Classification
A. D. Gordon

Monte Carlo Methods
J. M. Hammersley and D. C. Handscomb

Identification of Outliers
D. M. Hawkins

Distribution-free Statistical Methods
J. S. Maritz

Multivariate Analysis in Behavioural Research
A. E. Maxwell

Applications of Queueing Theory
G. F. Newell

Some Basic Theory for Statistical Inference
E. J. G. Pitman

Statistical Inference
S. D. Silvey

Models in Regression and Related Topics
P. Sprent

Sequential Methods in Statistics
G. B. Wetherill

(Full details concerning this series are available from the Publishers)

Applications of Queueing Theory

SECOND EDITION

G. F. NEWELL

Professor of Transportation Engineering
University of California, Berkeley

LONDON NEW YORK

CHAPMAN AND HALL

First published 1971 by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE
Second edition published 1982
Published in the USA by
Chapman and Hall
733 Third Avenue, New York NY 10017

Printed in Great Britain at the
University Press, Cambridge

ISBN 0 412 24500 0

British Library Cataloguing in Publication Data

Newell, G. F.
Applications of queueing theory.—2nd ed.—
(Monographs on statistics and applied probability)
1. Queueing theory
I. Title II. Series
519.8'2 TS7.9

ISBN 0–412–24500–0

Library of Congress Cataloging in Publication Data

Newell, G. F. (Gordon Frank), 1925–
Applications of queueing theory.

(Monographs on applied probability and
statistics)
1. Queueing theory. I. Title. II. Series.
T57.9.N48 1982 519.8'2 82-4423
ISBN 0–412–24500–0 AACR2

Contents

Preface to the first edition

The literature on queueing theory is already very large. It contains more than a dozen books and about a thousand papers devoted exclusively to the subject; plus many other books on probability theory or operations research in which queueing theory is discussed. Despite this tremendous activity, queueing theory, as a tool for analysis of practical problems, remains in a primitive state; perhaps mostly because the theory has been motivated only superficially by its potential applications. People have devoted great efforts to solving the 'wrong problems.'

Queueing theory originated as a very practical subject. Much of the early work was motivated by problems concerning telephone traffic. Erlang, in particular, made many important contributions to the subject in the early part of this century. Telephone traffic remained one of the principle applications until about 1950. After World War II, activity in the fields of operations research and probability theory grew rapidly. Queueing theory became very popular, particularly in the late 1950s, but its popularity did not center so much around its applications as around its mathematical aspects. With the refinement of some clever mathematical tricks, it became clear that exact solutions could be found for a large number of mathematical problems associated with models of queueing phenomena. The literature grew from 'solutions looking for a problem' rather than from 'problems looking for a solution.'

Mathematicians working for their mutual entertainment will discard a problem either if they cannot solve it, or if being soluble it is yet trivial. An engineer concerned with the design of a facility cannot discard the problem. If it is trivial, he should recognize it as such and do it. If he cannot solve it correctly, then he must do the best he can. The practical world of queues abounds with problems that cannot be solved elegantly but which must be analysed nevertheless. The literature on queues abounds with 'exact solutions,' 'exact bounds,' simulation models, etc.; with almost everything except common sense

methods of 'engineering judgment.' It is no wonder that engineers resort to using formulas which they know they are using incorrectly, or run to the computer even if they need only to know something to within a factor or two.

In the last 15 years or so, I have suffered many times the frustration of failing to solve elegantly what appeared to be a straightforward practical queueing problem, subsequently to discover that I could find very accurate approximations with a reasonable effort, and finally that I could obtain some crude estimates with almost no effort at all. There is no reason why students should suffer the same way. They should benefit from the mistakes of others and learn to do things in a sensible way, namely in the opposite order.

The following is an attempt to turn queueing theory around and point it toward the real world. It is, in essence, the fourth evolution of a series of lecture notes written for a course entitled 'Applications of Queueing Theory to Transportation.' The relevance of the subject to transportation, rather than to other possible fields of application, derives mostly from the fact that the course was given primarily for transportation engineering students and in a department of transportation engineering. The students had a diverse background, but the majority were graduate students with an undergraduate training in civil engineering. Most had just completed a one-quarter introductory course in probability theory at the level of Paul L. Meyer, *Introductory Probability and Statistical Applications* (Addison-Wesley, 1965) and were taking, concurrently, an introductory course in mathematical statistics. Most would not have had a course in advanced calculus and many would have forgotten much of their elementary calculus (students with a strong formal mathematics background usually had just as much difficulty with some of the graphical techniques as the engineering-oriented students had with the mathematics).

Whereas most of the queueing literature deals with equilibrium distributions of queue length, the main emphasis here is on time-dependent behavior, particularly rush hours in which the arrival rate of customers temporarily exceeds the service rate. The reason for this is that these are the situations which usually create large queues, and the most important practical problems are those in which the size of the queue is really a cause for concern. It turns out that these are among the simplest problems to solve approximately, but are so difficult to solve exactly that no one has yet solved a single special case, at least not in a form suitable for computation.

The techniques emphasized here are mainly 'fluid approximations' and 'diffusion approximations.' The former employs mostly graphical methods; the latter involves some elementary properties of partial differential equations but otherwise uses only a mixture of graphical methods and elementary analysis. Nowhere is use made of generating functions, characteristic functions, or Laplace transforms, which are the standard tools of analysis in conventional queueing theory methods.

No attempt is made here to construct any bibliography except for an occasional reference in the text to some particular paper. Although most of the methods described here have appeared in the literature before in the analysis of special problems, there does not appear to have been any systematic treatment of approximate methods in queueing theory. Some things here may be 'original' in the sense that no one has used a particular mathematical trick to solve a particular problem, but the techniques used are all basically very old, having been used in physics or engineering to solve other types of problems, long before anyone heard of 'queueing theory.'

Except for a short chapter on 'Equilibrium distributions' (Chapter 5), there is very little overlap between what is given here, and what is presented in other books on queueing theory. Although what follows is self-contained, it is not intended as a substitute for the more conventional treatments, but rather as a supplement to them. For further study of the more conventional aspects, it is recommended that a student read D. R. Cox and W. L. Smith, *Queues* (Chapman and Hall, 1961), for a very concise introduction; A. M. Lee, *Applied Queueing Theory* (Macmillan, 1966), for some interesting case histories of attempts to apply queueing theory to practical problems; and J. Riordan, *Stochastic Service Systems* (John Wiley, 1962) or N. V. Prabhu, *Queues and Inventories* (John Wiley, 1965), for a more complete introduction to the typical literature.

It is a pleasure to acknowledge the cooperation and assistance of the transportation engineering students at Berkeley who struggled with me through three preliminary versions of the class notes upon which this book is based. One of these, Brian Allen, was kind enough to help correct the final proofs. The typing and retyping of the notes was done by Phyllis De Fabio.

I would also like to express my thanks to Dr Arnold Nordseick and Professor Elliott Montroll who helped me, as a graduate student and post-doctoral fellow many years ago, to develop some of the attitudes which have influenced this book.

Special thanks go to my wife Barbara, who must endure the lonely life of a scientist's wife, and to my parents, who patiently guided me through my youth.

Berkeley, June 1970 G. F. Newell

Preface to the second edition

More than ten years have passed since the first edition of this book was published. It is interesting to look back now to see what has evolved during that time, and to look closely at the major improvements of this second edition.

The first edition had a title that promised 'applications' but the book contained only a few examples plus some hints on how one might attack certain applied problems, and the 'queueing theory' was not that which several generations of applied probabilists had developed. The style of the first edition, however, needs no defense; the 'theory' has, in fact, been applied to the analysis of a wide variety of practical problems which could not be solved by traditional methods of queueing theory and will continue to be used as a practical tool.

The first edition was essentially the lecture notes of a course which had evolved over a span of about four years. Variations of this course have been given almost every year since, but as it evolved further, more and more applications of the deterministic approximations were added until they consumed more than half of the course. For lack of time, the diffusion approximations were gradually squeezed out.

Although the mathematics of the deterministic approximations is elementary in the sense that it involves only graphs, algebra, and calculus, it requires skill and ingenuity to apply. Students certainly do not consider it easy. Without question, however, these deterministic approximations have found application to a much wider range of practical problems than the stochastic theory simply because the stochastic analysis of even the simplest systems which involve several servers or customer types is too tedious to be of much practical value.

Since a textbook must, of necessity, treat only simple illustrations which can be described in one or two lectures, this second edition still gives only some hints as to how one can analyze more complex problems. It is certainly inappropriate to try to describe in detail how,

for example, these methods can (and are) used for the analysis of traffic signal synchronization, production line design, bus dispatching policies, etc. The difficulty in these more complex applications comes not so much in the derivation of formulas for delays, queue lengths, etc. as in the interpretation of the results which typically contain many parameters. Each area of application involves special understanding of what questions one is trying to answer.

Although the text itself contains few references, I have added a bibliography of some of the applications of deterministic queueing with which I am familiar (mostly in the area of transportation engineering).

Even though most of the diffusion theory was eliminated from the original course, much of my own research during the last ten years has been in the area of stochastic approximations. When I started to write the present revision, I envisaged presenting an approximate stochastic version of most of the models described under the deterministic theory. I even taught an 'advanced' course as a means of testing some of the revised notes on stochastic approximations. This course, however, started from the beginning, and since most students of queueing theory are unfamiliar with properties of partial differential equations, this course barely covered some of the basic qualitative features of single-server queues. Much of what I had hoped to do must wait for some future occasion.

The first edition has been almost entirely rewritten, but the chapter titles remain almost the same, as does the general philosophy and style. The rejection of traditional approaches to queueing theory is perhaps even more emphatic.

Chapter 1 has changed little except that some notation has been revised and some other types of graphical representatives are discussed. The first problem set now introduces a hand simulation of a 'random walk' queue which, with little effort, gives students some preliminary feeling for the magnitude of statistical fluctuations. These simulations are very helpful in illustrating some of the effects discussed in later chapters.

The introduction to Chapter 2 on fluid approximations now discusses more thoroughly the qualitative features of some real queueing situations and questions of concern to engineers designing service systems. Two typical types of systems for which deterministic approximations are particularly useful are then analyzed, the rush hour with a steady service rate and a system with interrupted (pulsed) service (traffic signals and buses).

Chapter 3, describing the behavior of systems with several servers and/or customer types, has been considerably expanded to emphasize the benefits derived from drawing graphs of the cumulative arrivals and departures of anything which satisfies a conservation principle. Some of the illustrations are conventional (tandem queues or multiple-channel server queues) but many are designed simply to illustrate the art of modelling simple systems. The problem set contains other illustrations many of which are derived from real applications. It is the material in this chapter that has displaced much of the stochastic approximations in the course I have taught because this is the type of analysis that has proven to be most useful in the design of real systems. This chapter has become the main focus of the course in recent years.

The introductory chapter on stochastic models has been extensively rewritten to emphasize the theme that stochastic models should be chosen to represent the actual behavior of real systems not just to yield mathematically convenient formulas. It describes typical qualitative properties of arrival and departure processes, culminating in the argument that if one cannot find a convenient 'exact' model, one can usually evaluate the things one really wants simply by constructing a few hypothetical realizations of the cumulative arrivals and departures.

Chapter 5 on equilibrium distributions now includes a simple dimensional argument giving the typical magnitude of equilibrium queue lengths and relaxation times and discusses the question of how rapidly the traffic intensity can change if the queue distribution is to stay close to the equilibrium distribution. In the first edition these issues were postponed until Chapter 6 and obtained as a result of a rescaling of the diffusion equation (which makes the argument unnecessarily obscure). This chapter also contains a few more examples of traditional equilibrium queueing problems than the first edition, although the reader is still referred to other texts for a more detailed introduction to conventional methods.

Chapter 6, which is entirely new, deals with systems in which customers seldom interact because the server has a sufficiently large number of channels and/or the arrival rate is low. It includes the standard infinite-channel systems and loss systems but also some approximations for queueing systems in which customers are delayed only rarely. The behavior of queueing systems under light traffic has, for some reason, received little attention in the queueing theory literature. If, however, one can describe the system behavior for both

light and heavy traffic, it requires little imagination to guess how the system would behave for intermediate traffic (where exact results may be difficult to obtain). Perhaps this chapter will inspire further research on low traffic approximations.

Chapters 7, 8, and 9 are devoted to diffusion approximations. Chapter 7 starts with a fairly general stochastic process in two (or more) dimensions having the property that the state of the system changes by only a small amount in a short time. This is then specialized to treat properties of the joint arrival and departure processes, diffusion equations with state-independent coefficients, boundary conditions, and one-dimensional equations for the queue distribution. Chapter 8 deals with equilibrium and transient queue behavior for constant arrival and service rates, while Chapter 9 treats time-dependent queues, particularly the stochastic version of the rush hour and pulsed service problems introduced in Chapter 2.

Considerable work has been done in recent years on various queueing systems; multiple-channel servers, tandem queues and some more general networks of service systems. Although I had, at one time, planned to add perhaps two more chapters dealing with some of these results, I have (temporarily) abandoned the attempt because it would take too much time to put much of this in proper perspective. The complexity of the results in the existing literature is way out of proportion to its usefulness (including my own research). It will be some time before I can sift out from this material that which might be appropriate for an introductory book, but maybe there will be a third edition some day. Certainly the most difficult task in the analysis of any real system is the collection of relevant data to describe what is happening and the existing literature on queueing theory gives very little assistance to an engineer in deciding what to measure and how to interpret the results.

It is a pleasure to acknowledge the help I have received from the many students who have noted errors in preliminary revisions of these notes, suffered through unclear expositions, and tried to solve ill-posed problems. I have learned the most, however, from those students and colleagues who have actually collected data and analyzed real problems, particularly Van Olin Hurdle, now at the University of Toronto, who is a master at the use of graphical techniques.

Phyllis DeFabio has continued to type most of the multiple versions of notes from which this book derives.

CHAPTER 1

Introduction

1.1 Nature of the subject

Queueing theory is concerned, generally, with the mathematical techniques for analyzing the flow of objects through some network. The network contains one or more locations at which there is some restriction on the times or frequencies at which the objects can pass. A conservation principle applies; the objects do not disappear or disintegrate. Any object which cannot immediately pass some restriction is stored in some real or fictitious reservoir until it can. As long as there are objects in the reservoir waiting to pass, the facility will pass them as rapidly as the restriction will permit.

The objects could be anything which move from place to place (and satisfy a conservation principle), people, cars, water, money, jobs to be done, etc. The restrictions could be a service facility for people, a highway bottleneck for cars, a valve regulating the flow of water, a rule for money transactions, a finite labor supply for work to be done, or a finite speed with which a computer can handle calculations to be done.

One could consider all sorts of networks, but we will be concerned here mostly with the rather simple geometry in which all objects flow along some channel and all pass through the same restrictions as illustrated schematically in Fig. 1.1(a). In many (perhaps most) real systems, however, the objects are not identical. Although these objects may differ in many ways (color, size, name, etc.) the typical characteristics of these objects which are relevant to queueing analysis are:

(a) Different objects may take different lengths of time to pass the restrictions (there are long jobs and short jobs).
(b) Delays to different objects may be worth different amounts of money (to delay an aircraft carrying 400 passengers costs more than to delay a private aircraft).

