STOCHASTIC DELAYS IN TRANSPORTATION TERMINALS:
NEW RESULTS IN THE THEORY AND APPLICATION
OF BULK QUEUES

by

WARREN BUCKLER POWELL

B.S.E., Princeton University
(1977)

S.M., Massachusetts Institute of Technology
(1979)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January, 1981

Signature of author..
Department of Civil Engineering
January 27, 1981

Certified by...
Yosef Sheffi
Thesis supervisor

Accepted by...
Chairman
Departmental Committee on Graduate Students
of the Department of Civil Engineering

STOCHASTIC DELAYS IN TRANSPORTATION TERMINALS: NEW RESULTS IN THE THEORY AND APPLICATION OF BULK QUEUES

by

WARREN BUCKLER POWELL

Submitted to the Department of Civil Engineering
on January 27, 1981 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in
Civil Engineering

ABSTRACT

The incorporation of stochastic effects into models of terminal performance is traditionally accomplished through the use of simulation. In an effort to reduce the computational overhead of an analysis based on simulation, this research lays out an approach for studying stochastic delays in terminal operations based on theoretical results from the theory of bulk queues. After formulating different queueing problems as bulk queues, transforms of the queue length and waiting time distributions are derived for bulk arrival queues with service and random batch capacities. Also, the concept of a scheduled departure queue with cancellations is introduced and explored. In most cases, the relevant moment formulas for the mean and variance of both the length of the queue and the waiting time are derived.

Following the theoretical work, several numerical problems associated with the application of transforms are addressed. First, it is shown that the task of finding the zeroes of a particular function, required to solve the transforms, does not pose any significant numerical problems, contrary to several recent papers, and can in fact be performed very quickly. Next, it is shown that these zeroes can be used to perform a partial inversion of the queue length transform, but that methods for performing complete numerical inversions are extremely sensitive to roundoff errors. Finally, a method for approximating moments which eliminates the need of solving for zeroes is developed and shown to yield accurate results.

The last problem addressed is the validity of the assumption of a Poisson arrival process. It is shown that the Poisson arrival process is not a good approximation for general arrival processes which may in fact appear Poisson. With this in mind, a methodology is outlined for approximating general arrival processes.

Thesis supervisor: Yosef Sheffi
Title: Professor of Civil Engineering

Acknowledgements

The successful completion of any substantial piece of work inevitably involves the support of many individuals, and this thesis is no exception. While it is difficult to name veryone who has influenced the research, I would like to acknowledge the efforts of those who helped mold the final product. My first thanks go to my committee, Amedeo Odoni, Dick Larson, Paul Roberts, Pierre Humblet and in particular my chairman, Yossi Sheffi. Outside of a yeoman's job of reading and critiquing some terrible earlier drafts, Yossi has provided an atmosphere of encouragement throughout the research which helped me get through several difficult stages of this research. I will be forever grateful of the many afternoons I spent in his office discussing anything but my thesis. I would like to thank Amedeo for his many comments on several earlier drafts, and Pierre, for his meticulous checking of the theoretical work and for several observations which both simplified and extended the results. Dick Larson also provided valuable comments on the organization of the material in an earlier draft of the thesis.

The research presented in this thesis was intentionally confined to the development of analysis techniques from the esoteric theory of bulk queues. The motivation of the research, however, originated in the context of describing large scale, less-than-truckload trucking networks. In this respect, I would like to thank Paul Roberts for providing the "trucking" perspective of the research. I am also indebted to Mike McGee and John Terry of IU International for funding the first year of this research under the IU International Fellowship. This fellowship provided the much needed independence required to initiate such basic research. I would in particular like to thank Mike for setting up several interviews with terminal managers in Ryder and P.I.E., allowing me to see terminal operations first hand and to talk to shippers who use LTL services.

No list of acknowledgements would be complete without the list of people who have contributed indirectly, though no less significantly, by providing a congenial working environment. In this respect I would like to thank Hani and Millie, Mike, Cliff, Frank, Brendon, Eve, Vic, Fred, Oscar, Jim and Josue. I would also like to thank Woody Richardson for introducing me to the topic in the first place.

Finally, I would like to dedicate this thesis to my wife, Shari, whose enduring support and encouragement through the long haul has given this thesis purpose.

TABLE OF CONTENTS

List of figures

List of tables

Chapter 1 Modeling Stochastic Delays in Transportation Terminals

Over the last decade, planners have become increasingly aware of the importance of service reliability as a major factor in the mode choice decision. In both passenger and freight modes, users of the system are as sensitive to the variability in total travel time as they are to its expected value. Martland (1972), Reid et al. (1972), and Folk (1972) have looked at determinants of service reliability in the rail mode, and the subsequent effect on a shipper's logistics costs. This viewpoint contradicted the conventional wisdom that the rail mode offered a slightly lower level of service (measured in terms of average travel time) at a lower price. Several recent studies have focussed on similar problems in transit (see, for example, Turnquist and Bowman (1979), and Abkowitz (1980)), and a number of papers have appeared on the problem of controlling randomness in bus operations. Terziev et al. (1978) and Richardson (1979) have looked at the effect of variability in demand on service reliability in less-than-truckload (LTL) trucking networks.

Variability in the level of service offered can in most cases be attributed to the randomness in the demands placed on the system. Operators usually have very little control over the market (an exception being air markets, where reservation systems and discount fare restrictions help reduce some of the variability) and must design a set of services which strike an appropriate balance between costs and level of service, reflecting the needs of the user. In this respect, there are a number of strategies based on resource pooling and the level of system control which offer different choices in terms of cost and level of service. For example, vehicle pooling and increased flow consolidation over links and through terminals help to mitigate the relative variability in demands, but usually (though not always) at a lower level of service. Different levels of system control are found in the use of fixed departure schedules, set by a central scheduling department, and real time dispatching, exercised at the terminal level.

Thus, while an operator may not be able to control the market, there are a range of strategies available for responding to it. Unfortunately, the state of the art in analysis methods restricts the ability of a manager to effectively compare and evaluate these strategies in the context of realistic models. In particular, the analyst needs to be able to look at large networks in order to test routing policies, terminal location, fleet control strategies and vehicle dispatching strategies. At present, Monte Carlo simulation has been the only approach available for modeling complex systems with a stochastic component. While an extremely powerful tool for many situations, simulation is also notoriously slow computationally and poses additional statistical problems associated with analyzing the outputs. For example, a large motor carrier network might have over 400 terminals with over 1,500 individual queues. In most cases, extremely efficient deterministic network algorithms are used which model only the average flows on the system. These methods enable planners to estimate the capacity required at different points in the system and help locate potential bottlenecks. Because they ignore the distribution of demands at each link, however, deterministic models generally underestimate true delays. More importantly, the analyst cannot evaluate operating strategies aimed directly at controlling the effects of randomness in demand on service reliability and costs.

On the basis of these observations, it is clear that there is a definite need for a methodology for incorporating stochastic demand in the analysis of transportation networks. The level of detail required should be consistent with a large, strategic planning model and computationally fast enough to allow the evaluation of a large number of alternatives at reasonable cost. Such an objective has provided the general direction of this thesis. The goal here, however, is not in the actual development of such a model, but rather to first identify a basic methodology, and then to fill in many of the gaps needed to implement the approach. Such gaps include additional theoretical work on simplified problems and supporting numerical work required to implement the theory in a problem solving context.

The following sections provide a more detailed description of the specific set of topics that are covered in this research. The main focus is on the formulation of stochastic delays in steady state using the theory of bulk queues. Section 1.1 outlines the basic operations that a terminal performs, each of which represents a source of delay. Section 1.2 casts each operation as a specific queueing model and describes a classification scheme for the different types of queues encountered in terminals. Section 1.3 then highlights some of the important assumptions that are made in modeling the demands over the system. Finally, section 1.4 summarizes the chapter and presents the organization of the thesis.

1.1 Modeling delays in terminals

The total time required to transfer a passenger or shipment over a network can be divided into two components, linehaul time, and terminal transfer time. The relative sizes of each of these components may vary widely among models, but in most instances the principal source of randomness in the total time occurs at the terminals. For this reason, this research concentrates on finding the distribution of time from when a passenger or shipment arrives at a terminal until it departs. The first task, then, is to describe the basic transfer operations that take place in terminals and outline the delays that may occur at each point. This is done using the operations of a typical less-than-truckload (LTL) break-bulk terminal as an illustration, followed by brief descriptions of how the same concepts would apply to other modes.

Figure 1.1 shows a typical layout of an LTL break bulk terminal. These buildings resemble large warehouses with up to 150 doors which trucks back up to for loading or unloading. After an arriving truck first backs up to an empty door (usually preassigned), one or more men then unload the freight, using either a forklift or by placing the freight on a handcart. Each shipment is then loaded onto an appropriate outbound truck, or if no truck is waiting, placed on the floor and stored until a truck becomes available. When one does, then the freight waiting on the floor must be loaded onto the truck, which if full, may then be dispatched.

Dispatching in trucking terminals generally takes two forms. Trucks traveling from one break-bulk terminal to another are usually sent on a

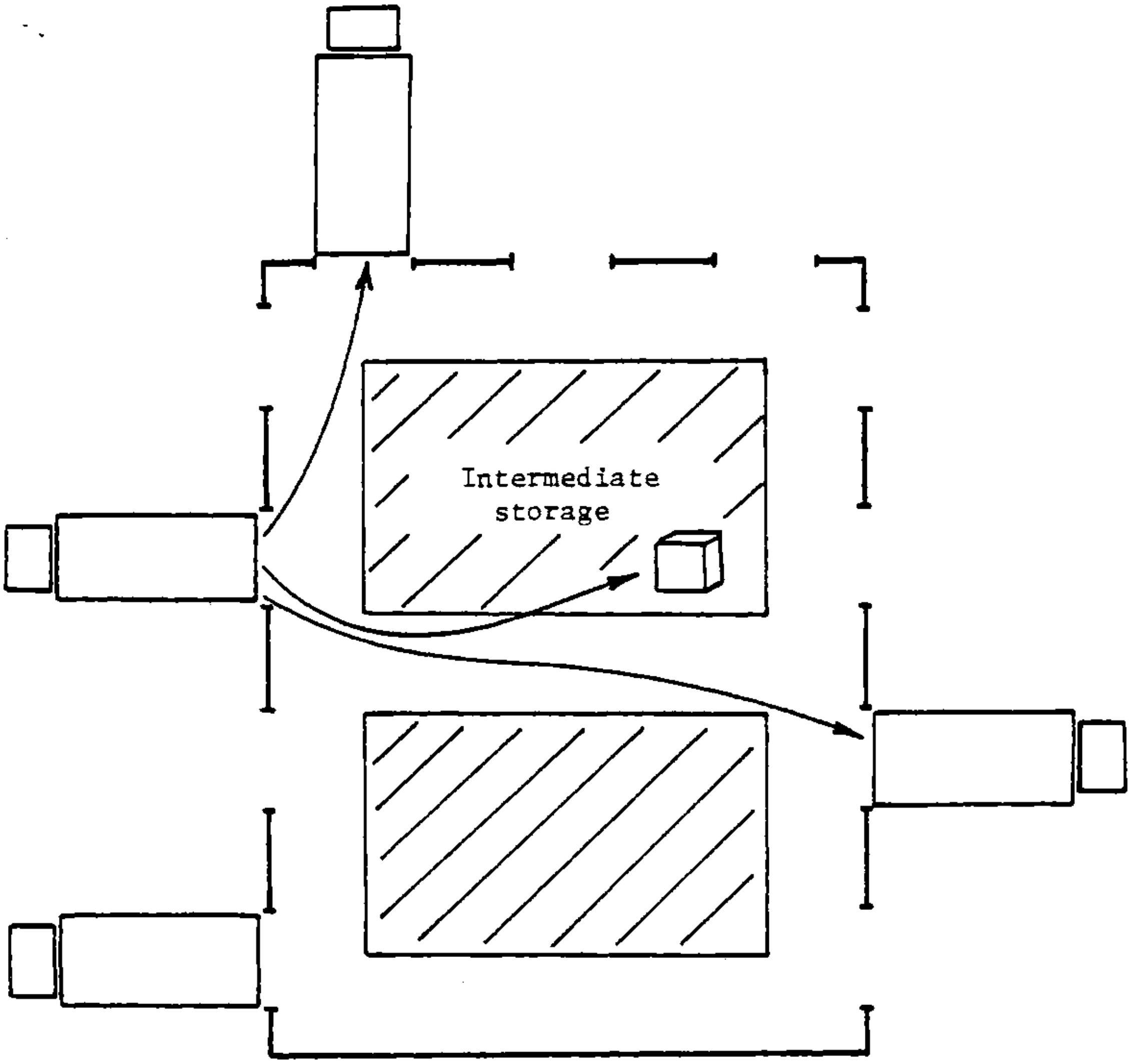


Figure 1.1

Layout of a less-than-truckload

consolidation terminal

go-when-filled strategy; since volumes between break-bulks are usually quite high, there is little chance that a shipment will wait more than a day before the truck is full. On the other hand, trucks moving between a break-bulk and an end-of-line terminal, which handles the pick-up and delivery operations with the shipper, tend to run on a once-a-day basis, even if the trucks are partially empty. Since the volume of freight into and out of end-of-line terminals is relatively low, this strategy guarantees a minimum level of service.*

The time required for a shipment to move through a terminal can be divided into three components, unloading, intermediate storage, and loading, as depicted in figure 1.2. The unloading time is that interval from when a vehicle first arrives in a terminal until the shipment in question is placed in intermediate storage. Intermediate storage generally refers to the terminal floor, but can also mean the outbound truck itself. Sorting the freight also occurs at this step. The time spent in intermediate storage is also referred to as connection delay and consists of the time from when a shipment is placed in intermediate storage until the decision to dispatch is made, at which point the loading process begins. In other words, a dispatch decision means that the vehicle should be sent as soon as the freight sitting on the terminal floor is loaded. The time required to complete this process is, of course, the loading time. Note that if all the freight has been loaded

* The two strategies also reflect differences in the nature of the two types of terminals. Break-bulk terminals operate continuously 24 hours a day; end-of-line terminals have a daily cycle with separate shifts since pick-ups and deliveries can only occur during business hours.

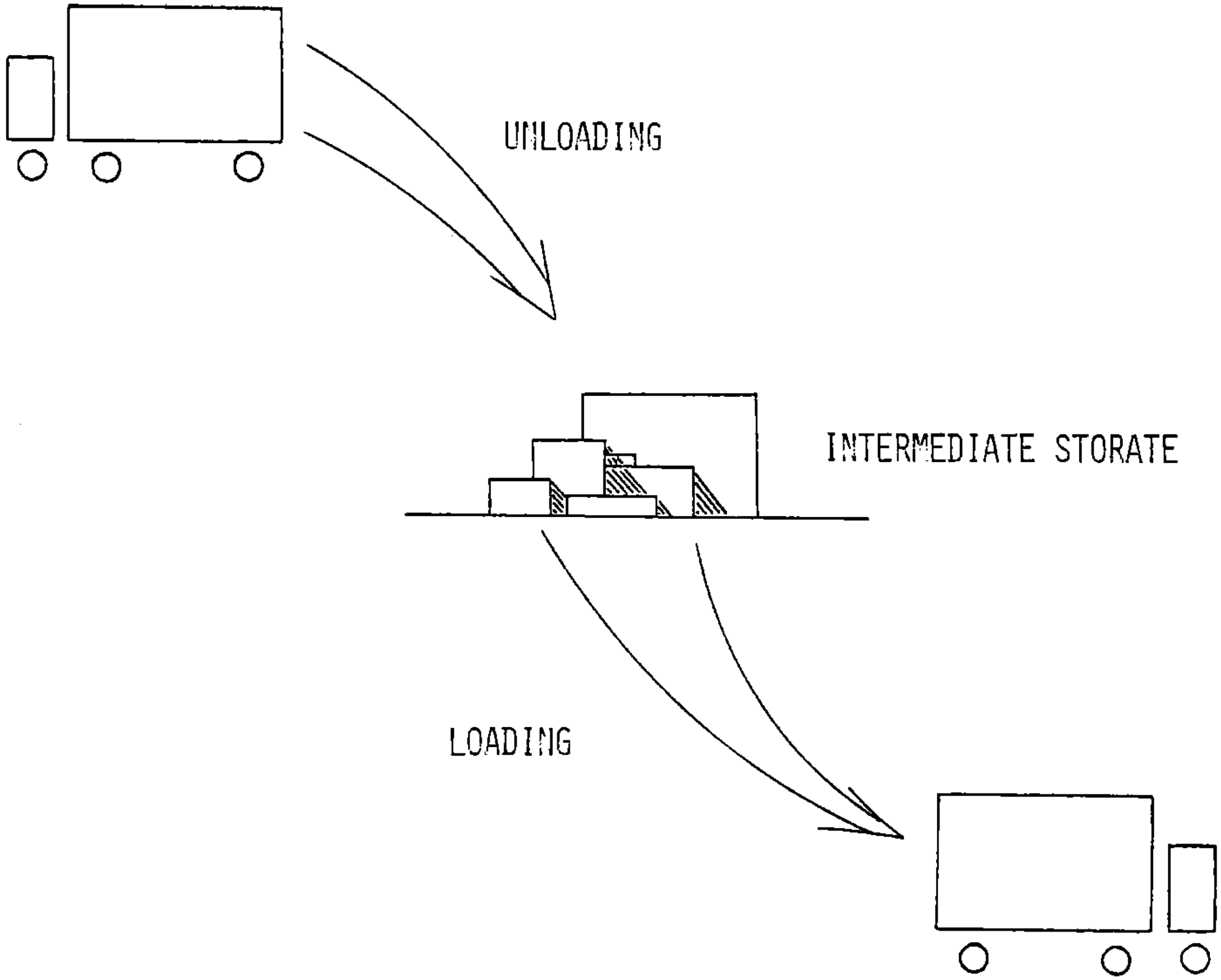


Figure 1.2

Sources of delays within a terminal

directly onto a waiting truck, then the loading time is zero.

The three step process of unloading, storage, and loading applies to every other mode as well, although in some cases one or two of the steps may require a negligible amount of time. For example, a bus stop is a very simple terminal where arriving passengers have no unloading step and loading time is often neglible. Rail classification yards, on the other hand, are relatively complex terminals where all three steps are important. Here, the "vehicle" is the locomotive consist (representing one or more locomotives), with the load being the set of freight cars being pulled. The unloading step is the classification process, where blocks of cars are disconnected from the train and placed on appropriate outbound tracks. Connection delay is then the time spent in the classification yard until the dispatch decision is made, at which point a crew begins to assemble the outbound train (the loading process).

Having described the principal operations of a terminal, the next task is to formulate each step as a queueing model. Section 1.2 provides these formulations and outlines a classification scheme for different types of queues arising in transportation applications. Then, section 1.3 discusses some of the major assumptions that are made regarding the demands on the system and the manner in which arrivals to a queue are being represented.

1.2 Formulating terminals as queues

The three steps in the terminal transfer process can each be modeled as a specific queue. For the moment, we are interested only in the existence of bulk arrivals, bulk service, or both. Later other assumptions regarding the nature of the arrival and service processes are presented.

1.2.1 The unloading queue

The unloading queue consists of vehicles, partially or completely full of passengers or freight, arriving at an unloading area. An unloading crew, once available, will unload the freight, placing each shipment in an intermediate storage location and then return for another. Such a system can be viewed as a bulk arrival system with service in single units, although it may be advantageous to assume bulk service as well. There are, however, instances in which the arriving vehicles are the "customers", and where the service time is the time required to unload the entire vehicle. An arriving vehicle, or group, is sometimes referred to as a "supercustomer" when viewed in this manner. This approach applies to problems where outbound departures from the storage queue might be delayed until the entire contents of a vehicle have been unloaded.

