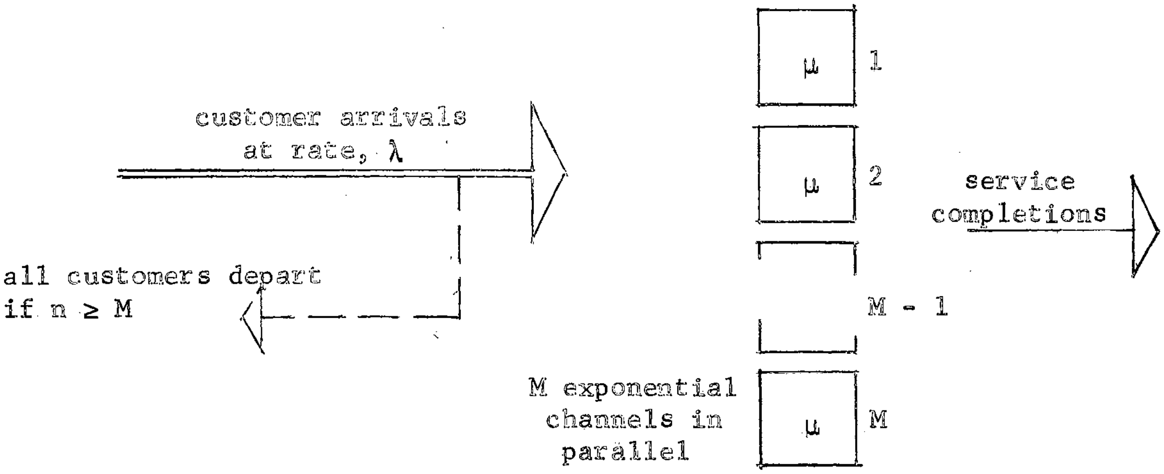


Figure 25. Multiple Channels, No Queue Allowed

The State Probabilities: P_0, P_n, P_M

By using the property that $\sum_{n=0}^{M} P_n = 1.0$, Equation (6.2a) may be solved for P_0 in terms of the functional variables as:

$$P_0 = e^{-\varphi}/E_M(\varphi) \tag{6.6}$$

P_0 indicates the proportion of time that there are no units in the system. System facility utilization, $(1 - P_0)$, indicates the proportion of time at least one service channel is occupied and is a measure of relative system efficiency. A family of curves for P_0 have been plotted in Figure 26 for several values of M.

Substituting Equation (6.6) into Equation (6.2a), the general state probability of exactly n units in the system may be expressed as:

$$P_n = e_n(\varphi)/E_M(\varphi). \tag{6.7}$$

The probability that all channels are occupied is the probability of exactly M units in the system. From Equation (6.7), P_M may be expressed as:

$$P_M = e_M(\varphi)/E_M(\varphi). \tag{6.8}$$

Since all customers arriving when the system is full will depart immediately, P_M indicates the proportion of lost customers. A family of curves for P_M have been plotted in Figure 27 for several values of M.

Measures of Effectiveness: $Q_{M,N}$, L

The probability of N or more units in a system of M service channels, $Q_{M,N}$, may be determined by summing the state probability of

Equation (6.7) from N to M. Expressed in terms of the functional variables:

$$Q_{M,N} = \sum_{n=N}^{M} P_n = \sum_{n=N}^{M} e_n(\varphi)/E_M(\varphi) \tag{6.9}$$

A family of curves for $Q_{M,N}$ have been plotted in Figures 28 and 29. In Figure 28, the ratio of N/M is 75% so that the graphs indicate the proportion of time that a multiple channel system of size M is three-quarters or more full. Similarly, in Figure 29, the ratio of N/M is 50% so that the graphs indicate the proportion of time that the system is one-half or more full. Thus, Figures 27 to 29 allow the investigation of system performance at the 100%, 75% and 50% levels of system capacity.

The mean number of units in the system is equivalent to the mean number of occupied channels and may be expressed as:

$$L = \sum_{n=0}^{M} n(P_n) = \frac{(\varphi)E_{M-1}(\varphi)}{E_M(\varphi)} \tag{6.10}$$

A family of curves for L have been plotted in Figure 30 for several values of M. The mean time spent in the system may be expressed in terms of L as:

$$W = L/\lambda_e \tag{6.11}$$

where λ_e has been previously defined as $\lambda(1.0 - P_M)$. Since no queue is allowed, L_q and W_q must equal zero. The use of Figures 26 to 30 are best illustrated through the presentation of the following example problems.

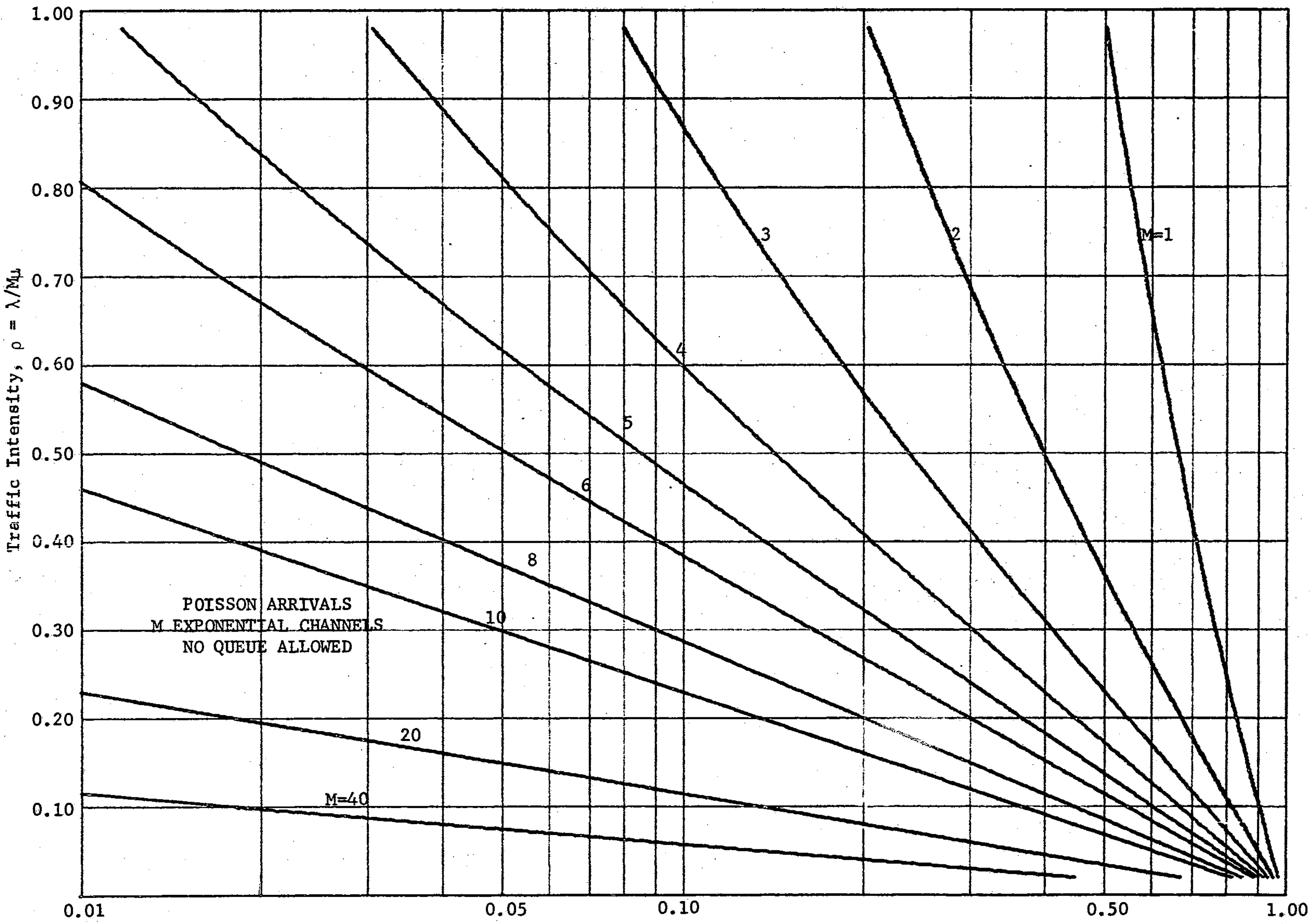


Figure 26. P_0, The Probability of No Units in the System, M Exponential Channels in Parallel, No Queue Allowed

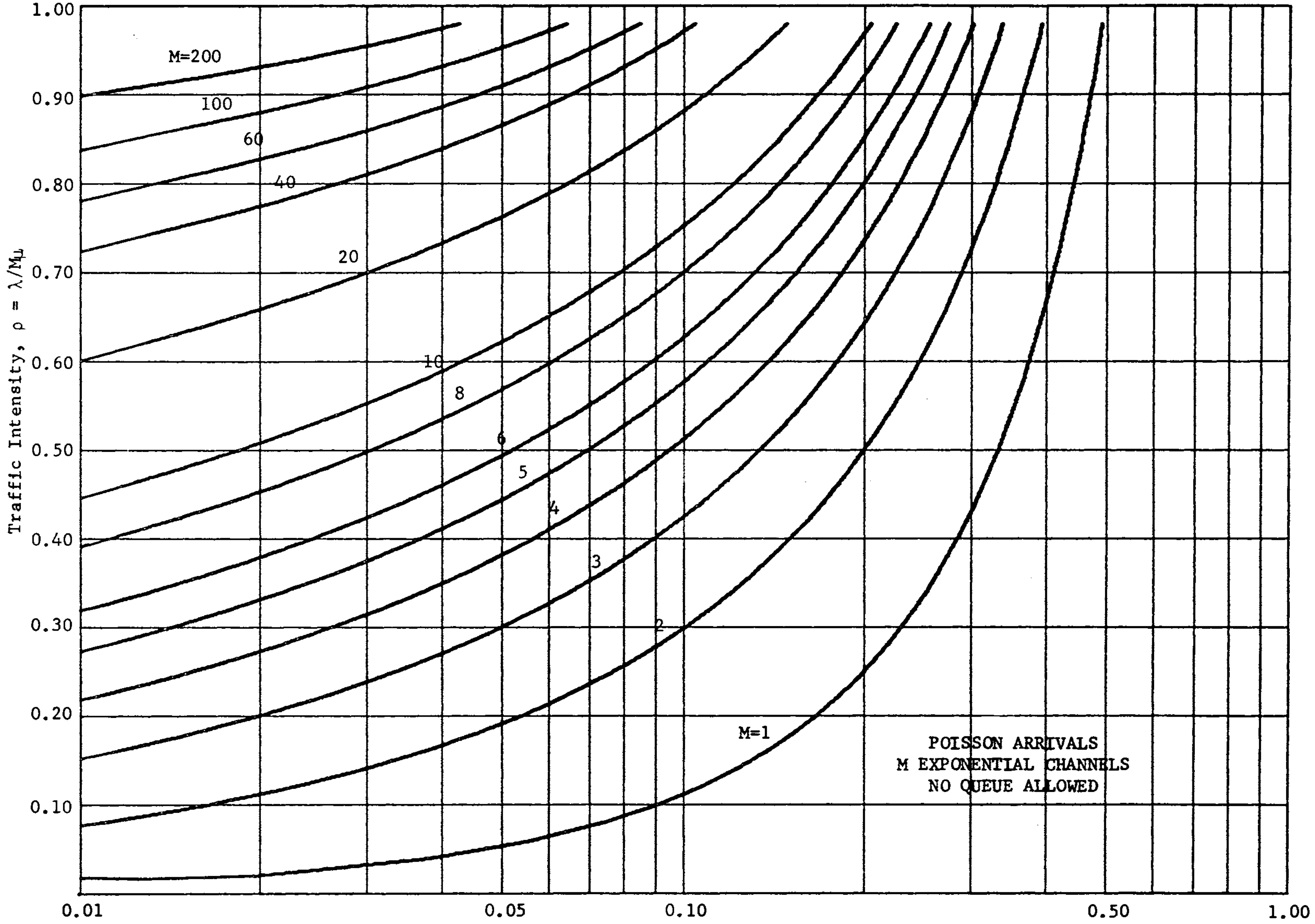


Figure 27. P_M, The Probability that the System is Full, M Exponential Channels in Parallel, No Queue Allowed