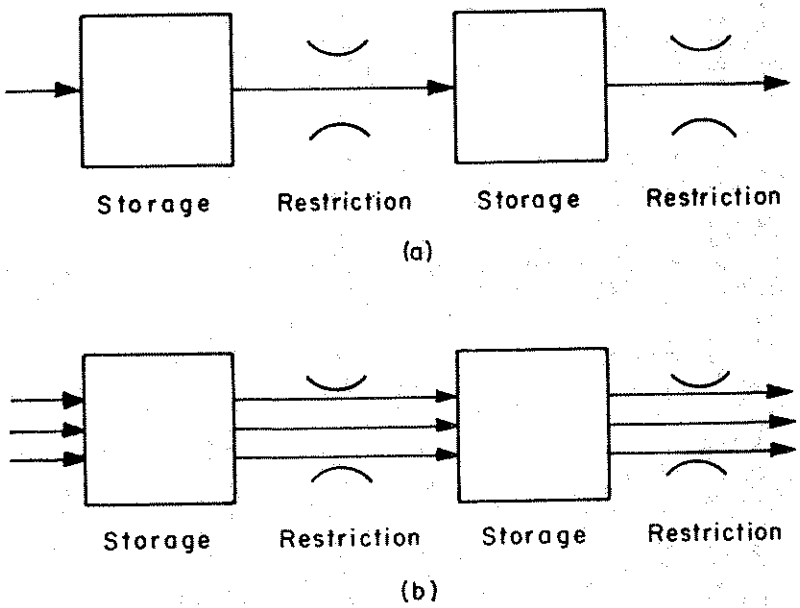

Figure 1.1 *Schematic picture of the flow of objects along a channel*

If one decomposes the objects into several categories, it is usually assumed that a conservation principle applies separately to each object type. Objects cannot disappear; nor can they change identity (a long job remains a long job, a commercial aircraft remains a commercial aircraft). A more realistic schematic picture of the system would be as shown in Fig. 1.1(b) with many streams passing the same restrictions. The restriction is usually a collective one restricting the rate at which objects can pass in various proportions.

The ultimate practical purpose of any theory is to make predictions of what will happen in some experiment that one has not yet done. The purpose of queueing theory is to provide a mechanism for predicting how some hypothetical or proposed system will behave. Sometimes one wishes to design a completely new facility where there was none before and would like to compare the predicted performances of various proposed systems. In this case one must usually make conjectures about the arrival rates of various objects and the consequences of various restrictions (facilities). The more common problem, however, is one in which a facility already exists. One can make observations on its present behavior; but, from these observations, one would like to predict how the system would behave if certain changes occur. The change might be an increased demand (arrival rate) as projected for some future time or it might be an improvement in the service rate of some facility or a change in strategy for sequencing the service of different object types.

The same type of mathematical techniques apply to a very wide variety of flow systems, but the measures of performance or goals may be quite different for different systems. In queueing theory one typically associates an implied cost with any delay and also a cost for providing a higher service rate at any restriction. The usual problem is to compare delays (and operating costs) for systems with different service components or strategies. Some typical systems of this type are:

(a) Objects move along a production line on which various tasks are performed at various rates but there is a penalty for storage of unfinished products. One might be able to decrease this storage cost by shifting some labor.
(b) Cars move along a highway having certain bottlenecks (such as traffic intersections) and there is an inconvenience associated with waiting. One might be able to decrease the delays by appropriate adjustment of the signal timing.
(c) Patients wish to enter a hospital which can handle only finitely many patients at a time, but delays may cause serious consequences, more so to some patients than to others.

The mathematical models of 'inventory theory' are quite similar; objects pass from a supplier to a reservoir to a customer. The objectives and strategies, however, are quite different from the above. There is usually a high penalty for an empty inventory (or a queue of unfilled orders). The strategy is to regulate the input (reorder stock) rather than the flow out of the inventory. In the 'theory of dams' one has an input (rain) to a reservoir and an output (consumption) but the regulation or strategy is applied to the output rather than the input. In insurance, gambling, banking, etc., one is concerned with the flow of money. The money is, in fact, only some numbers on an account book, but the rules of transfer are the same as if it were something physical. There is a flow of money into an account (an investment rate), and a flow out; and a resulting storage (balance). The strategy now may involve regulation of either the input or the output or both with potentially a rather complex set of objectives. In the university one has a flow of students and faculty into and out of the system with a

resulting population of each. There is a conservation principle: what comes in must go out, one way or another.

Since there is such a wide variety of possible applications of queueing theory or related theories, it is rather difficult to agree on some common terminology. Much of the conventional terminology has evolved from the following hypothetical system. The objects which move from place to place are called customers, which one typically imagines to be people. They arrive at some service point (a bank counter, a taxi stand, a highway intersection) at certain specified times. The service facility (the restriction) requires some time to serve each customer but is capable of serving only finitely many at a time (possibly just one). If customers arrive faster than the facility can serve them, they must wait in a queue (the reservoir).

Typically, both the customer arrival times and the service times are assumed to follow some specified stochastic behavior. One wishes to relate the delays to the customers, and the number of customers in the queue to the given properties of the arrival and service. In practical applications one usually wishes further to compare the operation of several possible modes of operation with respect to its type of service, cost, etc. Should there, for example, be a single queue for all bank tellers or separate queues for each?

In most of the following descriptions, we will also use the terms customer, server, and queue (except when it is clearly inappropriate) even though the terms object, restriction, and reservoir are more suggestive of the wide range of physical systems to which the same mathematics will apply.

What complicates the mathematical modelling of most real systems is that repetition of an experiment 'under identical conditions' does not usually yield exactly the same results every time. To make predictions of future behavior it is, generally, necessary to postulate some stochastic model and to estimate probabilities for certain events, i.e., fractions of times in which various events would happen over many repetitions of the observations. Unfortunately, in most applications of queueing theory, the observed properties of the stream of objects passing any point do not conform to any mathematically 'simple' stochastic model. To describe how the system behaves under repetitions of an experiment one must actually repeat the experiment to see what happens. It is usually rather dangerous to speculate on what would happen if the system were consistent with some hypothetical model. From repeated observations on an existing system one must, however, still make conjectures as to how the system would behave if certain changes were made in the system.

1.2 Mathematical and graphical representations of events

Since a study of any queueing system should start from some (real or implied) experimental observations, let us imagine that we station observers at various points in the system. For each service point we might place one observer just upstream of the server to record the times and identity of each customer that passes him. If customers travel with a finite speed and the queue has a positive physical length, we might ask this observer also to convert his observations into the times at which the customers would have reached the server if there were no queue (or if the queue occupied no space). We place a second observer at the server to record the times and identity of customers entering the server and possibly a third observer just downstream of the server to record the times at which customers leave the server (and their identity).

We will assume that at time 0, when the observations begin, the system is empty. If it is not, we can imagine that the system was empty for time $t < 0$ (whether it was or not) but that each observer records arbitrarily that any customer already downstream from him at $t = 0$ passed him at $t = 0$.

One could also imagine that the first observer assigns labels to each customer that passes him (for example, a number) and he asks the customer to keep the label with him at all times. All customers are now different by virtue of their labels (if not for other reasons) and each customer individually satisfies a conservation principle in that he does not disappear or change identity during the period of observations.

Since one may be interested in the possibility and possible consequences of the fact that customers might interchange positions in the queue or server (they pass each other), such a labeling will permit the downstream observers to detect any such rearrangement. We would, however, like to make a distinction between customers which differ only by virtue of having different labels and those which differ in some more significant way (the delay time of one is worth more than another or one is known to require a longer time in service) which may be relevant in selecting some service priorities. If it is relevant to treat the customer arrival as the superposition of several identifiably different streams, each satisfying a conservation principle, then we will ask each observer to record separately the times at which customers of each category pass.

For now, we will consider only the observations associated with customers within the same category, as if those of other categories (if any) were not there, or not observed.

If the first observer numbers the customers (of the same category) consecutively and assigns the numbers as labels let

$$0 \leq t_1 \leq t_2 \leq \ldots$$

represent the time at which customers 1, 2, . . . arrive at the server or would arrive at the server if the queue occupied no space. These times can be represented graphically as a sequence of points on the real line as in Fig. 1.2(a). It is more convenient, however, to represent these data by a graph of a function $A(t)$ which, for each t, represents the cumulative number of arrivals to time t:

$$A(t) = \text{number of } t_j \text{ with } t_j \leq t. \tag{1.1}$$

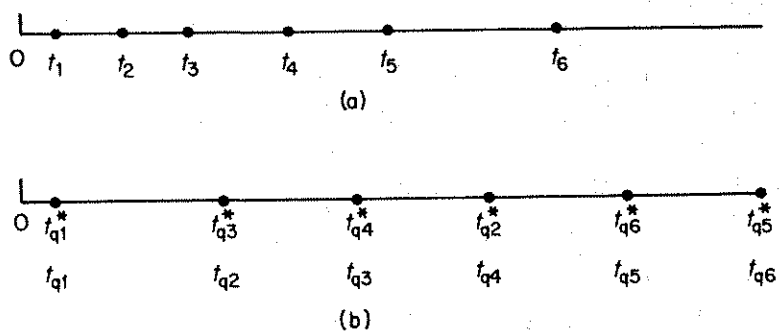

Figure 1.2 *Representation of arrival and departure times by points on a line*

This is a step function which increases by one at each time t_j as shown in Fig. 1.3.

One immediate advantage of this representation is that we will also have occasion to analyze arrivals and departures of quantities other than *numbers* of customers; for example, the cumulative value of products or the cumulative amount of work to be done. If the quantity in question is also conserved (what comes in must go out), then it is easy to generalize (1.1) to

$$A(t) = \text{cumulative quantity (or number) to arrive by time } t. \tag{1.1a}$$

This is also a monotone nondecreasing function of t, but it is not necessarily integer valued. It may or may not be a step function depending upon whether or not the arrivals are discrete. If they are discrete, the steps need not be equal.

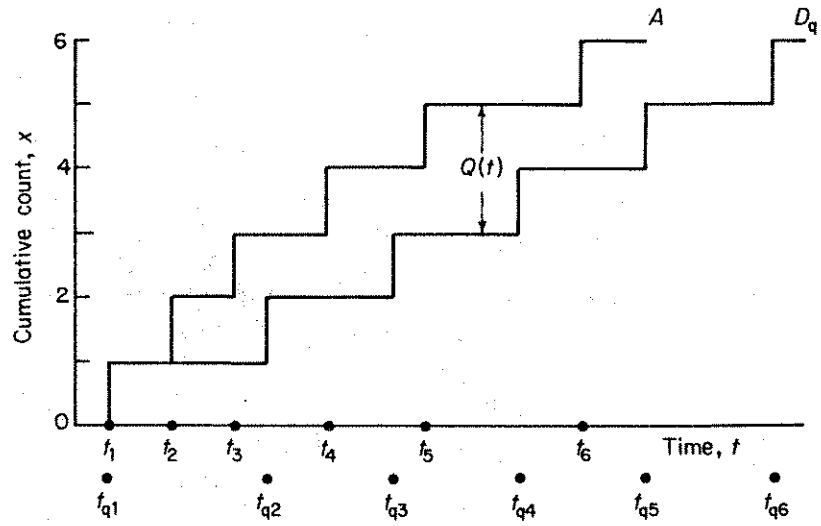

Figure 1.3 *Graphical representation of cumulative arrivals and departures from a queue*

For some purposes it is convenient to think of the graph in Fig. 1.3 as a graph of A versus t, i.e., the function $A(t)$, but for other purposes it is convenient to think of it as a graph of t versus A, i.e., the function $A^{-1}(x)$. If x is integer, we can consider it to be the label on the last arrival; for noninteger values we would interpret x as the cumulative number of arrivals or fractions thereof,

$$A^{-1}(x) = t_j \quad \text{for} \quad j-1 < x < j. \tag{1.2}$$

Perhaps it is better yet to consider this as simply a curve A in the t, x plane without specifying which variable is the 'independent' variable.

The second observer will record the times at which customers enter the server, along with the customer number assigned by the first observer. Let

t^*_{qj} = time customer number j leaves the queue and enters the service.

Whenever there is a queue of more than one customer, the order in which these customers enter the service need not be the same as the order in which they arrive. A rule describing how customers are selected from a queue is described in the queueing literature as 'queue discipline.' The discipline in which customers are served in order of their arrival is usually called 'first in, first out' or FIFO. This is the

simplest to describe mathematically because the times t^*_{qj} must now satisfy the conditions

$$0 < t^*_{q1} \leq t^*_{q2} \leq t^*_{q3}, \ldots, \quad \text{for FIFO.} \tag{1.3}$$

Some common examples of queue disciplines other than FIFO are:

(a) Last in, first out (LIFO). Suppose that letters to be typed or order forms to be processed accumulate in a pile, each new addition being placed on top. The typist or clerk now services the letters (the customers) by taking each new task from the top of the pile. A newly arriving task will be the next to be served provided it can be served before another arrives.

(b) Service in random order (SIRO). Passengers waiting to board a bus might appear to board in an order which bears no relation to the order in which they arrive. Random order of service is usually defined to mean that whenever a customer is selected from the queue, the selection is made in such a way that any customer in the queue at the time of selection is equally likely to be chosen.

(c) Priority service. Particularly if one has not initially decomposed customers into categories and considered each category separately, one might order the customers in queue according to some identifiable characteristic (length of job or value). The next customer to enter the service is then the one in queue with the highest ranking (top priority) at the time. There are several variations on this depending upon whether or not a high priority customer must wait until the next service completion to enter the service or if it can displace a lower priority customer from the service.

For some purposes (particularly if one is interested only in counts of customers but not their identity) it may be convenient for the second observer simply to record the ordered times at which customers enter the service even though the queue discipline is not FIFO. He, in effect, relabels the customers and defines t_{qj} as the time of the jth departure from the queue so that

$$0 < t_{q1} \leq t_{q2} \leq t_{q3} \leq, \ldots.$$

If one represents the times at which customers leave the queue by points on the real line as in Fig. 1.2(b), the set of times t_{qj} and t^*_{qj} represent simply two different labelings of the same set of points.

As with the times t_j, it is possible also to represent the t_{qj} by a graph

$$D_q(t) = \text{number of } t_{qj} \text{ with } t_{qj} \leq t, \tag{1.4}$$

the cumulative number of departures from the queue by time t, or more generally

$$D_q(t) = \text{cumulative quantity or number to leave the queue by time } t. \tag{1.4a}$$

The inverse of this, $D_q^{-1}(x)$, describes the ordered departure times

$$D_q^{-1}(x) = t_{qj} \quad \text{for} \quad j-1 < x \leq j. \tag{1.5}$$

If one draws both $A(t)$ and $D_q(t)$ on the same graph, as in Fig. 1.3, the curves cannot cross because, for any t, the number of customers which have left cannot exceed the number which have arrived. The vertical distance between the two curves at any time, representing the number of customers who have arrived but have not yet left the queue, is

$$\begin{aligned} &\text{quantity or number in the queue (queue length)} \\ &= Q(t) = A(t) - D_q(t) \geq 0. \end{aligned} \tag{1.6}$$

It is, of course, also true that

$$D_q^{-1}(x) - A^{-1}(x) \geq 0$$

because x customers cannot have left until at least x customers have arrived.