1.2.2 The storage queue

The storage queue represents passengers or freight waiting to depart on a specific outbound link, as illustrated in figure 1.3. Arrivals to the queue represent departures from the unloading queue, and here two possibilities arise. If the unloading process significantly spreads out the arriving groups (a group referring to a load of freight or passengers), then arrivals to the storage queue may be in single units, as illustrated in figure 1.4a. On the other hand, if unloading times are short relative to the interarrival times of groups to the unloading queue, then arrivals to the storage queue will also appear to be in groups as well as depicted in figure 1.4b. Service, however, is always in bulk, since departures from the queue are on vehicles.

1.2.3 The loading queue

Unlike the previous two cases where arrivals are in the form of vehicles or shipments, arrivals to the loading queue are represented by the sequence of decisions to load and send a vehicle. If departures are scheduled, then the schedule forms the arrival process to the queue. If departures are on a go-when-filled basis, then an arrival represents a queue becoming long enough to fill the vehicle. Service is rendered by the loading crew, and the service time is the time required to load all of the waiting shipments. Thus the queue has arrivals and departures in single units.

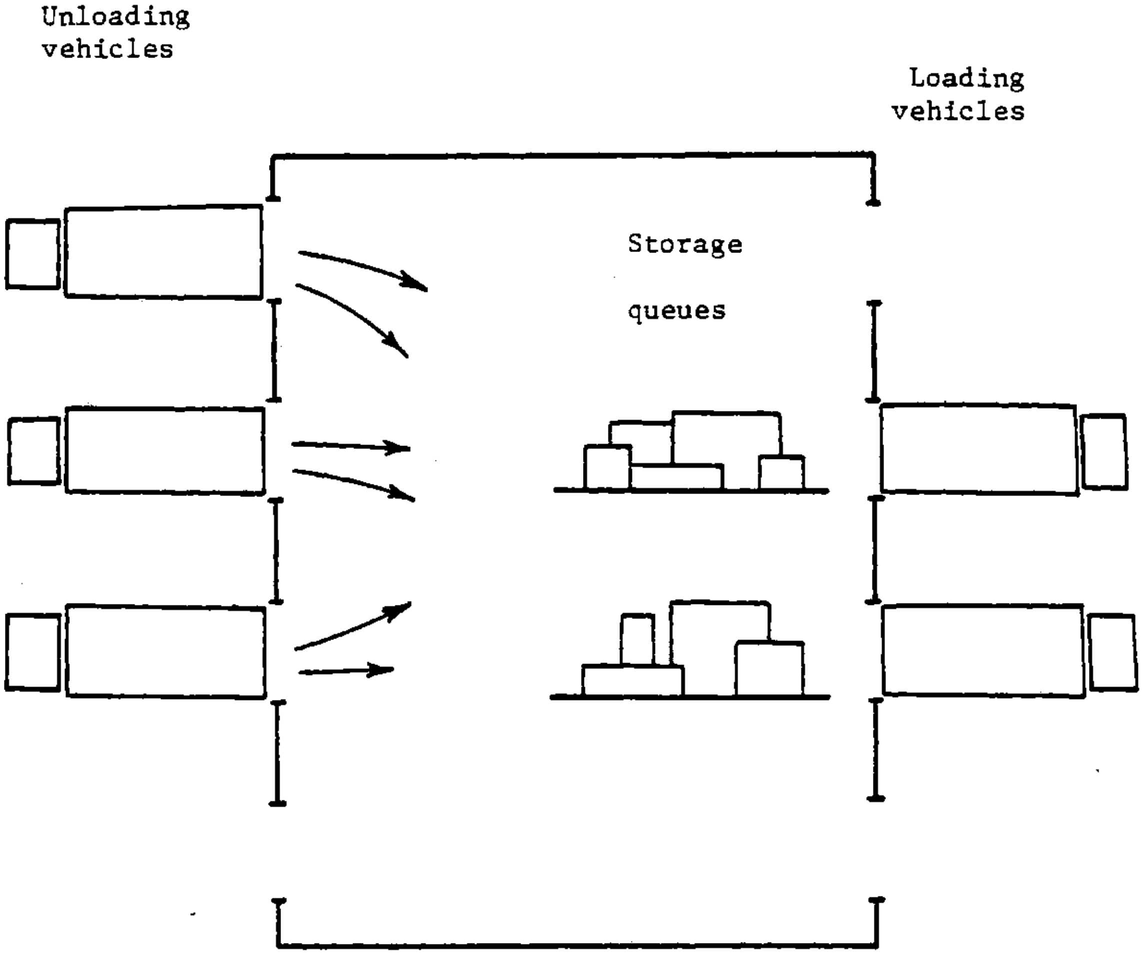


Figure 1.3

Illustration of storage queues

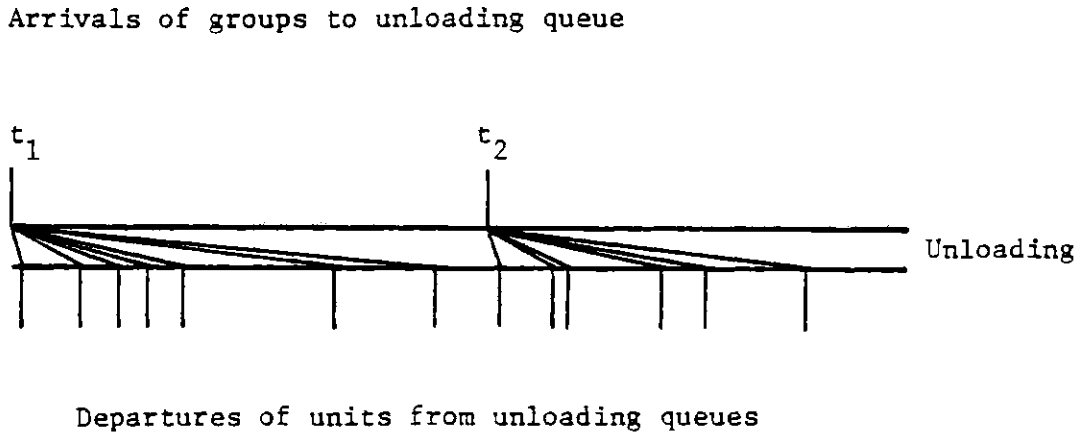

Arrivals of groups to unloading queue
t_1
t_2
Unloading
Departures of units from unloading queues

Figure 1.4a

Dispersed departure process from unloading queue

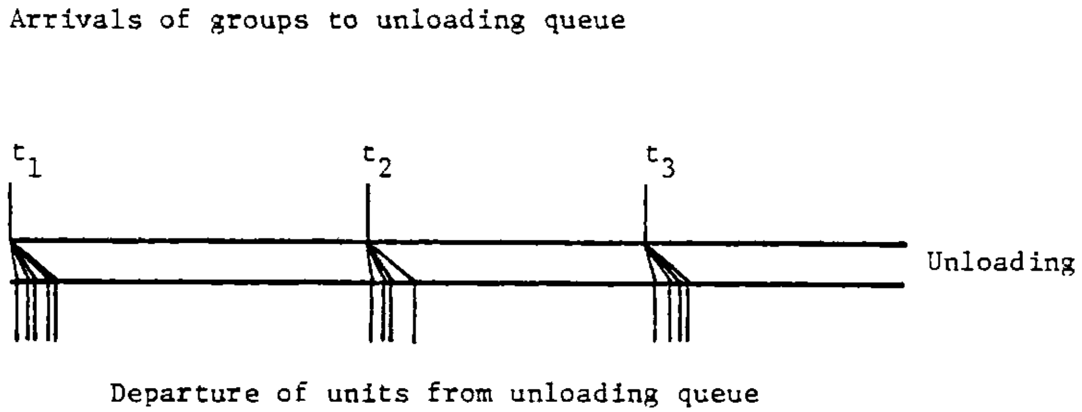

Arrivals of groups to unloading queue
t_1
t_2
t_3
Unloading
Departure of units from unloading queue

Figure 1.4b

Bunched departures from unloading queue

1.2.4 Other aspects of loading and unloading queues

The above descriptions identify where bulk arrivals and bulk service are likely to occur. The next question that naturally arises is the nature of the arrival and service processes. All the theoretical work presented in this research assumes simple or compound Poisson arrivals and general service times.* Notationally, the queue would be represented as $M^x/G^y/1$, meaning Poisson arrivals of groups of size x, general (independent) service times with bulk departures, where the capacity of the outbound vehicle is a random variable y. To obtain analytically tractible results, we must also assume a single server, which obviously poses some problems when there are, say, k loading and unloading crews. Such cases must be handled approximately, possibly by estimating bounds on delay obtained by first assuming k independent queues (an upper bound) and then by assuming a single server that operates k times as fast as a single one (a lower bound).**

Thus the general problem can be cast as an $M^x/G^y/1$ queue, which has as special cases queues with arrivals and for service in single units. There is, however, an additional and very important distinction separating loading and unloading queues from storage queues. The first two have a well defined service process, where the server is the loading/unloading crew and the service is the time required to move freight onto or off of a vehicle. Total loading or unloading time is then made up of the wait time in queue plus the service time. The service process for the storage

* Chapter 5 outlines methods for approximating general arrival processes.

** See, for example, Brumelle (1971).

queue, on the other hand, is more of an artificial construct. For example, if there are scheduled departures from a terminal once a day, then the service time is one day. This bears no relation to the service being performed, which is the movement of goods over the link. Hence, we would only be interested in statistics describing the length of and waiting time in the queue as opposed to the system including the queue plus the server.

The differences between loading/unloading queues and storage queues are not really fundamental, and in fact, the two types are actually closely related. The differences are, nonetheless, substantive in terms of the problems being described and should be distinguished. For this reason, the two types are termed here respectively service queues and dispatch queues, reflecting the nature of the underlying service process. Service queues represent the more conventional queues commonly studied in the literature, and can be distinguished partially by the fact that the server is idle if and only if the system is empty. Conventional notions of busy periods and idle periods also apply, concepts which become somewhat fuzzy in the context of dispatch queues.

Dispatch queues describe the time until a passenger or shipment finds available space on a departing vehicle, and are more difficult to characterize because the service processes are more complex. In service queues, for example, the server works as long as there are any demands waiting, and stops when the system is empty. Dispatch queues, on the other hand, may exhibit a variety of control strategies that determine when a vehicle leaves. These can be divided into three general categories based on the availability of vehicles and the level of control being exercised, as follows:

1) Scheduled departure queues assume that vehicles depart at predetermined instants that are independent of the arrival process or the state of the queue. The vehicle departs even if no-one is waiting. Examples include bus stops, subway stations, and trucking terminals with once-a-day departures.

2) Real time queues determine departures solely by the state of the queue. The most obvious example is the go-when-filled strategy, but also includes cases where there is a limit on the longest waiting time. The important feature here is that a vehicle must in theory always be available, regardless of the time of the last departure.

3) Quasi-real time queues are identical to real time queues with the exception that now there is a period of time following each departure during which no departures can occur, regardless of how long the queue might become. Once the vehicle becomes available, it may either leave immediately or be held until the dispatch criteria are satisfied.

One generalization of the scheduled departure queue is to allow for cancellations. In this case, we define a set of dispatch instants, t_o, t_1, t_2, . . ., which are independent of the arrival process or the state of the queue. At each instant, a decision must be made to send the vehicle or cancel the run, the latter choice occurring whenever the queue is less than some specified minimum. If the queue is too short, the departure must be cancelled until the next dispatch instant. Note that the scheduled departure queue without cancellations and real time queues can both be viewed as special cases of quasi-real time queues, whereas the scheduled departure queue with cancellations cannot.

The classification of queues presented here is summarized in table 1.1. The distinction between dispatch and service queues is an important one in the study of bulk queues, but is nonetheless ignored in a number of papers. This organization of the different types of queues proves useful in chapter 2, where the review of the literature is presented by identifying how the different contributions relate to each other. The theoretical work presented in chapter 3 focuses on the $M^x/G^y/1$ scheduled departure queue, with and without concellations, and the real time, go-when-filled queue, since these represent problems that need the most work. A detailed description of the problems to be solved is provided at the end of this chapter. The next section outlines some of the important assumptions regarding the demands on the system.

- I Dispatch queues
    - Scheduled departure queues
        - Without cancellations
        - With cancellations
    - Real time queues
    - Quasi-real time queues
- II Service queues

Table 1.1

Classification of queues

1.3 Modeling demand

In this section, an outline of some of the major assumptions is made in modeling the arrivals to a queue and the manner in which stochastic flows can be characterized. The two principal assumptions are:

1) the arrival and service processes are stationary, and the queue can be modelled in steady state, and
2) goods moving though the network can be described as discrete, identical units which are routed independently.

The first assumption is often made in queueing applications, and while it is probably not only the strongest one that is made, it is also fundamental to the methodology and therefore cannot be relaxed within the existing framework. One of the basic hypotheses of this thesis, however, is that the ability to incorporate randomness in demand represents a quantum improvement over deterministic methods, and that an improvement in this respect, which still produces a computationally feasible model, is more appropriate than one which is more accurate (e.g. simulation) but otherwise unusable.

The second assumption has two important components. The first requires that flows be discretized into homogeneous units and represents a natural outgrowth of the study of passenger systems, where each person is a separate unit. Rail freight systems also exhibit a natural discretization in terms of freight cars. In the case of LTL trucking, however, shipments are of varying size, and hence must be broken into some arbitrary unit, such as 1,000 pound blocks, for example. This leads into the second component which assumes independent routing of units. Violations of this assumption would occur, for instance, when families travel together, when a shipper sends several freight cars to the same destination, or when two 1,000 pound units actually represent a single 2,000 pound shipment.

At this point, it is useful to describe how stochastic flows over a transportation network can be characterized. In doing this, we use the notion of a compound arrival process. A simple arrival process, such as the Poisson, is used to describe the random arrivals of single customers. A compound arrival process occurs when groups of customers are arriving at the queue, where the arrival of groups might be given by a Poisson process, but where the size of each group is given by a completely independent distribution. The use of compound arrival processes is useful in characterizing not only the arrival of shipments of varying size from the shipper, but also the arrival of vehicles at a terminal with different sizes of loads.

It is important in the study of terminals to separate the arrival of vehicles to a terminal and the arrival of one or more units out of that vehicle to a particular queue. For example, the distribution of the number of units on an arriving vehicle might be given by a random variable F, but the distribution of the number of units out of that vehicle that, after being sorted, arrive to a particular departure queue might be given by a random variable G. It is the distribution of G that is needed to study the delays at the queue, and in chapter 3, we show how to find G given F, using the independence of routing assumptions.

1.4 Summary

This chapter serves as an introduction to the primary set of problems that are addressed in this research. The theoretical work focuses on the $M^x/G^y/1$ scheduled departure queue with cancellations, and therefore can be viewed as a contribution to the analysis of bulk arrival, bulk service queues in steady state. The discussion here is intended to illustrate in general how bulk queues can be used to study terminal operations, and in particular, the specific application of the research presented in subsequent chapters.

The outline of the research is as follows. First, chapter 2 provides a thorough review of the literature on bulk queues, using the classification scheme depicted in table 1.1 to help organize the many contributions. Also covered is some of the literature on queueing networks. Several recent papers on approximate methods for analyzing complex queueing networks are discussed in terms of applying the approach to transportation networks. These methods are interesting since, while the focus of this research is on single queues, they provide an approach for applying the results obtained here to the general study of transportation networks.

Next, chapter 3 presents a compenduem of theoretical results associated with the $M^x/G^y/1$ scheduled departure queue, with and without cancellations. The principal results are queue length and waiting time transforms and the formulas for the mean and variance of each. Part of the contribution here, in addition to several new results, is the simplicity of the analysis, which uses standard transform methods to solve for queues using the imbedded Morkov chain. It is then shown how the queue at regeneration points is related to the number of units in front of an arriving or behind a departing unit, and to the length of the queue at a random point in time.

Chapter 4 provides a considerable amount of numerical support work needed to bridge the gap between deriving transforms and using them. Solution of queue length transforms requires finding the zeroes of a particular function, a problem which in principle can pose serious computational problems. It is shown that, when set up in a particular way, the root finding problem does not in fact pose any serious difficulties and furthermore can be solved very efficiently. In addition, however, extremely accurate approximations are developed which give the mean and variance in closed form, thus eliminating the root finding procedure altogether. These are then used to fit approximate queue length distributions, which are compared against the true distributions and found to be quite accurate.

Chapter 5 then uses simulation to look at non-Poisson arrival processes that might occur in transportation networks. One conventional technique for determining if a given arrival process is Poisson and one new one are tested, and while the latter one, based on analysis of successive increments of the arrival process, is found to be much better, both have serious drawbacks. In particular, it is shown that assuming a Poisson arrival process can produce seriously biased estimates of the length of the queue even when the actual arrival process appears to be Poisson. Based on these results, a new approach for approximating bulk queues with general arrival processes is tested and shown to provide significantly better estimates of the mean queue length.

Finally, chapter 6 summarizes the major contributions of the research and outlines directions for further research.

Chapter 2 Review of the Literature on Bulk Queues and Queueing Networks

As is established in chapter 1, the principal focus of this research is the analysis of stochastic delays in terminals using the theory of bulk queues. The first step in this direction is a thorough review of the literature on the topic. To date, such an overview has been notably lacking and many papers can be somewhat ambiguous in terms of the specific problem being addressed. For this reason, the first goal of this chapter is to organize the many contributions in terms of the problem that is being considered, what results were obtained and the solution approach. Then, since the remainder of the thesis focuses on bulk arrival, bulk service, scheduled departure queues, the state of the art with respect to this particular problem is addressed and major gaps are highlighted. Filling these gaps then becomes the subject of chapter 3.

The chapter is organized into two sections. Section 2.1 concentrates solely on the bulk queueing literature and represents the relevant literature for the remainder of the thesis. Section 2.2, motivated by the original objective of applying the research to study large scale networks, briefly touches on some recent papers that have appeared in the area of approximate analysis of queueing networks. The purpose of this section is simply to indicate that a method for studying queueing in transportation networks exists. The topic is discussed again in chapter 6 where the mechanics of designing a method for analyzing large scale queueing networks are outlined as an area for further research.

2.1 The bulk queueing literature

The study of bulk queues originated with the pioneering paper by Bailey (1954), who considered the case of simple Poisson arrivals to a bulk queue with service capacity c. Service times were general and departures occurred regardless of whether anyone was waiting, a system which we refer to as a scheduled departure queue. This system has come to be known as the transportation problem of queueing theory. Since that time, papers have appeared which can be differentiated on the basis of 1) the type of queue (arrival process, service process, number of servers), 2) what is solved for (queues, waiting times, busy periods, etc), 3) the time domain of the solution (i.e. steady state or transient, and whether they apply to regeneration points or continuously over time), and 4) the method of solution (transforms or direct numerical methods). For simplicity, the review of the literature is organized in terms of the type of queue being solved, focusing specifically on the breakdown between systems with bulk arrival, bulk service, or both. The large majority of all papers assume Poisson or compound Poisson arrivals and general service times, although a few consider negative exponential, Erlang, or hyperexponential service distributions.

As discussed in chapter 1, there is also an additional question regarding whether the system being solved represents a dispatch queue or a service queue. For the most part, the papers with single arrivals and bulk service fall under the heading of dispatch queues, while the others (bulk arrival, single service, and bulk arrival, bulk service) are usually service queues. Exceptions do exist and these are noted as they occur.

The review is organized into three sections, which discuss, respectively, a) single arrival, bulk service queues, b) bulk arrival, single service queues, and c) bulk arrival, bulk service queues. These sections, however, are confined to papers which use standard transform techniques for solution. Section 2.1.4 reviews those papers which present numerical techniques for solving transient and steady state queues. Finally, section 2.1.5 summarizes the principal contributions and points out unsolved problems.