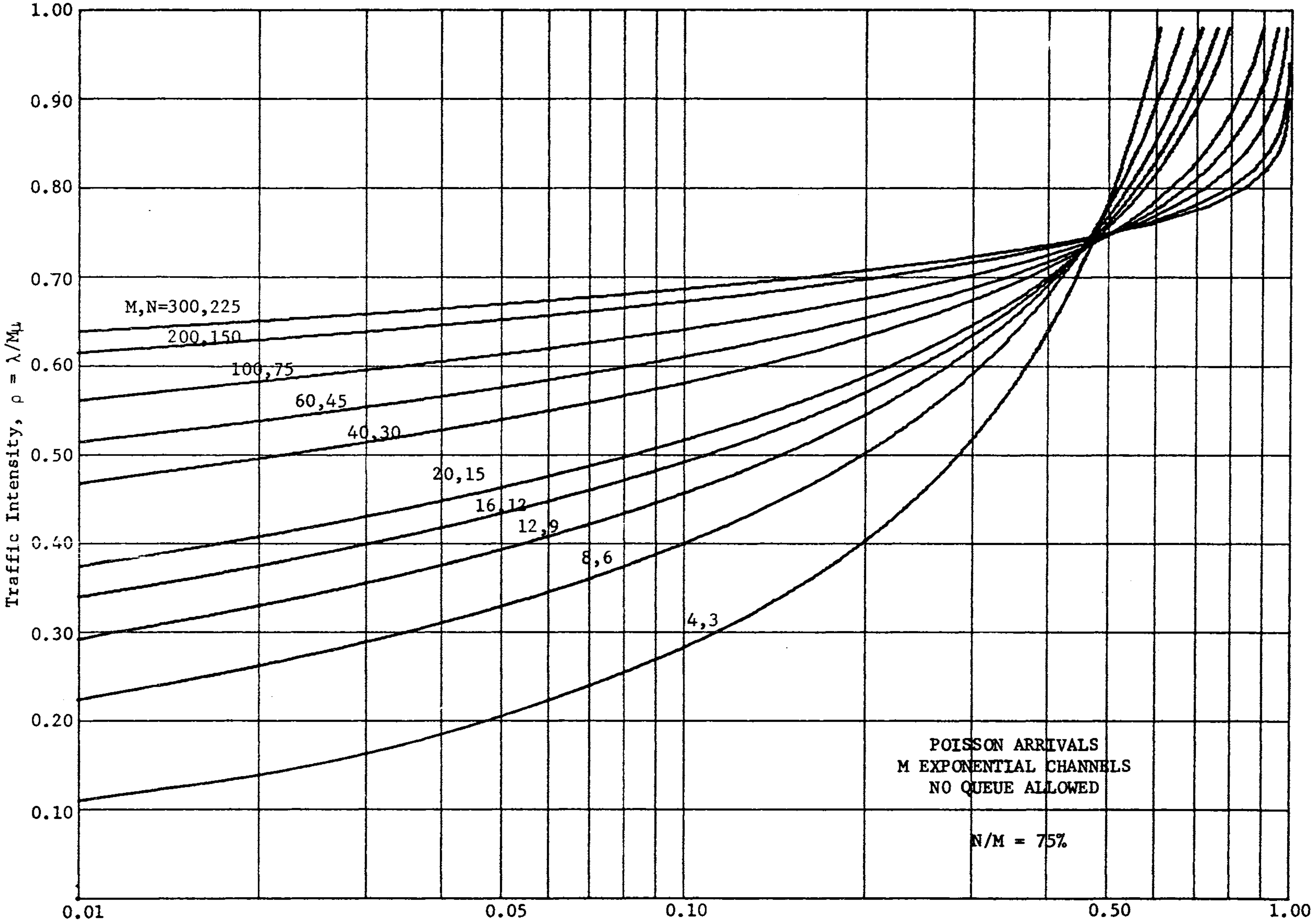


Figure 28. $Q_{M,N}$, The Probability of N or more Units in a System of M Exponential Channels in Parallel, No Queue Allowed, N/M = 75%

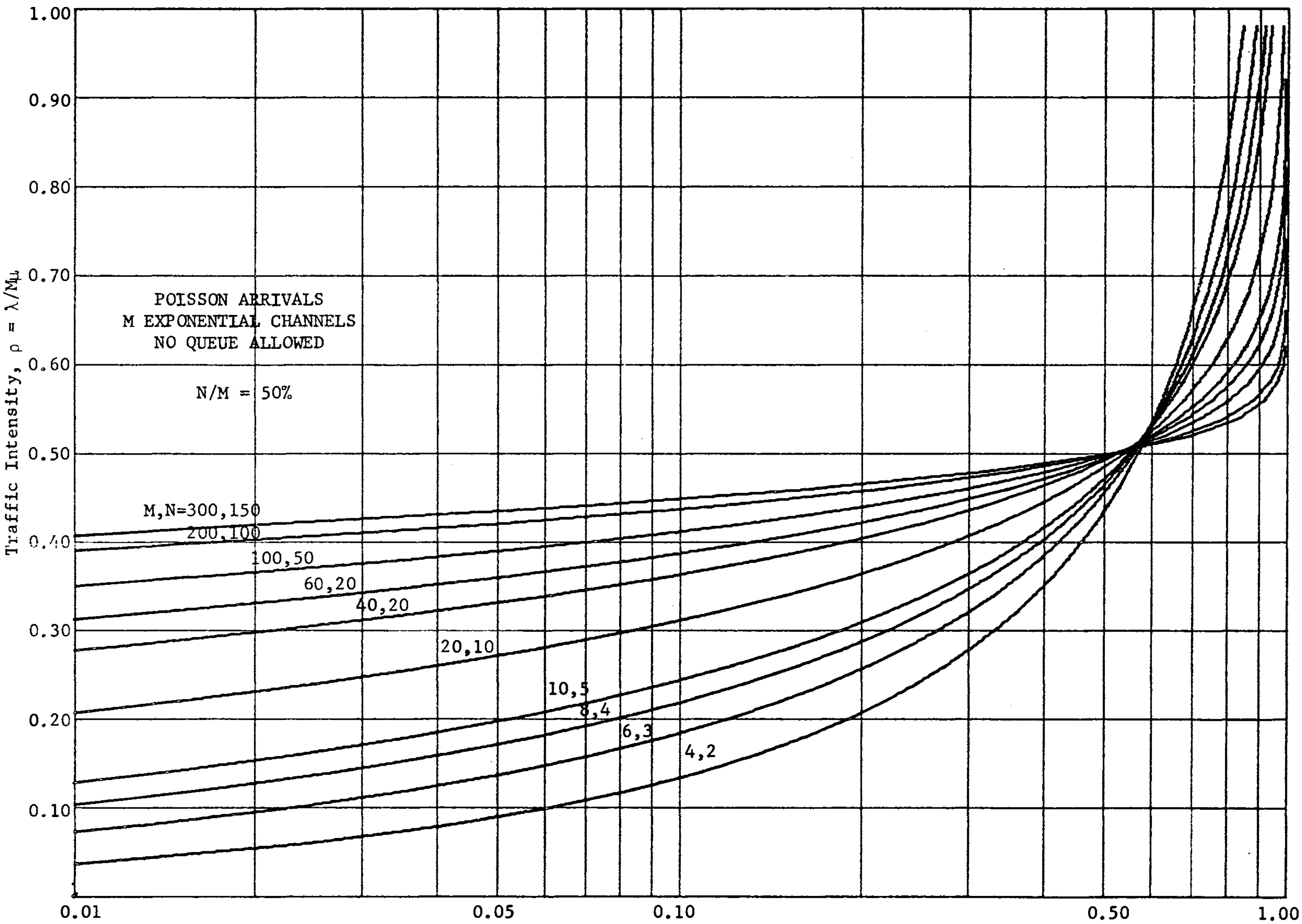


Figure 29. $Q_{M,N}$, The Probability of N or more Units in a System of M Exponential Channels in Parallel, No Queue Allowed, N/M = 50%

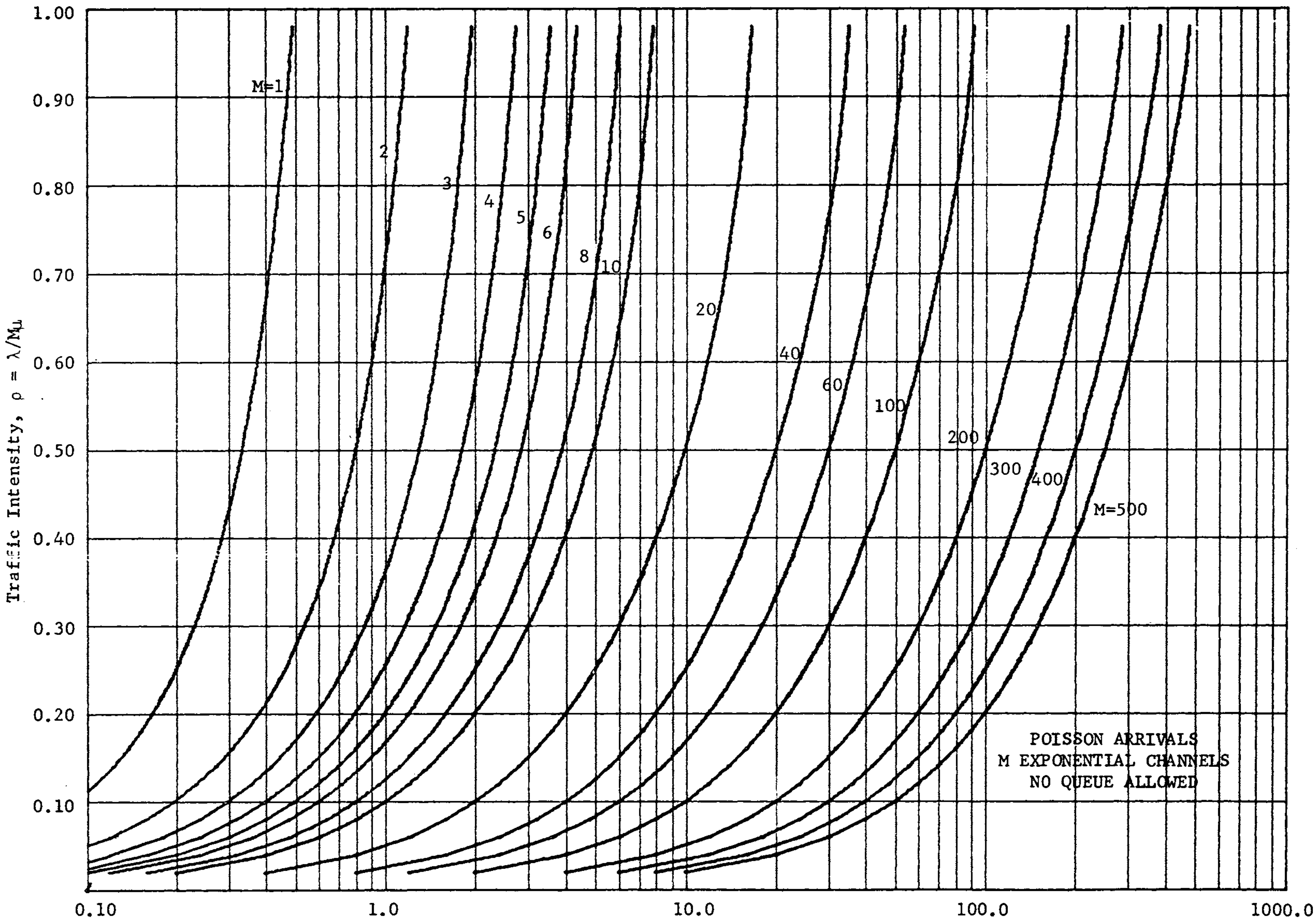


Figure 30. L, The Mean Number of Units in the System, M Exponential Channels in Parallel, No Queue Allowed

Problem No. 1: A small commercial establishment is to be provided with eight parking spaces. At the present time, it is estimated that vehicles will arrive at the rate of eight per hour and remain for an average duration of 30 minutes. The client advises that within the next two years, a 50% increase in business is anticipated. It is assumed that there will be a corresponding increase in the arrival rate to 12 vehicles per hour. Using Figures 26 to 30, what are the measures of effectiveness under the present conditions? To what extent will the anticipated increase in arrival rate influence system performance?

Solution: Enter all Figures using the appropriate value of ρ and the curve for M = 8. The results are listed below.

	Present	Future
arrival rate, λ	8.0	12.0
service rate, μ	2.0	2.0
system traffic intensity, $\rho = \lambda/M\mu$	0.50	0.75
P_0 (Figure 26)	0.018	$<<$ 0.01
P_M (Figure 27)	0.03	0.12
$Q_{8,6}$ (Figure 28)	0.02	0.48
$Q_{8,4}$ (Figure 29)	0.56	0.82
L (Figure 30)	3.80	5.20
$W = L/\lambda_e = L/\lambda_e(1.0 - P_M)$	0.490 hrs	0.492 hrs

Comments: When the combined service rate $M\mu$ is held constant, traffic intensity will increase as the arrival rate increases. Under this condition, Figure 26 logically indicates that P_0, the probability of no units in the system decreases as the arrival rate increases. With the present arrival rate, a completely empty lot is expected 1.8% of the time. As the arrival rate increases from eight to 12 vehicles per hour, P_0 becomes insignificant.

The most important measure of effectiveness for this problem is found in Figure 27. P_M indicates the proportion of arrivals who will find the system full and depart immediately. With the present arrival rate, 3% or three out of 100 arrivals will find all eight channels occupied and be lost to the system. With the future arrival rate, the number of lost customers will increase to 12%. Thus, a 50% increase in the arrival rate will result in a 300% increase in the total number of lost customers.

The proportion of time that 75% and 50% of system capacity is exceeded is given by $Q_{8,6}$ and $Q_{8,4}$ respectively. For example, with the present arrival rate, four or more vehicles are expected in the lot 56% of the time. With the future arrival rate, the same condition may be expected 82% of the time.

The average number of vehicles in the lot at any given time are 3.8 now and 5.2 in the future. Considering the system deterministically, that is, ignoring variation within the arrival and service rates, exactly eight arrivals and two service completions occur every hour. Thus, the mean number of units in a deterministic system would be 8.0/2.0 = 4.0 vehicles. Similarly, with the future arrival rate, the mean number of units in a deterministic system would be 12.0/2.0 = 6.0 vehicles. The difference between any probabilistic and deterministic analysis may be attributed to inherent variation in the arrival and service rates. In this instance, however, much of the difference involved may be attributed to the effect of lost customers. In the deterministic analysis, no customers are ever lost. The greater the proportion of lost customers, the greater the expected difference between probabilistic and deterministic analysis. With 3% lost customers, the difference is 4.0 - 3.8 = 0.2 while with 12% lost customers, the difference is 6.0 - 5.2 = 0.8.

Calculations for W show that unless a substantial proportion of customers are lost, the expected time spent in the system will be very close to the expected value of T_s.

Problem No. 2: The parking lot for a community medical center is to be designed so that no more than 5% of the arriving vehicles will find the lot full. It is estimated that arrivals will occur at 1 minute intervals and that each vehicle will remain an average of 30 minutes. How many spaces are required to satisfy the condition above? What proportion of time will the lot be more than 75% or 50% full?