The curve $D^{(t)}_q$ does not display the queue discipline. It gives only the count of departures but not the identity, and its inverse gives only the ordered departure times t_{qj}. It is possible, however, to draw a graph D^*_q which does display both the departure times and the queue discipline. If we consider x as the independent variable, we can draw a function of x having values

$$t^*_{qj} \quad \text{for} \quad j-1 < x \leq j$$

instead of the t_{qj} which defined the curve D_q. Whereas the curve D_q described a monotone nondecreasing function of x or t, the curve D^*_j will not be monotone unless the customers are served in the order in which they arrive ($t_{qj} > t_{qk}$ if $j > k$), consequently the curve D^*_q will not generally define a single-valued function of t.

If we draw both A and D^*_q on the same graph as in Fig. 1.4, the horizontal distance from $t = 0$ to A at height x, $j-1 < x < j$ is the time t_j at which customer j arrives, and the horizontal distance to D^*_q is the time he left the queue. The difference between them, the horizontal distance from A to D^*_q, is the time which the jth customer spends in queue

$$w_j = t^*_{qj} - t_j \geq 0. \tag{1.7}$$

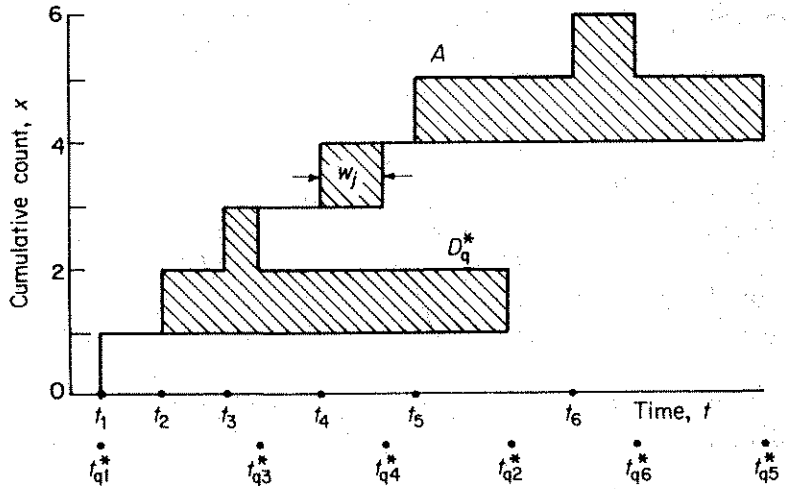

Figure 1.4 *Graphical representation of departure times*

This is also equal to the area of the rectangular strip between A and D_j^*, $i-1 < x < j$.

Whereas Fig. 1.3 gives a simple geometric interpretation of $Q(t)$, Fig. 1.4 gives a simple geometric interpretation of the w_j. One could also identify the w_j from Fig. 1.3 if this graph were supplemented with some scheme for identifying which step in D_q gives the departure time associated with the jth step of A.

Since D_j^* shows both the departure times and the order of customers, it must also define $Q(t)$. The curves D_q^* and A enclose an area, the locus of all horizontal lines from A to D_q^*. If one draws a vertical line at time t, it will slice this area in such a way that any point x in this area is identified with a customer who has arrived but has not yet left. The total length of vertical line between A and D_q^* is the number in the queue $Q(t)$. This is also true of Fig. 1.3, but in Fig. 1.3 this is a single line segment.

If the queue discipline is FIFO the curves D_q and D_q^* are, of course, the same curves.

The above definitions of A, D_q, etc., are simply a description of what two observers recorded and involve no 'theory' of what happened to the customers in the queue. If we were to place a third observer downstream of the server, he could record similar information independent of what happens in the server. From this we can define t_{sj}^* as the time at which customer j leaves the service, t_{sj} as the ordered

times at which customers leave, $D_s(t)$ as the cumulative number of customers to leave, and D_s^* as the curve defined by

$$t_{sj}^* \quad \text{for} \quad j-1 < x < j.$$

If we draw the curves A and D_s^* (instead of D_q^*) on the same graph, the horizontal distances would be interpreted as the times customers spend in the queue plus service (instead of just the queue) and a vertical slice of the area between A and D_s^* would determine the number of customers in the queue or in service (instead of $Q(t)$). Similarly, if we compare D_q^* and D_s^*, the horizontal distances would be the times customers spend in service and a vertical slice of the area between D_q^* and D_s^* would determine the number of customers in service at any time.

1.3 Modelling

In order to make predictions of what would happen in an experiment one has not done, one must relate the properties of D_q, D_s, etc., to any rules governing the behavior of the customers in the queue and the server. In a typical queueing problem one proposes one or more possible curves $A(t)$ or perhaps some probability distributions for $A(t)$, a description of the queue discipline (if relevant), and the manner in which the server operates. From this one wishes to evaluate, in effect, the curves D_q, D_s or any derived properties thereof, perhaps probability distributions of the w_j or $Q(t)$.

The rules governing the dynamics of the system could conceivably be quite complicated and involve interrelations between the service times, arrival times, customer types, etc., restricted only by the universal principle that a customer cannot leave before he has arrived or equivalently that the queue cannot be negative. Most systems which have been analyzed in the queueing literature, however, have rather simple rules. The mathematical complications are not directly associated with the queue dynamics, but with the stochastic analysis. Even though the postulated relations between arrival times and departure times appear quite simple, they lead to fairly complex relations between probability distributions for arrivals, departures, queue lengths, waits, etc.

A description of the server should at least define a relation between the curves D_q^* and D_s^*, i.e., between the t_{qj}^* and t_{sj}^*. The simplest rule is one in which the times each customer will be in service (or probability

distributions for the service times) are given, i.e.,

$$s_j = t^*_{sj} - t^*_{qj} \qquad \text{for all} \quad j.$$

From this one can, of course, immediately construct the D^*_s from the D^*_q.

In some situations, for example, customers being served by taxis, the customer might consider his service to start when he boards the taxi and to be completed when he reaches his destination, so that the s_j would be his trip time. For the next customer who is waiting to be served, however, the relevant 'service time' is the time from the start of the last service until the taxi accepts him. Alternatively, he might consider his service to start when the taxi has discharged the previous customer and accepted his order and to end when he is discharged.

In the analysis of most queueing problems it is usually implied that the 'service time' is the time from the start of one service until the server is available to accept the next customer, since this is the time which is relevant to the evolution of the queue.

If one has m taxis serving the same queue, they could be serving as many as m customers simultaneously. A server of this type consisting of m separate servers, each of which serves only one customer at a time, is called an m-channel server. For such a system, a new customer can enter service as soon as any channel is free. The order in which customers complete service is not necessarily the same as the order in which they started service (i.e., the service discipline is not FIFO). A server which can, at the start of any service, accept several customers at the same time (such as a bus or elevator) is called a bulk server.

The rules governing the server will generally also specify that upon completion of one service at time t another service should start immediately provided that $Q(t) > 0$. If the server is a bulk server, the rules will also specify how many customers the server can accept, given the value of $Q(t)$. If $Q(t) = 0$, the rules should specify when the next service starts, usually when the next customer arrives.

For most service systems that one encounters in queueing applications it is quite easy to follow the rules iteratively (either numerically or graphically) and construct D^*_q and D^*_s from a given curve $A(t)$ and the service times (i.e., perform a 'simulation').

If, for example, the server is a single-channel server one would specify $A(t)$ or equivalently the t_j, the service times s_j, and the queue discipline. The iterative rules describing the evolution of the D_q and D_s are that

$$t^*_{sj} = t^*_{qj} + s_j;$$

and, if the queue discipline is FIFO, the $j+1$th customer starts service at time t^*_{sj} if he has arrived, i.e. if $t_{j+1} < t^*_{sj}$, but, if not, he enters service as soon as he does arrive at time t_{j+1}. Thus, $t^*_{qj} = t_{qj}$, $t^*_{sj} = t_{sj}$ and

$$t_{qj+1} = \max(t_{j+1}, t_{sj}) = \max(t_{j+1}, t_{qj} + s_j). \tag{1.8}$$

Starting from an initial condition that the system is empty and $t_{q1} = t_1$, (1.8) determines each t_{qj+1} from the previous one, t_{qj}.

If we subtract t_{j+1} from both sides of (1.8) we can also write it in the form

$$t_{qj+1} - t_{j+1} = \max(0, t_{qj} - t_{j+1} + s_j). \tag{1.9}$$

For FIFO queue discipline

$$w_j = t_{qj} - t_j,$$

therefore, the w_j satisfy the equations

$$w_{j+1} = \max(0, w_j + s_j - (t_{j+1} - t_j)). \tag{1.10}$$

This describes the waiting times iteratively in terms of the service times and interarrival times $t_{j+1} - t_j$.

1.4 Averages

For most queueing systems, it requires only elementary mathematics to describe the detailed dynamics of the system. It is, essentially, the approximate theory of queues which is complicated. In analyzing the behavior of queues, one does not care to observe the arrival and departure times of every customer. On the one hand, this involves tedious manipulations. On the other hand, they are not very interesting data because many of them could not be reproduced if the experiment were repeated. One would prefer to specify only a few characteristics such as some average arrival rate or service rate, things which are nearly reproducible.

Even if one could describe in detail exactly how large the queue would be at every instant of time, one would probably disregard much of the detail. One would prefer to have only some approximate description or measure of performance. This is, in essence, why one treats queueing phenomena as stochastic processes. One only wishes to consider the average behavior of the system over a range of conditions, not the details of what happens in any particular experiment.

Whether one treats the system stochastically or deterministically,

there are certain gross properties one may wish to calculate; for example, the average wait in queue for a set of n customers or the time average queue length over some period of time.

The average time in queue for customers $j+1$ to $j+n$ inclusive is defined as

$$\langle w_k \rangle = \frac{1}{n} \sum_{k=j+1}^{j+n} w_k = \frac{1}{n} \sum_{k=j+1}^{j+n} (t_{qk}^* - t_k). \tag{1.11}$$

The w_k can also be interpreted as the area of a horizontal strip $k-1 < x < k$ between A and D_q^*. The sum of the w_k is, therefore, the area enclosed by A, D_q^* and two horizontal lines $x = j$ and $x = j+n$ as shown in Fig. 1.5(a).

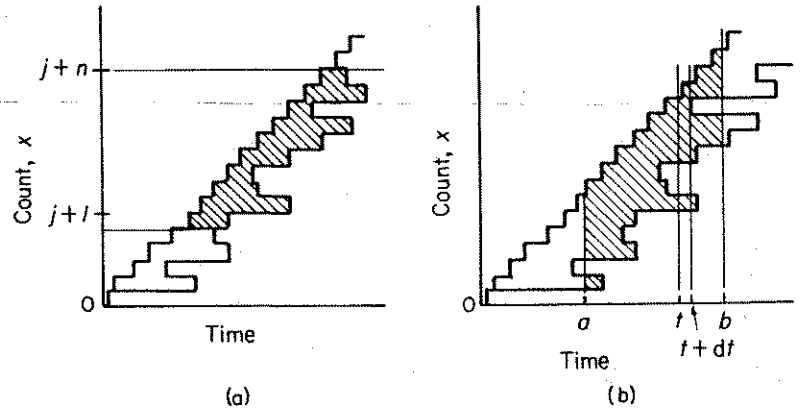

Figure 1.5 *Areas defining average wait and queue length*

The average queue length during some time interval (a, b) is defined as

$$\langle Q(t) \rangle = \frac{1}{(b-a)} \int_a^b Q(t)\,\mathrm{d}t. \tag{1.12}$$

Since $Q(t)$ can be interpreted as the length of cross section cut from the region between A and D_q^* by a vertical line at t, $Q(t)\,\mathrm{d}t$ represents the area cut from this region by a vertical strip between t and $t + \mathrm{d}t$. Thus the integral from a to b represents the area enclosed by A, D_q^* and two vertical lines at $t = a$ and $t = b$, as in Fig. 1.5(b).

If, in (1.12), we chose a and b as any times for which $Q(a) = Q(b) = 0$, and in (1.11) we chose $j = A(a), j+n = A(b)$, then the two areas in Fig. 1.5(a) and (b) would be the same areas. Both the horizontal and

vertical lines of Fig. 1.5(a) and (b) would cut the region between A and D_q^* at a single point. The total queueing time during the time interval (a, b) would be the same as the total queueing time for customers $j+1$ to $j+n$. In (1.11) this is represented as the sum of horizontal strips whereas in (1.12) it is represented as the sum of vertical strips. With the a, b, j, and n chosen in this way, it follows that

$$(b-a)\langle Q(t) \rangle = n \langle w_j \rangle = \text{total queueing time}$$

or

$$\langle Q(t) \rangle = \frac{n}{(b-a)} \langle w_j \rangle. \tag{1.13}$$

We can also define

$$\lambda_{ab} = n/(b-a) \tag{1.14}$$

as the average arrival rate during the time interval (a, b).

If the queue behavior is such that the queue vanishes repeatedly, every day at midnight or at other perhaps irregular (maybe stochastic) times with a finite spacing, the above relations would be valid for any or all choices of times a and b when the queue vanishes (not just consecutive times). Even if the queue does not vanish at times a and b, but it vanishes at many other times between a and b including some times close to a and b, the areas in (1.11) and (1.12) would differ only at the ends of the region between A and D_q^* and from a or b to the nearest time where $Q(t)$ does vanish. If these end areas are negligible compared with the total areas (1.13) is still approximately correct.

Although (1.13) is exactly correct if $Q(a) = Q(b) = 0$, essentially as a direct consequence of the definitions of the averages, and it may be approximately true under more general conditions, the $\langle w_j \rangle$, $\langle Q(t) \rangle$, and λ_{ab} will generally depend upon the particular choice of a and b.

Much of the mathematical literature on queueing theory deals with what is known as 'stationary arrivals' generated by a hypothetical source which operates from $t = -\infty$ to $t = +\infty$. The arrivals are assumed to have certain stochastic properties but of such a nature that for $a \to -\infty$ and $b \to +\infty$ each of the above averages, particularly the λ_{ab}, has a well defined limit. If such limits exist, (1.13) must, of course, be valid for $a \to -\infty$ and $b \to +\infty$. Theorems relating to the validity of (1.13) can be quite sophisticated,† but the complications are

† Little, J. D. C. (1961) A proof for queueing formula $L = \lambda W$. *Operations Research*, **9**, 383–7.
Jewell, W. S. (1967) A simple proof of: $L = \lambda W$. *Operations Research*, **15**, 1109–16.

mainly associated with the mathematical conditions which will guarantee the existence of the appropriate limits. From the point of view of practical applications, this is, however, rather academic since no real process runs from $t = -\infty$ to $t = +\infty$.

Equation (1.13) has an obvious generalization to cases in which the x coordinate measures some substance other than numbers of customers. Suppose, for example, that $A(t)$ measured the cumulative arrivals of a substance which arrives in discrete units of size a_k at times t_k, and leaves in the same units of size a_k at times t^*_{qk}. If $Q(t)$ now measures the amount of substance in the reservoir and $Q(a) = Q(b) = 0$, the area between A and D^*_q from a to b would be

$$\int_a^a Q(t)\mathrm{d}t = \sum_{k=j+1}^{j+n} a_k(t^*_{qk} - t_k),$$

with the sum extending over all arrivals between times a and b, or

$$\langle Q(t)\rangle = \frac{1}{(b-a)} \sum_{k=j+1}^{j+n} a_k(t^*_{qk} - t_k). \tag{1.15}$$

There are (at least) two possible interpretations of the right hand side of (1.15). If a_k represents the value of the kth arriving object or the cost per unit time for delay of the object (interest cost on the value), we could interpret

$$\frac{1}{n}\sum_{k=j}^{j+1} a_k(t^*_{qk} - t_k)$$

as the average cost of delay per arrival and write (1.15) as

$$\langle Q(t)\rangle = \lambda_{ab}\left[\frac{1}{n}\sum_{k=j}^{j+n} a_k(t^*_{qk} - t_k)\right] \tag{1.16}$$

with λ_{ab} the (average) arrival rate of objects. Alternatively we might interpret

$$\frac{1}{\sum_{k=j}^{j+n} a_k} \sum_{k=j}^{j+n} a_k(t_{qk} - t_k)$$

as the delay per unit of substance (for example, delay per person if the arrivals are buses with a_k passengers per bus) and write (1.15) as

$$\langle Q(t)\rangle = \left[\frac{\sum_{k=j}^{j+n} a_k}{(b-a)}\right]\left[\frac{1}{\sum_{k=j}^{j+n} a_k}\sum_{k=j}^{j+n} a_k(t^*_{qk} - t_k)\right] \tag{1.17}$$

= (arrival rate of substance) × (average delay per unit of substance).