2.1.1 Single arrival, bulk service queues

As mentioned earlier, Bailey (1954) originated the study of bulk queues by considering a system with simple Poisson arrivals to a server which takes, at particular points in time, all waiting customers up to a fixed capacity c. If no customers are waiting, an empty batch departs, implying that the server is never idle. The queue, denoted by $M/G^c/1$, is described using the embedded Markov chain defined at points of service completions. Bailey found the transform of the queue length distribution at these points, the derivation of which required the determination of the roots (zeroes) of a given function, which depend on the service time distribution. The solution was then demonstrated for deterministic and Erlangian service distributions.

Immediately following Bailey's paper, Downton (1955) derived the Laplace transform of the waiting time distribution in terms of the queue length and service time transform. The importance of this result is that the distribution of the waiting time (or at least its mean and variance) can be found with no additional numerical effort. Downton follows exactly the general outline of Bailey's paper, and derives the moments of the waiting time distribution for the same set of special cases. This paper, and that by Bailey, provide the basis for most of the subsequent literature on bulk queues. Both of these results have been extended by Peterson (1971) to the case where the capacity of the vehicle is random.

Motivated by the problem of having to solve for the roots of a function, Downton (1956) followed immediately with a paper that considered the limiting case where λ, $c \to \infty$, while holding the ratio λ/c constant. In the case of deterministic departures, this eliminated the need to solve for any roots; for an E_R (R stage Erlang) service distribution, the procedure still required the solution of R roots, although the function changed from a polynomial of c + R to a polynomial of order R with an exponential term, thereby simplifying the computations. Unfortunately, these results are felt to be of little practical use since the limiting process changes the nature of the arrival distribution. The coefficient of variation of the Poisson distribution with parameter λ is $1/\sqrt{\lambda}$, which of course becomes small as λ and c are increased. In the case of deterministic service times, the probability that the number of arrivals will exceed c becomes zero as λ, $c \to \infty$, implying that no one will ever miss a departure. This observation is borne out by Downton's numerical results which show both the average waiting time and its coefficient of variation decreasing significantly for all cases as c is increased.*

Using the method of phases, Jaiswal (1960a) develops the necessary equations for studying the queue length distribution over continuous time. This approach assumes that the service distribution of a randomly chosen

*Despite these problems, Downton's limiting results have been applied by Peterson (1977a), who apparently did not recognize that the limiting result implied no queueing, and hence could be solved much more simply using renewal theory.

customer is E_R with probability P_r, $r = 1, \ldots, j$ where $\sum_{r=1}^{j} P_r = 1$. Such a service distribution is fairly flexible and can be used to approximate a wide range of cases encountered in practice. Later, Jaiswal (1960b) applies this to the time dependent bulk queueing problem, although here he introduces the change that the server becomes idle if the system is empty following a service completion. In this case, an empty batch can never occur, since the server waits until at least one customer arrives. This case, then, would fall under the heading of service queues.

Neuts (1966) next extends the analysis to incorporate randomness in the capacity of the vehicle, a system which is denoted by $M/G^x/1$. Furthermore, to describe the system is continuous time, he formulates the problem as a semi-Markov process, using a bivariate sequence to reflect not only the length of the queue at regeneration points but also the time elapsed since the last service completion. Without making it explicitly clear, he also assumes the server becomes idle if the system is empty following a service completion. He then derives the transform of the queue length distribution, as well as the Laplace transform of the queue length over time, although these results were derived only for the case of fixed outbound capacity. He does not, however, discuss how to solve these equations numerically, and his results appear to be useful only for finding the limiting distribution as $t \to \infty$ (arrival and service rates are assumed constant over time).

Up to now, we have considered only service processes with no control over the size of the smallest batch to be served, with the possible exception of servers which become idle until at least one is in the queue.

To describe cases where the server may be held until a sufficient queue is waiting, Neuts (1967) introduces the concept of the (so-called) general bulk service rule,* where a server, on finishing one batch, may remain idle if there are fewer than m customers waiting for service. Thus all departing batches from the queue have at least m customers, although no more than c. This system, then, would fall under the heading of quasi-real time dispatch queues. Again formulating the problem as a semi-Markov process, he proceeds in much the same way as he did in his earlier paper, determining the transform of the queue length in discrete and continuous time.

Borthakur (1972) next studies the queue length for the simpler system with negative exponential service times operating under a general bulk service rule and derives the steady state queue length probabilities directly in terms of the root of a given polynomial. Following on this result, Medhi (1975) finds the waiting time density function in terms of the same root. In both these cases, the results were in terms of relatively simple expressions for the queue length and wait time distributions directly, and not their transforms.

Several authors have considered the multiple server case, beginning with Arora (1963) who studies the $M/M^c/2$ server queue. This is later extended by Ghare (1968) to the more general k server case, where he

*The scheduled departure queue with cancellations is not a special case of Neuts' general bulk service rule.

found the steady state queue length probabilities in terms of a single root. Medhi and Borthakur (1972) subsequently consider the 2 server queue under the general bulk service rule, which Medhi (1979) then extends to the k server case. Finally Roes (1966) has investigated the many server case with general input, $GI/M^c/k$.

2.1.2 Bulk arrival, single service queues

The literature on bulk arrival queues has proceeded for the most part independently of that on bulk service queues, using a more classical description of the server. Unlike bulk service queues, where the server may work continuously regardless of whether there is anyone in the system, bulk arrival, single service queues incorporate the more traditional concept of an idle period where the server stops if no one is in the system. Most of the work assumes compound Poisson input and general service, a system that is denoted $M^X/G/1$.

The earliest work on the topic is that by Gaver (1959) who, in a very thorough paper, finds transforms for the steady state queue length and waiting time* at a random point in time by modelling the queue as a semi-Markov process. Restrepo (1965) studies the system with Erlang service times using the method of stages. Foster (1961) and Connolly (1960) investigate systems where arriving batches are of fixed size;

*Unfortunately, there is an error in his result for the waiting time transform which is described in section 3.6.

Foster (1964) later relates the $G^c/M/1$ queue (arrivals of fixed size c with general interarrival times) to the $G/E_c/1$ queue (see also Kleinrock, vol. 1 (1975)). Gupta and Goyal (1965) consider the case with hyperexponential service, and Gupta (1964) studies the $M^x/E_r/1$ queue using the method of phases as devised by Luchak (1958), where service is a random number of exponential stages. Chaudry (1979) points out the relationship between $G^x/M/1$ and $G/E_x/1$, where E_x denotes a phased distribution where the distribution of the number of stages is the same as the distribution of the size of an arriving group in the associated bulk arrival queue. Chaudry also derives the relationship between the length of the queue at a departure instant and at a random point in time. Jenson and Paulson (1978) derive closed form expressions for the queue length distribution for the $M^x/M/1$ queue, where the arriving batches have a multinomial distribution. Finally, Harris (1970) investigates bulk arrival queues with state dependent service rates, considering, among others, the case where service is negative exponential at a rate linear with the number of customers in the system.

A separate group of papers have studied the $M^x/G/1$ (service) queue by first looking at the arrival and service of groups, also referred to as supercustomers, where the service time is that required to serve the entire group. Gaver (1959), uses this approach, but only finds the transform of the waiting time of the first customer in each group. Cohen (1969) also applies this concept, correctly differentiating between the time until the first customer in a group begins service, and the time required for other customers in the same group to be served before a particular, random customer reaches the server. However, Cohen uses the incorrect distribution for the number of customers in front of a randomly chosen customer which belong to the same group. This error is corrected by Burke (1975), who applies the same concept of working with supercustomers to find the true waiting time distribution for an individual customer.

2.1.3 Bulk arrival, bulk service queues

Unlike the previous two areas, relatively little attention has been devoted to the explicit problem of queueing systems where both arrivals and services are in batches. Miller (1959) was the first to consider the problem, assuming a service process which is idle if the system is empty, i.e. the $M^x/G^c/1$ service queue. Miller also introduces the concept of accessible and inaccessible batches, whereby the former allows customers who arrive during the service of a batch to join if there is empty capacity, while in the latter customers may enter a batch only at the beginning of the service. An example of an accessible batch would be the traffice light queue where cars that arrive at an intersection during a green may proceed on through. Inaccessible batches, on the other hand, would include vehicle departures from a terminal.* Miller finds the queue length transform in steady state but could not solve for the wait time transform except in the special case where arrivals or departures

*All queueing systems considered here are assumed to be of the inaccessible type.

were in single units. Cohen (1969) studies bulk arrival, bulk service queues with random departure capacities (which is denoted here by $M^x/G^y/1$, the x and y indicating that arrivals and server capacity are of variable size), assuming that the server will remain idle if no one is present. In this case, an idle period terminates with the arrival of the first group. He also studies the $M^x/G^y/1$ dispatch queue where the server never becomes idle. However, his results are expressed in terms of contour integrals rather than directly as transforms, and therefore are not in a form which lends itself readily to numerical analysis. Bhat (1964) studies the $M^x/G^y/1$ service queue using results from fluctuation theory, but he also does not solve explicitly for the queue length transform.

Teghem et al. (1969) and Borthakur and Medhi (1974) study the $M^x/G^c/1$ queue operated under the general bulk service rule and find, respectively, the number in the system and the number in the queue. In both cases, the problem is studied using the theory of semi-Markov processes, and the Laplace transform (over time) of the transform of the number in the system is obtained. This result could in principle be used to find the steady state transform of the number in the system, although this is not done in the latter paper.

2.1.4 Numerical methods in bulk queues

Every transform result in problems with bulk service queues requires for solution the determination of the c-1 roots of a particular function.

This solution technique is used widely in queueing theory and to date is the only known method for finding certain transforms. Several authors, however, have questioned the feasibility of actually implementing the technique. The major criticism is that is some cases the determination of roots can be numerically hazardous, as is pointed out by Page (1965) and demonstrated by Dahlquist _et al_. (1974, p. 246). Neuts also makes the comment that, "From an aesthetic viewpoint it is unattractive, as it involves several steps without a clear probabilistic significance", (Neuts, 1979, p. 767). Furthermore, transform analysis does not in general yield readily to inversion and therefore often does not provide an explicit description of the actual distribution. These problems have given rise to several iterative numerical procedures which yield the desired probabilities without resorting to transforms and complex analysis.

The first and simplest of the procedures is one used by Hirasawa (1971) in the analysis of elevator systems which is referred to here as the method of numerical convolutions. The procedure, studied in much greater depth by Bagchi and Templeton (1972) who provide an abstract formalism for it, relies on the concept of the imbedded Markov process and uses a recursive formula to describe the behavior of the system from one regeneration point to the next. The mechanics of the alogrithm require little more than the repeated numerical convolution of two probability vectors, and can be used to study a system under time varying or steady state conditions, although the transient analysis uses the number of regeneration points as the time variable.* For example, it is

*This limitation of the result is not discussed by the authors.

necessary to know the distribution of arrivals between the n^{th} and $n+1^{st}$ service completions, even though we do not know when these completions occurred. On the whole, the method is extremely powerful and is discussed in greater detail in chapters 3 and 4.

A different procedure is presented by Wijngaard (1978) for finding the stationary distribution of queue length for state dependent bulk service queues where customers arrive singly and the service distribution is negative exponential. The algorithm uses the approximate triangularity of the transition matrix to find the mean recurrence times of states, and from this derives the steady state probabilities.

By far the most concerted effort at developing a numerical procedure for analyzing bulk arrival, bulk service is that due to Neuts. In a series of papers (1974, 1976, 1977a, 1977b) culminating in a final (at the moment) expository paper (1979), Neuts outlines an algorithm for calculating the stationary queue length distribution which involves the manipulation of vectors and matrices of dimension c (the capacity of the vehicle). Unfortunately, the method (the mechanics of which are outlined in appendix A) is extremely involved and is very sensitive to the value of c (he does not discuss its sensitivity to the utilization ratio ρ). Some computational results are reported in Neuts (1976) for c equal to 5 and 10 which demonstrates this sensitivity; CPU times on a CDC 6500 jumped from approximately .7 seconds for c = 5 to 4 seconds for c = 10. Thus while the algorithm does have some advantages over that of the numerical convolutions method (in particular, the dimension of the vectors and matrices in Neuts' procedure are insensitive to the level of congestion) it is also fairly complex and computationally appears to be quite slow.

2.1.5 Summary

It is useful at this point to refer back to the classification of queues given in table 1.1. To understand approximately where the state of the art now stands, this listing is shown again in table 2.1, with a cross-classification of five important queueing systems, differentiated in terms of whether they are bulk arrival, bulk service, or both and whether the cases of bulk service have a fixed capacity c or random capacity Y. In each cell, it is indicated whether the steady state queue length or waiting time transforms have been found. Of particular interest here is the fact that no results have been obtained thus far for bulk arrival, bulk service scheduled departure queues and no waiting time results are available for any bulk arrival, bulk service queue. Also, no papers were found considering any real time policies such as go-when-filled, although it should be pointed out that the $M/G^c/1$ go-when-filled queue is trivial to solve, and the extension to $M/G^y/1$ is probably quite simple. Also, no-one has considered the scheduled departure queue with cancellations, which is discussed briefly in chapter 1. With the exception of the general bulk service rule, all of the untouched cases are considered in chapter 3.

This concludes the review of the bulk queueing literature. The next section looks at some of the queueing network literature and discusses a possible methodology for approximately analyzing networks of queues in transportation.

Service control policy		Queueing Configuration				
Dispatch queues		$M/G^c/1$	$M/G^y/1$	$M^x/G/1$	$M^x/G^c/1$	$M^x/G^y/1$
	Scheduled departure	1,2	1,2			
	Real-time					
	Quasi-real time (general bulk service rule)	1,2			1	
Service queues		1,2	1,2	1,2	1	1

1: queue length transform
2: waiting time transform

Table 2.1

Summary of Contributions

2.2 The queueing network literature

In chapter 1, the desire to study stochastic delays in large transportation networks is cited as one of the primary reasons for turning to queueing theory for computationally efficient solution methods. The approach that is being taken to study these delays, however, focuses on the study of single, isolated queues. It is useful, then, to take a moment to review the current literature in queueing networks to see if such an approach can be extended to study delays in transportation networks.

The large majority of papers in the queueing network literature has centered around the pioneering paper by Jackson (1957), who showed that networks of M/M/k queues could be solved exactly by studying each queue in isolation. Since then, a number of papers have generalized Jackson's result by showing that the same solution technique applies to open and closed networks with feedback, state dependent arrival rates, and different classes of customers (see chandy (1972) and Gordon and Newell(1967)). Baskett *et al* (1975) consider networks were the only restriction on service time distributions is that they have rational Laplace transforms and where the service discipline is processor sharing or last come, first served, or where the number of servers meets or exceeds the waiting space. They explicitly exclude first come, first served systems except where the service time distribution is negative exponential. (For a review of these papers see Lemoine (1977, 1978) and Kelly (1979). All of these papers, by one means or another, show that the steady state solutions of the joint probability vector for the network as a whole can be expressed as a product of the state probability functions for each queue in isolation (see Chandy et al (1977) for a discussion of the different approaches used to prove product form solutions).

Despite the generalizations, no paper has been able to demonstrate a product form solution for a system with anything other than M/M/k servers if the service discipline is first come, first served. The simplicity of such systems arises from the fact that the output process of an M/M/k queue is also Poisson, as shown by Burke (1956). At the same time, the output of an M/G/k system, for any $k < \infty$, is renewal if and only if G = M, in which case the output is Poisson (Daley (1973)). Hence, the output process for an M/G/1 queue is not even renewal. The only exception to the M/G/k case is the M/G/∞ queue, which also features a Poisson output process. Since the output process of one queue may be the input process for another, a single M/G/1 queue in a network destroys the analytical tractibility of the system. In addition, since the arrival process to a queue may be the superposition of two or more departure processes, it is also true that the superposition of two or more processes is Poisson if and only if all the processes are also Poisson.

The net result of this is that networks of queues with non-exponential service times (and FCFS service disciplines) cannot be solved exactly, either in closed form or numerically. As might be expected, service times which can be approximated well by a negative exponential distribution are extremely rare in actual systems and almost non-existent in transportation applications. The question then shifts to the possibility of approximate solutions of more general problems. Toward this goal, several recent papers have appeared which attempt to model a system by approximate decomposition of the network into isolated queues or subsystems.

As mentioned above, such an approach is exact only when the system is completely Markovian. Kuehn (1979) uses this methodology to study networks of servers with general service time distributions, which consists of three principal steps. First, modeling each queue as a GI/G/1 system, he approximates the departure process as a general renewal process and fits an interdeparture time distribution using the first two moments of the actual interdeparture times. Next, since the arrival process to a queue may be the superposition of two or more departure processes, which would not be renewal, arrival processes are replaced with approximate renewal processes by matching the first and second moments of the interarrival time distributions. Finally, delays for each GI/G/1 queue are estimated using an approximate formula.

Following Kuehn's work, Whitt (1979a,b) provides a careful discussion on the replacement of general arrival processes with renewal ones. He also introduces a new approach which is discussed in chapter 5, and then compares his procedure with that by Kuehn in the context of a GI/M/1 queue. Without going into the advantages and disadvantages of each approach, the important point is that researchers have begun to study general queueing networks by first approximating the arrival process and then approximating the performance of the queue. In chapter 5, a method for estimating queue lengths for $G^x/G^y/1$ systems is proposed, introducing the possibility of studying transportation networks in the same manner being used for communication networks.

2.3 Summary

This chapter provides an outline of the state of the art in bulk queues and queueing networks. The subject of queueing networks is not discussed again until chapter 5, when the types of arrival processes likely to be encountered in transportation networks are described. Such applications are used to motivate the approximate analysis of bulk queues with general arrival processes.

Based on the review of the bulk queueing literature, chapters 3 and 4 focus almost exclusively on bulk arrival, bulk service, scheduled departure queues, with special emphasis on the option to cancel departures if the queue is too short. Chapter 3 presents the needed theoretical work, bulkding off the methodology described in the original papers by Bailey (1954) and Downton (1955). Since virtually all the results are expressed in terms of transforms, chapter 4 describes the techniques required to solve these transforms numerically, and discusses some of the difficulties that may be encountered.

Chapter 3 The $M^x/G^y/1$ Scheduled Departure Queue

In the preceding chapters, sources of delays in transportation terminals are formulated as queueing problems and the relevant bulk queueing literature is reviewed. In this chapter, we focus on the dispatch queue with compound Poisson arrivals and scheduled bulk departures, which is denoted here the $M^x/G^y/1$ scheduled departure queue. Standard transform methods are used for solution and all the results assume steady state conditions prevail.

To date, Cohen (1969) is the only author to have studied the $M^x/G^y/1$ scheduled departure queue, a system which he refers to as the transportation problem. His analysis, however, is extremely complex and he does not explicitly solve for the queue length transform. Peterson, however, has studied the $M/G^y/1$ scheduled departure queue and some of his results apply to the problem at hand. Several other authors have studied the $M^x/G^y/1$ service queue, where the server becomes idle if the system is empty, including Teghem et al (1969), Bhat (1964) and Borthakur and Medhi (1977). Not withstanding the difference in the type of problem solved (i.e. the presence of idle periods), the methodology used here and the presentation of the results is much simpler than that of other papers. In addition, practical problems associated with obtaining numerical results is touched on at several points in the discussion. We also introduce and analyze a new queueing system referred to as the scheduled departure queue with cancellations.

This chapter summarizes a number of new results which apply to scheduled departure, bulk arrival, bulk service queues. Of these three are felt to be of particular significance, namely:

1) formulas for the mean and variance of the length of the queue,
2) the queue length transform for the scheduled departure queue with cancellations,
3) a method for finding the moments of the wait time distribution with bulk arrivals.

The first result is a generalization of Bailey's moment formulas to allow for bulk arrivals. The second represents a new problem which has not been previously studied. The third extends the moment formulas obtained by Downton to allow for bulk arrivals, representing the first time delays have been solved for in any bulk arrival, bulk service queue.