Solution: $T_a = 1$ minute; $T_s = 30$ minutes

$\lambda = 1/T_a = 1/1 = 1$ arrival per minute

$\mu = 1/T_s = 1/30 = 0.033$ departures per minute

$\varphi = \lambda/\mu = 1/0.033 = 30$

$\rho = \varphi/M = 30/M$

The value of M must be known or assumed before entering any figure for multiple channel systems. Since only conditions of $\rho < 1$ are considered, and $\rho = \varphi/M$, the numeric value of M must be greater than the numeric value of φ. The most logical approach to problem solution is to assume a value of M just greater than φ, test the required condition, and increment M to repeat the process if necessary. Using Figure 27, this process is illustrated below.

try $M = 32$; $\rho = 30/32 = 0.94$; $P_{32} = 0.11$; new trial required

try $M = 34$; $\rho = 30/34 = 0.87$; $P_{34} = 0.07$; new trial required try $M = 36$; $\rho = 30/36 = 0.83$; $P_{36} = 0.05$; condition satisfied. Entering Figures 28 and 29 with $\rho = 0.83$ and $M = 36$, $Q_{36;27} = 0.67$ and $Q_{36;18} = 0.92$.

Comments: Summarizing the results, 36 spaces are required if no more than 5% of the arriving vehicles are to find the lot full. If 36 spaces are provided, 67% of the time, the lot will contain 27 or more vehicles and approximately 92% of the time, the lot will contain 18 or more vehicles. To give a useful description of system performance, the probability statements have been made at the 100%, 75%, and 50% capacity levels. In this particular example, the governing relationship was set at the 100% capacity level. The subsequent values of M and ρ automatically determines the probability of occurrence for the remaining two levels of capacity, that is, the probability of occurrence at all three levels of capacity are mutually interdependent. It should be understood that any one of the three levels of capacity may be used as the governing relationship.

The opportunity to evaluate these types of systems at three capacity levels have been provided because different systems are best evaluated at different levels. Small systems of less than 20 channels are generally most critical at the 100% capacity level. In some systems, notably school, hospital, and employee parking lots, lost customers will seek a parking space elsewhere and return to the desired destination. In these instances, an analysis at the 100% capacity level would be appropriate.

For very large parking lots, a sizeable proportion of the spaces must necessarily be located long distances from the desired destination. Especially where the facility is highly competitive, observation will show that customers will be lost even though the system is only 75% or 50% full. In these instances, it would be more appropriate to make the analysis at one of the lower levels of system capacity.

It is emphasized that Figures 28 and 29, the 75% and 50% capacity levels, are cumulative probabilities. $Q_{M,N}$ is the sum of individual state probabilities of from N to M units in the system. It is often difficult to comprehend the relationships between cumulative probabilities. Therefore, it is unwise to arbitrarily select either the 75% or 50% capacity level as the governing relationship without consideration of system behavior at the other levels of capacity. The best approach will be to make several analyses with incremental values of M at all three levels of system capacity. Several complete descriptions of the system will then be available for evaluation from which one may be selected as the final solution. This method is demonstrated in the following example problem.

Problem No. 3: In the preliminary design for a large shopping center complex, it is estimated that customers will enter the parking lot at 15 second intervals. It is further estimated that each parking space will be occupied for an average of 30 minutes. Market analysis indicates that the complex will be highly competitive in a suburban, private transportation-oriented area. Subjectively, what is the preferred range of parking spaces for the conditions above?

Solution: λ = 240 arrivals/hour; μ = 2 departures/hour; φ = 120; ρ = 120/M. Using the iterative procedure introduced in Problem No. 2, the results are listed in Table VII.

TABLE VII

ALTERNATIVE SOLUTIONS TO PROBLEM NO. 3

Analysis	Assumed Number of Spaces, M	Proportion of Time Capacity Level is Exceeded		
		Capacity Level		
		100%	75%	50%
1	140	1%	90%	100%
2	160	<1%	50%	100%
3	180	<<1%	12%	98%
4	200	<<1%	1%	96%
5	220	<<1%	<1%	85%
6	240	<<1%	<<1%	75%

Comments: If 140 spaces are provided, approximately 1% of the arriving vehicles will find the system full. The lot will be 75% or more full 90% of the time and almost always be more than 50% full. If the appearance of a full lot will discourage the entry of customers, 140 spaces will not be an adequate solution.

At the other extreme, the provision of 240 spaces will result in a system that is seldom more than 75% full. Only 75% of the time will the lot be more than 50% full. As the number of spaces provided increases beyond 160, the probability of a full lot becomes progressively more insignificant. Similarly, as the number of spaces provided increases beyond 200 spaces, the probability that the lot is more than 75% full becomes progressively more insignificant. In addition, all systems of more than 200 spaces are meaningful only at the 50% capacity level.

A preferred solution lies between 180 and 220 spaces. Even with the relatively small incremental values of M, substantial differences in the probable appearance of the lot are noted. The most meaningful indicators of system performance within this range are the proportion of times that the system is 50% or more full. The ultimate solution, in most cases, will be governed by costs and the geometry of site lay-out. However, using the procedures above, an architect should have an extremely useful picture of the probable behavior of the system.

Infinite Queue Allowed

This section considers systems of multiple exponential channels in parallel where the formation of an infinite queue is allowed. The infinite queue condition assumes that all arriving customers enter and remain in the system until service is completed, regardless of the system condition upon their arrival. In other words, all customers are patient and none are lost.

Literally interpreted, the condition that no customers are lost is too rigid for most architectural systems. Since all customers independently exercise the option to join or not join the system according to their personal needs, a few customers will generally always be lost. However, in satisfying the condition that system traffic intensity be less than unity, queues of excessive length will generally not occur. Consequently, the proportion of lost customers will be small relative to the total number of arrivals.

If the condition beyond which lost customers will occur can be identified, the proportion of time that the condition occurs must be made small if a minimum number of lost customers is desired. Procedures by which the preceding level of system performance may be achieved are considered in this section. When the proportion of lost customers is small, the infinite queue condition may be assumed. The loss of accuracy will generally be no greater than the error incurred by estimating rather than measuring the arrival and service rates.

The measures of effectiveness developed in this section assume that a single queue is formed when all channels are occupied. In contrast, most architectural systems will consist of separate queues for each service channel. When faced with the alternative of joining any one of several queues, the individual customer will almost always choose the shortest. In addition, each channel is assumed to operate at the same rate of service. Thus, multiple queues will almost always be of equal length. Although deviating from the exact theoretical structure of the model, the length of multiple queues may be considered as approximately equal to the length of a theoretical single queue divided by the number of available channels. For example, if there are two service channels and the theoretical single queue contains six customers, the system may be thought of as two separate queues with approximately three customers in each. The model for multiple exponential channels in parallel with an infinite queue allowed is schematically illustrated in Figure 31.

State Probabilities: P_0, P_n, Q_M

By using the property that $\sum_{n=0}^{\infty} P_n = 1.0$, Equations (6.2a) and (6.2b) may be solved for P_0 and expressed in terms of the functional variables as:

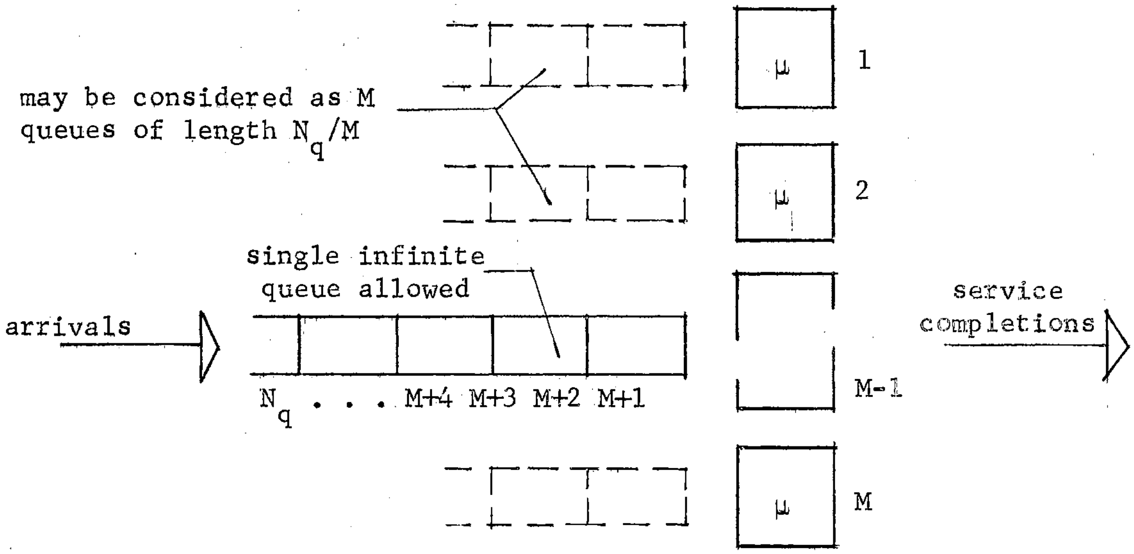

may be considered as M queues of length N_q/M
single infinite queue allowed
arrivals
N_q . . . M+4 M+3 M+2 M+1
μ
1
2
M-1
M
service completions

Figure 31. Multiple Channels, Infinite Queue Allowed

$$P_0 = \frac{(1 - \rho)e^{-\varphi}}{D_{M-1}(\varphi)} \tag{6.12}$$

P_0 is the probability of no units in the system, for which a family of curves have been plotted in Figure 32 for several values of M. P_n, the probability of exactly n units in the system may be determined by substituting Equation (6.12) into Equations (6.2a) or (6.2b). By summing Equation (6.2b) from M to infinity, the probability of M or more units in the system, Q_M, may be determined. When $n \geq M$, all channels are occupied. Q_M therefore represents the proportion of arrivals that are required to join a queue and may be expressed in terms of the functional variables as:

$$Q_M = \sum_{n=M}^{\infty} P_n = \frac{e_M(\varphi)}{D_{M-1}(\varphi)} \tag{6.13}$$

A family of curves for Q_M have been plotted in Figure 33 for several values of M.

Using previous definitions, the quantity $(1 - P_0)$ is the facility utilization and represents the proportion of time at least one service channel is occupied. The quantity $(1 - Q_M)$ is a valuable measure of effectiveness as it indicates the proportion of time that instantaneous service is available, that is, there is no queue and at least one service channel is unoccupied.

For single channel systems in which an infinite queue is allowed, traffic intensity, $\rho = \lambda/\mu$, was shown to represent the proportion of time that the channel is occupied.[1] Thus, in single channel systems, ρ may be thought of as the mean fraction of channel occupied. For multiple channel systems, $\rho = \lambda/M\mu$ has been defined as the system traffic intensity where $M\mu$ represents the rate of service for all M channels combined. Hence, where infinite queues are allowed, system traffic intensity similarly represents the mean fraction of channels occupied.

The state probabilities for multiple channel systems have been defined using two relationships, Equations (6.2a) and (6.2b), because multiple channel systems operate under two, distinct, operational conditions. In the first condition, $(0 \le n \le M)$, no queue exists as all units present are in the process of service. Units in the system advance at a rate equal to μ, the mean service rate of an individual channel. In the second condition, $(M \le n \le \infty)$, a single infinite queue has formed as all M channels are occupied. While in the queue, units will advance at a rate of $M\mu$. Upon entering the service channel, the

[1] See Chapter V, pp. 48-49.

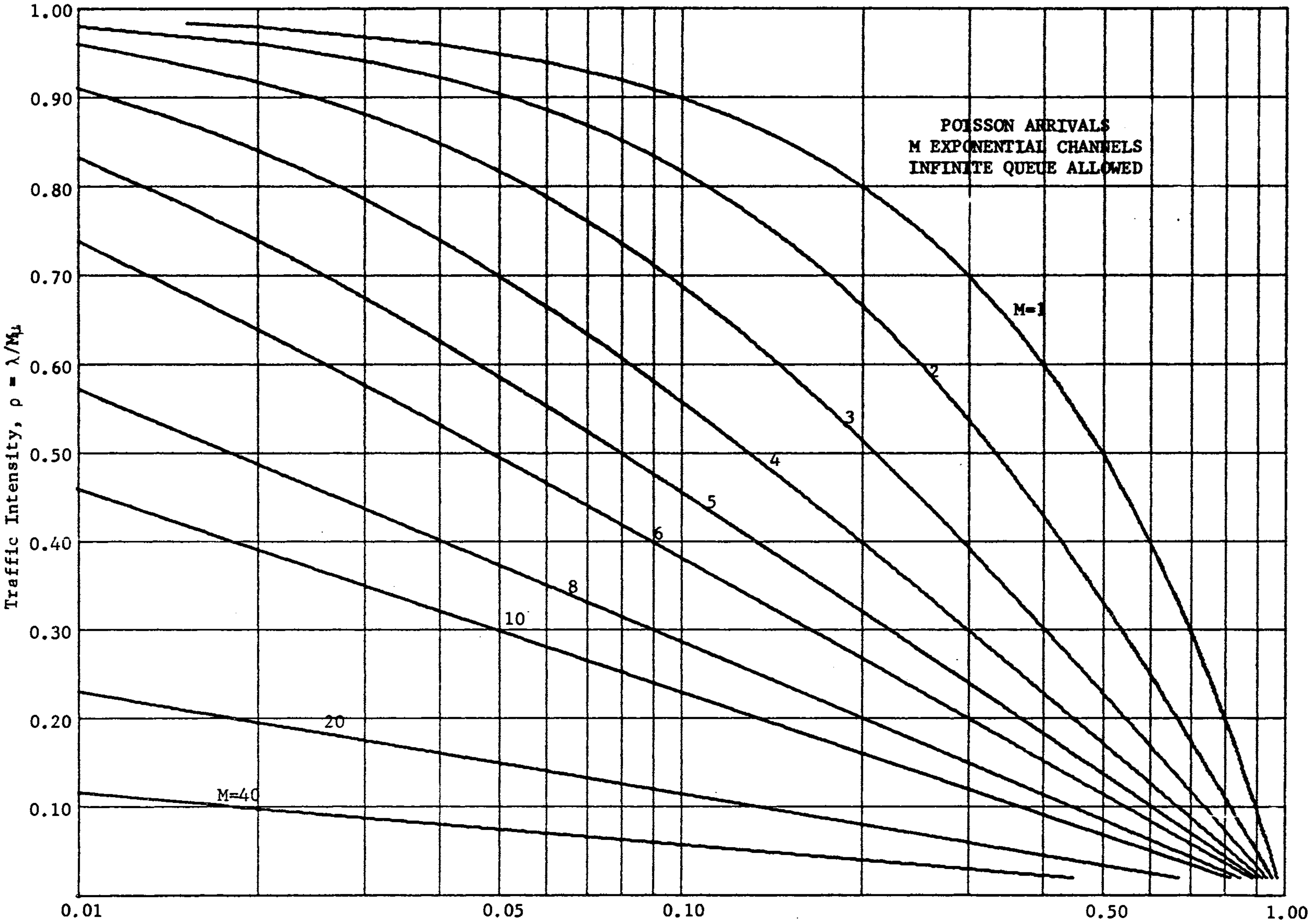


Figure 32. P_0, The Probability of No Units in the System, M Exponential Channels in Parallel, Infinite Queue Allowed