The above formulas were derived from the geometric properties of two curves (A and D^*_q) but did not depend upon how the curves were generated. If the curves had been A and D^*_s instead of A and D^*_q, and the system was empty at times a and b, we would interpret (1.13) or (1.17) as

$$\langle Q_s(t)\rangle = \text{average substance in the system}$$
$$= \text{(arrival rate of substance)} \times \text{(average time in system per unit of substance).} \tag{1.18}$$

Similarly, if the curves were D^*_q and D^*_s but there was nothing in the server at times a and b, (1.13) or (1.17) could be written as

$$\text{average substance in service} = \text{(arrival rate of substance)} \times \text{(average time in service per unit of substance).} \tag{1.19}$$

Furthermore, if a and b were times when the system is empty, the arrival rates in (1.17), (1.18), and (1.19) would be all the same and the average delay times refer to the same set of objects.

The above formulas are true regardless of the queue or server discipline, but the average queue length, which does not recognize the identity of customers, could have been evaluated from the curves A and D_q. If a change in the queue discipline does not change the curve D_q, i.e., the times at which customers enter the server, then the average wait per customer is independent of the queue discipline. The D_q will be independent of queue discipline, thus unaffected by an interchange of customers, if and only if the service times of all customers are equal (or if the service times are random, they are 'interchangeable').

If this is true, the advantage of FIFO discipline over other types of queue disciplines is related not to the average delay but to the variations in delay about the average. Obviously, last in, first out

discipline gives a high proportion of very short delays (less than one service time) but also some very long delays.

1.5 Applications of $L = \lambda W$

In a typical queueing problem, one specifies the arrival rate of customers, λ_{ab}, and the service times. One wishes to evaluate (among other things) the average queue length $\langle Q(t) \rangle$ and/or the average delay per customer $\langle w_j \rangle$. Equation (1.13) relates these two unknowns in a simple way, so it suffices to determine either one or the other.

Equation (1.19), however, has some more direct applications because both factors on the right hand side are usually given. Consequently, one can immediately evaluate the average number of customers in service, provided, of course, that the long time arrival rate of customers is sufficiently low that the system will empty occasionally, i.e., the server is capable of serving customers fast enough eventually to keep up with the arrivals.

If the server is an m-channel server, with each channel serving at most one customer at a time, the number of customers in service is the same as the number of busy servers. Thus (1.19) determines also the (time) average number of servers that are busy, which is obviously a lower bound on the number of servers m which one needs to keep up with the arrivals. A telephone company, for example, might know the frequency of calls (arrival rate) between two cities and the average duration of a call (service time). It wishes to know the minimum number of channels it must provide to handle the traffic. An airport designer may know how many aircraft arrivals are expected each day and the average 'turn around time', i.e., the average time an aircraft occupies a gate position (service time). He wishes to know the minimum number of gate positions he must build.

In each case, the designer would also like to know something about the peaking of demand and the delays that would result from various choices of m, but (1.19) will not determine that. In fact, if all channels are identical, presumably the service time of a customer will be independent of which channel it uses and, consequently, also independent of the number of channels. Thus the right hand side of (1.19) does not depend upon m (provided m is large enough eventually to serve the arrivals) and the average number of busy servers is independent of the number of channels.

The difference between a service with many channels and one with only the minimum number is that the former can serve customers with

less delay in queue. If one has an arbitrarily large number of channels, many channels will be used during temporary surges in the arrivals but during lulls relatively few will be used. If one has only the minimum number of servers, a queue forms during the surges but the customers in queue are served during the lulls; the servers are kept busy all of the time serving either new arrivals or those in queue. The time average number of busy servers is the same in both cases, but the former has larger fluctuation.

Equation (1.19) also gives some interesting information for a single-channel server. Again one typically knows the arrival rate and average service time, so (1.19) determines the time average number of customers in service. For a single-channel server, however, the number in service at any time can be only 0 or 1. The time average number in service is, therefore, the same as the fraction of time the one server is busy. The right hand side of (1.19), $\lambda_{ab} \langle s_j \rangle$, is, in this case, called the 'traffic intensity' usually denoted by ρ:

$$0 \leq \rho = \lambda_{ab} \langle s_j \rangle \leq 1. \tag{1.20}$$

The quantity $1 - \rho$ is the fraction of time the server is idle. If a customer arrives at a random time uniformly distributed over the time interval (a, b), $1 - \rho$ can also be interpreted as the probability that the customer finds the server idle and can enter service with no delay.

1.6 Other graphical representations

In Section 1.2 we represented $A(t)$ and $D_q(t)$ as two curves on the same graph, i.e., as the locus of points in the (x, t) plane with coordinates $(A(t), t)$ and $(D_q(t), t)$ respectively, or equivalently $(x, A^{-1}(x))$ and $(x, D_q^{-1}(x))$. This is the most common way of representing these quantities graphically because it shows very conveniently most of the quantities one wishes to observe, particularly if the queue discipline is FIFO and $D_q^* = D_q$. The advantage of this type of graphical representation of the data t_j, t_{qj}, etc., over other possible schemes derives from the fact that one can easily visualize the geometrical addition or subtraction of line segments or areas. The graphs conveniently show geometrically the subtraction $A(t) - D_q(t)$ to give $Q(t)$ and, for FIFO queue discipline, the geometric subtraction of line segments in a different direction $D_q^{-1}(x) - A^{-1}(x)$ to give the wait $w(x)$. Furthermore, it conveniently shows addition of areas to give the total waiting time. If these are the features of primary interest,

it is difficult to imagine how one could find any better way to show all of these on the same graph.

If one wishes to compare the behavior of the curves $A^{(1)}(t), D_q^{(1)}(t)$ as observed on one day with a new pair of curves $A^{(2)}(t), D_q^{(2)}(t)$ observed on another day, the fact that one must draw two curves for each day and then compare the *pairs* of curves may be awkward. It is possible, however, to show the evolution of both $A(t)$ and $D_q(t)$ simultaneously by a single curve if one goes to a three-dimensional space. In an (x, y, t) space one can draw a curve $(A(t), D_q(t), t)$ or in a (t, t_q, x) space one can draw a curve $(A^{-1}(x), D_q^{-1}(x), x)$.

Each of these is a step function curve. The former moves a unit step in the x direction at each time t_j and in the y direction at each time t_{qj}, whereas the latter curve jumps from (t, t_q) coordinates (t_j, t_{qj}) to $(t_{j+1}, t_{q,j+1})$ when x passes j. The two curves $(A(t), t)$ and $(D_q(t), t)$ are the projections of $(A(t), D_q(t), t)$ onto the (x, t) and (y, t) planes, respectively. Correspondingly, the curves $(A^{-1}(x), x)$ and $(D_q^{-1}(x), x)$ are the projections of $(A^{-1}(x), D_q^{-1}(x), x)$ onto the (t, x) and (t_q, x) planes, respectively. We have, of course, previously drawn these projections on the same graph. We could have drawn them on separate graphs but would then lose the simple geometric interpretation of queue length and wait in queue which resulted from subtracting distances between the curves A and D_q.

There is still a third projection of each of these three-dimensional curves, the projections on the (x, y) or (t, t_q) planes. These are curves having a parametric representation $(A(t), D_q(t))$ and $(A^{-1}(x), D_q^{-1}(x))$, respectively, as shown in Fig. 1.6. Actually, the latter 'curve' is only a sequence of points since it moves only at integer x, but we can arbitrarily join these points by a piecewise linear curve. One may also label the time t or count x as a parameter along the curve, particularly since the only relevant parameter values are the discrete times t_j and t_{qj} or the integer values of x.

On the $(A(t), D_q(t))$ curve, a horizontal step of the curve has a length equal to the number of successive arrivals before a departure, and a vertical step has a length equal to the number of successive departures before the next arrival. If a t_j should be equal to some t_{qk}, we would have simultaneous horizontal and vertical steps. This would certainly occur at times when a customer arrives and finds the server empty, since then $t_j = t_{qj}$. It seems most natural to represent simultaneous arrivals and departures by a single line of slope 1.

That $A(t) \geq D_q(t)$ and $D_q^{-1}(x) \geq A^{-1}(x)$ was displayed in Fig. 1.3 by the D_q curve always being below or to the right of A. One needed to

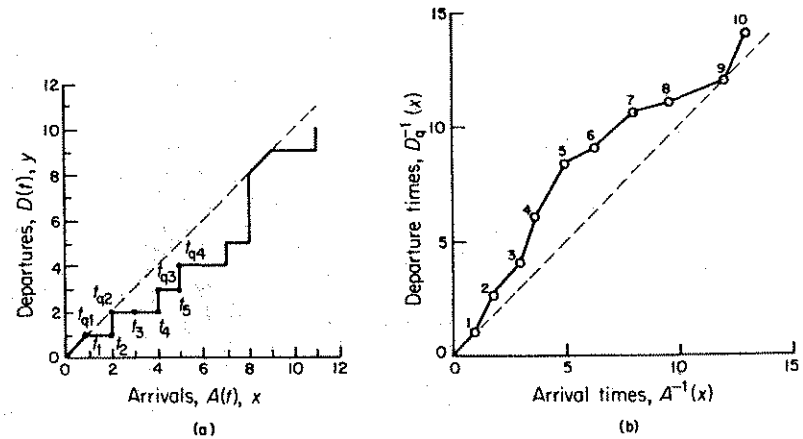

Figure 1.6 *Parametric representations of $D(t)$ versus $A(t)$ and $D_q^{-1}(x)$ versus $A^{-1}(x)$*

compare two curves to see this. In Fig. 1.6, however, this important property is shown by the one curve $(A(t), D_q(t))$ which must lie always on or below the line $x = y$, or by $(A^{-1}(x), D_q^{-1}(x))$ which must lie on or above the line $t = t_q$. One is still comparing two curves, but the straight line does not depend upon the evolution of the system. If one superimposes curves obtained on several days, the boundary line is the same for all days.

A graph such as Fig. 1.6 displays quite different aspects of the system behavior from Fig. 1.3. The former shows in a more convenient way comparative properties of the arrival and departure times (for FIFO), but it does not show conveniently any comparative properties of counts and times because one or the other is represented only as a parameter along the curve. Fig. 1.6(a) does not show waiting times conveniently because these involve the time parameter t; Fig. 1.6(b) does not show queue lengths conveniently because these involve the customer count parameter x. Neither graph shows the arrival rate, departure rate. $\langle w_j \rangle$, or $\langle Q(t) \rangle$ since these all involve both counts and time.

Fig. 1.6(a) does show the queue length identified with any point on the curve having integer coordinates; it is either the horizontal or vertical distance from the curve to the 45° line. Similarly, Fig. 1.6(b) shows the waiting time associated with any point (t_j, t_{qj}) as the horizontal or vertical distance to the 45° line.

If, in addition to the constraint that $Q(t) \geq 0$ and $w_j \geq 0$, one were to impose a restriction, that the queue could not exceed some number c (a storage capacity) or that the wait could not exceed some bound, these restrictions could also be represented conveniently in Fig. 1.6(a) or (b) respectively by drawing another 45° line; for example, $x - y = c$ in Fig. 1.6(a). The curve $(A(t), D_q(t))$ would then always lie between the lines $x - y = 0$ and $x - y = c$.

Problems

1.1 From a sequence of 150 random digits $X_i = 0, 1, \ldots, 9$, generate a sequence of numbers

$$Y_i = \begin{cases} 1 & \text{if } X_i = 0, 1, 2, \text{ or } 3 \\ 0 & \text{if } X_i = 4, 5, 6, 7, 8, \text{ or } 9 \end{cases} \quad i = 1, \ldots, 150.$$

From a different sequence of 150 random digits $X_i' = 0, 1, \ldots, 9$ generate a sequence

$$Y_i' = \begin{cases} 1 & \text{if } X_i' = 0, 1, 2, 3 \text{ or } 4 \\ 0 & \text{if } X_i' = 5, 6, 7, 8, \text{ or } 9 \end{cases} \quad i = 1, \ldots, 150.$$

The X_i, X_i' may be obtained from tables of random digits or one may use last digits of consecutive telephone numbers from a telephone book (excluding any numbers which may be listed twice, because a person may have both a business and personal listing, for example).

Consider a queueing system for which customers arrive and leave only at integer values of time $t = i$. One customer arrives at time i if $Y_i = 1$, none if $Y_i = 0$. If a customer is in service at time $i - 1$, it will leave the service at time i if $Y_i' = 1$, otherwise no customer leaves at time i. The server serves only one customer at a time.

Draw graphs of $A(t)$, $D_s(t)$, and $Q_s(t) = A(t) - D_s(t)$ on 10 squares to the inch graph paper with a scale of 20 time units to the inch and 10 customers to the inch. Also draw graphs of $(A(t), D_s(t))$ and $(A^{-1}(x), D_s^{-1}(x))$. Interchange the sequences Y_i and Y_i' and draw the corresponding graphs.

Note: This may be assigned as a class exercise. Each student should select a different set of random digits. One can then imagine that the different curves represent the results of some experiment which was repeated many times. The class should then compare the graphs obtained by different students.

1.2 If on a graph of $Q(t)$ versus t one sees that an arriving customer causes $Q(t)$ to increase from k to $k+1$, how would one identify when he left the queue if the queue discipline is last in, first out? How would one identify the same thing from graphs of $A(t)$ and $D_q(t)$?

1.3 (a) Let $0 = t_1 < t_2 < \ldots t_n$ be ordered arrival times and $0 < t_{q1} < t_{q2} < \ldots t_{qn} = t_n$ ordered departure times from a queue which vanishes at time 0 and t_n. If $t_{qj}^* = t_{qn_j}$ is the departure time of customer j, show that the sum of the squares of the delays

$$\sum_{j=1}^{n} (t_{qn_j} - t_j)^2$$

is least if $n_j = j$, i.e., for FIFO service. The t_{qj} are assumed to be independent of the order of service.

(b) As a generalization of (a), suppose the cost of dealy to the $j$the customer is a function $P(w_j)$ of the delay $w_j = t_{qn_j} - t_j$ with the function $P(x)$ the same for all customers. The total cost of delay to all customers is

$$\sum_{j=1}^{n} P(w_j).$$

If the marginal cost per unit delay $p(x)$,

$$p(x) = \mathrm{d}P(x)/\mathrm{d}x, \quad P(x) = \int_0^x p(x')\,\mathrm{d}x',$$

is a monotone increasing function of x, show that the total cost of delay is least for FIFO.