In addition, a variety of other results have also been obtained, all of which apply to bulk arrival systems. These are:

4) the queue length transform when vehicle capacities are random,
5) a light traffic approximation for queues with cancellations,
6) the queue length transform for queues where the sequence of service times (departure headways) forms an alternating renewal process (this is used to describe the bus bunching problem),
7) the transform of the size of an arriving group for a queue imbedded in a network,

8) the relationship between the distribution of the number of units:

a) at a dispatch instant,

b) at a random point in time,

c) in front of an arriving unit,

d) behind a departing unit.

The presentation of these results is organized as follows. Section 3.1 briefly reviews the notation and conventions used throughout the chapter. Section 3.2 derives the queue length transform for the $M^x/G^y/1$ queue* through a straightforward extension of Bailey's original paper. This result is then extended still further to allow for cancellations when the length of the queue at a dispatch instant is less than some value m. Next, section 3.3 presents several variants of these problems along with other related results, described as topics 5,6 and 7 above. Section 3.4 then derives a method for finding the moments of the waiting time for $M^x/G^y/1$ queues and provides the formulas for the first two moments. Section 3.5 discusses topic 8, contributing several new insights regarding the length of the queue from different perspectives and at different points in time. Finally, section 3.6 synthesizes several known results for the $M^x/G/1$ <u>service</u> queue. The justification for this final block of material is its importance in describing the unloading queue.

* Unless specifically stated otherwise, all queues are of the scheduled departure type.

3.1 Notation

Perhaps the single biggest difficulty with reading theoretical presentations is becoming comfortable with the notation. This section is therefore intended as a quick introduction to the symbols and conventions that are used throughout the thesis, and should serve as a convenient reference for the reader who forgets what a symbol means and cannot find where it was defined. We do not attempt to go into any great detail describing each letter as each is introduced and defined as it is needed. Instead, all the symbols are listed in table 3.1 with brief definitions as a reference guide only, and not as an introduction that should be covered on first reading.

More important at this point is an outline of special conventions that are used. Isolated capital letters, such as X, will always denote random variables. The subscripted lower case versions of the same letter, such as x_i, are the associated probabilities. In other words, $x_i = \text{Prob}\{X = i\}$. Lower case letters in brackets $\{x\}$, refer to the entire probability vector, as in $\{x\} = x_o, x_1, \ldots$. Finally, capital letters expressed as functions always represent transforms of the associated random variables. For discrete random variables, we use the z transform defined as $X(z) = \sum_{i=0}^{\infty} x_i z^i$. Density functions for continuous random variables, such as B, are written using lower case functions, such as b(t), with cumulative distribution functions $B(t) = \int_o^t b(t)dt$. In this case, the transform is defined as the Laplace transform denoted by a capital letter, superscripted by an asterisk, with the argument s, e.g.

Table 3.1

Notation

Q	length of the queue immediately prior to a departure
Q^m	length of the queue prior to a dispatch instant when a minimum load constraint m is imposed
$\overset{n}{Q}$	number of units in front of a randomly arriving unit, including other units in the same arriving group
$\hat{Q}$	number of units behind a random departing unit, including others in the same departing batch
$\tilde{Q}$	total number of units in queue as seen by a random arriving unit at a dispatch instant
Q_t	length of the queue at a random point in time
R	number of units remaining following a dispatch instant
Y	number of units arriving during a service period (the time between successive departures)
$\hat{Y}$	number of units arriving during a service period when sampling is performed over the units
$\tilde{Y}$	number of units in front of a random arriving unit that arrived during the same service interval, including units arriving in the same group
F	number of units on a random incoming vehicle
N	number of vehicles arriving during a service period
G	size of a group arriving at a queue; number of units out of an incoming vehicle which will depart through a given outbound queue
$\tilde{G}$	number of units in front of a randomly chosen unit arriving in the same group
V	(random) capacity of an outbound vehicle

Table 3.1 (cont'd)

H	number of arrivals during the wait time of a randomly chosen unit
U	number of units already on an outbound vehicle before it arrived at a terminal
W	waiting time of a randomly chosen unit
c	(fixed) capacity of an outbound vehicle
m	minimum load constraint for an outbound vehicle

$B^*(s)$, defined as $B^*(s) = \int_0^\infty e^{-st} b(t)dt$. We generally use the term Laplace-Stieltjes (L.S.) transform, defined by $B^*(s) = \int_0^\infty e^{-st} \, dB(t)$, which is a slightly more general definition. In all cases, the letter z is used as the argument for transforms of discrete random variables while s is used for continuous random variables, where both z and s are defined over the complex plane.

Finally, since these results are not confined to freight or passenger systems, the more generic term "unit" is used to refer to flows through the system. This will help emphasize the discretized nature of the flows, regardless of whether it is composed of individual passengers, freight cars, or a continuum of shipment sizes broken into one ton increments.

3.2 Queue length transforms for the $M^X/G^Y/1$ queue

In this section we consider the scheduled departure queue with compound Poisson arrivals with a general service time distribution and constant vehicle capacities. In section 3.2.1, vehicles are assumed to depart irrespective of the length of the queue, implying that a departing vehicle may be partially or completely empty. Next, section 3.2.2 introduces a control policy where a departure may be cancelled if the length of the queue is less than some minimum.

3.2.1 The $M^x/G^c/1$ queue without cancellations

The queue length transform when arrivals are described by a compound Poisson process can be obtained as a direct extension of Bailey's original work. The derivation used here, however, is somewhat simpler and more clearly identifies how bulk arrivals are incorporated. We begin in the usual manner by setting up a recursive formula that describes the state of the system from one regeneration point to the next. Thus, let t_n denote the time of the departure of the n^{th} vehicle. Define Q_n to be the number of units waiting immediately prior to t_n and R_n to be the number left over immediately after t_n. The sequence of interdeparture times (service times) given by $T_n = t_n - t_{n-1}$ are assumed to be independently and identically distributed (i.i.d.) with density function $b(t)$ and L. S. transform $B^*(s)$. If Y_n is the number of units arriving in the interval (t_n, t_{n+1}) then we have the following relationship:

$$Q_{n+1} = R_n + Y_n \qquad 3.1$$

We may also define the dispatch operator $D\{\cdot\}$ which maps the units waiting at t_n^- to those remaining at t_n^+, implying that $R_n = D\{Q_n\}$ and hence:

$$Q_{n+1} = D\{Q_n\} + Y_n \qquad 3.2$$

Other authors generally use the convention $R_n = [\,Q_n - c]^+$, where $[x]^+ = \max(o, x)$. The use of the operator notation is made clear when scheduled departures with cancellations are considered. If the $\{Y_n\}$ form a set of i.i.d. random variables, then the sequence $\{Q_n\}$ constitutes a first order Markov chain. This condition is satisfied if a) arrivals occur according to a simple or compound Poisson arrival procees or b) if the number of groups arriving during the interval is deterministic and the size of each group forms a sequence of i.i.d. random variables. The latter case is especially important in transportation applications where schedule coordination between arriving and departing vehicles may produce the required conditions.

We may also introduce the interdepature times $T_n = t_n - t_{n-1}$, whereby the bivariate sequence $\{Q_n, T_n\}$ forms a semi-Markov process, enabling us to study the process between departures, an approach which has been taken by several other authors (e.g. Neuts (1966), and Borthakur and Medhi (1974)). It is shown in section 3.6 that all the needed information regarding the system (the queue as seen by a randomly chosen arriving unit and the wait time of that unit) can be derived from the simpler sequence $\{Q_n\}$.

Before proceeding to derive the transform, it should be pointed out that the steady state distribution of Q can be found directly from 3.2 using the method of numerical convolutions. We have by assumption that the Y_n are distributed according to some random variable Y. Define the probability vector $\{y\}$ by:

$$y_i = \text{Prob}\{Y = i\} \quad i = 0, 1, 2, \ldots$$

where it is assumed that the y_i can be computed directly. Similarly define $q_i^n = \text{prob}\{Q_n = i\}$, and $r_i^n = \text{prob}\{R_n = i\}$, where we now have from 3.2:

$$q_i^{n+1} = \sum_{j=o}^{i} r^n_j \; y^n_{i-j} \qquad 3.3$$

where:

$$r_o^{\;n} = \sum_{j=o}^{c} q_j^{\;n}$$

$$r_i^{\;n} = q_{i\;+c}^{\;n} \qquad\qquad i = 1, 2,\ldots$$

Starting with an initial solution $r_o^{\;o} = 1$, $r_i^{\;o} = 0$, $i = 1, 2, \ldots,$ we may use (3.3) recursively to solve for the steady state distribution of Q (assuming it exists). Because of its relative simplicity, this procedure is used as a basis for comparison when the evaluation of the transform approach is undertaken in chapter 4.

Returning to equation 3.1, we define the following transforms:

$$Q_n(z) = \sum_{i=o}^{\infty} q_i^{\;n} z^i \qquad 3.4$$

$$R_n(z) = \sum_{i=o}^{\infty} r_i^{\;n} z^i \qquad 3.5$$

$$Y(z) = \sum_{i=o}^{\infty} y_i z^i \qquad 3.6$$

Since Y_n and R_n are independent, we have, using the basic properties of transforms:

$$Q_{n+1}\,(z) = R_n(z)\;\; Y(z) \qquad 3.7$$

Expressing $R_n(z)$ in terms of $\{q_i^n\}$ gives:

$$R_n(z) = \sum_{i=o}^{c-1} q_i^{\ n} + \sum_{i=c}^{\infty} q_i^{\ n} \ z^{i-c}$$

$$= \sum_{i=o}^{c-1} q_i^{\ n} + z^{-c} \ [Q_n(z) - \sum_{i=o}^{c-1} q_i^{\ n} z^i]$$

$$= z^{-c} \left\{ \sum_{i=o}^{c-1} q_i^{\ n} (z^c - z^i) + Q_n(z) \right\} \qquad 3.8$$

Assume for the moment that the steady state queue length distribution exists and is unique. Then, taking the limit as n→ ∞ , and denoting $\lim Q_n(z) = Q(z)$, we may substitute 3.8 into 3.7 and solve for Q(z),giving:

$$Q(z) = \frac{\sum_{i=o}^{c-1} q_i (z^c - z^i)}{\frac{z^c}{Y(z)} - 1} \qquad 3.9$$

Equation 3.9 still has c unknowns, $q_o, \ldots, q_{c-1}$, on theright hand side. The classical approach to solving for these remaining quantities is through an application of Rouche's theorem which reads as follows (Churchill _et_ _al_., p. 300):

Theorem: Let f and g be functions which are analytic inside and on a closed contour C. If $|f(z)| > |g(z)|$ at every point z on C, then the functions f(z) and f(z) + g(z) have the same number of zeroes* (counting multiplicities) inside C.

Using this theorem, it is possible to show (see Appendix B) that the function $z^c - Y(z)$ has c zeroes on or within the unit circle. Since

* A point $\bar{z}$ is a zero of f(z) if $f(\bar{z}) = 0$.

Q(z) must be absolutely convergent in this region (i.e. it cannot contain any poles), the numerator must have the same set of zeroes. Denoting these zeroes by $z_o, z_1, \ldots, z_{c-1}$, we also note that $z = 1$ will always be one of the zeroes, and hence we adopt the convention that $z_o = 1$. This particular zero does not give us any information regarding the unknown probabilities, but we do have the additional fact that:

$$\lim_{z \to 1} Q(z) = 1 \tag{3.10}$$

We now consider the conditions required to guarantee the existence and uniqueness of Q(z). Let S denote the (possibly infinite) set of states making up the Markov process. Necessary and sufficient conditions for existence and uniqueness is that there exist a single communicating class of states $C \subseteq S$ in which the states are aperiodic and positive recurrent. A slightly stronger assumption is that the chain also be irreducible, whereby C=S, which implies that the process be ergodic. For problems in queueing theory, it is almost impossible to have more than one communicating class and hence the stationary distribution, if it exists, is always unique. For reasons outlined in appendix B, however, the problem is simplified somewhat if we assume the stronger condition of ergodicity, and thus we assume throughout the remainder of the research that the process is in fact ergodic. To guarantee existence, or, equivalently, to ensure that the states within C are positive recurrent, we must assume that $\rho = \overline{Y}/c < 1$, which is the usual condition for steady state analysis.

An alternative set of conditions for existence and uniqueness deals directly with the task of solving for the unknown probabilities. As described above, and outlined in greater detail in appendix E.2, these unknowns can be determined from a set of simultaneous linear equations. A necessary and sufficient condition for existence and uniqueness, then, is that these equations be consistent. Bailey shows that this is true if and only if all the zeroes are distinct, a condition that depends on Y(z) and must be satisfied on a case by case basis. In appendix B, a simple test is provided for determining whether all the zeroes are distinct and is then illustrated for several cases.

Returning to the problem of finding Q(z), it is possible to solve for Q(z) directly in terms of the zeroes. We just note that the polynomial in the numerator is of order c, enabling us to write:

$$\sum_{i=0}^{c-1} q_i(z^c - z^i) = \left(\sum_{i=0}^{c-1} q_i\right) \prod_{i=0}^{c-1} (z-z_i) \qquad 3.11$$

Now we have the single unknown $\sum_{i=0}^{c-1} q_i$ which may be evaluated using 3.10. Doing this, we have:

$$\lim_{z\to 1} Q(z) = 1 = \lim_{z\to 1} \frac{\left(\sum_{i=o}^{c-1} q_i\right)(z-1) \prod_{i=1}^{c-1} (z-z_i)}{\frac{z^c}{Y(z)} - 1}$$

Applying l'Hopital's rule yields:

$$1 = \frac{\left(\sum_{i=0}^{c-1} q_i\right) \prod_{i=1}^{c-1} (1-z_i)}{c - Y'(1)} \qquad 3.12$$

where we have adopted the shorthand notation

$$\frac{d}{dz} Y(z)\Big|_{z=1} = Y'(1) = \overline{Y}$$

Thus we have:

$$\sum_{1=0}^{c-1} q_i = (c-\overline{Y})(z-1) \prod_{i=1}^{c-1} \left(\frac{1}{1-z_i} \right) \tag{3.13}$$

where $\overline{Y} = Y'(1)$ = the expected number of arrivals during a service interval. Note that 3.13 gives us the probability that a randomly chosen vehicle will not be completely full. Substituting 3.13 and 3.11 into 3.9 gives:

$$Q(z) = \frac{(c-\overline{Y})(z-1) \prod_{i=1}^{c-1} \left(\frac{z-z_i}{1-z_i} \right)}{\frac{z^c}{Y(z)} - 1} \tag{3.14}$$

Equation 3.14 is the same as that obtained by Bailey, although the derivation is somewhat different. What is significant about the analysis here is the ease with which compound arrival processes are incorporated. Bailey, assuming a simple Poisson arrival process, noted that $Y(z) = B^*(\lambda - \lambda z)$, where λ is the arrival rate of units to the queue. To allow for a compound Poisson arrival process, let G_m be the size of the m^{th} arriving group with transform $G(z)$. It is well known, then, that the transform of the distribution of the total number of arrivals during a service interal is given by $Y(z) = B^*(\lambda - \lambda G(z))$, where λ is now the arrival rate of groups to the queue.

At this point it is possible to obtain a number of important results regarding the behavior of the queue at departure instants. Using 3.14, it is possible to find the mean and variance of the queue distribution using the following standard formulas:

$$E\,[Q] \quad = \quad Q'(1) \qquad 3.15$$

$$\mathrm{Var}\,[Q] \quad = Q''(1) + Q'(1) - Q'(1)^2 \qquad 3.16$$

Although very straight forward, the algebra required to use these relationships can be extremely tedious and is therefore left in appendix C. The results of this exercise are the following formulas, reported here for the first time:

$$E[Q] = \frac{\bar{\bar{Y}} + (c-\bar{Y}) - (c-\bar{Y})^2}{2(c-\bar{Y})} + \sum_{i=1}^{c-1} \frac{1}{1-z_i} \qquad 3.17$$

$$\mathrm{Var}[Q] = \frac{4\bar{\bar{\bar{Y}}}(c-\bar{Y}) + (1+6\bar{\bar{Y}})(c-\bar{Y})^2 + 3(\bar{\bar{Y}})^2 - (c-\bar{Y})^4}{12(c-\bar{Y})^2}$$

$$-\sum_{i=1}^{c-1} \frac{z_i}{1-z_i} \qquad 3.18$$

where $\bar{Y}$, $\bar{\bar{Y}}$ and $\bar{\bar{\bar{Y}}}$ are respectively the mean, variance and third moment about the mean of the number of arrivals during a service interval. For simple Poisson input, $\bar{Y} = \bar{\bar{Y}} = \bar{\bar{\bar{Y}}} = \rho c$. Substituting these into 3.17 and 3.18 yields the formulas that were originally derived by Bailey:

$$E[Q] = \frac{1-c(1-\rho)^2}{2(1-\rho)} + \sum_{i=1}^{c-1} \frac{1}{1-z_i} \qquad 3.19$$

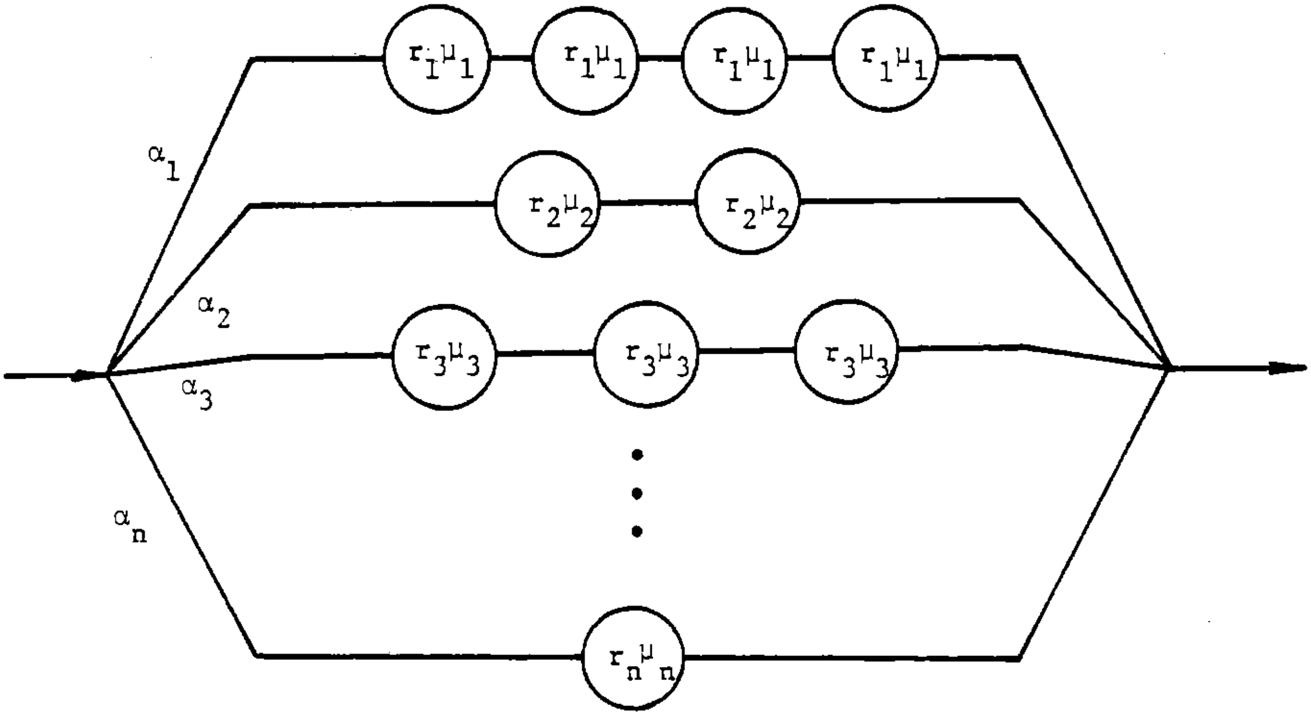


Illustration of hyperstage distribution

Figure 3.1

$$\mathrm{Var}[Q] = \frac{1 + 2\rho + 6\rho c(1-\rho)^2 - c^2(1-\rho)^4}{12(1-\rho)^2} - \sum_{i=1}^{c-1} \frac{z_i}{1-z_i} \qquad 3.20$$

The moment formulas can be of some use by themselves, or, as is shown in section 4.3, can be used to fit approximate distributions if other statistics are required. It is also possible to develop a set of equations that allow numerical inversions to be performed, enabling us to explicitly calculate the probabilities q_o, q_1, ... This topic is pursued in some depth in section 4.2.