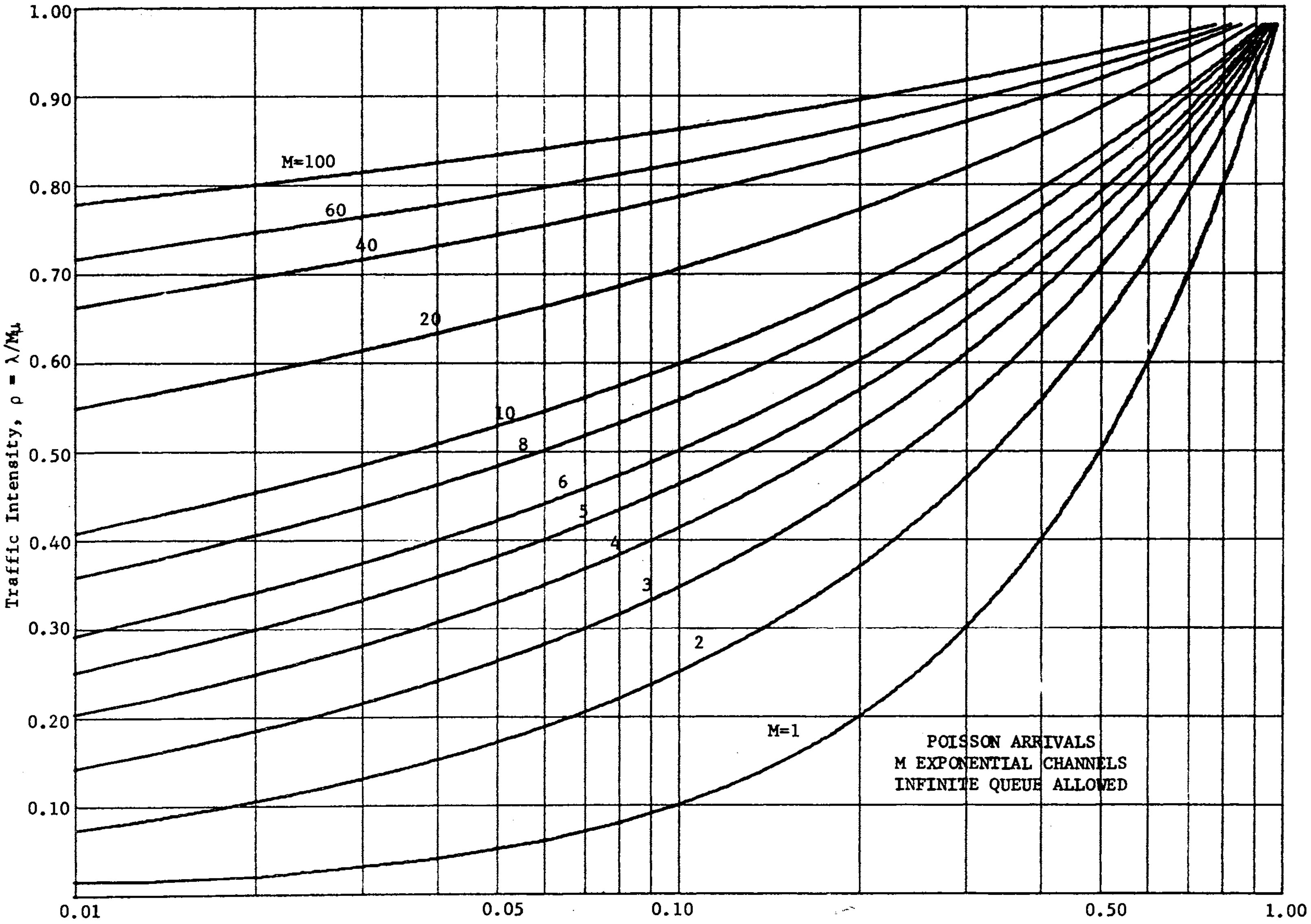


Figure 33. Q_M, The Probability that All Channels are Occupied, M Exponential Channels in Parallel, Infinite Queue Allowed

advancement rate for any particular unit will decrease to μ.

Because measures of effectiveness must consider every probable state of the system from 0 to ∞, and two operational conditions exist, the derivation of any subsequent system parameter becomes highly complex. The derivation and expression of measures of effectiveness, particularly those concerning individual units, are simplified greatly if limited to the condition of $M \leq n$. Hence, most of the following discussion will be concerned with the second condition of $(M \leq n \leq \infty)$.

From these concepts of state probability, several broad generalizations concerning the relative merits of fewer, high-speed channels as opposed to a greater number of slower channels can be made. In addition, most architectural systems are subject to eventual increases in arrival rates due to expansion or growth. Therefore, methods by which system capacity may be increased are of great importance. Two commonly applied options are available. First, the service rate of existing channels may be increased; or second, retaining the same service rate, the number of channels may be increased. Inherent to each scheme are several advantages and disadvantages.

For a given number of channels, an increase in service rate will involve a corresponding increase in the proportion of time that any one channel is completely idle.[a] Since $\rho = \lambda/M\mu$ indicates the mean fraction of occupied channels, a decrease in relative efficiency will also occur. However, as the capacity and service rate are increased, the number of customers serviced and the speed by which they are serviced may be substantially increased. The advantages are therefore mostly to the

[a] See Chapter V, Problem No. 3, pp. 56-57.

customer.

As the number of channels is increased in proportion to the increase in arrival rate, no loss of system efficiency will result if the same service rate is retained. Since the number of units that may be simultaneously serviced has increased, the number of units in the queue relative to the number of units in the system will decrease. With units advancing in the queue at rate $M\mu$, the mean delay in the queue will be correspondingly reduced. As shown in Figure 32, the probability of instantaneous service increases as M increases for a given value of ρ. Most important, the system may be operated nearer to full utilization, that is, ρ nearer to unity, before queues of excessive length occur. However, the speed of the service channel has not been increased so that time spent in service remains unchanged. Thus, the advantages are mostly to the efficiency of the system rather than the customer.

<u>Measures of Effectiveness</u>: L_q, L, $Q_{M,Nq}$, $G_{qM}(T_s)$

The mean number of units in the queue may be determined by substituting Equation (6.12) into Equation (6.2b) and summing from M to ∞. Alternate terms in the summation will cancel out so that L_q may be expressed in terms of the functional variables as:

$$L_q = \sum_{n=M}^{\infty} (n - M)P_n = \frac{\rho e_M(\varphi)}{(1 - \rho)D_{M-1}(\varphi)} \tag{6.14}$$

Since ρ represents the mean fraction of channels occupied, the product ρM gives the mean number of units in service. Hence, the mean number of units in the system may be expressed as:

$$L = L_q + \rho M \tag{6.15}$$

L has been plotted for several values of M as a function of system traffic intensity in Figure 34.

As developed for previous models, the mean time spent in the system, $W = L/\lambda$; the mean time spent in the queue, $W_q = L_q/\lambda$. Figure 34 demonstrates clearly that for a given value of ρ, an increase in M results in a corresponding increase in L and decrease in L_q. Expanding the concepts introduced in the previous section, assume that the initial costs and operating costs of individual channels are proportional to the rate of service. Consequently, two slow channels will cost as much as one channel operating twice as fast. Whenever the most important requirement is to minimize the length of queue, it is advantageous to have many channels operating at slower rates. However, whenever it is more important to minimize total delay time W, it is better to have fewer, high-speed channels.

Q_{M,N_q} is defined as the probability of N_q or more units in the queue. Since all channels must be occupied before the formation of a queue, Q_{M,N_q} is also the probability of $(N_q + M)$ or more units in the system. Q_{M,N_q} may be expressed in terms of the functional variables as:

$$Q_{M,N_q} = \sum_{n=N_q}^{\infty} P_n = \sum_{n=0}^{\infty} P_{M+N_q+n} = \frac{\rho^{N} e_M(\varphi)}{D_{M-1}(\varphi)} \tag{6.16}$$

Four basic variables are involved in the expression of Q_{M,N_q}. They are Q_{M,N_q}, M, N_q and ρ. Since two-dimensional graphs are limited to three variables, any one of the variables must be assumed or held constant. Consequently, Figures 35 to 39 are plotted separately for values of

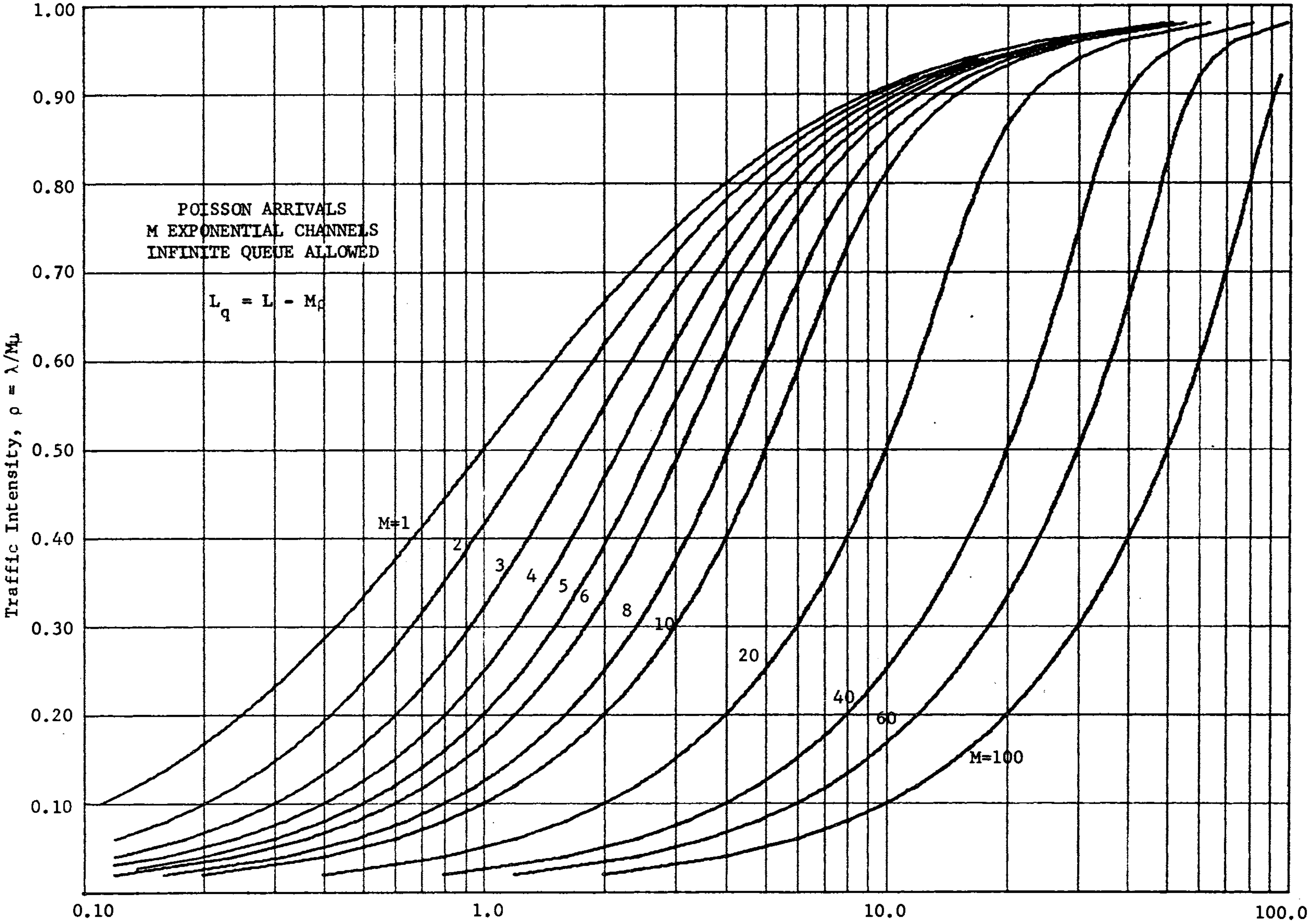


Figure 34. L, The Mean Number of Units in the System, M Exponential Channels in Parallel, Infinite Queue Allowed

$M = 2$, 3, 4, 5, and 6, and 8 and 10, respectively. In each figure, curves for N_q are given in even increments of M. For example, in Figure 35 where $M = 2$, curves for $N_q = 2$, 4, 6, 8, 10, and 12 have been provided. N_q represents the length of a single infinite queue. However, it has been established that the single queue may be thought of as M queues, each approximately N_q/M in length. Hence, the same curves above may be thought of as two separate queues of length 1, 2, 3, 4, 5, or 6, respectively.

A related set of graphs are provided in Figures 40 to 44 for $G_{qM}(T_s)$, which represents the probability that time spent in the queue by an individual unit exceeds a multiple of T_s. The derivation of $G_{qM}(T_s)$ is well beyond the scope of this study. However, it may be shown that $G_{qM}(T_s)$ may be expressed in terms of the functional variables as:

$$G_{qM}(T_s) = Q_M e^{-[M(1-\rho)cT_s]} \tag{6.17}$$

where Q_M is defined in Equation (6.13) and c represents a multiple of T_s.

The use of Figures 32 to 44 is best illustrated through the use of the following example problems.