CHAPTER 2

Deterministic fluid approximation – single server

2.1 Introduction

To analyze the behavior of some existing service facility which serves a single category of customers, imagine that one were to record the arrival and departure times of all customers over some very long period of time, possibly several years. Such data do exist for some real systems. For example, any computer controlled traffic signal system is connected with many permanently installed vehicle detectors. At any particular traffic signal (server) there is likely to be one or more detectors upstream of the signal and also detectors downstream (perhaps near the next downstream signal). These detectors transmit an electrical pulse to the computer every time a vehicle passes. Most of these data are discarded after use, but they could be kept on a magnetic tape. Any airport also maintains a record of all aircraft movements including arrivals and departures from the runway and gate positions.

In the last chapter we discussed some of the microscopic properties of the $A(t)$, $D_q(t)$ curves; that they have integer steps at each arrival time. If one were to draw a graph of $A(t)$ and $D_q(t)$ for several years on a scale such that one could see each arrival, it might require a square mile of graph paper. If, however, one were to rescale the graph, one would see different types of time-dependent phenomena depending upon the scales of counts and time.

If you were counting cars on a highway, for example, which at various times might carry flows of the order of 1000 cars per hour, one could see the integer steps in $A(t)$ if the scales of time and count were comparable with, say, 10 seconds to the cm and 3 counts per cm. If, however, one were to choose a scale of about 1 min per cm and 20 counts per cm, the integer steps would be too fine to show very clearly. The $A(t)$ and $D_q(t)$ curves would likely appear as a nearly smooth

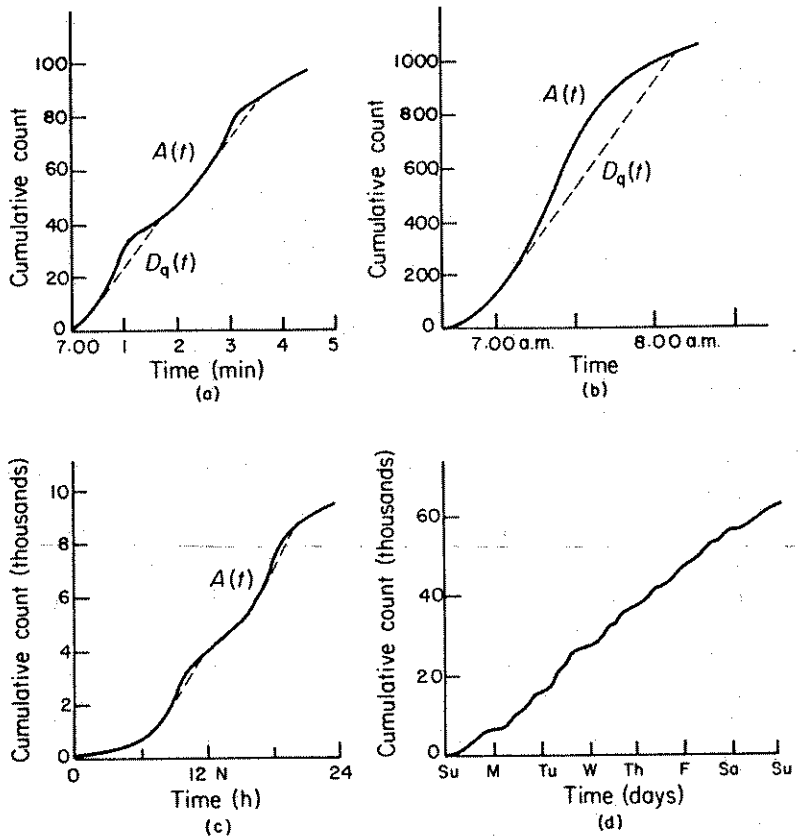

Figure 2.1 *Cumulative arrivals on various time scales*

curve, as illustrated in Fig. 2.1(a), having some wiggles of rather erratic form caused by 'random surges'.

Depending upon how steep the curves become as a result of the wiggles and how fast the server operates, the surges may or may not cause a queue. If the server has nearly equal service times for all cutomers, the $D_q(t)$ curve might look like the broken line curve of Fig. 2.1(a), which shows a queue whenever the surges generate a slope temporarily exceeding the service rate.

If we draw the curve on a still coarser time scale of say 20 minutes per cm and 200 counts per cm, the wiggles in the $A(t)$ might not show very clearly, but one would see the 'peak demands' or 'rush hours'. For example, between 7.00 a.m. and 8.00 a.m. of some day, the curve for

$A(t)$ might be of the type shown in Fig. 2.1(b). The $A(t)$ curve may be quite smooth, but not linear. If there is a portion of the curve where the arrival rate exceeds the service rate, the $D_q(t)$ could deviate appreciably from $A(t)$ showing queues possibly of the order of 100 cars.

On a scale of a few hours per cm and several thousand counts per cm, the difference between $A(t)$ and $D_q(t)$ would likely be too small to measure (the queues are not measured in thousands), but over a reasonable width of paper one sees the 24-hour flow pattern. It might show a morning and evening peak but a very low arrival rate between 2.00 a.m. and 5.00 a.m. with no queue (even on the scale of Fig. 2.1(a)) as in Fig. 2.1(c).

If the graph is continued for a week as in Fig. 2.1(d), we would probably see different patterns on different days, particularly Sunday, Monday, Friday, and Saturday. If it is drawn on a scale of a week per cm, we would no longer see the daily rush hours. We might see some variation in the daily counts within the week (particularly the weekend) but over many weeks there is likely also to be some seasonal variation. Finally, if one draws the graph on a scale of several months per cm, one would no longer see the daily variations. The seasonal variations would still be visible, but, in addition, one might see a gradual growth of traffic from one year to the next.

To analyze queueing delays, it is obviously not very helpful to draw $A(t)$ and $D_q(t)$ on such a scale that one cannot measure the difference between the curves. One would not ordinarily draw a graph on a scale such as Fig. 2.1(d) (or a coarser scale), but would observe that the queue (almost always) vanishes around 2.00 a.m. to 5.00 a.m. Instead of continuing the curve $A(t)$ for several days, one could reset the counters to zero at the same time each day and draw separate curves $A^{(j)}(t)$ and $D_q^{(j)}(t)$ for each jth day, i.e., a set of curves of the type shown in Fig. 2.1(c).

If we compare the curves on various days, we should, of course, find that some days have similar patterns (successive Fridays, for example). We would first try to classify the days in some systematic way so that all curves in the same class can be compared as being nearly 'equivalent'. No matter how one does this, however, there are likely to be some curves which are unlike any others. There are days when there were failures in the service, unusual patterns caused by a snow storm or whatever, which occur only once or twice in several years and possibly never in quite the same way again.

In the analysis of most practical problems, one would probably not

try to analyze the behavior of the system for all time periods. Presumably, there are certain things about the system behavior which are undesirable and which one wishes to correct through some modification in design, strategy of operation or whatever. Since it is expensive to collect and analyze data, one must first decide 'what was the problem?' or 'what are the most important of several possible problems?'

The type of techniques one will use to analyze the system depends upon whether one wishes to investigate the queueing due to the wiggles in $A(t)$ as illustrated in Fig. 2.1(a), the rush hours as in Fig. 2.1(b), or the unusual events. If it is the rush hour, one must further decide if the problem is primarily the weekday rush, the Monday–Friday rush or the weekend demands (as for recreational facilities).

To design a facility to accommodate the unusual events often means merely that one has some emergency procedures for diverting the customers (aircraft are sent to another airport in a snow storm or an announcement is made that some facility will be closed). The evaluation does not typically involve conventional queueing methods because the inconvenience may not be the usual delay associated with customers waiting to be served. The inconvenience may be difficult to quantify, yet many facilities (particularly public facilities) are designed in response to complaints about service during unusual situations. If the performance of a system is mainly a 'political' issue, there is no point in making an economic evaluation of delays during typical days. (People may want a rapid transit system no matter what it costs.)

We will be concerned here, by implication at least, primarily with patterns of system behavior that recur many times or are (partially) predictable. If it costs more to build a facility with a larger service rate, an efficiently designed system will always cause some delays, because there is no benefit associated with any excess capacity which is never used. The benefit associated with an increment of capacity that is used for only a short period of time or infrequently is also very small. Consequently, the capacity should always be somewhat less than the maximum demand during some time period. In principle, the proper choice of capacity should involve a compromise between the cost of a large facility and the inconvenience of delays for a smaller facility.

Even after one has classified the days into well defined categories, one will still find that the $A^{(j)}(t)$ curves within the same class are different in various ways. The arrival times of the customers are not likely to be exactly the same. Furthermore, although the curves on different days may have nearly the same form when drawn on a scale

such as Fig. 2.1(b) or (c), the wiggles on a scale corresponding to Fig. 2.1(a) do not occur at the same times or have the same shapes.

Most of the literature on queueing theory deals with the analysis of queueing on a scale corresponding to Fig. 2.1(a). It is generally assumed that if one takes the arithmetic average of the $A^{(j)}(t)$ over many days,

$$\langle A^{(j)}(t)\rangle = \frac{1}{n}\sum_{j=1}^{n} A^{(j)}(t), \tag{2.1}$$

that this averaging will smooth out most of the wiggles of the individual curves (for sufficiently large n) giving a curve for $\langle A(t)\rangle$ that is nearly linear over some appropriate time interval. Actually most of the theory deals with 'stationary processes' corresponding to some hypothetical process that runs from $t = -\infty$ to $t = +\infty$. It is usually further assumed that the arrival rate, i.e., the slope of the linear $\langle A^{(j)}(t)\rangle$, is less than the (average) service rate while the server is busy.

Whereas $Q^{(j)}(t) \geq 0$ on every day and, consequently, the average of $Q^{(j)}(t)$ over n days is positive, if one had a hypothetical process which had an arrival curve $\langle A^{(j)}(t)\rangle$ and a server which, while busy, serves at a constant average rate larger than the arrival rate, there would be no queue. Thus, the average $\langle Q^{(j)}(t)\rangle$ is not the same as the queue generated by the average arrival curve. The former is always larger than the latter because the queues generated by random surges are not compensated by a negative queue during the lulls. Despite the fact that the $Q^{(j)}(t)$ are not the same on different days, one will likely find that the total delay over some sufficiently long time (or the time average of $Q^{(j)}(t)$ or the average wait of many customers on the jth day) is nearly the same on all days.

In practical applications, however, many (perhaps most) queueing problems are of a type analogous to that shown in Fig. 2.1(b). If one compares these curves on different days, one still sees that the curves have wiggles in different places on different days but the amplitude of the wiggles is small compared with a typical queue length during the rush. Except for effects caused by wiggles near the beginning or end of the queueing period, the $\langle Q^{(j)}(t)\rangle$ is nearly the same as would be generated by a hypothetical arrival process $\langle A^{(j)}(t)\rangle$ served by an average server.

In analyzing any queueing problems in which queues form systematically at nearly the same time on all days (of the same category) it is convenient artificially to separate the queue length into two parts, a part which would be generated by a hypothetical arrival

process $\langle A^{(J)}(t)\rangle$ served by an average server, and the excess queue caused by the daily variations about these averages. The former part is called the 'deterministic queue'. If the latter can be described by some probability model, it will be called the 'stochastic queue'.

In the following, we will first consider a variety of problems which can be analyzed approximately from the evaluation of the deterministic queues. We will later describe some of the stochastic effects associated with some relatively simple types of systems.

In addition to the examples described in Fig. 2.1 primarily for a system having predictable rush hours but a steady server, deterministic approximations are also useful for describing any system for which a queue forms for predictable reasons. There are many systems in which the service is interrupted for specified times long enough for a sizeable queue to form. The service may be interrupted for one class of customers because the server is being used to serve another class of customers or perform some other function. At a highway traffic signal, for example, the signal turns red while a traffic intersection is used to serve another traffic stream. A bus will interrupt the loading of passengers who have waited for the bus because the bus must be used to transport the passengers somewhere. A server may be interrupted also for repairs.

The objective in modeling any such system is, of course, to relate the behavior of the system to various parameters associated with the arrivals or the server so that one can predict how some hypothetical system with different parameters will perform.

2.2 A rush hour

To analyze a rush hour of the type shown in Fig. 2.1(b) and to estimate what queues would exist for various service rates, one must first collect data to estimate or predict a possible arrival curve $A(t)$ or an average curve $\langle A^{(J)}(t)\rangle$.

In practical applications, as for example in counting cars on a highway where there is no automatic recording equipment, the data which are often recorded consist only of the counts of arrivals during consecutive time intervals of, say, 5, 10, or 15 minutes. These data are also commonly displayed as a histogram of counts as in Fig. 2.2 which one interprets as some step function approximation to a smooth function $\lambda(t)$. These observations might not be repeated on another day, it being assumed that the results would be reproducible within the typical range of statistical fluctuations, and that the hypothetical

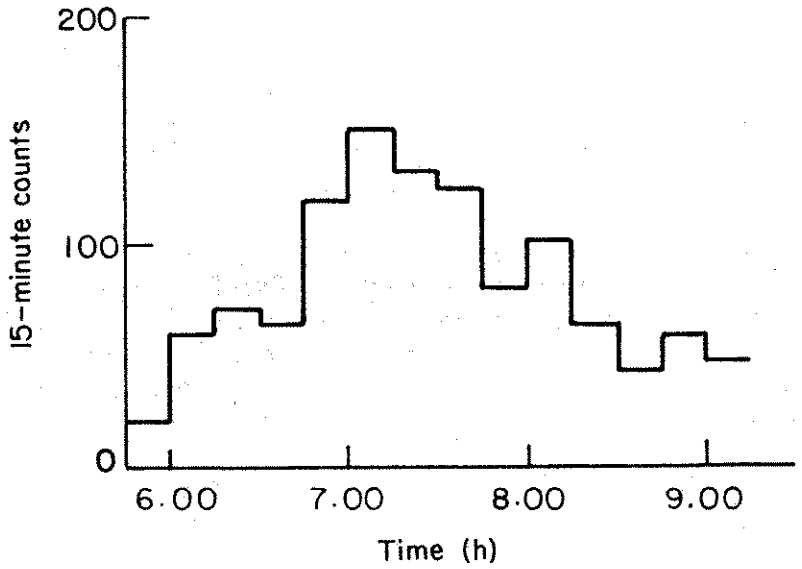

Figure 2.2 *Histogram of 15-minute counts*

$\lambda(t)$ which one would presumably obtain by averaging over many days is some suitable smoothing of the histogram.

From the histogram one can, of course, evaluate $A(t)$ at the discrete times corresponding to the ends of the counting interval. Lacking a statistical model, one would probably assume that the counts within the time intervals were nearly uniformly distributed over the interval, that the $A(t)$ therefore increases nearly linearly between the ends of the time intervals, that the $\langle A^{(J)}(t)\rangle$ would be a smoothing of $A(t)$ on some appropriate time scale, and that $\lambda(t)$ is some suitable smoothing of the histogram.

Since the questions we will be asking relate directly to the curve $A(t)$ and areas between it and some proposed $D_q(t)$, it is more important that one has an accurate estimate of $A(t)$ (for all t) than $\lambda(t)$. If one is smoothing 'by eye', it is generally better to smooth the $A(t)$ directly and evaluate $\lambda(t)$, if relevant, as the derivative of the smoothed $A(t)$ than to estimate $\lambda(t)$ from a smoothing of the histogram and evaluate $A(t)$ by integrating the smoothed $\lambda(t)$. In doing the latter, one might, in an attempt to follow the histogram 'locally', make systematic errors which would accumulate in the integration to obtain $A(t)$, and therefore cause inaccurate estimates of the $Q(t)$.