Before moving on to the problem of scheduled departure queues with cancellations, there is one special case of the service time distribution which deserves mention because of its simple analytical properties. The distribution, termed here the hyperstage distribution, is actually a very general class of distributions whose principle feature is that it has rational Laplace transforms. It is best described by Kleinrock (1975 p. 144) as the distribution of time required by a customer to pass through the network depicted in figure 3.1. The customer has a choice of i=1, ..., N branches, each of which may be chosen with probability α_i. To move over branch i, the customer must pass through r_i stages, each with a processing time given by a negative exponential distribution with parameter $r_i\mu_i$. The distribution includes, as special cases, the Erlang and hyperexponential distributions.

The importance of this distribution is that it can be used to approximate virtually any distribution encountered in practice while possessing very simple analytical properties. A number of authors have used it in the bulk queueing literature, including Bailey (1954), Downton (1955), Gupta (1964), Gupta and Goyal (1965), Foster (1961, 1965), Jaiswal (1960a, b) and Luchak (1958). In most cases, the distribution was used to formulate a queue as a completely Markovian system using the well known method of phases. Our interest, however, is in bringing out its implications with regard to solving the transform in equation 3.14 as was originally noted by Bailey, whose presentation we now follow.

Assume we are describing the $M/E_r^c/1$ queue, where E_r denotes the r stage Erlang distribution. In this case, Y(z) is given by:

$$Y(z) = \left(\frac{r\mu}{r\mu + \lambda - \lambda z}\right)^r \qquad 3.21$$

The denominator of 3.14 now becomes:

$$\left(\frac{1}{r\mu}\right)^r \left(z^c(r\mu + \lambda - \lambda z)^r - (r\mu)^r\right) \qquad 3.22$$

Equation 3.22 has c + r zeroes, of which c are on or within the unit circle and r are outside. Denoting these zeroes $z_o, \ldots, z_{c-1}, \ldots, z_{c+r-1}$, we may write 3.14 as follows:

$$Q(z) = \frac{K \prod_{i=0}^{c-1} (z-z_i)}{\prod_{i=0}^{c+r-1} (z-z_i)} \qquad 3.23$$

where K is a constant of proportionality. Cancelling common zeroes and evaluating K yields:

$$Q(z) = \prod_{i=c}^{c+r-1} \frac{1-z_i}{z-z_i} \qquad 3.24$$

Q(z) may now be inverted by expanding the right hand side by partial fractions. From a computational perspective, there are several features of this problem, discussed in section 4.1.1, which suggest that the necessary zeroes may be difficult to find. Also, for bulk arrival systems where the maximum size of an incoming group is $\hat{c}$, the number of zeroes located outside the unit circle that must be found jumps to $\hat{c} \cdot r$. If the service distribution is a hyperstage type with N branches and r_i stages in branch i, the number of zeroes outside the unit circle increases still further to $\hat{c} \sum_{i=1}^{N} r_i$. This may be considerably more difficult than finding the c zeroes within the unit circle. On the other hand, the number of zeroes within the unit circle remains at c regardless of the functional form of Y(z).

This section introduces the basic methodology for solving for queue length transforms for bulk arrival, bulk service queues. The most important results are the transforms in equations 3.9 and 3.14, and the moment formulas in equations 3.17 and 3.18. In the sections which follow, these results are extended to allow for cancellations and vehicles with random capacities.

3.2.2 The $M^X/G^c/1$ queue with cancellations

The scheduled departure queue with cancellations is defined as follows. Let t_n be the time of the n^{th} dispatch instant, representing the points in time where departures may occur. If at t_n the length of the queue is at least m, then the vehicle will depart carrying up to c units. If the length of the queue is strictly less than m, then the run is cancelled and the time until the next dispatch instant is drawn from the service time distribution. The service time is defined as the time between dispatch instants and is independent of whether a departure actually occurred or not.

Examples of scheduled departure queues arise frequently in freight applications where a vehicle might depart from a terminal once a day. Ifthere is too little traffic to economically justify sending the vehicle, the departure can be cancelled and the traffic held over to the next day. It is important to understand the difference between this kind of control strategy, which has not been dealt with before in the open literature, and the general bulk service rule proposed by Neuts (1967) which has been studied by a number of authors (Tegham et al (1969), Borthakur and Medhi (1974), Borthakur (1971), and Medhi (1975, 1979). In both cases, there is a period immediately following a departure during which no departures may occur and arrivals must queue up. Where the two models differ is when the length of the queue is less than m. Under the general bulk service rule, the vehicle is simply held until the length of the queue reaches m, at which point the vehicle departs, whereas in our case the departure is cancelled until the next dispatch instant. The length of the next service time is independent of whether the vehicle has been held or not.

The system is easiest to describe at dispatch instants. Let Q_n^m be the number of units waiting prior to the n^{th} departure, where the superscript m is used to denote the presence of a minimum load constraint. Proceeding in much the same way as we did for the case without cancellations, we observe:

$$Q_n^m = D_m \{Q_n\} + Y_n \qquad 3.25$$

where the regeneration points now occur at dispatch instants. The dispatch operator $D_m \{\cdot\}$ is defined as:

$$D_m \{x\} = \begin{cases} x & x < m \\ 0 & m \leq x < c \\ x-c & x \geq c \end{cases} \qquad 3.26$$

Proceeding as before we find the transform of R_n as follows:

$$R_n(z) = \sum_{i=0}^{m-1} q_i z^i + \sum_{i=m}^{c-1} q_i + \sum_{i=c}^{\infty} q_i \; z^{i-c}$$

$$= \sum_{i=o}^{m-1} q_i z^i + \sum_{i=m}^{c-1} q_i + z^{-c} \; [Q(z) - \sum_{i=0}^{c-1} q_i z^i]$$

$$= z^{-c} \left[(z^c-1) \sum_{i=0}^{m-1} q_i z^i + \sum_{i=m}^{c-1} q_i (z^c - z^i) + Q(z) \right] \qquad 3.27$$

Taking the limit as n → ∞ gives:

$$Q^m(z) = R(z) \cdot Y(z)$$

$$= \frac{(z^c-1) \sum\limits_{i=0}^{m-1} q_i z^i + \sum\limits_{i=m}^{c-1} q_i (z^c - z^i)}{\frac{z^c}{Y(z)} - 1} \qquad 3.28$$

Note that the denominator in (3.28) is the same as that in 3.9, enabling us to make the same observation that there must be c zeroes on and within the unit circle. The numerator, on the other hand, is a polynomial of order c + m -1, containing not only the same c zeroes as in the denominator but m - 1 additional ones as well. Unfortunately, we cannot locate these new zeroes in the same manner as the previous ones using the known function in the denominator. Two alternative approaches can be used to resolve this problem. The first involves using the c zeroes in the denominator to solve directly for the unknown probabilities. In other words, denoting as before the zeroes within the unit circle by z_o, z_1, ..., z_{c-1}, we have the following system of equations* (obtained by substituting the zeroes into the numerator of 3.28):

$$(z_j^c-1)\sum_{i=0}^{m-1} q_i z_j^i + \sum_{i=m}^{c-1} q_i (z_j^c - z_j^i) = 0 \qquad j = 0,1,\ldots,c-1 \qquad 3.29$$

In this case, the transform is as shown in 3.28.

The development of the second method is motivated by the observation that it seems unnecessary to have to solve for c unknowns when in fact only m-1 unknowns are added by the minimum load constraint (in most cases, m < c/2). To begin, let z_c, . . ., z_{c+m-1} denote the new zeroes of the numerator, enabling us to express $Q^m(z)$ as follows:

*Setting up these equations is actually somewhat more involved; see appendix E.2.

$$Q^m(z) = \frac{K \cdot \prod_{i=0}^{c+m-1} (z-z_i)}{\frac{z^c}{Y(z)} - 1}$$

Evaluating the constant yields:

$$Q^m(z) = \frac{(c-\overline{Y})\ (z-1) \prod_{i=1}^{c+m-1} \left(\frac{z-z_i}{1-z_i} \right)}{\frac{z^c}{Y(z)} - 1} \qquad 3.30$$

Now define the polynomials $A^c(z)$ and $B^m(z)$ as follows* (the superscripts c and m are dropped when writing out the coefficients):

$$A^c(z) = \sum_{i=0}^{c} a_i z^i$$

$$= (c-\overline{Y})(z-1) \prod_{i=1}^{c-1} \frac{z-z_i}{1-z_i} \qquad 3.31$$

$$B^m(z) = \sum_{i=0}^{m-1} b_i z^i$$

$$= \prod_{i=c}^{c+m-1} \frac{z-z_i}{1-z_i} \qquad 3.32$$

Our approach is, rather than to solve for the remaining zeroes, to solve directly for the polynomial B(z) in terms of its unknown coefficients using the known coefficients $a_0, a_1, \ldots, a_c$. To do this, define $\Psi(z) = A^c(z) \cdot B^m(z)$, where:

*$B^m(z)$ should not be confused with the L.S. transform $B^*(s)$ of the service time distribution.

$$\Psi(z) = \sum_{i=0}^{c+m-1} \psi_i z^i = (z^c-1) \sum_{i=0}^{m-1} q_i z^i + \sum_{i=m}^{c-1} q_i (z^c - z^i) \qquad 3.33$$

Rewriting the right hand side of 3.33 in terms of increasing powers of z gives:

$$\Psi(z) = -\sum_{i=0}^{m-1} q_i z^i - \sum_{i=m}^{c-1} q_i z^i + (q_o + \sum_{i=m}^{c-1} q_i) z^c + \sum_{i=1}^{m-1} q_i z^{i+c} \qquad 3.34$$

Matching the coefficients of like powers of z gives the following expressions:

$$\psi_k + \psi_{k+c} = 0 \qquad k = 1, 2, \ldots, m-1 \qquad 3.35$$

Expressing the coefficients ψ_k in terms of the polynomials $A^c(z)$ and $B^m(z)$ gives:

$$\sum_{i=0}^{k-1} a_{k-i} b_i + (a_o + a_c) b_k + \sum_{i=k+1}^{m-1} a_{c+k-i} b_i = 0 \qquad k = 1, \ldots, m-1 \quad 3.36$$

The polynomial $B^m(z)$, however, has m unknowns while 3.36 provides only m-1 equations. The final equation is found by noting from 3.32 that:

$$B^m(z)\Big|_{z=1} = \sum_{i=0}^{m-1} b_i = 1 \qquad 3.37$$

The system of equations given by 3.36 and 3.37 requires that we solve for only m unknowns, a problem that requires $m^3/3$ additions and multiplications. To this we must add the $c^2/4$ operations required to solve for the polynomial $A^c(z)$. If we assume $m = c/2$ (typically m will be less than half the capacity of the vehicle) then the total is $\frac{c^2}{4} \left(\frac{c}{6} + 1 \right)$ additions and multiplications, or approximately an order of magnitude faster (for large c) than if we solved the full c × c system of equations. Once we have both polynomials $A_c(z)$ and $B^m(z)$, then the probabilities $q_0, \ldots, q_{c-1}$ are given by:

$$q_i = - \sum_{j=0}^{i} a_{i-j} \, b_j \qquad i = 0, \ldots, m-1 \qquad 3.38$$

$$= - \sum_{j=0}^{m-1} a_{i-j} \, b_j \qquad i = m, \ldots, c-1 \qquad 3.39$$

At this point, it is useful to look more closely at $Q^m(z)$. Letting $Q(z) = Q^0(z)$ as given in (3.14), then, since $A^c(z)$ is unaffected by the minimum load m, we may write:

$$Q^m(z) = Q(z) \cdot B^m(z) \qquad 3.40$$

Using 3.40, it is easy to compute the moments of Q^m using equations 3.17 and 3.18 and the coefficients of the polynomial $B^m(z)$.

We have now completed the basic derivations for the scheduled departure queue with and without cancellations. The following section considers the special case m = c corresponding to the go-when-filled, scheduled departure queue. This result is then used to solve the go-when-filled real time queue, where departures occur as soon as the queue will fill the vehicle. (Remember that for scheduled queues, departures may occur only at predetermined dispatch instants).

3.2.3 Scheduled and real-time queues under a go-when-filled policy

An interesting special case of the scheduled departure queue with cancellations is the go-when-filled (GWF) policy. As one might expect, the transition from scheduled departures to real time dispatching under the GWF policy is a minor one, and hence the latter case is covered here as well. For the moment, however, assume that we have a scheduled departure queue with m=c (i.e., the vehicle leaves only if it is full). Equation 3.28 now becomes:

$$Q^c(z) = \frac{(z^c - 1) \sum_{i=0}^{c-1} q_i z^i}{\frac{z^c}{Y(z)} - 1} \qquad 3.41$$

As before, the numerator must have the same zeroes on and within the unit circle as the denominator, where these zeroes are unaffected by the minimum load constraint. In the numerator, however, we see that the additional c-1 zeroes introduced by the minimum load are evenly distributed around the unit circle, and can be solved for by inspection. In fact, the polynomial $B^m(z) = B^c(z)$ for m=c can now be written:

$$B^c(z) = \frac{1}{c} \; \frac{z^c - 1}{z-1} = \frac{1}{c} \sum_{i=0}^{c-1} z^i \qquad 3.42$$

An interesting feature of $B^c(z)$ is that it is independent of the arrival process, implying that the distribution of units left over is independent of the utilization ratio. Put differently, the increase in the length of the queue produced by switching from m=0 to m=c is independent of ρ. By inspection, we can see that the distribution of units left over is exactly the (discrete) uniform distribution between 0 and c-1, and hence has mean $\frac{c-1}{2}$ and variance $\frac{c^2-1}{12}$. Thus we may find the moments of Q^c as follows:

$$E(Q^c) = E(Q) + \frac{c-1}{2} \qquad 3.43$$

$$Var(Q^c) = Var\ (Q) + \frac{c^2-1}{12}$$

The go-when-filled, scheduled departure queue can be used to approximate a real-time GWF policy in <u>continuous time</u>, as opposed to simply at regeneration points. Assume the time between dispatch instants is some fixed interval Δt**. Remembering that N(z) is the transform of the number of groups to arrive between dispatch instants, we have, for a Poisson process:

$$N(z) = n_0 + n_1 z \quad + 0(\Delta t^2) \qquad 3.44$$

where $n_1 \cong \lambda\Delta t$ and $n_o \cong 1 - \lambda\Delta t$.

**The size of Δt may be determined by practical considerations. For example, the dispatcher may only check on the status of a queue once an hour.

Assume that Δt is sufficiently small. Then

$$Y(z) \cong 1 - n_1 + n_1 G(z) \qquad 3.45$$

As $\Delta t \to 0$, $n_1 \to 0$ and $Y(z) \to 1$, leaving:

$$Q^c(z) = B^c(z) \qquad 3.46$$

Thus, in the limit, the steady state queue length in continuous time is simply the discrete uniform distribution between 0 and c-1. Although the queue will of course exceed c at certain points in time (as a result of bulk arrivals), once this occurs the vehicle is dispatched within Δt. Thus as $\Delta t \to 0$, the probability of ever seeing the queue longer than c-1 vanishes.

The same queue can also be modelled in real time by defining each arrival instant as a dispatch instant, and then finding the queue length at each dispatch instant. In this case we have by definition one arriving group between dispatch instants, or $Y(z) = G(z)$, and all our previous results apply. This is true irregardless of the arrival process as long as the size of each group forms a sequence of i.i.d. random variables. The only exception occurs when the size of an incoming group may be larger than c. In this case, we must allow for the possibility of two or more departures taking place at the same instant. For all practical purposes, however, this possibility may be ignored.

3.2.4 Vehicles with random capacities

As is pointed out earlier, the bulk arrival queue with random service capacities, denoted the $M^x/G^y/1$ queue, has been studied by several authors. Miller (1959) and Bhat (1964) both study the $M^x/G^y/1$ service queue, where the server becomes idle if and only if the system is empty. Miller uses the conventional imbedded Markov chain approach but does not fully simplify his results which is also left in terms of the unknown probabilities $q_o,\ldots,q_{c-1}$. Bhat presents a very complex analysis using results from fluctuation theory, but does not explicitly solve for the queue length transform. Cohen (1969) studied the problem addressed here (i.e. the scheduled departure $M^x/G^y/1$ queue) but he also presents a very complex analysis without explicitly solving for the queue length transform. Peterson (1971) considers the same problem assuming simple Poisson arrivals; the extension to compound Poisson arrivals, however, is straightforward. The methodology used here, however, is substantially different and represents a new approach to deriving queue length transforms. Following the derivation, we discuss several interesting applications of the concept of random capacities.

The derivation presented here uses somewhat different arguments than those used previously for the case with fixed vehicle capacities. The approach taken brings out some of the underlying relationships more clearly and represents an interesting alternative to Bailey's derivation. We begin by assuming the capacity of the vehicle is given by a random variable V where, following our usual convention, v_i=Prob[V=i]. We also define a new random variable W where:

$$W = Q - V = W^{+} + W^{-} \qquad 3.47$$

where $W^{+} = [Q\text{-}V]^{+}$ and $W^{-} = [Q\text{-}V]^{-}$. Taking transforms of both sides of 3.47 gives:

$$W(z) = Q(z)\ V(\tfrac{1}{z}) = W^{+}(z) + W^{-}(z) - 1 \qquad 3.48$$

where:

$$W(z) = \sum_{i=-c}^{\infty} w_i z^i$$

$$W^{+}(z) = \sum_{i=-c}^{0} w_i + \sum_{i=1}^{\infty} w_i z^i$$

$$W^{-}(z) = \sum_{i=-c}^{0} w_i z^i + \sum_{i=1}^{\infty} w_i$$

Clearly $W^{+} = R$, the number of units left over following the departure of a vehicle. W^{-}, on the other hand, is minus the amount of empty space on a departing vehicle. Using the fact that $W^{+}(z) = R(z) = Q(z)Y(z)$ and solving for $Q(z)$ gives:

$$Q(z) = \frac{1 - W^{-}(z)}{\frac{1}{Y(z)} - V(\frac{1}{z})}$$

$$= \frac{z^c - z^c W^{-}(z)}{\frac{z^c}{Y(z)} - z^c V(\frac{1}{z})} \qquad 3.49$$

The reason for multiplying and dividing by z^c is that now we may apply the same arguments used before to find the remaining unknowns, which in this case are $w_{-c},\ldots,w_{-1}$. Noting that the numerator is a polynomial of order c and that the denominator must have c zeroes inside and on the unit circle (if $\rho = \overline{Y}/\overline{V} < 1$), we may express Q(z) as:

$$Q(z) = \frac{(\overline{V} - \overline{Y})\ (z-1) \prod_{i=1}^{c-1} \frac{z-z_i}{1-z_i}}{\frac{z^c}{Y(z)} - z^c V(\frac{1}{z})} \qquad 3.50$$

Unlike the previous case with fixed capacity, we do not recover the probabilities $q_o, \ldots, q_{c-1}$, but rather the probabilities w_{-i} = the probability of i units of empty space on an outbound vehicle. However, we also have that:

$$w_{-i} = \sum_{\ell=0}^{c-i} q_\ell v_{\ell+i} \qquad 3.51$$

which forms a system of linear equations which uniquely determine $q_o, \ldots, q_{c-1}$. Furthermore, the equations are triangular, and hence may be solved directly using the recursion:

$$q_i = \frac{1}{v_c} [w_{-i} - \sum_{\ell=0}^{i-1} q_\ell v_{\ell-i+c}] \qquad 3.52$$

As a brief aside, it is interesting to note that the recursion relating Q_n and Q_{n+1} can be replaced by one in terms of R_n and R_{n+1}:

$$R_{n+1} = [R_n + Y_n - V_n]^+ \ . \qquad 3.53$$

Equation 3.53 is a standard relationship in queueing theory and applies to other problems such as describing the unfinished work in a G/G/1 queue. This observation allows us to use the bounds and approximations developed for other applications in the context of bulk queues.