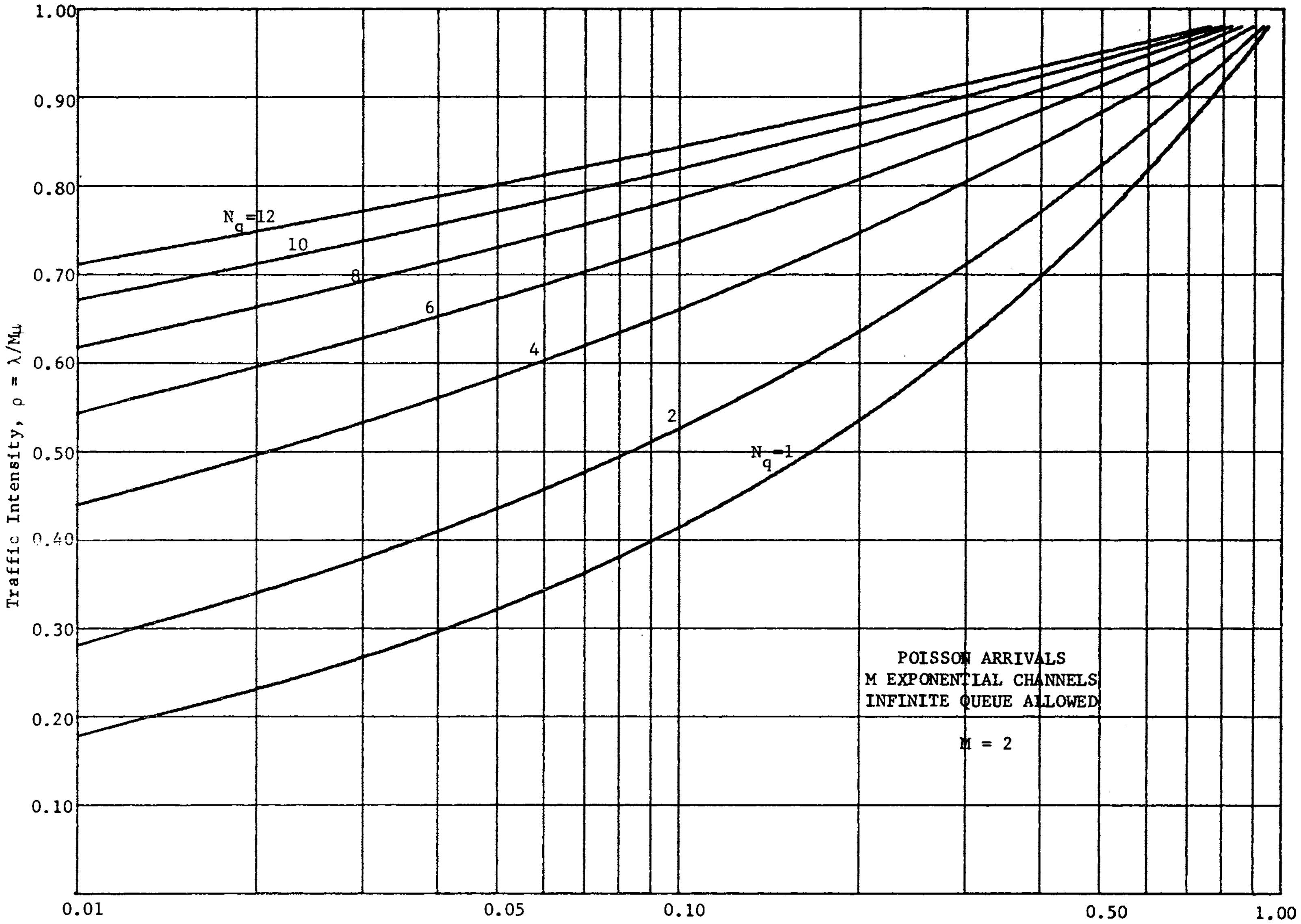


Figure 35. Q_{M,N_q}, The Probability of N_q or More Units in the Queue, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 2

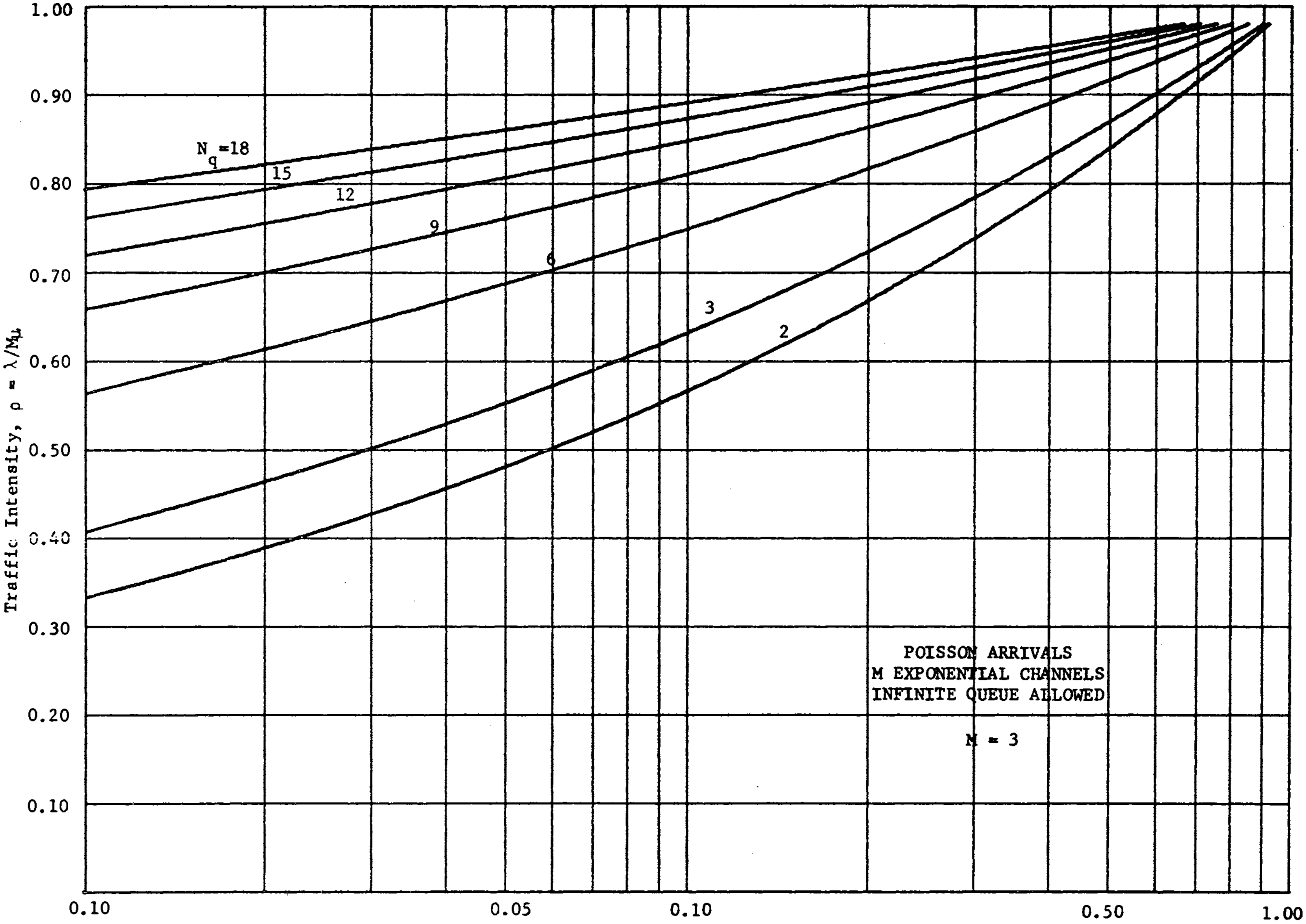


Figure 36. Q_{M,N_q}, The Probability of N_q or More Units in the Queue, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 3

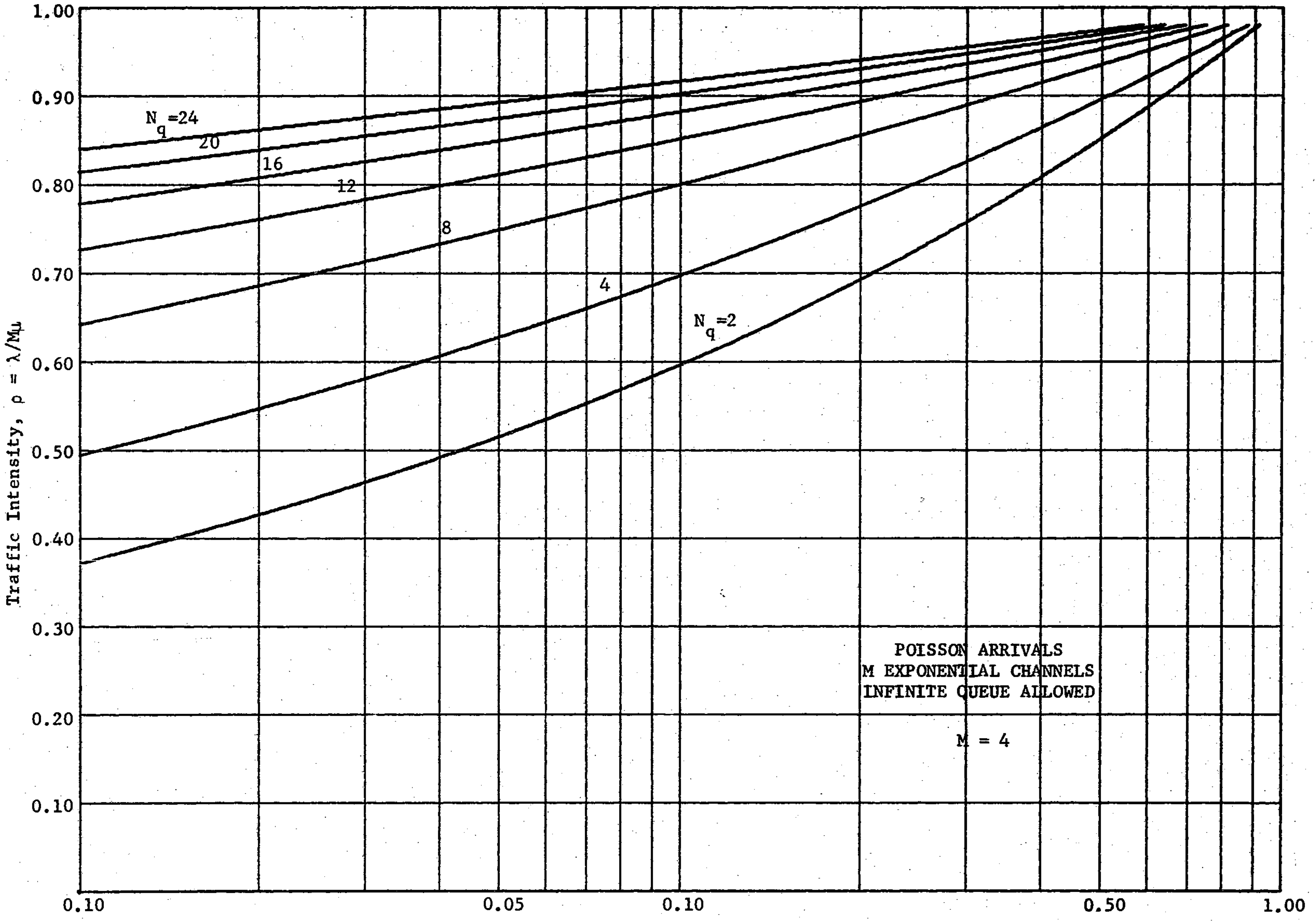


Figure 37. Q_{M,N_q}, The Probability of N_q or More Units in the Queue, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 4

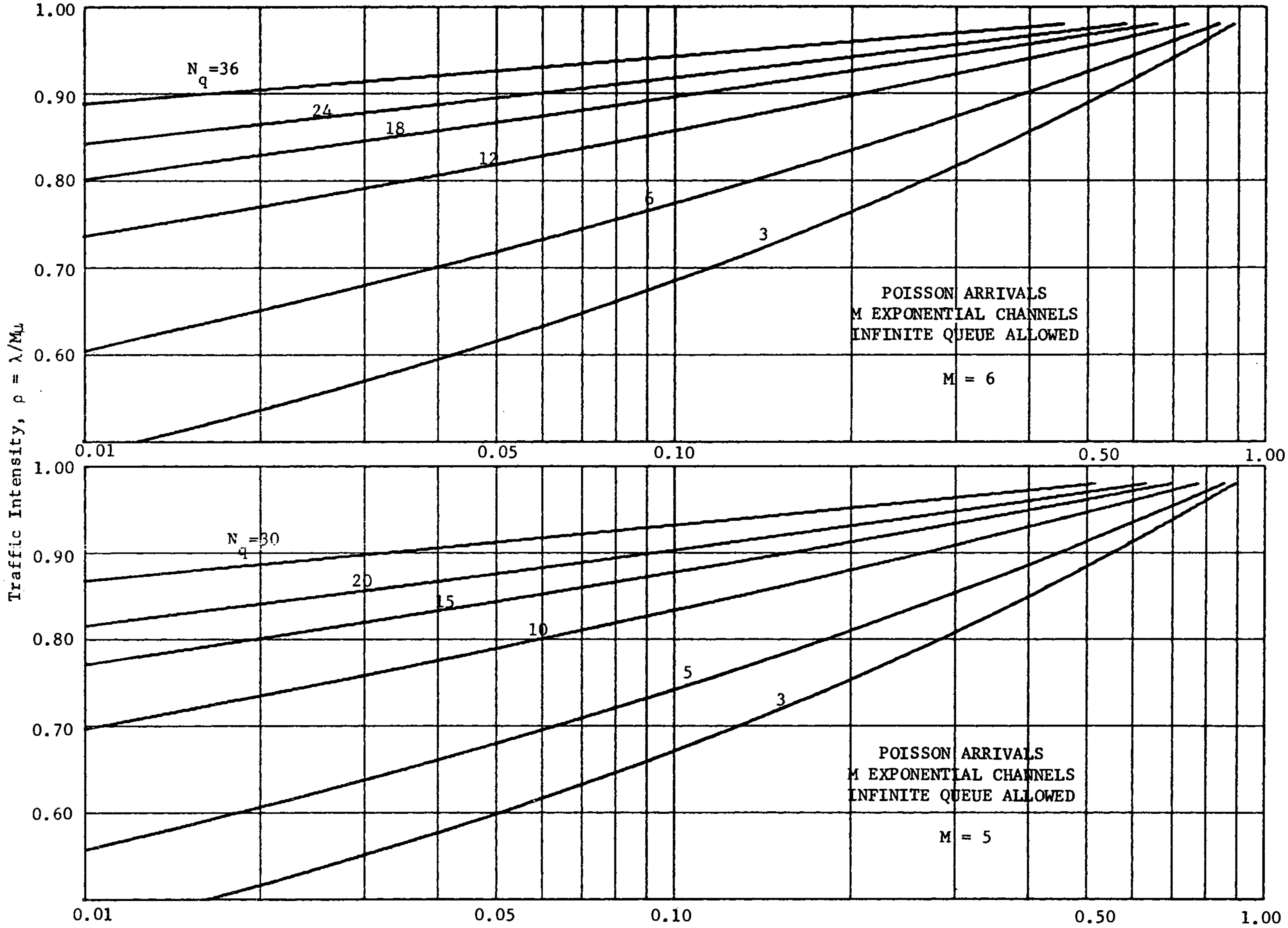


Figure 38. Q_{M,N_q}, The Probability of N_q or More Units in the Queue, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 5 & 6