Having obtained an approximate deterministic $A(t)$, we might now imagine that we have a server which operates at some constant average rate μ when busy. The construction of the curve $D_q(t)$ is illustrated in Fig. 2.3(a) for a typical rush hour specified by a given arrival curve $A(t)$ and service rate μ. The curve $D_q(t)$ follows $A(t)$ very closely (essentially zero queue) until a time to when the arrival rate $\lambda(t)$ is equal to μ. If for some range of t with $t > t_0$, $\lambda(t) > \mu$, a queue starts to form at time t_0 and the service rate remains constant at the value μ. Thus the departure curve becomes a straight line of slope μ tangent to $A(t)$ at t_0 and extending from t_0 until some time t_3 when the arrival rate has been below μ long enough for the service to have caught up with the arrivals, i.e., $D_q(t_3) = A(t_3)$. After time t_3, the queue stays zero, i.e., $D_q(t) = A(t)$ for $t > t_3$ until such time as the $\lambda(t)$ again exceeds μ.

Note that the graphical construction of $D_q(t)$ is very easy for any given μ. One merely pushes a straight edge at slope μ against the curve $A(t)$ to form the tangent at t_0 and draws the line from t_0 until it meets $A(t)$ again.

One can see immediately from Fig. 2.3(a) that, for any smooth $A(t)$, $Q(t)$ grows quadratically in time from time t_0 but vanishes linearly in t near t_3.

It is often convenient also to draw graphs of $\lambda(t)$ and the actual departure rate $\mu(t)$ as in Fig. 2.3(b). Until time t_0, $\mu(t) = \lambda(t) \leq \mu$. For some time after time t_0, however, $\mu(t) = \mu \leq \lambda(t)$ as shown by the broken line. If $\lambda(t)$ reaches a maximum at time t_1 and decreases, it will equal μ again at some time t_2.

The length of the queue at time t

$$Q(t) = A(t) - D_q(t) = \int_{t_0}^{t} [\lambda(\tau) - \mu(\tau)]\, d\tau \tag{2.2}$$

is represented in Fig. 2.3(b) by the shaded area between $\mu(\tau)$ and $\lambda(\tau)$. It reaches a maximum at time t_2 when $\lambda(\tau) = \mu$. After time t_2, the queue decreases until time t_3 when the area between $\lambda(t)$ and μ of Fig. 2.3(b) from time t_2 to t_3 is equal to the corresponding area between times t_1 and t_2. Note that $\mu(t)$ will, generally, have a discontinuity at time t_3 when, as the queue vanishes, $\mu(t)$ drops from μ to $\lambda(t_3)$.

Although Fig. 2.3(b) shows clearly the evolution of the $\lambda(t)$ and $\mu(t)$, it does not show conveniently such things as waits, queue lengths, etc., as does Fig. 2.3(a).

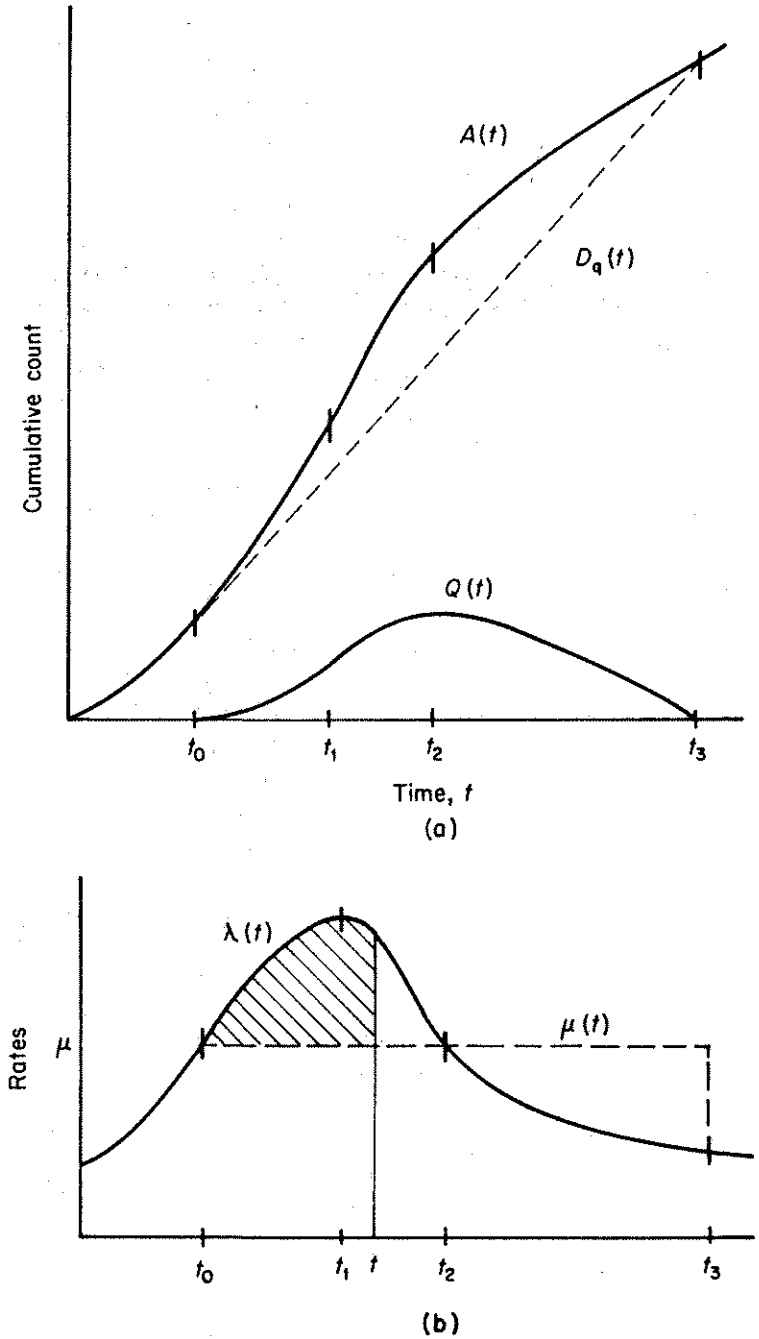

Figure 2.3 *Graphical construction of queue evolution*

2.3 A slight overload

In most engineering applications, the evaluation of delays is only the first step in an analysis, the final result of which is a decision as to what to build. If costs of delays are large (in some sense) relative to cost of construction, one should build a facility with a μ very close to the peak arrival rate, thereby keeping the delays low. If, however, construction costs are high, one builds a facility only large enough to serve all customers eventually, but certainly not to serve them with no delay. The second step in such an analysis is to see how the total delay over the rush hour depends upon the service rate μ.

For any given $A(t)$, the queue lengths, delays, etc., are quite sensitive to the service rate. One can see this immediately by observing how $D_q(t)$ would change if its slope were changed.

If the period of time over which a queue exists (t_0 to t_3 of Fig. 2.3(a)) is so large that $A(t)$ cannot be approximated by any simple formula, the easiest way to evaluate the total delay over the rush hour as a function of μ is simply to draw several curves of $D_q(t)$ for a reasonable selection of μ values. Then evaluate the area between $A(t)$ and each $D_q(t)$ using a planimeter or by counting squares on the graph paper.

If, however, the costs of delays are sufficiently large, one will eventually build a facility such that $D(t)$ is rather close to $A(t)$ with some small value of $t_3 - t_0$. Graphical methods are not very convenient in this case unless one can somehow make a preliminary guess of the likely range for the final choice of μ and therefore t_0 and t_3. To obtain any accuracy from graphical methods, one must draw the graph so as to magnify the range of t from t_0 to t_3. Instead of doing this, however, one might consider an analytic approach.

Suppose that $\lambda(t)$ rises to a maximum at a time t_1 as shown in Fig. 2.3(b). In most practical situations, it would be reasonable to assume that $\lambda(t)$ has a Taylor series expansion in $t - t_1$, or is at least twice differentiable near $t = t_1$, so that $\lambda(t)$ can be approximated by a quadratic function.

$$\lambda(t) = \lambda(t_1) - \beta(t - t_1)^2 \tag{2.3}$$

for some constant β;

$$\beta = -\frac{1}{2}\frac{d^2\lambda(t)}{dt^2}\bigg|_{t=t_1},$$

at least over some sufficiently small range of $t - t_1$ (presumably for $t_0 < t < t_3$).

If μ is sufficiently close to $\lambda(t_1)$, the time t_0 of Fig. 2.3(b), where $\mu = \lambda(t_0)$, can be estimated from (2.3); also the time t_2 where μ is again equal to $\lambda(t)$.

$$\begin{aligned} \mu &= \lambda(t_1) - \beta(t_0 - t_1)^2, \\ t_0 &= t_1 - \left[\frac{\lambda(t_1) - \mu}{\beta}\right]^{1/2}, \\ t_2 &= t_1 + \left[\frac{\lambda(t_1) - \mu}{\beta}\right]^{1/2}. \end{aligned} \tag{2.4}$$

It is convenient now to write $\lambda(t) - \mu$ in the factored form

$$\lambda(t) - \mu = \beta(t - t_0)(t_2 - t). \tag{2.4a}$$

This representation of $\lambda(t) - \mu$ follows directly from the postulates that it is quadratic in t, it has zeros at $t = t_0$ and $t = t_2$, and the second derivative with respect to t is -2β.

The queue at any time $t_0 < t < t_3$ is obtained by substitution of (2.4a) into (2.2),

$$Q(t) = \beta(t - t_0)^2\left[\frac{(t_2 - t_0)}{2} - \frac{(t - t_0)}{3}\right]. \tag{2.5}$$

From this we see that $Q(t)$ grows quadratically in $t - t_0$ for t near t_0 and has a maximum at t_2,

$$Q(t_2) = \frac{\beta}{6}(t_2 - t_0)^3 = \frac{4[\lambda(t_1) - \mu]^{3/2}}{3\beta^{1/2}} \tag{2.6}$$

proportional to the 3/2 power of the oversaturation $\lambda(t_1) - \mu$. The queue vanishes again at time t_3 given by

$$t_3 = t_0 + (3/2)(t_2 - t_0) = t_0 + 3(t_1 - t_0). \tag{2.7}$$

Thus (2.5) can also be written as

$$Q(t) = \frac{\beta}{3}(t - t_0)^2(t_3 - t). \tag{2.5a}$$

Despite all the parameters, this function has a universal shape. If we were to translate the time and rescale the graph so that time was measured from t_0 in units of $t_3 - t_0$, and $Q(t)$ were measured in units of $Q(t_2)$, we could define

$$t' = (t - t_0)/(t_3 - t_0), \quad Q'(t) = Q(t)/Q(t_2),$$

and obtain

$$Q'(t) = (27/4)t'^2(1-t') \quad \text{for} \quad 0 \le t' \le 1.$$

This contains no parameters; it is a function only of t' which vanishes quadratically at $t' = 0$, linearly at $t' = 1$ and has a maximum of $Q'(t) = 1$ at $t' = 2/3$.

Finally, the total delay W over the rush hour is obtained from the area between $A(t)$ and $D_q(t)$, i.e., by integration of (2.5a).

$$\begin{aligned} W &= \int_{t_0}^{t_3} d\tau Q(\tau) = \frac{\beta}{3}\int_{t_0}^{t_3} d\tau(\tau - t_0)^2(t_3 - \tau) \\ &= \frac{\beta}{3}(t_3 - t_0)^4 \int_0^1 du\, u^2(1-u) = \frac{\beta}{36}(t_3 - t_0)^4 \\ &= \frac{9[\lambda(t_1) - \mu]^2}{4\beta}, \end{aligned} \tag{2.8}$$

which is proportional to the square of the amount of oversaturation, $\lambda(t_1) - \mu$, or the fourth power of the duration of the queue $t_3 - t_0$. The β represents the curvature of $\lambda(t)$; a large β means a sharp peak for $\lambda(t)$, a small β a flat peak. A sharp peak, for fixed $\lambda(t_1)$ and μ, implies from (2.4) a short duration of the queue, a small maximum queue (2.6), and a small total delay (2.8).

The estimation of total delay is one of the most common problems to arise in practical applications. Some further refinements of it will be discussed again in Chapter 8. Any conclusions obtained here are tentative and subject to unknown errors arising from the use of deterministic approximations. One must be particularly cautious of the possibility that the queue lengths calculated here may be overshadowed by queues generated by stochastic effects.

2.4 Delays over many years

For many types of service facilities, the arrival rate of customers shows rush hours each day, variations in 24-hour arrival patterns throughout the week, seasonal variations, and a general growth in demand from year to year as discussed in Section 2.1. A question that often arises is the following: one has an estimate of the annual growth of demand over the next several years and one knows the costs of construction of facilities of various service rates μ; when should one expand the service?

We are concerned here mainly with the method of evaluating the

delays associated with various strategies rather than with the final problem of selecting an optimal strategy. In principle, one can draw a graph of $A(t)$ for a time range of 5 years, but if one draws it on a 5-year time scale, one does not see the individual daily rush hours which are the source of the delays; a 5-year plot will show the annual growth, possibly the seasonal oscillation, but little else.

Since the queue is likely to vanish at the same time each day, one could draw separate graphs of the $A^{(j)}(t)$ for the jth day, and represent the total delay as the sum over j of the areas between the $A^{(j)}(t)$ and $D_q^{(j)}(t)$. Although one could, in principle, draw some 10^3 such graphs, and evaluate them separately, it should be possible to classify the days into weekends, weekdays, etc. Days of the same classification are still likely to show trends or growth, but the shapes of the $A^{(j)}(t)$ are likely to be nearly the same, differing only in the total count (scale). Formally, for days of the same class, we might assume that

$$A^{(j)}(t) = A^{(j)}F(t) \tag{2.9}$$

in which $A^{(j)}$ is independent of t and $F(t)$ is independent of j. One could, for example, let $A^{(j)}$ be the 24-hour count so that $F(24) = 1$.

If we draw a graph of the function $F(t)$ as in Fig. 2.4(a), we can

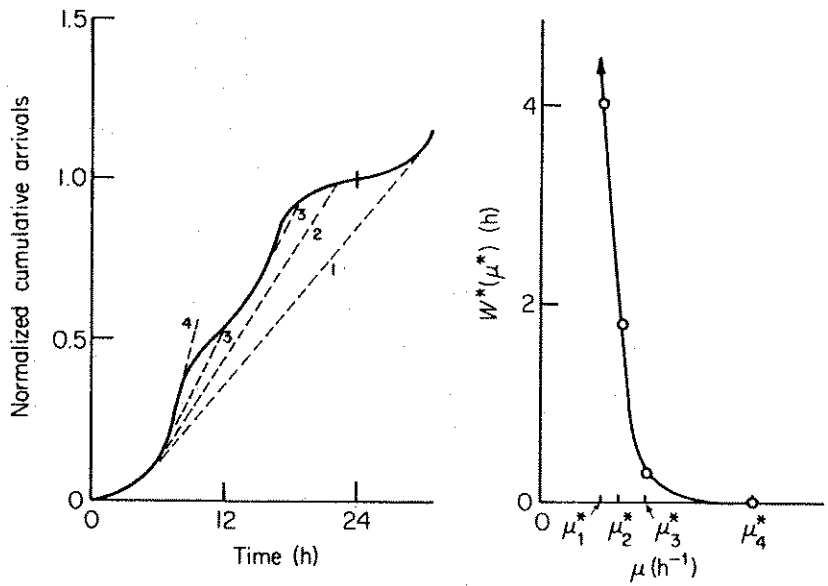

Figure 2.4 *Evaluation of a rescaled wait for various service rates*

evaluate the total wait (area) $W^*(\mu^*)$ that would exist for a hypothetical cumulative arrival curve $F(t)$ and service rate μ^*. If we do this for many value of μ^*, we can construct a graph of $W^*(\mu^*)$ versus μ^* as in Fig. 2.4(b).