Turning now to the case with cancellations, we again begin with 3.47 as a starting point. This time, however, we no longer have that

$R = W^+$. Instead, we note that we may write:*

$$R = [Q-V]^+ + [Q-m]^- + m\ u(m-Q) \tag{3.54}$$

where:

$$u(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

To take transforms of 3.54, temporarily define the random variable A as follows:

$$A = [Q-m]^- + m\ u(m-Q) \tag{3.55}$$

Taking transforms of 3.55 gives:

$$\begin{aligned} A(z) &= \sum_{i=0}^{m-1} a_i z^i \\ &= \sum_{i=m}^{\infty} q_i + \sum_{i=0}^{m-1} q_i z^i \end{aligned} \tag{3.56}$$

Again defining $W^+ = [Q-V]^+$, we now wish to find R(z) in terms of $W^+(z)$ and A(z). Letting r_k = Prob[R=k], we have that:

$$r_k = \quad \text{Prob } [W^+ = \ell, A = k - \ell] \tag{3.57}$$

Observing that A=0 if $W^+ > 0$ and $W^+=0$ if $A > 0$, we find, for $k > 0$:

$$\begin{aligned} r_k &= \text{Prob}[W^+=k, A=0] + \text{Prob}[W^+=0,\ A=k] && 3.58 \\ &= \text{Prob}[W^+=k] + \text{Prob}[A=k] \\ &= w_k^+ + a_k\ . && 3.59 \end{aligned}$$

Of course, $a_k=0$ if $k > m$. To find r_o, we use the relation:

$$\begin{aligned} r_o &= \text{Prob}[W^+=0, A=0] \\ &= \text{Prob}[W^+=0 | A=0]\ \text{Prob}[A=0] && 3.60 \end{aligned}$$

* We implicitly assume here that V is greater than m.

To find the conditional probability on the right hand side of 3.60, we have:

$$\text{Prob}[W^{+}=0] = w_o^{+} = \sum_{i=m}^{c} v_i \sum_{j=0}^{i} q_i \tag{3.61}$$

Conditioning on the event A=0 gives:

$$\text{Prob}[W^{+}=0 \ A=0] = \frac{\sum\limits_{i=m}^{c} v_i \sum\limits_{j=0}^{i} q_j - \sum\limits_{j=1}^{m-1} q_j}{q_o + \sum\limits_{j=m}^{\infty} q_j}$$

$$= \frac{w_o^{+} - (1 - a_o)}{a_o} \tag{3.62}$$

Thus:

$$r_o = w_o^{+} + a_o - 1 \tag{3.63}$$

We can now find R(z) to be:

$$R(z) = W^{+}(z) + A(z) - 1 \tag{3.64}$$

Combining 3.48 and 3.64 gives:

$$Q(z)\ V(\tfrac{1}{z}) = R(z) + 1 - A(z) + W^{-}(z) - 1$$

$$= R(z) + W^{-}(z) - A(z) \tag{3.65}$$

Using R(z) = Q(z)/Y(z) and solving for Q(z) yields:

$$Q(z) = \frac{A(z) - W^{-}(z)}{\frac{1}{Y(z)} - V(\frac{1}{z})}$$

$$= \frac{z^{c}[A(z) - W^{-}(z)]}{\frac{z^{c}}{Y(z)} - z^{c}V(\frac{1}{z})} \tag{3.66}$$

As occurred when the vehicle capacity is deterministic, the minimum load constraint does not affect the denominator. At the same time, the numerator is now a polynomial of order c+m-1. The unknowns $w_{-c},\ldots,w_{-1},a_o-w_o,a_1,\ldots,a_{m-1}$ can be determined in the same manner as was used for the case of fixed capacity.

The following examples demonstrate how randomness in capacity can arise (the first example is reported in Peterson (1971)).

Example 3.1: The multi-stop problem

Often in scheduled networks a particular vehicle will make several stops along a route, picking up and dropping off passengers along the way. Units on board a vehicle coming into a terminal which are not getting off then have the effect of reducing the capacity of the vehicle for those trying to get on. Let U be the random variable describing the number of units already on a vehicle which are not getting off at a given terminal with transform U(z). It is easily verified that:

$$V(z) = z^c\, U\left(\frac{1}{z}\right) \qquad 3.67$$

Or equivalently:

$$U(z) = z^c\, V\left(\frac{1}{z}\right) \qquad 3.68$$

Substituting 3.68 into 3.50 gives the expression:

$$Q(z) = \frac{(c-\overline{U}-\overline{Y})(z-1)\prod_{i=1}^{c-1}\left(\frac{z-z_i}{1-z_i}\right)}{\frac{z^c}{Y(z)} - U(z)} \qquad 3.69$$

where we have substituted $\overline{V} = c - \overline{U}$.

Example 3.2: Priority queues

In many cases in freight applications a carrier will offer two levels of service which separates standard and high priority traffic, where the latter category is loaded first on any outbound vehicle with the remaining capacity then allotted to the lower priority traffic. By virtually guaranteeing the high priority traffic space on the first outbound vehicle, the carrier can sell the service at a higher rate, and offer the other traffic (typically low valued commodities) a reduced rate with some reduction in service. Of particular interest to the carrier is being able to estimate a) the probability the high priority traffic will exceed the capacity of the vehicle, requiring the operator to make a costly additional run or the embarrassment of a late delivery and b) the level of service differential which would then assist in the setting of prices (of course this is part of a larger equilibration problem).

The priority queueing problem is easily conceptualized as a multistop vehicle routing problem. Assume there are $p = 1, \ldots, P$ priorities with $p = 1$ being the highest. We can now think of P separate queues, where the vehicle stops first at $p = 1$, then $p = 2$ and so on. The capacity available to a given priority level is simply whatever is left over after all the higher priority traffic has been loaded on. The problem therefore has already been solved in the previous example.

Example 3.3: Weight versus volume

A problem frequently encountered in LTL trucking is that a trailer may be full yet still be carrying well below its maximum weight limit if the freight is of sufficiently low density. For this reason, a linehaul trailer capable of carrying up to 45,000 pounds will often fill up with less than 35,000 pounds. One approach to modeling this would be to specify two capacities, and attach to each unit of freight another variable describing its volume. The queue would then be described by considering both weight and volume, each with its own upper limit. A simpler and more useful approach would be to estimate a distribution describing the number of tons a full truck might carry. The transform of this distribution could then be used directly in equation 3.49.

The weight and volume constraint, of course, is not unique to trucking. A similar problem arises in rail freight where a train is usually pulling a mixture of fulls and empties. Again, the maximum number of cars that can be pulled is determined by their combined weight, where we may estimate directly the distribution of the maximum number of cars before the weight constraint is met. Trains also have the added dimension of locomotive availability, where the possibility of insufficient locomotives contributes to the randomness of the total capacity of the train. This problem can be modeled by first estimating the distribution of the maximum number of cars F a single locomotive can pull, with transform $F(z)$. Now let $G(z)$ be the transform of the distribution of the number of locomotives that will be available. The transform of the total capacity of the train is then $V(z) = G(F(z))$.

Example 3.4 Random shipment size

Up to now, we have discretized shipments into specific weight units, allowing the capacity of the vehicle to be represented by a maximum allowable number of units. Alternatively, we could let each shipment be an individual customer and then use the concept of random vehicle capacities to reflect the random size of each shipment.

This concluces the most important results regarding queue length transforms. The next section reports on several related results, followed by section 3.4 which addresses the problem of finding waiting time transforms.

3.3 Extensions of the scheduled departure queue

Having solved for the queue length transforms for the $M^x/G^y/1$ queue with and without cancellations, it is now possible to consider three related results. The first is a light traffic approximation for queues with cancellations. Second, the queue length transform is found when the sequence of departure headways forms an alternating renewal process. Third and last is the transform of the size of a group arriving to the queue, where the group has just come off a vehicle arriving from an upstream queue. Such a problem would arise when studying a queue imbedded in a network. All the results here are new.

3.3.1 A light traffic approximation for scheduled queues with cancellations

The scheduled queue with cancellations as presented in section 3.2.2 can be applied in situations with any level of demand. In some instances, however, it may be possible to simplify the calculations by introducing approximations. For example, in heavy traffic (ρ close to 1) the probability of a cancellation will be negligible, enabling us to ignore the minimum load constraint. Of greater interest is the light traffic case where the minimum load is more likely to have an effect. Since the capacity of the vehicle is less likely to be constraining, we can consider an approximation which ignores the constraint altogether, assuming, in effect, an infinite capacity vehicle. Proceeding as before to find the distribution of units left over, we now have:

$$\begin{aligned} R(z) &= \sum_{i=0}^{m-1} q_i z^i + \sum_{i=m}^{\infty} q_i \\ &= \sum_{i=0}^{m-1} q_i\,(z^i - 1) + 1 \qquad 3.70 \end{aligned}$$

Using $Q(z) = R(z) \cdot Y(z)$ gives

$$Q(z) = \left[\sum_{i=0}^{m-1} q_i (z^i - 1) + 1 \right] \cdot Y(z) \qquad 3.71$$

Unlike the earlier problems, we no longer need to solve for any zeroes. Instead, 3.71 can be used to set up an m x m system of equations for the remaining unknowns. First, we must write the right-hand side of 3.71 in terms of simple powers of z which, after some manipulation, yields:

$$Q(z) = \sum_{i=0}^{m-1} \left[(1 - \sum_{\ell=0}^{m-1} q_\ell) \; y_i + \sum_{j=0}^{i} q_j \; y_{i-j} \right] z^i$$

$$+ \sum_{i=m}^{\infty} \left[(1 - \sum_{\ell=0}^{m-1} q_\ell) \; y_i + \sum_{j=0}^{m-1} q_j y_{i-j} \right] z^i \qquad 3.72$$

Matching the coefficients of the first m-1 powers of z on both sides gives the following system of equations:

$$\sum_{\ell=0}^{i-1} (y_i - y_{i-\ell}) \; q_\ell + (1 - y_o) \; q_i + \sum_{\ell=i+1}^{m-1} y_i \; q_\ell = y_i \qquad i=0, \ldots, m-1 \qquad 3.73$$

In writing 3.73 we use the convention that $\sum_{i=m}^{n}(\cdot) = 0$ if $m > n$. Once $q_o, \ldots, q_{m-1}$ have been found, the rest of the probability vector is easily computed by matching the coefficients of like powers of z on both sides of 3.72. This approach to computing the probability vector $\{q\}$ is outlined in further detail in section 4.2.

3.3.2 Scheduled departures with correlated headways

Let t_n be the time of departure of the nth vehicle and define $\tau_n = t_{n+1} - t_n$ to be the headway separating the nth and n+1st vehicles. Thus far, we have assumed the departure times of vehicles to form a renewal process in which case the sequence $\{\tau_n\}$ forms a set of i.i.d. random variables. Now assume that each vehicle has a scheduled departure time T_n with an actual departure time given by $t_n = T_n + \varepsilon_n$, where

ε_n is an arbitrary random variable where $E(\varepsilon_n) = 0$, $Var(\varepsilon_n) = E(\varepsilon_n^2) = \sigma^2$ and Cov $(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$. Now we have $\tau_n = T_{n+1} - T_n + \varepsilon_{m+1} - \varepsilon_m$, Cov $(\tau_n, \tau_{n+1}) = -\sigma^2$, and a correlation coefficient given by Cov $(\tau_n, \tau_{n+1})/$ $Var(\tau_n) = -.5$. In this case successive headways are negatively correlated, implying that the number of arrivals Y_n between successive departures is correlated as well, violating one of the basic assumptions of the model. Unfortunately, this problem arises often in practice, one of the most obvious examples being the bus bunching phenomenon where buses serving the same route tend to form pairs, resulting in a sequence of short and long headways.

A simple approximation for dealing with such problems is to define a new sequence of headway intervals $\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \ldots$, where $\hat{\tau}_{2n+1} \sim \hat{\tau}_1$ and $\hat{\tau}_{2n} \sim \hat{\tau}_2$ and where $\hat{\tau}_1$ and $\hat{\tau}_2$ are independent but with different distributions. In order words, we are replacing a general sequence of headways with an alternating renewal proce-s in order to capture the first order correlation structure. The correlation coefficient C_R for such a process is given by:

$$C_R = \frac{E(\hat{\tau}_{2n+1} - \bar{\tau})(\hat{\tau}_{2n} - \bar{\tau})}{[Var(\hat{\tau})]^{1/2}}$$

$$= \frac{-(\bar{\tau}_1 - \bar{\tau}_2)^2}{E(\tau_1^2) + E(\tau_2^2) - 2\bar{\tau}^2} \qquad 3.74$$

where $\bar{\tau} = (\bar{\tau}_1 + \bar{\tau}_2)/2$. Corresponding to such a sequence of headways is a sequence Y_n of the number of arrivals during each headway. Analogously, we have $Y_{2n+1} \sim Y_1$ and $Y_{2n} \sim Y_2$. The following recursions may now be defined:

$$Q_{2n+1} = R_{2n} + Y_{2n}$$

$$Q_{2n+2} = R_{2n+1} + Y_{2n+1} \qquad 3.75$$

we can now assume that Y_{2n} and Y_{2n+1} are conditionally independent. Letting n → 0, we may write:

$$Q_1(z) = R_2(z) \cdot Y_2(z)$$

$$Q_2(z) = R_1(z) \cdot Y_1(z) \qquad 3.76$$

where Q_1, R_1 and Y_1 have the steady state distributions corresponding to Q_{2n+1}, R_{2n+1}, and Y_{2n+1}, and Q_2, R_2 and Y_2 correspond similarly to Q_{2n}, R_{2n} and Y_{2n}. The transform $R_1(z)$ is given by:

$$R_1(z) = z^{-c}\left[\sum_{i=0}^{c-1} q_i^1\,(z^c - z^i) + Q_1(z)\right] \qquad 3.77$$

Hence:

$$Q_2(z) = z^{-c}\left[\sum_{i=0}^{c-1} q_i^1\,(z^c - z^i) + Q_1\,(z)\right] Y_1\,(z) \qquad 3.78$$

Similarly:

$$Q_1(z) = z^{-c}\left[\sum_{i=0}^{c-1} q_i^2\,(z^c - z^i) + Q_2\,(z)\right] Y_2(z) \qquad 3.79$$

Substituting 3.79 into 3.78 gives:

$$Q_2(z) = z^{-c}\left[\sum_{i=0}^{c-1} q_i^1\,(z^c - z^i) + z^{-c}\left(\sum_{i=0}^{c-1} q_i^2(z^c - z^i) + Q_2(z)\right) Y_2(\,z)\right] Y_1(z)$$

$$= \frac{z^c \sum_{i=0}^{c-1} q_i^1\,(z^c - z^i) + \sum_{i=0}^{c-1} q_i^2(z^c - z^i) \cdot Y_2(z)}{\frac{z^{2c}}{Y_1(z)} - Y_2(z)} \qquad 3.80$$

An analogous expression can be found for $Q_1(z)$, although 3.80 is all we require for solution. The denominator in 3.80 has 2c zeroes on and within the unit circle which can be used to solve a 2c × 2c system of equations for the unknowns. Since the numerator is a polynomial of infinite order, it does not appear possible to take advantage of the shortcuts used earlier (i.e. expressing the numerator directly in terms of the zeroes). The potentially large set of simultaneous equations casts some doubt on our ability to accurately solve for the unknown probabilities. No attempt is made in this research to solve 3.80, but chapter 4 presents some tentative results using Gaussian elimination to solve equation 3.9 which is somewhat simpler than 3.80.

3.3.3 The transform of the size of an arriving group for a queue imbedded in a network

In the previous sections we concentrate on finding queue length transforms for different problems. In this section, we derive several results that are of use in describing the arrival process to a queue. Using the transform of the length of the queue, the transform of the size of a departing batch is derived. This is necessary when studying networks of queues where the departures from one queue became the arrivals to another.

Consider the problem depicted in Figure 3.2. Vehicles depart from node A with a load of passengers which are dropped off at node B. Each passenger then heads to a specific departure point i, i=1, ..., I, and

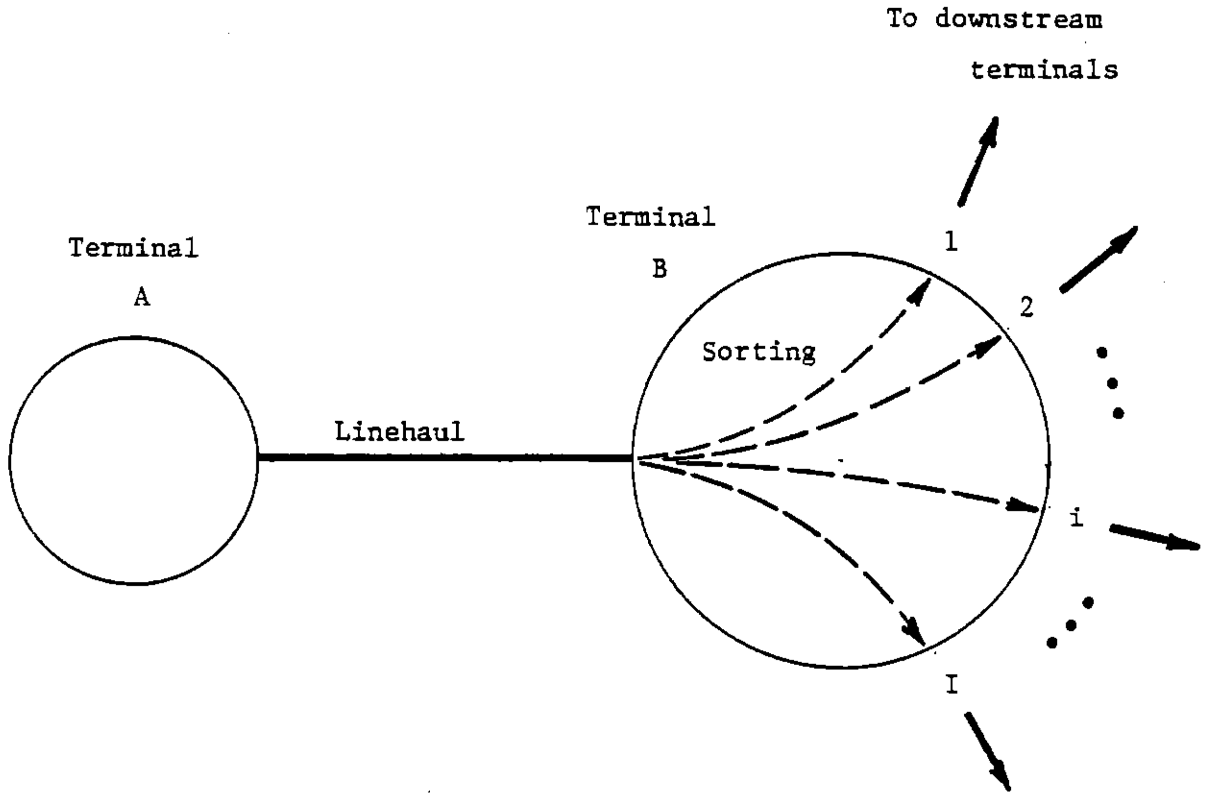


Illustration of sorting process for a terminal imbedded in a network

Figure 3.2

waits for the next appropriate outbound vehicle. Each of the points i=1, ..., I represents a queue with bulk arrivals, where the size of each group arriving to queue i is determined by the size of the load arriving from terminal A and the fraction of passengers who then depart from queue i. What we wish to do now is derive the transform of the size of a load departing from A, and from this find the transform of the size of a load joining queue i.