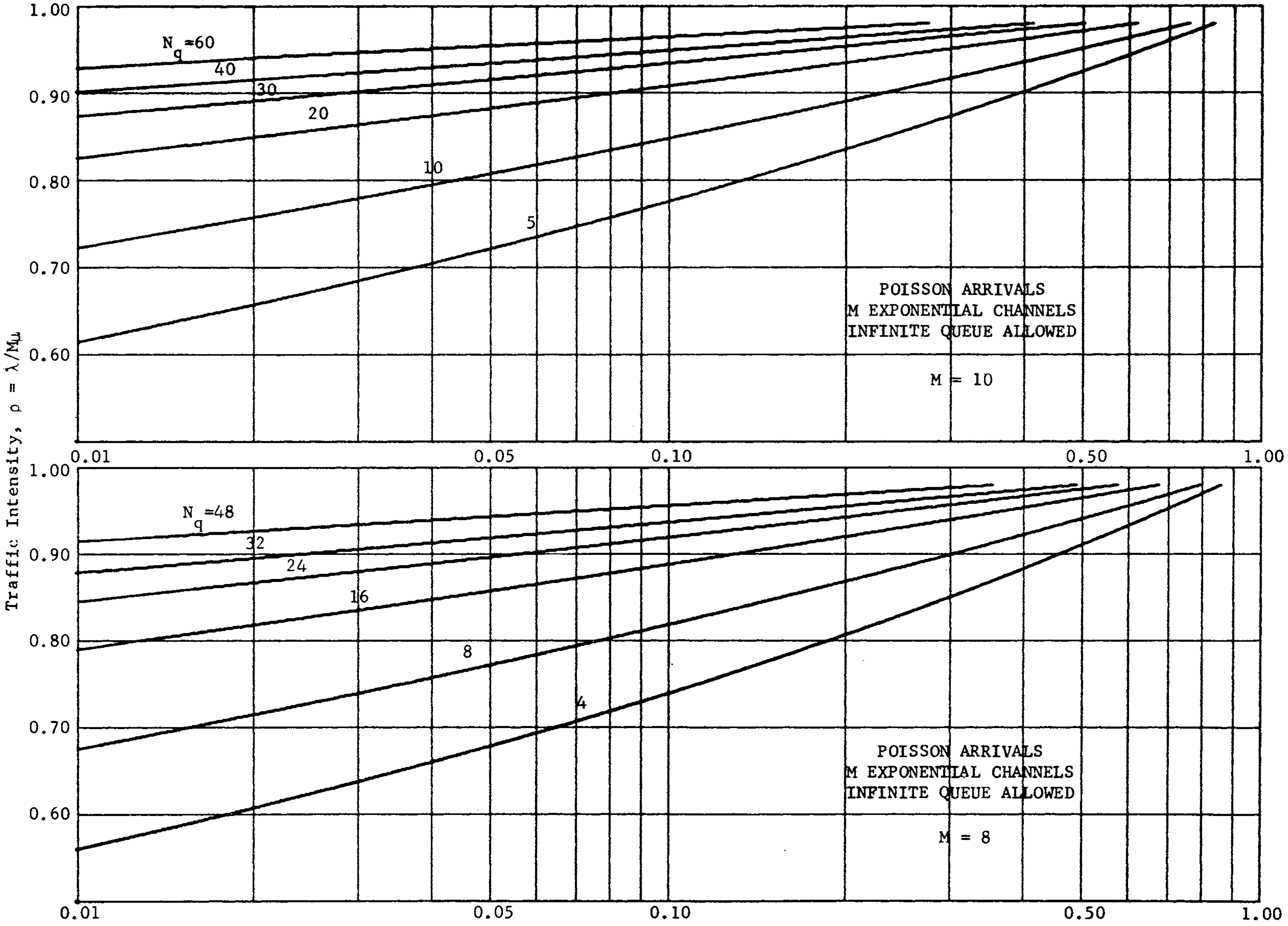


Figure 39. Q_{M,N_q}, The Probability of N_q or More Units in the Queue, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 8 & 10

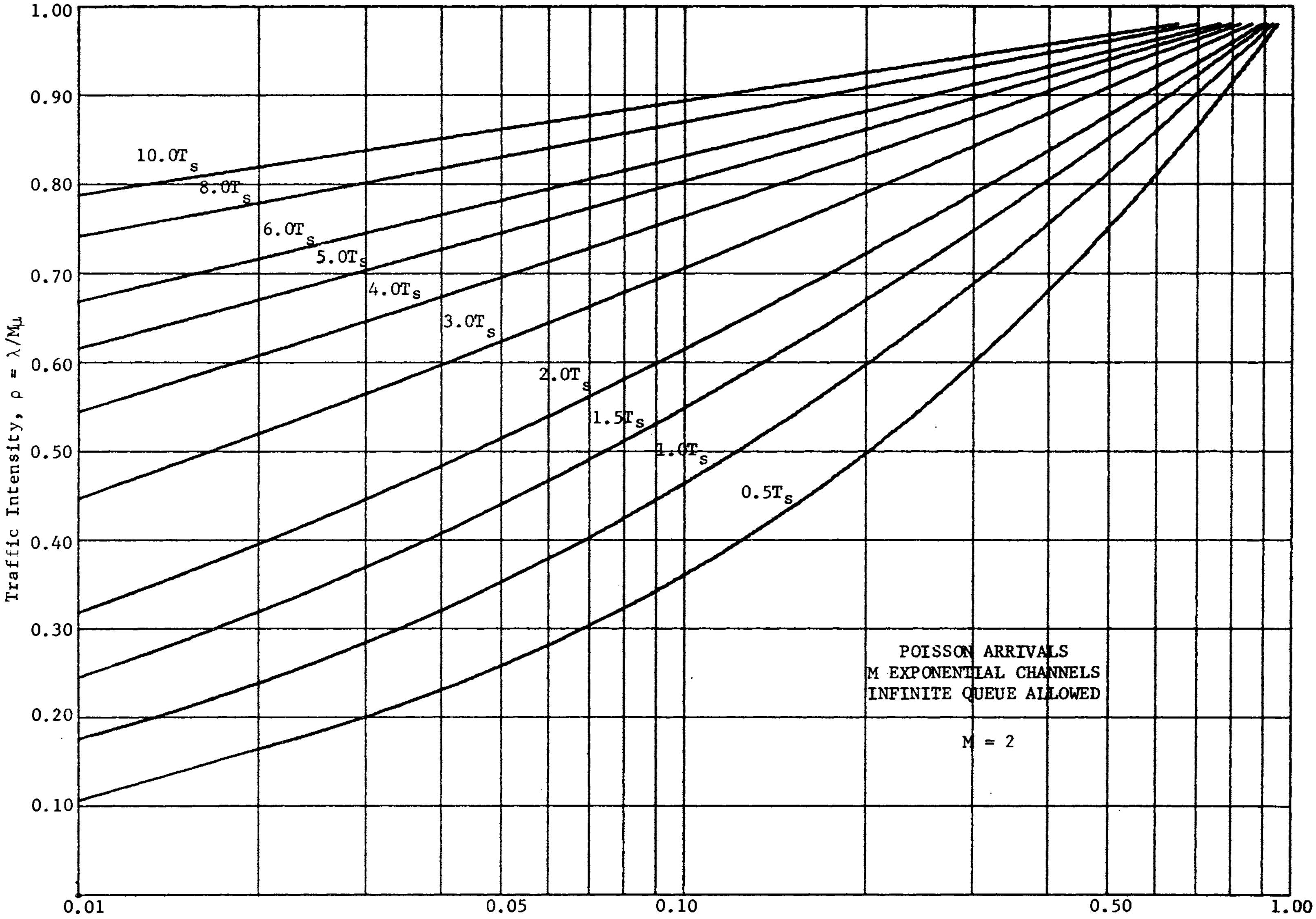


Figure 40. $G_{qM}(T_s)$, The Probability that Time Spent in the Queue Exceeds a Multiple of T_s, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 2

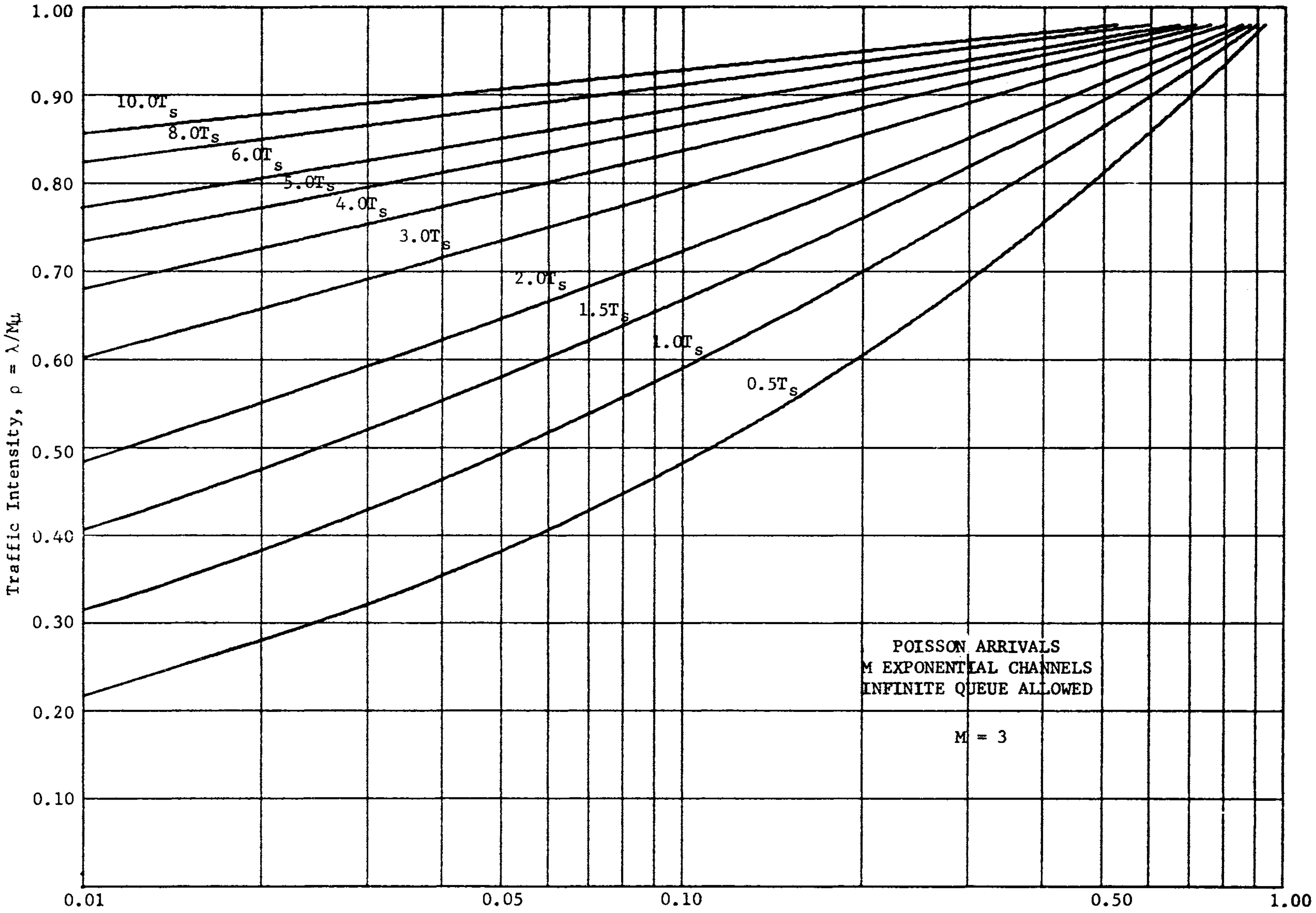


Figure 41. $G_{qM}(T_s)$, The Probability that Time Spent in the Queue Exceeds a Multiple of T_s, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 3