The delay that would occur for an arrival curve $A^{(j)}F(t)$ and service rate μ can now be obtained simply by rescaling coordinates in Fig. 2.4. The delay $W^{(j)}$ is just $A^{(j)}$ times the W^* evaluated at a service rate $\mu^* = \mu/A^{(j)}$, i.e.,

$$W^{(j)} = A^{(j)} W^*(\mu/A^{(j)}). \tag{2.10}$$

Actually, it may be more advantageous to draw a graph of the function

$$H(s^*) = s^* W^*(1/s^*) \tag{2.11}$$

rather than $W^*(\mu^*)$, because (2.10) can then be written as

$$W^{(j)} = \mu H(A^{(j)}/\mu). \tag{2.12}$$

The total delay over many days (years) having the same $F(t)$ is obtained by adding the $W^{(j)}$ for each day,

$$\text{total delay} = \mu \sum_j H(A^{(j)}/\mu). \tag{2.13}$$

If $A^{(j)}$ is slowly varying with j, one would not evaluate the H for every day but would group together the days with nearly the same $A^{(j)}$. Once the curve $H(s^*)$ has been drawn, it is a simple exercise to observe the H evaluated at $A^{(j)}/\mu$, i.e., the $A^{(j)}$ measured in units of μ, and evaluate (2.13). A change in μ involves only a repetition of the calculation (2.13) with new units.

2.5 Queueing to meet a schedule

In most models of queueing it is customary to imagine that the arrival pattern is given or observed and that it will not change if the service rate changes. The objective is to serve these arrivals as early as possible, i.e., to maximize the number of service completions by time t subject to the constraint that the service rate shall not exceed some rate μ and the number of service completions shall not exceed the number of arrivals (an upper bound on $D(t)$). There are other situations, however, in which one might specify a lower bound on the number of service completions by time t and one wishes to minimize the number of service completions by time t subject to the same

constraint that the service rate not exceed μ and that the number of service completions stays above the lower bound.

Suppose, for example, that a factory can manufacture goods at some maximum rate μ. The future demand for these goods is predictable; $D_d(t)$ goods should be produced by time t (the cumulative demand). The demand rate $dD_d(t)/dt$ is time dependent (rush hours, seasonal demand, etc.) and may at time exceed μ, although the long time average demand rate is less than μ (i.e., for sufficiently large t, $D_d(t) < \mu t$).

To meet this demand it is necessary to stockpile goods ahead of the demand surges, but one does not wish to store any more than necessary. The number of goods produced by time t, $D(t)$, should therefore be the minimum number such that $D(\tau) \geq D_d(\tau)$ and $dD(\tau)/d\tau \leq \mu$ for all values of τ including $\tau > t$.

The method of constructing $D(t)$ is similar to the construction of $A(t)$ in the conventional queueing problem except that one starts at some time in the distant future and draws a line of slope μ backwards in time whenever $dD_d(t)/dt$ exceeds μ as illustrated in Fig. 2.5. It is in fact the same type of construction as for $A(t)$ if, for some future time t^* when the demand will certainly be satisfied, one draws a graph of $D_d(t^*) - D(t^* - \tau)$ versus $t^* - \tau$, the future work to be done (by time t^*) versus the time remaining to do it, $t^* - \tau$.

Of course, vertical and horizontal distances between $D(t)$ and $D_d(t)$ have the obvious interpretations as the stockpile of goods and the time any object remains in inventory (if goods are used in the order they are produced).

Another example of the same type of theory relates to the morning commuter rush hour. Suppose that a transportation system has a bottleneck which can accommodate a maximum flow of μ. Let $D_d(t)$ represent the number of commuters (customers) who must be at work by time t, or actually the number that must have passed the bottleneck by time t in order to be at work on time, and suppose that $dD_d(t)/dt > \mu$ during some time interval. In order for everyone to be at work on time it is necessary that some people arrive at work ahead of schedule.

If some transportation manager could assign arrival time reservations, he would presumably tell the persons who should be at work at time t to arrive at time τ such that $D(\tau) = D_d(t)$ as in Fig. 2.5 much as a factory manager would produce $D(\tau)$ goods by time τ in order to satisfy the demand $D_d(t)$ at time t.

Unfortunately, most transportation facilities do not have a reservation system and a traveler whose aim is to maximize his own benefit

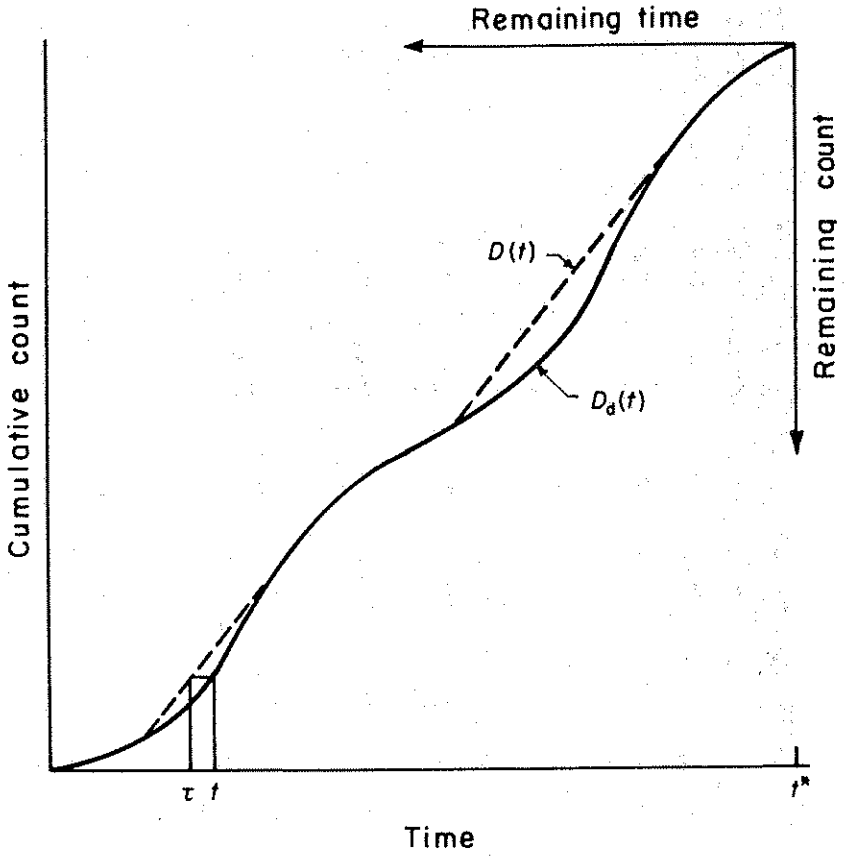

Figure 2.5 *Cumulative counts to meet a scheduled demand*

is not likely to cooperate. If there were no queue behind the bottleneck, a person who must be at work by time t would see that he could be at work on time even if he arrived at late as time t, but someone else would be late for work. This is one of many examples in transportation and elsewhere in which minimizing some global objective is not achieved by each person doing what is optimal for himself.

One might now ask if there is some arrival curve $A(t)$ which results if every persons tries to do what is optimal for himself. An arrival pattern which identifies an arrival time with each individual would be described as *stable* if, for that pattern, no individual can find a better arrival time than the one he has. An interesting feature of this

situation is that there is *no* arrival pattern with finite arrival rate and FIFO queue discipline that is stable with respect to the objective that each individual arrives as late as possible so as still to be at work on time.

To prove this, one need only observe that any assignment giving an arrival curve A with finite slope as in Fig. 2.6 and guaranteeing that each person is at work on time will necessarily cause some people to be at work early. However, any such person who must be at work by time t can determine from the curve A an arrival time t' as shown in Fig. 2.6 such that he will be at work exactly at time t, an arrival time which, for some people at least, must be later than the one assigned (thus the assignment is not stable).

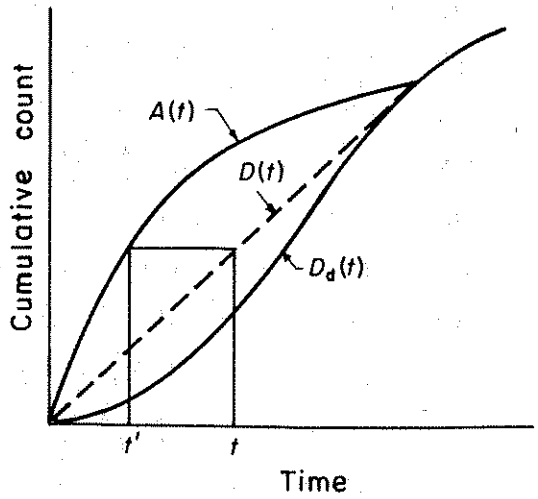

Figure 2.6 *Arrivals who must be at work on time*

To allow simultaneous arrivals (infinite slope for A) does not resolve the difficulty unless there is some mechanism by which people who arrive simultaneously can be served in the order of their work schedules. Neither is it helpful to recognize that there may be some restriction upstream of the bottleneck which limits the rate of arrivals, because the curve A actually refers to the expected times of arrivals, the times people would arrive in the absence of a queue. It makes no difference where they wait.

To obtain a stable assignment, one must postulate a different type of individual objective than arriving as late as possible. Stable

assignments do exist under the hypothesis that each individual assigns a price p per unit of delay in queue and a price $p' < p$ per unit of time by which he arrives to work early. He then selects an arrival time so as to minimize his cost. An assignment is stable if no individual can find a less costly arrival time than the one he has.

2.6 Pulsed service

For many types of service facilities, service occurs in pulses. A traffic signal at a highway intersection passes cars at a fairly constant rate for a certain time while the signal is green, but then provides no 'service' while the signal is red. Any form of public transportation providing service at a single terminal (an airplane, train, bus, elevator, etc.) will load passengers during a certain time interval after a vehicle has arrived at the terminal; but after the vehicle leaves, there is no service until the next vehicle arrives. Pedestrians wishing to cross a highway will queue until a sufficiently long gap appears in the traffic stream of cars. Mail in a post office is stored in sorting bins until, at certain discrete times, the accumulated mail is passed to the next sorter or put on a truck, train, or whatever. Service at some facilities may also be interrupted for repairs or maintenance.

In many of these situations, the arrival curve $A(t)$ is smooth in the sense that $\lambda(t)$ is nearly constant over time intervals of duration comparable with the time between service pulses and each service pulse is sufficient to exhaust the queue of waiting customers. For the traffic signal this means that the signal is 'isolated', i.e., the arrivals are not themselves pulsed by an upstream signal, and the traffic is light enough that the queue clears during the green time. For the public transportation example, the vehicles have sufficient capacity to serve all waiting customers. The pedestrians cross a street in a pack whenever there is a gap, and mail pick-ups take all the mail which has accumulated.

Deterministic fluid approximations are particularly useful for a crude analysis of these types of situations. The queues may well become sufficiently large between service pulses to justify use of such approximations even though, over a long time period, the system is 'undersaturated' in the sense that the queue does not continue to grow from one cycle to the next.

A typical graph of arrivals and departures for the above type of system is shown in Fig. 2.7. Suppose, as suggested by the traffic signal

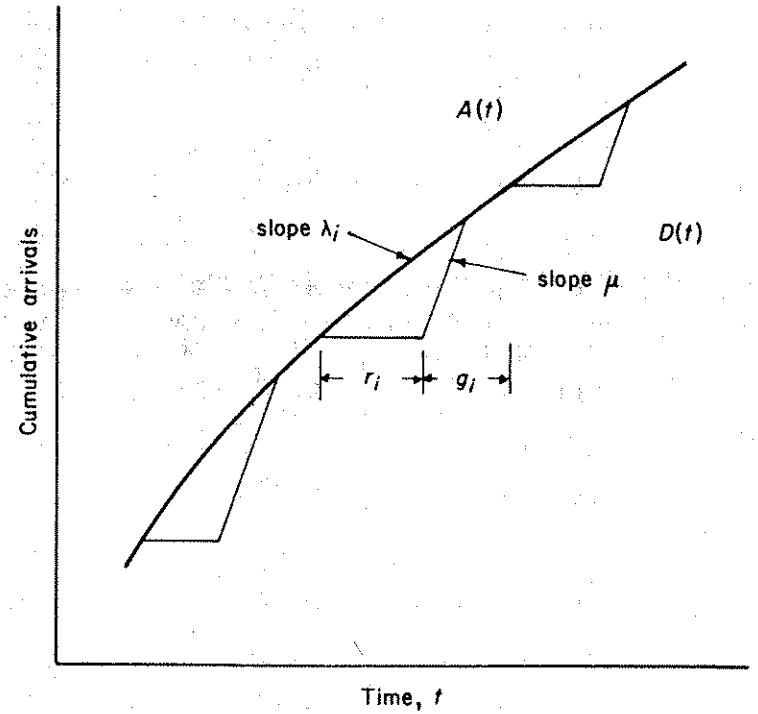

Figure 2.7 *Arrivals and departures for pulsed service*

application, we let

r_i = red time of the ith cycle
g_i = green time of the ith cycle
λ_i = arrival rate during the ith cycle
μ = service rate during green.

Since $\lambda_i < \mu$, the maximum wait in the ith cycle is r_i. If N_i is the number of customers delayed in the ith cycle, then the queue clears at time

$$N_i/\lambda_i = N_i/\mu + r_i,$$

so

$$N_i = r_i\lambda_i/(1-\lambda_i/\mu).$$

The total wait in the ith cycle, the area of the ith triangle in Fig. 2.7, is therefore $r_iN_i/2$, i.e.,

$$\text{wait in } i\text{th cycle} = \frac{1}{2}\frac{r_i^2\lambda_i}{(1-\lambda_i/\mu)}. \tag{2.14}$$

We have assumed here that $\lambda(t)$ is nearly constant over the ith cycle with a value λ_i. It should also be nearly the same for adjacent cycles, but we do not rule out the possibility that λ_i may change appreciably over a sufficiently long time. We are making no restrictions, however, on how the g_i and r_i vary from cycle to cycle, as long as g_i is large enough for the queue to clear in each cycle, i.e.,

$$\lambda_i(r_i+g_i) \leq \mu g_i. \tag{2.15}$$

According to the definitions of Section 1.4, the average wait per customer during the ith cycle is the total wait during the ith cycle divided by the number of customers in the ith cycle:

$$\text{average wait per customer in the } i\text{th cycle} = \left[\frac{r_i^2\lambda_i}{2(1-\lambda_i/\mu)}\right]\left[\frac{1}{\lambda_i(r_i+g_i)}\right]$$

$$= \frac{r_i^2}{2(r_i+g_i)(1-\lambda_i/\mu)}. \tag{2.16}$$

Correspondingly, the average wait per unit time, i.e., the average queue length is

$$\text{average queue during the } i\text{th cycle} = \frac{r_i^2\lambda_i}{2(r_i+g_i)(1-\lambda_i/\mu)}. \tag{2.17}$$

Over n cycles, the average wait is again the total wait during n cycles divided by the number of arrivals over n cycles.

$$\text{average wait per customer over cycles } j+1 \text{ to } j+n$$

$$= \langle w \rangle = \frac{\frac{1}{2}\sum_{i=j+1}^{j+n} r_i^2\lambda_i/(1-\lambda_i/\mu)}{\sum_{i=j+1}^{j+n}(r_i+g_i)\lambda_i}$$

$$= \frac{\frac{1}{n}\sum_{i=j+1}^{j+n}\frac{1}{2}r_i^2\lambda_i/(1-\lambda_i/\mu)}{\frac{1}{n}\sum_{i=j+1}^{j=n}(r_i+g_i)\lambda_i}. \tag{2.18}$$

We have divided numerator and denominator by n, because the numerator has the interpretation as the arithmetic average wait per cycle and the denominator the arithmetic average number of arrivals per cycle time.