Addressing the first problem, let F denote the size of the incoming load from the queue at terminal A whose steady state queue distribution has transform Q(z). Then under a scheduled departure dispatching policy without cancellations we have:

$$\begin{aligned} f_i &= q_i \qquad\qquad i = o,\ldots, c-1 \\ f_c &= \sum_{i=c}^{\infty} q_i \end{aligned} \qquad 3.81$$

Taking the appropriate transforms and reducing gives:

$$F(z) = z^c - \sum_{i=o}^{c-1} q_i (z^c - z^i) \qquad 3.82$$

Using 3.9 we may rewrite 3.32 as

$$F(z) = z^c - Q(z) \left[\frac{z^c}{Y(z)} - 1\right]$$

$$= z^c \; [\; 1 - R(z)] + Q(z) \qquad 3.83$$

Alternatively, we may use 3.14 and express F(z) in terms of the zeroes required to find Q(z), as follows:

$$F(z) = z^{c} - (c - \overline{Y})(z - 1) \prod_{i=1}^{c-1} \frac{z-z^{i}}{1-z_{i}} \qquad 3.84$$

If a minimum load constraint of m is imposed, then, letting F^{m} be the corresponding load size, we find:

$$F^{m}(z) = z^{c} - [(z^{c}- 1) \sum_{i=0}^{m-1} q_{i} + \sum_{i=m}^{c-1} q_{i} (z^{c} - z^{i})] \qquad 3.85$$

Using 3.28 we may simplify this to:

$$F^{m}(z) = z^{c} [1 - R(z)] + Q^{m}(z) + (z^{c}-1) \sum_{i=o}^{m-1} q_{i} (z^{i}-1) \qquad 3.86$$

In this case we are forced to leave $F^{m}(z)$ in terms of the probabilities $q_{o},\ldots, q_{m-1}$, which are easily determined using the methods outlined in section 4.2. For the special case of m = c, it is easily verified that $\sum_{i=o}^{c-1} q_{i} = 1 - \rho$, where $\rho = \bar{Y}/c$. This gives:

$$F^{c}(z) = 1 - \rho + \rho z^{c} \qquad 3.87$$

$F^{c}(z)$ of course is simply the transform of a binomial random variable which takes on values of 0 and c with probabilities 1 - ρ and ρ. It is interesting to note that $F^{c}(z)$ does not depend on the characteristics of the queue itself. We can interpret ρ now as the probability a vehicle will be dispatched (with load c) at a given (scheduled) dispatch instant, although it is important to realize that the probability of a departure from one instant to the next is not independent.

In deriving 3.82,3.86 and 3.87 the assumption is made that an empty load could depart from the queue. This serves not only to simplify the expression for the size of each load but also the analysis of the departure process. To allow for the presence of empty loads, we introduce the concept of a virtual departure which would occur, for example, when a minimum load constraint produced a run cancellation. Using this notion, we find that the departure process of vehicles is independent of a minimum load constraint since a real or virtual departure occurs at each dispatch instant regardless of the length of the queue.

With the total size of an inbound load determined, we turn now to the problem of finding the distribution of units out of an inbound load headed for an outbound link. This is solved easily as follows. Let Θ be the fraction of units out of an inbound load headed for a given outbound link, and as before, let G be the size of the resulting group. We assume now that all units are travelling independently. Thus if an inbound load has exactly k units, then G has a binomial distribution with parameters k and Θ, with transform:

$$G(z|k) = (1 - \Theta + \Theta z)^k \qquad 3.88$$

Taking expectations of both sides gives:

$$G(z) = \sum_{k=o}^{c} f_k \, G(z|k)$$

$$= \sum_{k=o}^{c} f_k \, (1 - \Theta + \Theta z)^k$$

$$= F(1 - \Theta + \Theta z) \qquad 3.89$$

Equation 3.89 gives a simple relationship between the total size of an incoming load and the size of the group arriving at a particular queue.

This completes the major results relating to queue length transforms, although we return briefly to the topic in section 3.5. The next section looks at the important problem of waiting times in bulk arrival, bulk service queues.

3.4 Waiting time transforms for bulk arrival, bulk service queues

Until now our attention has focused on studying the length of the queue at dispatch instants. These results can be used to determine a limited set of level of service parameters such as the probability of missing the first departure. In this section, we consider the problem of waiting times for bulk arrival, bulk service queues. To date, waiting time transforms have been found only for single arrival, bulk service queues (Downton(1955)) and bulk arrival, single service queues (Burke(1975)). We present here a methodology for finding all the moments of the waiting time distribution, but only in certain examples can we explicitly find the transform itself.

One such example is where customers arrive singly, and hence Downton's result is obtained as a special case. The derivation used here, however, is considerably different than his.

The approach we use begins by proving that the distribution of the number of units in front of a randomly chosen arriving unit, including other units in the same arriving group, is the same as the distribution of the number of units behind a departing unit, including others in the same departing group. We denote the number in front of an arriving unit by $\tilde{Q}$ and the number behind a departing unit $\hat{Q}$. This observation simplifies the derivation of the distribution of $\hat{Q}$. Then, by relating this distribution to the waiting time distribution we can find the moments of the latter.

We begin by proving that the distributions of $\hat{Q}$ and $\tilde{Q}$ are the same. This is analogous to the well known fact that the number of units _at an arrival instant_ is the same as that _at a departure instant_ for any single arrival, single service queue. It is equally well known that this statement is not true for queues with bulk arrivals or bulk service. The important difference between this statement and the one we are trying to prove is that the number of units at an arrival instant is the same as the number of units at an arrival instant is the same as the number of units in front of the _first_ unit in each arriving group. However, the proof we need for bulk systems is a simple extension of that used for single arrival, single service systems (see, for example, Gross and Harris (1974), pp. 235-236). Referring to figure 3.3, define $Q(t)$ as the number of units in queue at time t, and let each circle (denoting an individual unit) be the points in time

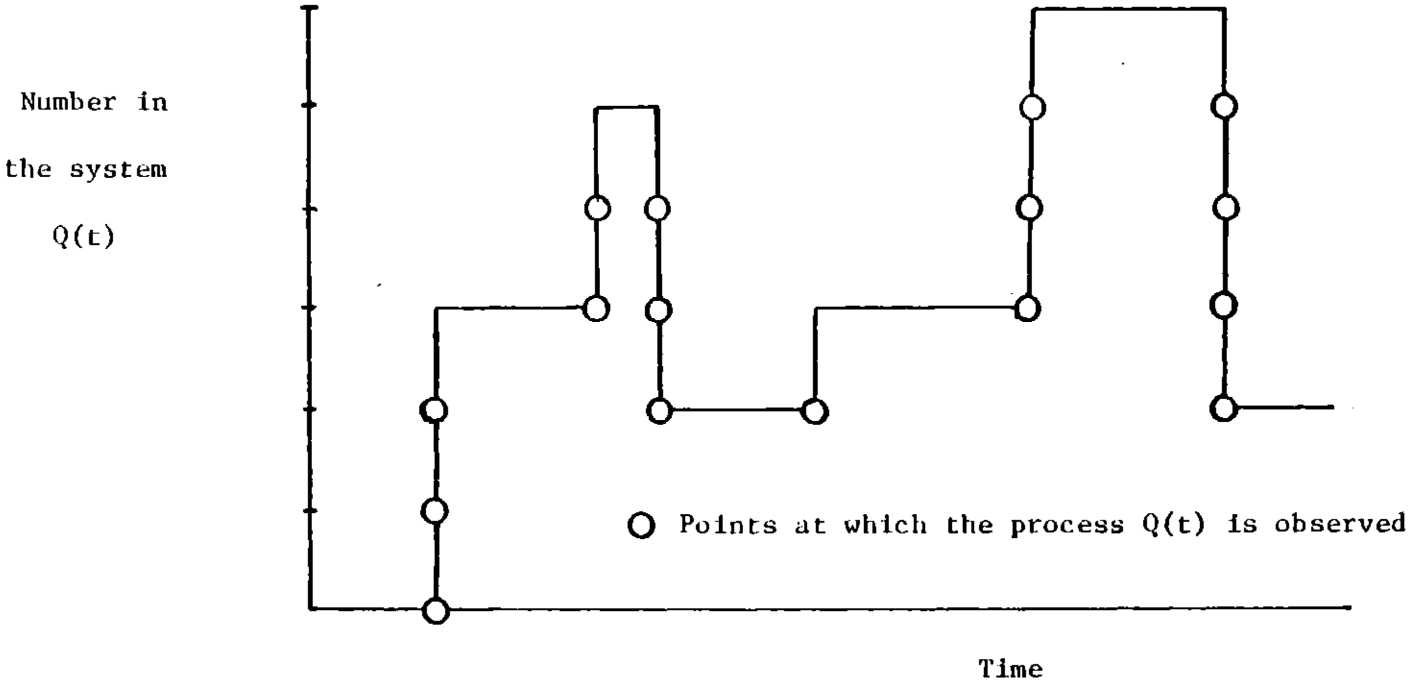


Figure 3.3

Illustration of arrivals and departures
from the system

at which the process is observed. Following the derivation by Gross and Harriss, define $A_n(t)$ to be the number of unit upward jumps from state n in the interval (0,t), representing the number of times an arriving unit sees n customers in front. Similarly, define $D_n(t)$ to be the number of downward jumps through state n. It is clear from figure 3.3 that:

$$|A_n(t) - D_n(t)| \leq 1 \qquad 3.90$$

Let A(t) and D(t) be respectively the total number of arrivals to and departures from the system. We may now write:

$$\lim_{t\to\infty} \frac{A_n(t)}{A(t)} = \tilde{q}_n \qquad 3.91$$

and

$$\lim_{t\to\infty} \frac{D_n(t)}{D(t)} = \hat{q}_n. \qquad 3.92$$

Using arguments identical to those given in Gross and Harris, it is possible to show that $\tilde{q}_n = \hat{q}_n$.

With this result in hand, we address the problem of finding $\tilde{Q}(z)$. Let $\tilde{Y}$ be the number of units in front of a random unit which arrived during the same headway interval, where $\tilde{Y}$ may include units from the same arriving batch. As before, let R be the number left over following the last dispatch instant (we are allowing the possibility of cancellations). Then $\tilde{Q} = R + \tilde{Y}$, or, since R and $\tilde{Y}$ are independent:

$$\tilde{Q}(z) = R(z)\ \tilde{Y}(z) \qquad 3.93$$

The quantity $\tilde{Y}$ is analogous to the forward recurrence time in a discrete renewal process where the length of time between renewals is Y. It is well known, then, that:

$$\tilde{Y}(z) = \frac{1 - Y(z)}{\overline{Y}\ (1-z)} \qquad 3.94$$

Combining 3.93 and 3.94 gives:

$$\tilde{Q}(z) = \frac{1 - Y(z)}{\overline{Y}\ (1-z)} R(z)$$

$$= \frac{1 - Y(z)}{\overline{Y}\ (1-z)} \frac{Q(z)}{Y(z)} \qquad 3.95$$

Now consider the perspective of a departing unit. On leaving the sytem, there will be a total of Q units behind a departing unit. Out of these $\hat{Q}$ units, some will have arrived in the same group and the rest will have arrived afterward. Let $\tilde{G}$ denote the first quantity and H denote the latter. $\tilde{G}$ is the number of units behind a random unit which arrived in the same group, and is similar to $\tilde{Y}$ in that it can be thought of as the backward recurrence time in a discrete renewal process. Thus we have:

$$\tilde{G}(z) = \frac{1 - G(z)}{\overline{G}\ (1-z)} \qquad 3.96$$

We also note that $\hat{Q} = \tilde{G} + H$, or, since $\tilde{G}$ and H are independent:

$$\hat{Q}(z) = G(z)\ H(z) \qquad 3.97$$

Now using the fact that $\tilde{Q}(z) = \hat{Q}(z)$, and combining 3.95, 3.96 and 3.97, we can solve for H(z), giving:

$$H(z) = \frac{\overline{G}\ (1-z)}{1 - G(z)} \frac{1 - Y(z)}{\overline{Y}\ (1-z)} \frac{Q(z)}{Y(z)}$$

Or:

$$H(z) = \frac{\overline{G}}{\overline{Y}} \frac{1 - Y(z)}{1 - G(z)} \frac{Q(z)}{Y(z)} \qquad 3.98$$

Equation 3.98 is a very important results, and is quite general as well. For example, at no point did we ever assume that vehicle capacities were constant, and hence it applies to the case with random vehicle capacities. It also applies to the case where cancellations are allowed, although a few words of explanation are in order. When deriving $\tilde{Q}(z)$, we used Q(z) at each dispatch instant. $\hat{Q}(z)$, however, is defined only at departure instants, since downward transitions occur only when a vehicle leaves. Thus H is the number of units behind a departing unit, even though Q(z) in 3.98 is defined at dispatch instants.

We still have not discussed the topic of waiting times. To do this, let W*(s) be the Laplace transform of the waiting time distribution. If we have compound Poisson arrivals, then $W^*(\lambda - \lambda G(z))$ is the transform of the distribution of the total number of units arriving during the waiting time of a randomly chosen unit. Clearly:

$$W^*(\lambda - \lambda G(z)) = H(z) \qquad 3.99$$

Equation 3.99 is the equation needed to relate the (unknown) moments of W to the (known) moments of H. Depending on the functional form of G(z), however, it will not always be possible to solve explicitly for W*(s). One special case where this can be done is when the arrival process is simple, whereby G(z) = z. Letting $z = 1 - s/\lambda$, we obtain:

$$W^*(s) = H(1 - s/\lambda)$$

$$= \frac{\lambda(1 - Y(1 - s/\lambda))}{\overline{Y}\, s} \frac{Q(1 - s/\lambda)}{Y(1 - s/\lambda)} \qquad 3.100$$

Equation 3.100 is the result obtained by Downton (1955). In general, however, it will not be possible to invert the function s(z) = $\lambda - \lambda G(z)$ to find z(s). Instead, we must express the moments of W in terms of the moments of H, as follows:

$$E(W) = \frac{\bar{H}}{\lambda\bar{G}} \qquad 3.101$$

$$\mathrm{Var}(W) = \frac{1}{\lambda^2\bar{G}^2}\,[\bar{\bar{H}} - \bar{H}(\frac{\bar{\bar{G}}}{\bar{G}} + 1)] \qquad 3.102$$

where $\bar{H}$ and $\bar{\bar{H}}$ are the mean and variance of H. $\bar{H}$ and $\bar{\bar{H}}$, in terms of the moments of G, Y and Q, are found to be (see appendix D):

$$E(H) = \bar{\bar{H}} = \frac{1}{2}(\frac{\bar{\bar{Y}}}{\bar{Y}} + \bar{Y} - 1) - \frac{1}{2}(\frac{\bar{\bar{G}}}{\bar{G}} + \bar{G} - 1) + (\bar{Q} - \bar{Y})$$

$$= \frac{1}{2}[\frac{\bar{\bar{Y}}}{\bar{Y}} - \bar{Y} - \frac{\bar{\bar{G}}}{\bar{G}} - \bar{G}] + \bar{Q} \qquad 3.103$$

$$\mathrm{Var}(H) = \bar{\bar{H}} = (\frac{4\bar{Y}\bar{\bar{\bar{Y}}} + 6\bar{Y}^2\bar{\bar{Y}} - \bar{Y}^2 + \bar{Y}^4 - 3\bar{\bar{Y}}^2}{12\bar{Y}^2})$$

$$-(\frac{4\bar{G}\bar{\bar{\bar{G}}} + 6\bar{G}^2\bar{\bar{G}} - \bar{G}^2 + \bar{G}^4 - 3\bar{\bar{G}}^2}{12\bar{G}^2})$$

$$+(\bar{\bar{Q}} - \bar{\bar{Y}}) \qquad 3.104$$

The derivations of equations 3.103 and 3.104 were simplified by using the relationship:

$$H + \tilde{G} = \tilde{Y} + R \qquad 3.105$$

where H and $\tilde{G}$ are independent and $\tilde{Y}$ and R are independent. Thus

$$E(H) = E(\tilde{Y}) - E(\tilde{G}) + E(R) \qquad 3.106$$

$$\mathrm{Var}(H) = \mathrm{Var}(\tilde{Y}) - \mathrm{Var}(\tilde{G}) + \mathrm{Var}(R) \qquad 3.107$$

Also, $Q = R + Y$, where R and Y are independent, and thus $E(R) = E(Q) - E(Y)$ and $\mathrm{Var}(R) = \mathrm{Var}(Q) - \mathrm{Var}(Y)$.

As with equations 3.17 and 3.18 for the mean and variance of the queue length, 3.103 and 3.104 are expressed in terms of the moments of the number of arrivals between service instants, allowing the equations to be used as approximations for more general arrival processes than the Poisson. The equations are exact, however, only when arrivals are in fact Poisson (or more precisely, when the sequence Y_n is i.i.d.). If the time between departures is exactly T, then:

$$Y(z) = \exp\{-\lambda T(1-G(z))\} \qquad 3.108$$

where λ is the arrival rate of groups. From 3.108 we find:

$$\overline{Y} = \lambda T\overline{G} \qquad 3.109$$

$$\overline{\overline{Y}} = \lambda T(\overline{\overline{G}}+\overline{G}^2) \qquad 3.110$$

$$\overline{\overline{\overline{Y}}} = \lambda T[\overline{\overline{\overline{G}}}+3\overline{\overline{G}}(\overline{G}-1)+\overline{G}^3-3\overline{G}^2+2\overline{G}] \qquad 3.111$$

For single arrivals, $(G(z)=z)$, 3.103 and 3.104 reduce to;

$$E(V) = \overline{Q} - \frac{1}{2}\lambda T \qquad 3.112$$

$$\mathrm{Var}(V) = \frac{(\lambda T)^2 - 6\lambda T}{12} + \overline{\overline{Q}} \qquad 3.113$$

The mean and variance of the wait time are now given by: (substituting 3.112 and 3.113 into 3.101 and 3.102)

$$E(W) = \frac{\bar{Q}}{\lambda} - \frac{T}{2} \tag{3.114}$$

$$\mathrm{Var}(W) = \frac{T^2}{12} + \frac{\bar{\bar{Q}} - \bar{Q}}{\lambda^2} \tag{3.115}$$

To check our results, assume we now have an infinite capacity vehicle, in which case $Q = Y$, and hence $\bar{Q} = \bar{\bar{Q}} = \lambda T$. Equations 3.114 and 3.115 then reduce to:

$$E(W) = \frac{T}{2} \tag{3.116}$$

$$\mathrm{Var}(W) = \frac{T^2}{12} \tag{3.117}$$

This is exactly what we would expect, since wait times would be uniformly distributed between 0 and T.

These results for bulk arrival, bulk service are new, as the current literature has not yet addressed the problem of finding the waiting time distribution for such systems. As this presentation has demonstrated, by developing a basic understanding of the problem, it is possible to show how Downton's earlier result can be extended to allow for the more general case.

The derivation presented in this section is based on the careful distinction between the number of units at a departure instant and the number behind a departing unit. In the course of conducting this investigation, a number of interesting relationships were uncovered which have not been discussed in the literature. These observations are now reported on in the following section.

3.5 Relationship between the length of the queue from different perspectives

In sections 3.2 and 3.3, the transform of the length of the queue at dispatch instants is obtained. In section 3.4, it is shown that the distributions of $\tilde{Q}$ and $\hat{Q}$ are the same, and the transform of this distribution is found using basic concepts in renewal theory. In this section, we apply similar concepts to find the transform of the length of the queue at a random point in time, denoted by Q_t, and summarize the relationships between Q, $\tilde{Q}$, $\hat{Q}$ and Q_t.