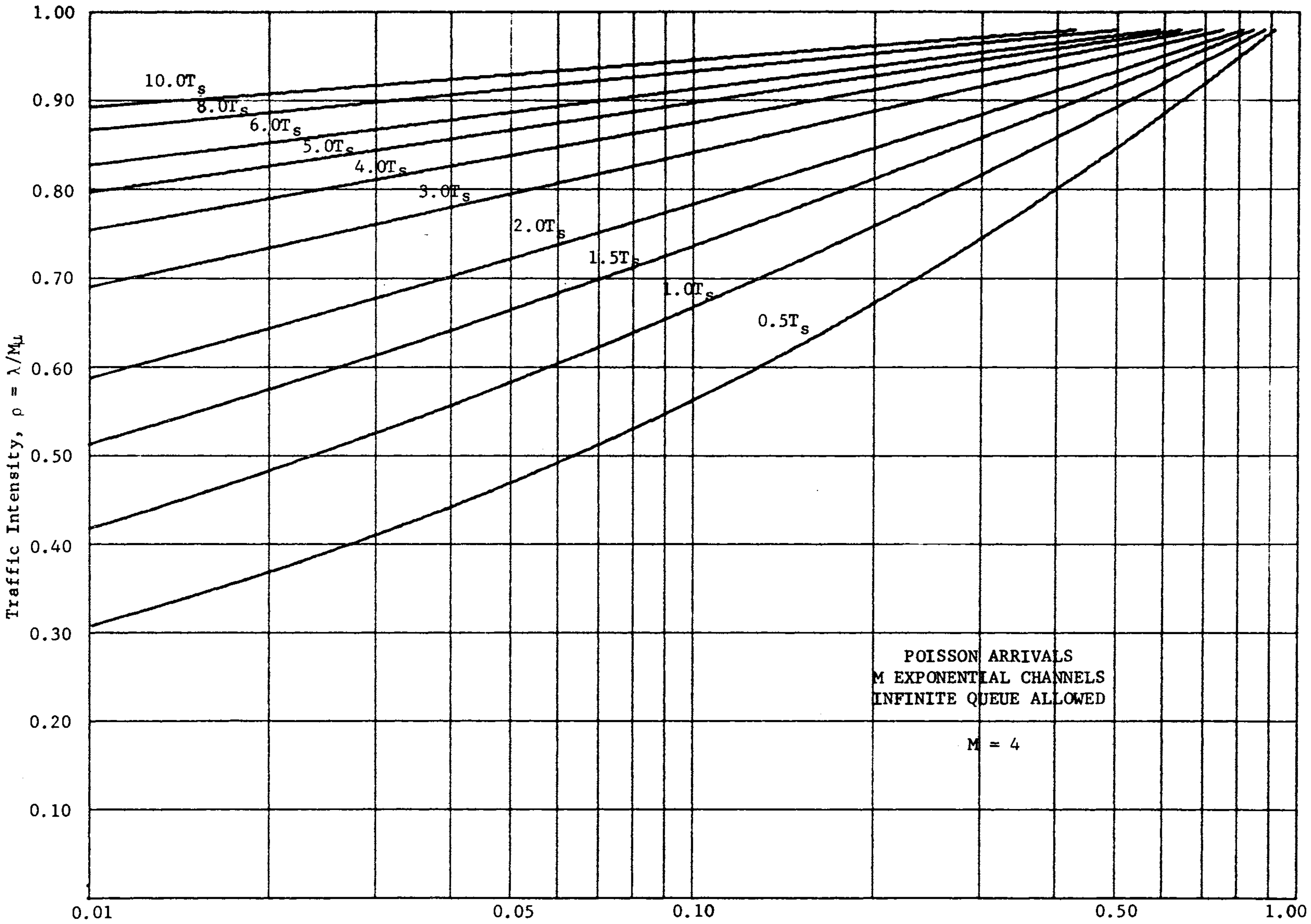


Figure 42. $G_{qM}(T_s)$, The Probability that Time Spent in the Queue Exceeds a Multiple of T_s, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 4

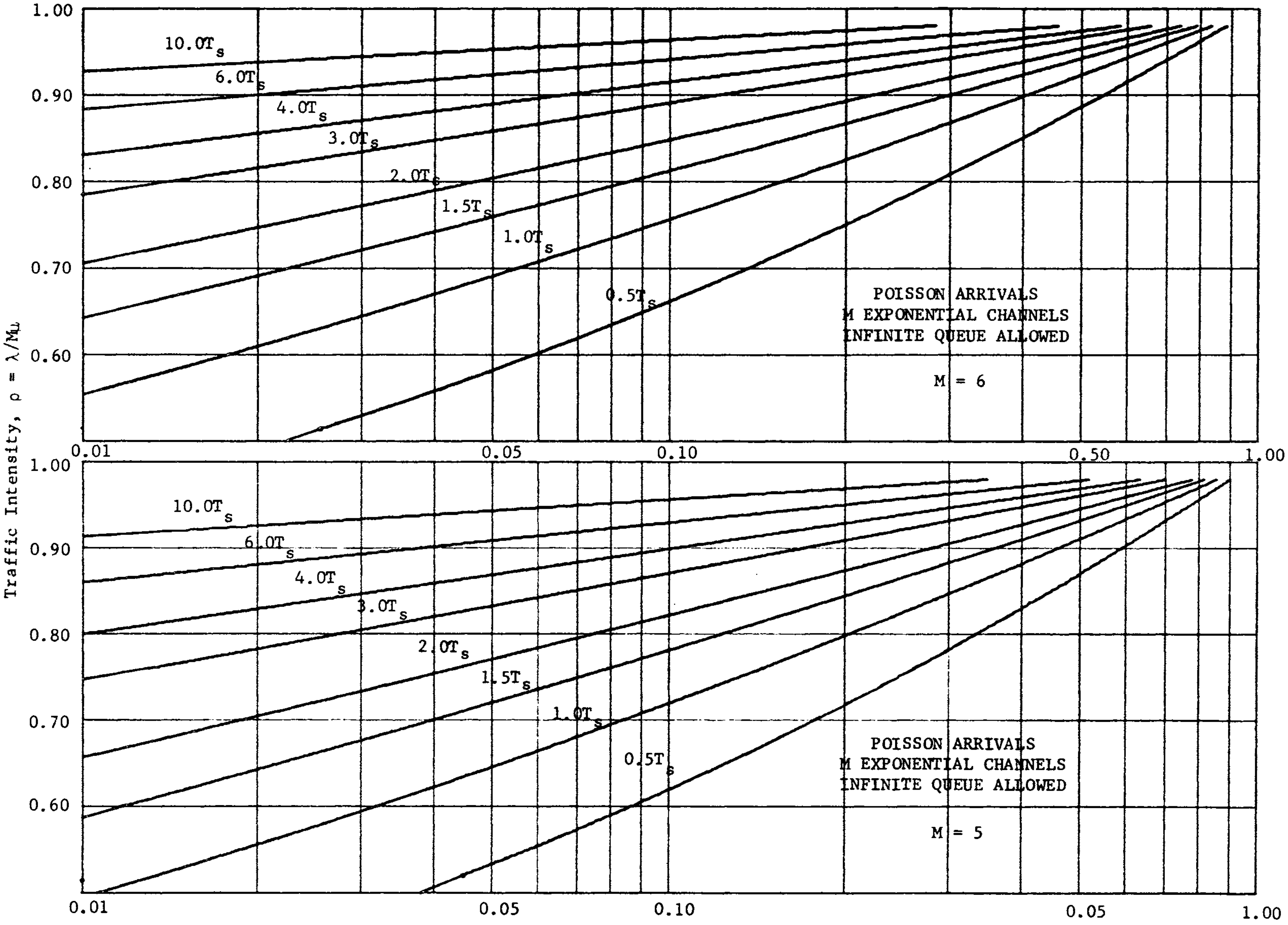


Figure 43. $G_{qM}(T_s)$, The Probability that Time Spent in the Queue Exceeds a Multiple of T_s, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 5 & 6

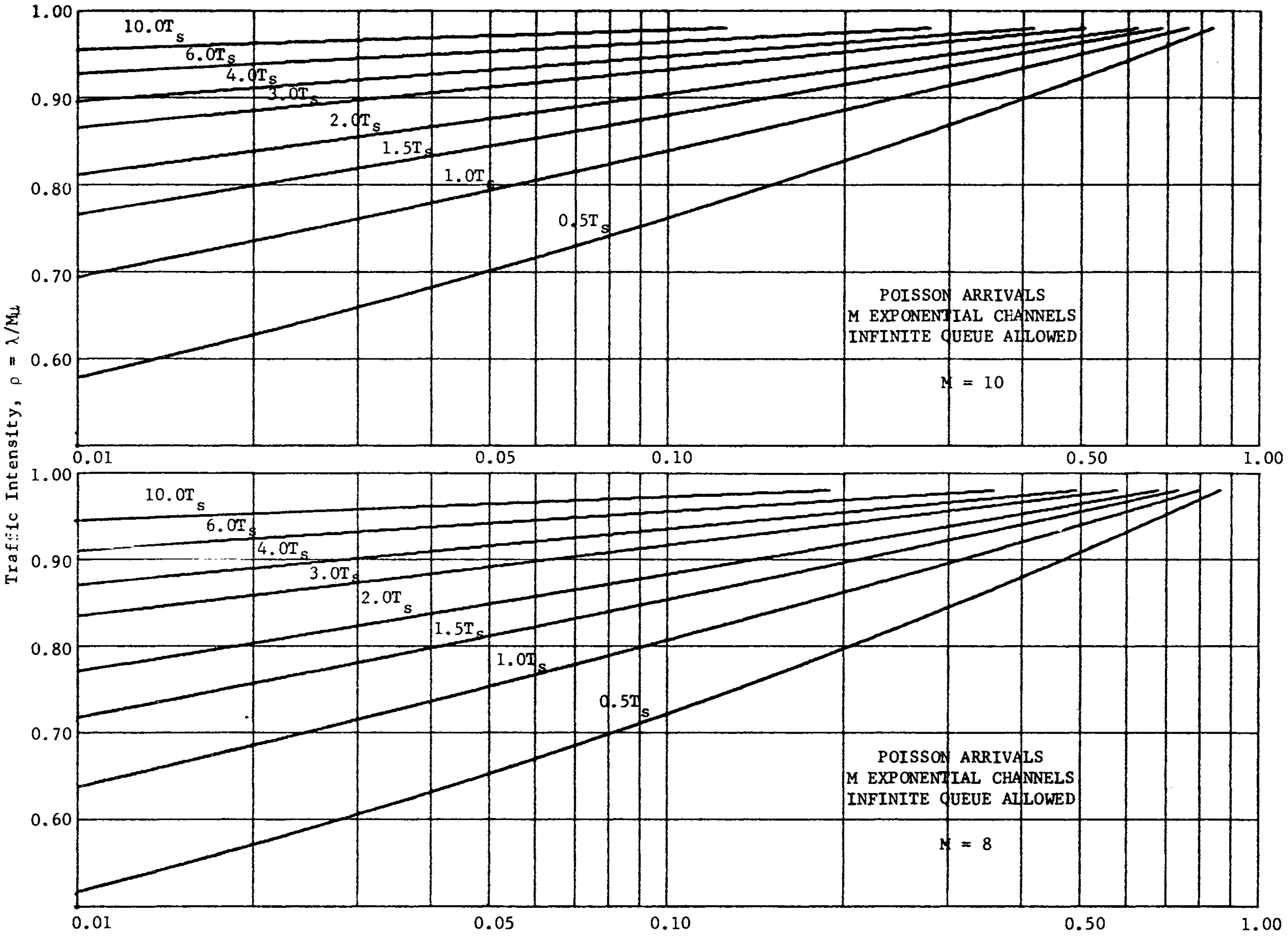


Figure 44. $G_{qM}(T_s)$, The Probability that Time Spent in the Queue Exceeds a Multiple of T_s, M Exponential Channels in Parallel, Infinite Queue Allowed, M = 8 & 10

Problem No. 4: An airline ticket counter has been provided with accomodations for two clerks, in front of which, separate queues will form. The average interval between arrivals and average duration of service are estimated as 5.0 and 7.5 minutes, respectively. Using Figures 32 to 44, describe the expected system performance.

Solution: $T_a = 5.0$ minutes; $T_s = 7.5$ minutes

$\lambda = 1/5.0 = 0.20$; $\mu = 1/7.5 = 0.133$

$\varphi = 0.20/0.133 = 1.5$; $\rho = 1.5/2 = 0.75$

Since there are two channels, enter all figures with $M = 2$. With $\rho = 0.75$, the appropriate figure and results are:

(Figure 32) $P_0 = 0.15$

(Figure 33) $Q_M = 0.63$; $(1.0 - Q_M) = 0.37$

(Figure 34) $L = 3.5$; $L_q = L - \varphi = 3.5 - 1.5 = 2.0$

$W = L/\lambda = 3.5/0.20 = 17.5$ minutes;

$W_q = L_q/\lambda = 2.0/0.20 = 10.0$ minutes.

(Figure 35)

$N_q/M =$	1	2	3	4	5	6
$Q_{M,N_q} =$	.36	.20	.12	.06	.035	.02

(Figure 40) G_{qM}(7.5 minutes) = 0.38 $(1.0T_s)$

G_{qM}(15.0 minutes) = 0.23 $(2.0T_s)$

G_{qM}(30.0 minutes) = 0.084 $(4.0T_s)$

G_{qM}(45.0 minutes) = 0.030 $(6.0T_s)$

Comments: P_0 indicates that 15% of the time, the system will contain no units. Consequently, both clerks will be completely idle. Facility utilization, $(1 - P_0)$, indicates that 85% of the time, at least one of the two clerks will be busy. Q_M indicates that 63% of the time, two or more units will be present, and therefore represents the proportion of time a queue will have formed. The complementary statement, $(1 - Q_M)$, indicates that 37% of the time, at least one channel is unoccupied, and therefore represents the proportion of time that instantaneous service is available.

From Figure 35, the expected proportion of time that N_q/M or more customers are waiting in each of the separate queues has been determined. For example, when $N_q/M = 1$, $Q_{M,N_q} = 0.36$. This indicates that queues of one or more in length may be expected 36% of the time. Since there are two queues and two channels, this also represents the proportion of time four or more customers may be expected in the system.

The set of values for $G_{qM}(T_s)$ give the proportion of time that customers are delayed in the queue longer than a specified duration. For example, 7.5 and 45.0 minutes in the queue are exceeded 38% and 3% of the time respectively.

<u>Problem No. 5</u>: A franchise for additional routes has been granted to the airline in Problem No. 4. The increase in traffic is reflected in customer traffic which has doubled. Thus, the average interval between arrivals has decreased to 2.5 minutes. If the airline wishes to maintain the same level of customer service, system capacity must be increased. With computerized ticket handling aids, the airline estimates that the average duration of service may be reduced to 3.75 minutes. However, sufficient space and personnel are available to simply increase the number of channels to four. What description of system performance may be given for each of the alternatives above?