The important thing to observe here is that (2.18) is not generally the same as the arithmetic average of (2.16).

If λ_i varies sufficiently slowly with i that it is nearly constant over n cycles, (2.18) simplifies to

$$\langle w \rangle \simeq \frac{(1/n)\sum_{i=j+1}^{j+n} r_i^2}{2(1-\lambda_j/\mu)(1/n)\sum_{i=j+1}^{j+n}(r_i+g_i)}. \tag{2.19}$$

In many of the above applications, the service in each cycle ceases as soon as the queue vanishes. This is approximately true for a vehicle-actuated signal and for passengers boarding a public transportation vehicle. In this case, the equality holds in (2.15) for all i. One can exploit this to eliminate one of the variables r_i or g_i or to express both in terms of the total cycle time (r_i+g_i). If we do the latter (2.14) becomes

$$\text{wait in } i\text{th cycle} = \frac{1}{2}(1-\lambda_i/\mu)(r_i+g_i)^2\lambda_i$$

and (2.19) becomes

$$\langle w \rangle \simeq \frac{1}{2}(1-\lambda_i/\mu)\frac{(1/n)\sum_{i=j+1}^{j+n}(r_i+g_i)^2}{(1/n)\sum_{i=j+1}^{j+n}(r_i+g_i)}. \tag{2.20}$$

In some applications, the cycle time varies appreciably from cycle to cycle (even though λ_i does not). A vehicle-actuated signal, for example, may have a red time determined by fluctuating traffic from a side street, so that (r_i+g_i) is interpreted as a random variable with some specified probability distribution. For a public transportation system, the time (r_i+g_i) is interpreted as the 'headway' between dispatches which may also have random fluctuations. For pedestrians crossing a street, (r_i+g_i) is the time interval between acceptable gaps in the traffic stream.

If we interpret (r_i+g_i) as an ith observation of a random variable $(R+G)$, then the arithmetic mean of many observations is (by definition) interpreted as the expectation. Thus

$$(1/n)\sum_i (r_i+g_i)^2 \simeq E\{(R+G)^2\} = E^2\{R+G\} + \text{Var}\{R+G\},$$

$$1/n\sum_i (r_i+g_i) \simeq E\{R+G\}$$

and

$$\frac{(1/n)\sum_i (r_i+g_i)^2}{(1/n)\sum_i (r_i+g_i)} \simeq \frac{E^2\{R+G\}+\mathrm{Var}\{R+G\}}{E\{R+G\}}$$

$$= E\{R+G\}\,[1+C^2(R+G)]$$

in which C is the coefficient of variation

$$C^2(R+G) \equiv \mathrm{Var}\{R+G\}/E^2\{R+G\}. \tag{2.21}$$

In many cases (particularly for loading bus passengers), the service rate is large compared with λ_i, i.e. $\lambda_i/\mu \ll 1$ and $g_i/r_i \ll 1$. Thus (2.21) simplifies still further to

$$\langle w \rangle \simeq \frac{1}{2} E\{R\}\,[1+C^2(R)]. \tag{2.22}$$

If headways between buses, for example, are all equal (so that $C^2(R) = 0$), it is obvious that a person who arrives at a random time unrelated to any possible bus schedule will wait, on the average, a half a headway, i. e., $\frac{1}{2}E\{R\}$. If, however, the headways are irregular, it is more likely that a person will arrive during a long headway than during a short headway; therefore, the average wait will be larger than that associated with the average headway. In the absence of any control to keep buses on schedule, there is, typically, a tendency for the headway distribution to become exponential. For an exponential distribution $C^2 = 1$, i.e., the average wait for buses with an exponential headway distribution is twice that of a regular schedule with the same average headway.

2.7 Applications

In most applications of the above formulas, the choice of the time between pulses is subject to certain constraints. Clearly, if one could freely choose any cycle time or headway, one would choose it to be arbitrarily small so as to minimize delay.

(a) *A fixed-cycle traffic signal*

An isolated traffic signal actually serves two (or more) traffic streams, each of which receives pulsed service. During part of the red time for one stream, the signal is green for the other traffic stream. There is, however, an effective lost time when neither traffic stream flows.

If r_i' and g_i' represent the red and green times for the second traffic stream, we can write

$$r_i = g_i' + L \quad \text{and} \quad r_i' = g_i + L,$$

in which L is the effective lost time per cycle. The cycle time is given by

$$r_i + g_i = r_i' + g_i' = g_i + g_i' + L.$$

For just two traffic streams with arrival rates λ_i and λ_i' and service rates μ and μ', the wait per cycle is, according to (2.14)

$$\text{wait in } i\text{th cycle} = \frac{(g_i'+L)^2\lambda_i}{2(1-\lambda_i/\mu_i)} + \frac{(g_i+L)^2\lambda_i'}{2(1-\lambda_i'/\mu')}$$

provided

$$\lambda_i(g_i+g_i'+L) < \mu g_i \quad \text{and} \quad \lambda_i'(g+g_i'+L) < \mu' g_i'. \tag{2.23}$$

If the arrival rates are constant, $\lambda_i = \lambda$, $\lambda_i' = \lambda'$, and the signal is periodic, $g_i = g$, $r_i = r$, etc., then the wait per unit time is

$$\frac{(g+L)^2\lambda}{2(g+g'+L)(1-\lambda/\mu)} + \frac{(g'+L)^2\lambda'}{2(g+g'+L)(1-\lambda'/\mu')} \tag{2.24}$$

provided (2.23) is true; otherwise the queues grow from cycle to cycle. In (2.23). λ/μ can be interpreted as the fraction of the cycle time necessary to serve the flow λ. A necessary condition for (2.23) to hold is that

$$\frac{\lambda}{\mu} + \frac{\lambda'}{\mu'} < 1 - \frac{L}{(g+g'+L)}. \tag{2.25}$$

Typically a traffic engineer has the option of choosing the g and g' within some practical range of values satisfying (2.23). He might choose them so as to minimize (2.24). For most reasonable values of the parameters, the minimum of (2.24) occurs at the boundary (2.23), i.e., for the shortest possible g and g' which can accommodate the flows. The issues of traffic signal setting are more complex than this, however, because stochastic effects become very important as the g and g' approach their minimum values. If one includes the delays due to stochastic effects, the minimum delay per unit time actually occurs for a cycle time approximately twice that predicted above. The deterministic theory, however, describes at least a first (but very crude) approximation to the delays. Some of the stochastic effects will be discussed in Chapter 9.

In practice, the values of g and g' are usually constrained also by the time it takes a pedestrian to cross the road.

(b) *Bus dispatching*

In selecting the headways between buses, it is clearly advantageous to make $E\{R\}$ and $C^2(R)$ as small as possible. Since buses are subject to various random disturbances generated by traffic congestion, signals, loading times, etc., it is usually necessary to introduce some types of control in order to keep $C^2(R)$ small. The most common scheme of regulation is to impose a schedule.

It is easy to prevent buses from running ahead of schedule but more difficult to prevent them from running late. To obtain a two-sided control, one must provide some slack time in the schedule so that buses can gain on the schedule once they have fallen behind. The more slack one introduces, however, the slower the speed and, for a fixed number of buses, the longer is the average headway. The minimum average wait per passenger involves a compromise between a small $E\{R\}$ (loose control) and a small $C^2(R)$ (tight control). The best strategy will not generally involve exactly equal headways, $C^2(R) = 0$.

If one controls headways so that $C^2(R) \ll 1$, the issue of selecting an optimal $E\{R\}$ usually centers around the fact that there is a cost (per unit time or per trip) associated with each bus dispatched. In principle, this cost should be balanced against the cost or inconvenience of delay. If it costs γ to despatch a vehicle and it costs p per unit of wait, the cost per unit time of bus operation is $\gamma/E\{R\}$ and the total cost per unit time (for $C^2(R) \ll 1$) is approximately

$$\frac{\gamma}{E\{R\}} + p\frac{E\{R\}}{2}\lambda(t). \tag{2.26}$$

If $\lambda(t)$ varies only slightly during a headway, the minimum total cost over a long period of time can be achieved by choosing $E\{R\}$ so as to minimize the cost rate (2.26) at every t. Thus the optimal $E\{R\}$ at time t is

$$E\{R\} = \left[\frac{2\gamma}{p\lambda(t)}\right]^{1/2}. \tag{2.27}$$

With this choice of $E\{R\}$, the costs of delay and of operation are equal.

The number of passengers on each bus is

$$\lambda(t)\,E\{R\} = \left[\frac{2\gamma}{p}\lambda(t)\right]^{1/2}. \tag{2.28}$$

If over some long period of time $\lambda(t)$ should increase by a factor of 4 (but γ and p remain constant), the increased passengers would be accommodated by dispatching twice as many buses with each bus carrying twice as many passengers.

If buses have limited capacity and (2.28) should exceed the capacity of the bus, the optimal strategy is to dispatch the buses so that they are barely full. If one were to dispatch at any longer headway, passengers would be left behind and the delays would grow arbitrarily large. It is generally true that once a vehicle is full, it should be dispatched immediately. Nothing can be gained by having it sit waiting for passengers who cannot board.

(c) *Queueing for gaps*

For pedestrians crossing a highway or cars which must yield to another traffic stream, the delays to these customers depend upon the time interval between acceptable gaps. The time between services is, of course, quite sensitive to the arrival rates of the opposing traffic stream and its headway distribution.

The typical question which arises here is whether or not one should install a traffic signal which, in effect, changes the headway distribution on the opposing stream. As in example (a), the issue now becomes a balance between the delays for the two traffic streams.

Problems

2.1 A service facility (highway) is capable of serving customers (cars) at a constant rate of μ customers per hour. Customers arrive at a constant rate $\lambda_1 < \mu$ until some time $t = 0$ (7.00 a.m.), but from $t = 0$ until some time τ (9.00 a.m.) they arrive at a rate $\lambda_2 > \lambda_1$. After time τ the arrival rate returns to the value λ_1 and remains there until time τ' (7.00 a.m. the next day) when the pattern repeats itself. If $\lambda_2 > \mu$, customers who cannot be served immediately form a queue and are served first-in, first-out.

Draw curves for the cumulative arrivals and departures of customers starting at time $t = 0$ when there is no queue. Evaluate and identify geometrically

(a) the maximum queue length
(b) the longest delay to any customer
(c) the duration of the queue
(d) the total delay to all customers during the time 0 to τ'.

2.2 As in Section 2.2, let $A(t)$ represent the cumulative number of customers to arrive by time t and $D_q(t)$ the cumulative number

to enter the service. Suppose, however, that there is a storage space for only c customers (enough to keep the server busy at all times when $\lambda(t) > \mu$); anyone who arrives when the storage is full goes away (he may be served elsewhere). If the server can serve at a maximum rate μ and $A(t)$, $\lambda(t)$ are as shown in Fig. 2.3(a), (b), determine $D_q(t)$ and $\mu(t)$.

2.3 In selecting a facility to serve the arrival pattern of problem 2.1, a designer has the option of choosing any values of μ. The cost per unit time of providing a service rate μ (labor, interest on investment, etc.), however, is proportional to μ, i.e.,

$$\text{service cost during time } \tau' = \alpha\mu\tau',\ \alpha = \text{constant}$$

regardless of whether or not the facility is fully utilized. The designer proposes that the value of a customer's time is worth p per unit time ($ per hour), i.e., the total cost of delay is p times the total delay.

Determine the choice of μ which the designer would select if his objective is to minimize the sum of service cost and delay costs.

2.4 Suppose that on day j the arrival curve of customers to a facility with fixed service rate μ has the form

$$A_j(t) = A_j F(t)$$

with $F(t)$ independent of j. On day 0 the arrival rate has a single maximum of the type

$$\lambda_0(t) = \mu - \beta(t - t_1)^2$$

i.e., the maximum arrival rate on day 0 is just equal to the service rate.

If the demand increases at a constant fractional rate of α per day

$$A_j = A_0[1 + \alpha j] \quad \text{for} \quad j \geq 0$$

how will the total delay W_j on day j increase with j.

2.5 An automobile assembly plant can assemble cars only at a single rate of μ cars per day or close down. It costs $\$p$ per day to store an assembled car. There is a fixed cost per day that is independent of whether the factory is operating or not. In addition, there is a cost of $\$\alpha$ per day (labour) to operate at rate μ and the equivalent of 2 weeks of operating cost to close the factory and start again (no matter how long it is closed).

What strategy of operation should be used to minimize the long time average cost per day of operation if the factory must satisfy a steady demand of λ cars per day, $\lambda < \mu$?

2.6 A vehicle-actuated traffic signal serves two traffic streams with cumulative arrival curves $A_1(t)$ and $A_2(t)$. Show how one could graphically construct departures curves $D_1(t)$ and $D_2(t)$ for the two traffic streams if the queue discipline alternates as follows: stream 1 is served at a rate μ until the queue in stream 1 vanishes, then there is a lost time L during which no one is served, then stream 2 is served at a rate μ until that queue vanishes, then is another lost time L followed by service to stream 1, etc.

2.7 Two bus routes, one with headways of 10 minutes, another with headways of 20 minutes, merge along a common section of route. The schedules are synchronized so as to create headways in the sequence 5, 5, 10, 5, 5, 10, ... If a passenger who wishes to travel along the common route arrives at a random time, what is the probability that he must wait for a time greater than t? What is his average waiting time?

2.8 Each of two buses carries passengers from a depot to various destinations and return for another trip with a round trip time very nearly equal to T. The buses are run by independent drivers, however, who make no attempt to coordinate their schedules. Actually, one bus runs slightly faster than the other so that over many trips the fraction of trips that the second bus leaves within a time t after the first bus is t/T, $0 < t < T$. In effect, the times between departures of the buses are random with a uniform distribution over the interval $0 < t < T$.

If passengers arrive at the depot at a constant rate, what is the average time that a passenger must wait for the next bus? Compare this with the wait if the headways are controlled so as to be $T/2$.

2.9 Two shuttle buses, each of which can carry c passengers, serve the same bus depot. The time between dispatches of the same bus is 15 minutes, but one of the buses is dispatched 10 minutes after the other so as to create headways 10, 5, 10, 5, etc. Passengers arrive at the depot at a constant rate of λ per minute.

If any passenger who arrives at the depot and finds a queue of c passengers goes away, what is the long time average number of passengers served per unit time by the buses, as a function of λ? What is the average wait per passenger for those passengers who actually board a bus?