<u>Solution A</u>: $T_a = 2.5$ minutes; $T_s = 3.75$ minutes; $M = 2$

$$\lambda = 1/2.5 = 0.40; \quad \mu = 1/3.75 = 0.267$$

$\mu = 0.40/0.267 = 1.5; \; \rho = 1.5/2 = 0.75$

Since M and ρ remain unchanged, the description of system performance remains identical to that of Problem No. 4 with two exceptions. The mean time spent in the system and in the queue are functions of the arrival rate. Thus, $W = L/\lambda = 3.5/0.40 = 8.75$ minutes; and $W_q = 2.0/0.40 = 5.0$ minutes. Even though the arrival rate has doubled, the reduction in service time has more than compensated for its effect on system performance. Since T_s is now 3.75 minutes, $G_{qM}(1.0T_s) = 0.38$ now represents the proportion of time delay in the queue exceeds 3.75 rather than 7.5 minutes.

<u>Solution B</u>: $T_a = 2.5$ minutes; $T_s = 7.5$ minutes; $M = 4$

$\lambda = 1/2.5 = 0.40; \; \mu = 1/7.5 = 0.133$

$\varphi = 0.40/0.133 = 3.0; \; \rho = 3.0/4 = 0.75$

Since there are four channels, enter all figures with M = 4. With $\rho = 0.75$, the appropriate figure and results are:

(Figure 32) $P_0 = 0.04$

(Figure 33) $Q_M = 0.50; \; (1.0 - Q_M) = 0.50$

(Figure 34) $L = 4.5; \; L_q = L - \varphi = 4.5 - 3.0 = 1.5$

$W = L/\lambda = 4.5/.40 = 11.25$ minutes;

$W_q = L_q/\lambda = 1.5/.40 = 3.75$ minutes.

(Figure 37)

$N_q/M =$	1	2	3	4	5	6
$Q_{M,N_q} =$	0.16	.048	.017	<.01	<.01	<.01

(Figure 43) $G_{qM}(7.5 \text{ minutes}) = 0.18 \quad (1.0T_s)$

$G_{qM}(15.0 \text{ minutes}) = 0.11 \quad (2.0T_s)$

$G_{qM}(30.0 \text{ minutes}) = <0.01 \quad (4.0T_s)$

<u>Comments</u>: By doubling the speed of the existing service channels,

almost all benefits are directed to the customer. Total time spent in the system is substantially reduced. However, improvement of parameters concerning the queue are not as great when compared to the second alternative. By doubling the number of channels to four, the mean time spent in the queue is reduced from 10.0 to 3.75 minutes. Queue lengths of greater than three seldom occur and the probability of spending more than 15 minutes in the queue is reduced from 0.23 to 0.11. The proportion of time that instantaneous service is available has increased by 13%; facility utilization has improved by 11%. However, since the service rate has not been improved, the total time spent in the system is 11.25 minutes. Hence, most of the benefits are directed to the system rather than the customer.

If the cost of both alternatives is the same, the decision must be based upon the relative merits of fast customer service or slow customer service with relatively shorter queue lengths and delays. The appearance of short queues is usually more important where the facility is highly competitive. In addition, the number of lost or reneging customers is generally smaller if the queue has the appearance of advancing rapidly. In this case, however, assuming that customer service is of greater importance, increasing the speed of existing service channels would generally be the preferred solution.

Problem No. 6: An architect has been commissioned to design a branch bank for a large financial concern. The bank president, a pioneer in the suburban banking business, demands and insures excellence in customer service by making unannounced visits to his branch facilities. Through these experiences, he has established a firm, organizational policy. In a brief interview with the architect, the president had stated, "We are in business for our customers and only their satisfaction will keep us in business. Whenever I enter any of my branches as a routine customer, I do not expect to find more than one customer preceding me in line. In fact, even having to wait in line upsets me a great deal. If I must wait, nine times out of ten I do not expect a delay greater than one-half the time it would ordinarily take to complete my transaction."

From the president's staff, the architect is advised to expect a maximum of 80 routine customers during a peak hour period. During these hours, bank research indicates that transactions average 3.0 minutes. The staff also advises to design for peak conditions. During lulls, unoccupied tellers are busy at their stations with other tasks essential to the organization. With this information, the architect must present his preliminary design. How many tellers should he provide? What statements concerning system performance may the architect make to enhance the acceptability of his proposal?

Solution: The manner in which this problem is presented is typical of those encountered by practicing architects. A feasible solution may be obtained in three basic steps. First, assumptions must be made so that the concepts of this study may be applied. Second, the data and constraint relationships provided by the president and his staff must be interpreted or translated into terms of queueing variables. And third, using the variables of the second step, inference is made of the probable behavior of the proposed solution.

Assumptions, (Step No. 1): a) the customer population is infinite or very large; b) the arrival rate follows a Poisson's distribution; c) the service times follow an exponential distribution; d) customers receive service on a first come, first served basis; e) all arriving customers remain in the system until service is completed, that is, an infinite queue is allowed; and f) a finite number of service channels are to be arranged in parallel.

Interpretation, (Step No. 2):

T_s = 3.0 minutes; $\mu = 1/3.0 = 0.333$ service completions/minute;

λ = (80 customers/hour)(1/60 hour/minute) = 1.333 customers/minute;

$T_a = 1/1.333 = 0.75$ minutes;

$\varphi = \lambda/\mu = 1.333/0.333 = 4.0$; $\rho = \varphi/M = 4.0/M$.

The probability of M customers in the queue, $N_q = M$ or $N_q/M = 1.0$, is to be made insignificant. (". . . .I do not expect to find more than one customer preceding me in line.") Insignificance is a relative term which may be evaluated after analysis on several solutions is completed. The proper relationships to be used are found in Figures 35 to 39.

The time spent in the queue should not exceed $T_s/2$ more than 90% of the time. (". . . nine times out of ten I do not expect a delay greater than one-half the time it would ordinarily take to complete my transaction.") Translated into terms of Figures 40 to 44, the probability that time spent in the queue exceeds $0.50T_s$ or 1.5 minutes is to be made ≤ 0.10.

Solution, (Step No. 3): Since $\varphi = 4.0$, the number of channels provided must be greater than four if ρ is to be less than unity. Thus, the minimum number of channels required is five. Using the iterative procedure previously introduced:

Try M = 5; $\rho = 4.0/5 = 0.80$;

from Figure 38, $Q_{M,N_q} = Q_{5,5} = 0.18$;

from Figure 43, $G_{qM}(T_s) = G_{q5}(0.5T_s) = 0.33 > 0.10$; new trial required.

Try $M = 6$; $\rho = 4.0/6 = 0.67$;

from Figure 38, $Q_{M,N_q} = Q_{6,6} = 0.027$;

from Figure 43, $G_{qM}(T_s) = G_{q6}(0.5T_s) = 0.11 \simeq 0.10$; condition satisfied.

$Q_{6,6}$ indicates that 27 out of 1000 customers will arrive to find six or more customers in line or the equivalent of one or more customer waiting at each teller's station. $Q_{6,12}$, where $N_q/M = 2.0$, is less than 0.01 which indicates that two or more customers in each line will occur infrequently. Hence, the probability that a customer will be preceded by two customers in line may be considered insignificant.

With M established as six and $\rho = 0.67$, these additional statements may be made:

from Figure 33, $Q_6 = 0.28$; $(1.0 - Q_6) = 0.72$.

from Figure 34, $L = 4.6$; $L_q = L - \varphi = 4.6 - 4.0 = 0.6$;

$W = 4.6/1.33 = 3.45$ minutes; $W_q = 0.6/1.33 = 0.45$ minutes.

As the president has emphasized short queues and delays, he should be impressed with the statistics above. Instantaneous service, that is, no wait in a queue, is available 72% of the time. The average time spent in the queue for all customers is less than 30 seconds.

Comments: In many instances, the problem may be complicated by the potential of variable service rates. For example, the installation of sophisticated communications systems or similar equipment could substantially reduce service time. If the cost of time spent in the system, both in waiting and in service can be economically evaluated, a system resulting in minimum cost at a stated level of service may be determined. Methods of optimization have been developed in the operations research field. However, the simplicity of techniques presented in this study allow the rapid investigation of several systems. Consequently, a trial and error process of analysis will not be overly tedious, which again illustrates the versatile usefulness of the concepts presented.

CHAPTER VII

SUMMARY AND CONCLUSIONS

The purpose of this study has been to provide the architect with a usable set of graphically presented relationships with which to analyze the performance of elementary queueing systems. Written for the purpose of applications in the field of architectural design, it was assumed that the reader was unfamiliar with the topics of statistics, probability, random variation, and the theory of queues. Hence, discussion was initiated with an introduction to the fundamentals of statistics for the measurement of arrival and service rates.

Probability distributions were shown to represent populations of arrival and service times subject to random variation. Since architectural queueing systems are subject to random variation, evaluation was made in probabilistic terms. Probability was established as a concept which indicates the proportion of time a stated event is expected to occur or not occur.

The five basic elements of all queueing systems were introduced with the use of Moore's Organization Chart. The elements were customer population, number of channels, queue discipline, arrival distribution, and service distribution. From these elements, several models were constructed, representative of many systems commonly found in architecture.

Associated with each model, measures of effectiveness were presented which provided a means for evaluating system performance. Measures of effectiveness were considered as any relationship which expressed in numeric terms an indication of the long-run behavior of the system. The determination of measures of effectiveness from their graphic presentation was explained in detail by the use of several example problems. Through appropriate comments, their interpretation and use in the making of inference towards predicted system behavior were discussed subjectively.

In reviewing Moore's Organization Chart, it will be seen that less than one-half of the elements have been considered. Consequently, it is obvious that this study is incomplete. Many elements commonly found in real-world architecture have been neglected. Some of the more important, in order of their frequency of occurrence, may be listed as follows: a) arrival and service distribution other than exponential; b) systems with time-dependent arrival and service rates; c) systems with reneging customers; d) bulk queue disciplines; e) multiple channel systems in series; and f) systems with priorities.

Any further extensions of this study should continue with graphic presentations as many of the relationships are highly complex. However, further studies will also increase the number of graphs until they become too numerous for practical usefulness. Hence, consolidation of relationships in the form of more involved nomographs, even if they require additional or simplifying assumptions, is highly recommended.

The importance of queueing theory applications to architectural design is emphasized by quoting a portion of Chapter I.

> Queueing problems abound in architectural design. Buildings are not inanimate objects, but dynamic systems of traffic flow in which queueing situations are the rule rather than the exception. As the complexity of architectural structures increase, there is a resultant increase in the number of causes

> for waiting. As waiting increases, the necessity for the architect to satisfy service demands also increases . . . It is the responsibility of the architect to evaluate the demand, establish the appropriate level of service, estimate the various costs associated with the satisfaction of demand, and determine the optimum level for system capacity.

This study, in an attempt to expand the technical capabilities of the architect, is "A Graphic Introduction to Problems in Queueing Theory for Architects and Engineers."

A SELECTED BIBLIOGRAPHY

Cochran, William G. Sampling Techniques. New York: John Wiley and Sons, Inc., 1953.

Cox, D. R. and Walter L. Smith. Queues. London: Methuen and Company, 1961.

Fabrycky, W. J. and Paul E. Torgersen. Operations Economy. New Jersey: Prentice-Hall, Inc., 1966.

Feller, W. An Introduction to Probability Theory and Its Applications, Vol. I. New York: John Wiley and Sons, 1957.

Fry, Thornton C. Probability and Its Engineering Uses. New York: D. Van Nostrand Company, Inc., 1928.

Grant, Eugene L. Statistical Quality Control. New York: McGraw-Hill Book Company, 1964.

Hillier, Frederick S. and Gerald J. Lieberman. Introduction to Operations Research. San Francisco: Holden-Day, Inc., 1967.

Morse, Philip M. Queues, Inventories, and Maintenance. New York: John Wiley and Sons, Inc., 1963.

Prabhu, N. U. Queues and Inventories. New York: John Wiley and Sons, 1965.

Ruiz-Pala, Ernesto, Carlos Avila-Beloso, and William W. Hines. Waiting-Line Models. New York: Reinhold Publishing Corporation, 1966.

Saaty, Thomas L. Elements of Queueing Theory. New York: McGraw-Hill Book Company, Inc., 1961.

Steel, Robert G. D. and James H. Torrie. Principles and Procedures of Statistics. New York: McGraw-Hill Book Company, Inc., 1960.

Votaw, David F. and Herbert S. Levinson. Elementary Sampling for Traffic Engineers. Connecticut: The Eno Foundation for Highway Traffic Control, 1962.

VITA

Arthur Yukio Kishiyama

Candidate for the Degree of

Master of Architectural Engineering

Thesis: A GRAPHIC INTRODUCTION TO PROBLEMS IN QUEUEING THEORY FOR ARCHITECTS AND ENGINEERS

Major Field: Architectural Engineering
Minor Field: Industrial Engineering and Management

Biographical:

Personal Data: Born in Pomona, California, the son of Kosaku K. and Ellen T. Kishiyama.

Education: Graduated from San Luis Obispo High School, San Luis Obispo, California, in June, 1959; received the Bachelor of Science degree with major in Architectural Engineering from California State Polytechnic College, San Luis Obispo, California, in June, 1963; completed the requirements for the Master of Architectural Engineering at Oklahoma State University in May, 1968.

Professional Experience: Architectural designer, William Kubach Associates, San Mateo, California, 1963; entered the United States Air Force, 1964 with current rank of Captain; Project Design Engineer, Engineering and Construction Branch, Base Civil Engineers, Sewart Air Force Base, Tennessee, 1964 to 1966.

Professional Organizations: Scarab, National Professional Architectural Fraternity; Tau Sigma, National Honorary Engineering Fraternity; Society of American Military Engineers; Phi Kappa Phi, National Honor Society.