For compound arrival processes, the distribution of the length of the queue at a random point in time is the same as the number of units in front of the first unit in an arriving group. We have already found the number in front of a random arriving unit, $\tilde{Q}$, which includes others in front in the same arriving group plus those already in the queue. The first quantity is denoted $\tilde{G}$, since this is analogous to the number behind a departing unit which arrived in the same batch, and has the same distribution. The second quantity is just Q_t, thus:

$$\tilde{Q} = \tilde{G} + Q_t \qquad 3.118$$

Since $\tilde{G}$ and Q_t are independent, we find:

$$Q_t(z) = \frac{\tilde{Q}(z)}{\tilde{G}(z)} \qquad 3.119$$

$$= \frac{\overline{G}}{\overline{Y}} \frac{1 - Y(z)}{1 - G(z)} \frac{Q(z)}{Y(z)} \qquad 3.120$$

Comparing 3.98 and 3.120, we find that $Q_t(z) = H(z)$. From before, we have:

$$\tilde{Q}(z) = \hat{Q}(z) = \frac{1 - Y(z)}{\overline{Y}\ (1-z)} \frac{Q(z)}{Y(z)} \qquad 3.121$$

Thus 3.120 and 3.121 quickly summarize the relationship between Q, $\tilde{Q}$, $\hat{Q}$ and Q_t, and apply to any $M^x/G^y/1$ queue, with or without cancellations. Of course, the fact that $\tilde{Q}(z) = \hat{Q}(z)$ is completely general and applies to any $G^x/G^y/1$ system. Equation 3.119 was previously obtained by Chaudry (1979) for the $M^x/G/1$ service queue using the method of supplementary variables, without noticing that the difference between Q_t and $\tilde{Q}$ was simply the number in front of a random unit that arrived in the same group.

3.6 The $M^x/G/1$ service queue

In this final section we report on several known results for what has been termed here the unloading queue, consisting of bulk, Poisson arrivals to a queue which serves units singly with a general service time distribution. Other authors who have considered the same problem include Gaver (1959), Cohen (1969), Chaudry (1977) and Burke (1975).

The problem is discussed here because of its importance in transportation and to add several observations which have not been made in the literature. The expressions for the queue length and waiting time transforms can be obtained in a manner similar to that used before for the scheduled departure queue.

To proceed, we begin by establishing a recursion similar to that in equation 3.2. However, whereas in the scheduled departure queue no distinction was made between the beginning and ending of a service period, the unloading queue becomes idle when the system is empty and hence the end of one service period may be separated from the beginning of the next. We adopt the usual convention of defining the imbedded Markov chain at the end of each service period. Also, we now define $\tilde{Q}$ to be the total number in the system and W the total system time since the distinction between the queue and the system is meaningless for freight and unimportant for passengers. As before, we define $\tilde{Q}_n$ as the number remaining in the system following the departure of the n^{th} unit and Y_n the total number of units arriving during the service of this unit, and hence:

$$\tilde{Q}_{n+1} = \begin{cases} G_n - 1 + Y_n & \tilde{Q}_n = 0 \\ \tilde{Q}_n - 1 + Y_n & \tilde{Q}_n > 0 \end{cases} \qquad 3.122$$

where G_n is the size of the group the n^{th} unit arrived in. Solving by the usual transform method gives us:

$$\tilde{Q}_{n+1}(z) = \sum_{i=0}^{\infty} \tilde{q}_i^{n+1} z^i = \left[\frac{1}{z} \cdot G(z)\ Y(z)\right]\tilde{q}_o + \left(\frac{1}{z}\,\frac{\tilde{Q}_n(z)-q_o}{1-q_o} \cdot Y(z)\right) \cdot (1-q_o) \qquad 3.123$$

Taking the limit as n → ∞ and solving for $\tilde{Q}(z)$ yields:

$$\tilde{Q}(z) = \frac{\tilde{q}_o(G(z) - 1)}{\frac{z}{Y(z)} - 1} \qquad 3.124$$

We can now find q_o using $\lim_{z \to 1} \tilde{Q}(z) = 1$, which gives:

$$q_o = \frac{1 - \overline{Y}}{\overline{G}} \qquad 3.125$$

Defining $\rho = \overline{Y} = \lambda\overline{G}\overline{B}$, we obtain:

$$\tilde{Q}(z) = \frac{(1-\rho)(G(z) - 1)}{\overline{G}\left[\frac{z}{Y(z)} - 1\right]} \qquad 3.126$$

Equation 3.126 gives the transform of the distribution of the number of people behind a randomly departing unit, and hence is analogous to Q found earlier for the scheduled departure queue. As before, this is equivalent to the number in front of a randomly arriving unit, where again we include other units in the same arriving group. Thus we may easily find the length of the queue at a random point in time (or equivalently, as seen by the first unit in a group), which we define as $Q_t(z)$ (the subscript t denoting the average over time). Remembering that $\tilde{G}$ is the number of units in front of a random unit which arrived in the same time group, we have:

$$\begin{aligned}\tilde{Q}(z) &= \tilde{G}(z) \cdot Q_t(z) \\ &= \frac{1 - G(z)}{\overline{G}(1-z)} \quad Q_t(z) \qquad 3.127\end{aligned}$$

Using 3.126 and 3.127 to solve for $Q_t(z)$ gives:

$$Q_t(z) = \frac{(1-\rho)(z-1)}{\frac{z}{Y(z)} - 1} \qquad 3.128$$

Equation 3,128 has been found previously by Gaver (1959) and Chaudry (1979) using the method of supplementary variables.

Interestingly, $Q_t(z)$ is the same as that found for the scheduled departure queue at departure instants when the capacity of the outbound vehicle is unity. This equivalence does not appear to have a simple, intuitive explanation.

The transform of the total waiting time (including service time) can be found by using results known for the M/G/1 queue with single arrivals, and brings out the relationship between the two queueing systems. We proceed by finding the waiting time until the first unit begins service (this can be thought of as the time until a crew begins to unload a vehicle), denoted W_1. This can be found by treating each group as a supercustomer with service time transform $G(B^*(s))$ (see Cohen (1969)); the concept of a supercustomer in bulk arrival queues is implicit in a number of papers, including Gaver (1959) and Jaiswal (1960a,b). Thus we have (Kleinrock (1975):

$$W_1^*(s) = \frac{s(1-\rho)}{s-\lambda+\lambda G(B^*(s))} \qquad 3.129$$

This is the result obtained by Gaver, who incorrectly identifies it as the wait time for an individual unit. This is found by adding the service time for the other units in the same group which are served first, denoted by W_2, with transform:

$$W_2^*(s) = \tilde{G}(B(s))$$

$$= \frac{1 - G(B(s))}{\overline{G}\ (1-B(s))} \qquad 3.130$$

Finally, we must add on the service time of the unit itself. The total wait time is now:

$$W^*(s) = W_1^*(s)\ W_2^*(s)\ B^*(s)$$

$$= \frac{s(1-\rho)}{s-\lambda+\lambda G(B(s))} \cdot \frac{1 - G(B(s))}{\overline{G}(1-B(s))} \cdot B(s) \qquad 3.131$$

Equation 3.131 was obtained recently by Burke (1975), who points out that several other authors provided incorrect derivations (Cohen (1969), Keilson (1962)) by neglecting to account for the difference between the distribution of the size of a group when sampled over groups and when sampled over units.

3.7 Summary

This chapter has provided the important foundations required before any numerical work can be attempted. The material presented here does not, however, represent an exhaustive treatment of the theory of bulk queues since our focus has been on reporting new contributions to the field rather than summarizing old ones. The most important of the new results are 1) extension of Bailey's work to bulk arrivals with the new set of moment formulas, 2) derivation of the transform of the waiting time distribution for bulk arrivals to scheduled departure queues and the associated moment formulas, and 3) the introduction and analysis of scheduled departure queues with cancellations. In addition, the presentation has emphasized not only deriving new results but interpreting them as well, and several observations have been made (e.g. regarding the distribution of units in front of and behind a random unit) that clarify some of the results.

While the motivation of this chapter is to fill in the gaps in the literature, many other results have been reported in the literature which are not touched on here. Section 3.6 briefly touched on some important results for the unloading queue with a general service time distribution. As pointed out in 2.1.2, several authors have been able to analyze the problem using variants of the hyperstage distribution which enables a continuous time formulation of the problem and hence avoids the necessity of using Kendall's concept of the imbedded Markov chain. Considerable attention has been devoted to the general bulk service rule (see 2.1.1 and 2.1.3) which may prove useful in some applications, although it is not considered in this research.

As is pointed out in the beginning of the chapter, the remainder of the thesis can be regarded as supplementing the theoretical work reported here. First, in chapter 4, several numerical problems associated with solving the transforms are addressed. These include a) locating the necessary zeroes needed to solve for the transform, b) inverting the transform and c) approximations for both the queue length distribution and its moments. Then, chapter 5 considers the Poisson arrival process as an approximation for more general processes, and suggests several methods for approximating $G^x/G^y/1$ queues.

Chapter 4 Numerical Analysis of Transforms

The value of transform methods, notwithstanding their theoretical elegance, depends in a large part on their ability to provide useful numerical results. Having completed the theoretical groundwork in chapter 3, it is now necessary to turn to the equally important probelm of obtaining numerical solutions. As a topic which has received almost no attention in the literature (inasmuch as bulk queues are concerned), the research here focuses on two common problems encountered in the application of transform results, namely a) finding the zeroes needed to solve the queue length transform and b) inverting the transform. The first problem, discussed in section 4.1, is frequently cited as a major criticism of bulk queueing theory (see e.g. Neuts(1979) and Bagchi and Templeton (1972)) on the basis of the (well-known) fact that solving for the zeroes of a function can be numerically hazardous. Other authors have in fact applied the theory to specific problems (e.g. Peterson (1977a), Novaes (1963), and Groninger (1966)) but none have reported on a careful, systematic analysis of the topic. Also, chapter 1 discussed the problem of applying the results to large scale networks, where the issue of computational speed becomes important. This aspect of the problem is also considered here.

The second problem, inverting the transform, is one that is normally not addressed in either the theoretical or the applied literature. In the case of bulk queues, the problem breaks into parts, the first consisting of finding the set of probabilities q_o, ..., q_{c-1}, and the second consisting of finding the remaining probabilities, q_c, q_{c+1}, ... Two methods are outlined in section 4.2 for finding the first c probabilities. A separate procedure is then suggested for finding the rest of the probability vector, and all three procedures are tested numerically.

Finally, section 4.3 moves in a somewhat different direction and presents approximations for the mean and variance of the queue length for a particular set of queueing problems. The motivation for this work is the desire to have estimates of the moments in closed form, thereby avoiding the problem of solving for zeroes.

4.1 Finding the zeroes

The technique of finding zeroes to solve transforms has been used in countless papers in the queueing literature and constitutes what now might be considered the classical approach for solving many problems in queueing theory. As is pointed out in section 2.1.4, however, finding zeroes can, from a practical standpoint, be difficult to do within the confines of a computer. Considering the number of papers which depend on the technique together with its critics (e.g. Neuts (1979) and Bagchi and Templeton (1972)), it is worthwhile to spend some time studying the problem carefully. This section looks at three issues relating to the topic, namely numerical stability, accuracy and efficiency. The first of these addresses the question of whether the search procedure will converge on the desired set of zeroes for a range of different problems and different parameter values. The second, accuracy, is used here to describe the sensitivity of the results (i.e. the values computed for the zeroes) to the stopping criterion used in the search procedure.

Finally, efficiency refers to the speed with which the calculations can be performed.

The methodology that is used consists of finding zeroes for three different problems corresponding to a) simple Poisson arrivals to a queue with deterministic departures headways, b) simple Poisson arrivals with Erlang distributed headways, and c) compound Poisson arrivals with deterministic headways. The last of these was constructed by assuming a Poisson arrival stream of vehicles from 10 different upstream terminals, each operating with simple Poisson arrivals and deterministic departures. Each of these upstream queues were solved to find the transform of the distribution of the number of units on each vehicle as described in section 3.3.4.

In all three cases, the root finding problem reduces to finding c zeroes, z_i, $i = 0, \ldots, c-1$, such that:

$$f(z_i) = z_i^c - Y(z_i) = 0 \quad i=0, \ldots, c-1 \qquad 4.1$$

We note at this point that $c = 20$ is used throughout unless specifically stated otherwise. The three cases are differentiated in the form of $Y(z)$, as follows:

Case I: Simple Poisson arrivals, deterministic headways;

$$Y(z) = \exp\{-\lambda T(1 - z)\} \qquad 4.2$$

$\lambda =$ arrival rate

$T =$ headway

Case II: Simple Poisson arrivals, Erlang headways:

$$Y(z) = (1 + \theta - \theta z)^{-k} \qquad 4.3$$

$$\theta = \frac{\lambda\bar{T}}{k}$$

$\bar{T}$ = average headway

k = parameter of the headway distribution

Case III: Compound Poisson input, deterministic headways:

$$Y(z) = \exp\{ -T \sum_{i=1}^{10} \lambda_i(1 - F_i(1 - \theta_i + \theta_i z))\} \qquad 4.4$$

λ_i = arrival rate of vehicles from terminal i

$F_i(z)$ = transform of the distribution of the number of units on a vehicle arriving from terminal i (equation 3.84)

θ_i = fraction of units arriving from terminal i which will head outbound over the queue in question

Cases I and II are of interest because they represent respectively irrational and rational functional forms for Y(z). Case III is of interest simply because of the complexity of the functional form for Y(z) which will certainly have an impact on the speed with which equation 4.1 can be solved and may affect the stability of the root finding procedure.

The procedure used to find the zeroes is a straightforward application of the Newton-Raphson algorithm and is described briefly in appendix E.1. There it is also shown that cases II and III must be solved as a two dimensional problem, while case I can be reduced to one dimension and solved much more efficiently. Some care was exercised in the coding of the algorithm, but it is possible that further improvements in computational performance may be realized with the use of either a more sophisticated search procedure or more careful programming. For this reason, the experiments reported here must be viewed as simply indicative of what can be expected when implementing the procedures.

As a final comment, it is useful to point out that the problem of finding roots in queueing theory often reduces to finding the roots of the type of function shown in equation 4.1. The same equation occurs in virtually every other bulk queueing problem including, for example, the general bulk service rule (see section 2.1.1 and Neuts (1967), and Borthakur and Medhi (1974)). It also arises, for c=1, in some cases with arrivals and service in single units, (in particular the G/M/1 queue), as shown by Kleinrock (1975, pp. 132 and 249).

4.1.1 Numerical stability

Numerical stability is an issue that is encountered frequently in iterative numerical methods and is concerned with the simple question of whether an algorithm will converge for all reasonable values of the parameters of the problem. In most cases the problem reduces to one of finding a good starting point, since virtually all techniques have excellent convergence properties in the vicinity of the solution. When a good starting point is not known a priori, the simpler first order methods such as the Newton-Raphson procedure used in this research can be extremely unpredictable if the function is ill behaved.

There are three characteristics of the function in equation 4.1 that suggest that the usual difficulties with finding roots will not occur. First is the fact that all the roots lie within the unit circle, providing not only a good starting point but also a bound on all the other roots, guaranteeing that we will not have to search over an arbitrarily large portion of the complex plane.* The second characteristic is that rather than having an arbitrary function with c-1 roots, equation 4.1 can be expressed in terms of c-1 equations, each with a single root. To do this, we write:

* This would not be true for case II if we decided to solve for the K-1 roots outside of the unit circle as was described in section 3.2, eq. 3.24.

$$\tilde{f}(z) = \frac{z^c}{Y(z)} - 1 = 0 \qquad 4.5$$

Or:

$$\frac{z^c}{Y(z)} = \exp\ \{2\pi i\} \qquad 4.6$$

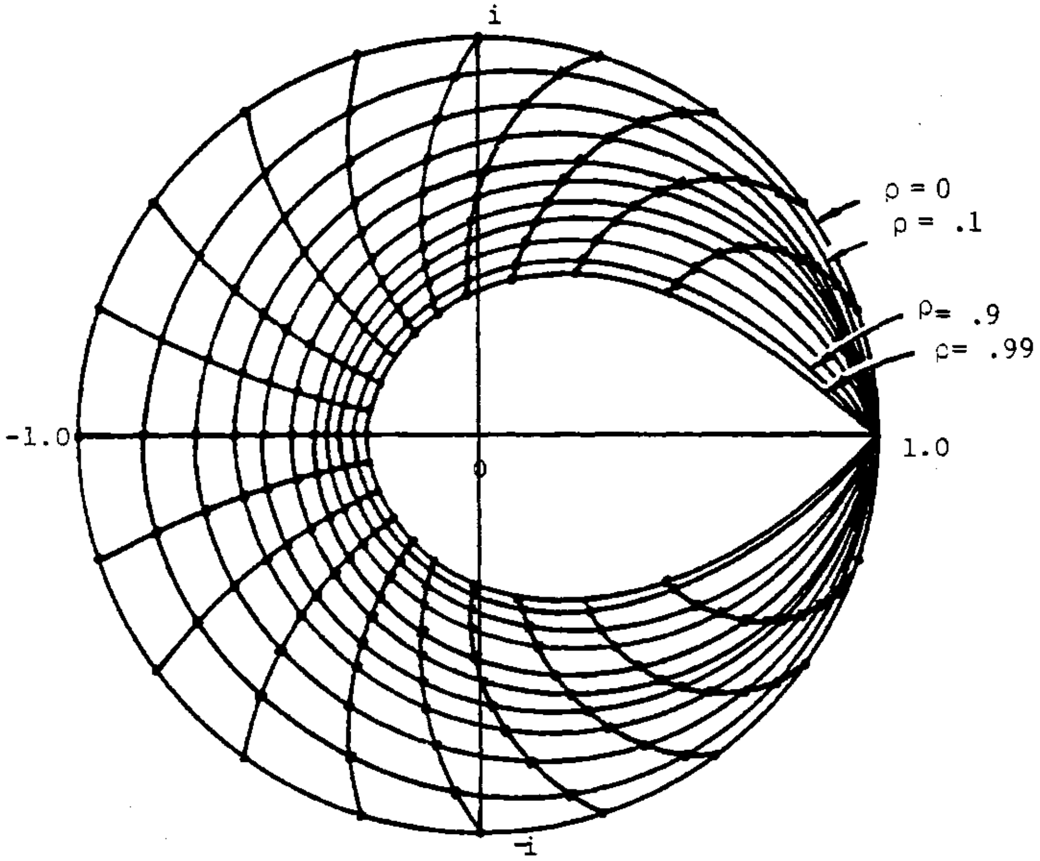
where $i = \sqrt{-1}$. Taking the c^{th} root of both sides gives:

$$z\ Y(z)^{-1/c} = \exp\ \{\frac{2\pi k i}{c}\}\ k = 1,\ 2,\ \ldots,\ c-1 \qquad 4.7$$

Thus equation 4.7 gives c-1 equations, each of which can be solved for a single root (using arguments similar to those used to show that equation 4.1 has c zeroes on or within the unit circle, it is possible to show that equation 4.7, for each value of k, has a single root on or within the unit circle). This is considerably simpler than finding c-1 roots out of a single equation, which poses the problem of trying to avoid finding the same root twice. We should point out that this is not the only way to solve for the desired roots and in some cases, not even the most obvious way. For example, case II produces in equation 4.5 a polynomial of order K + c and can thus be approached using standard procedures for finding the roots of polynomials, as is usually suggested in the literature. This creates the problem of distinguishing between roots inside and outside the unit circle and is felt to be a much more difficult task. Equation 4.7 suggests a general method which will always produce the desired roots inside the circle regardless of the functional form of Y(z).

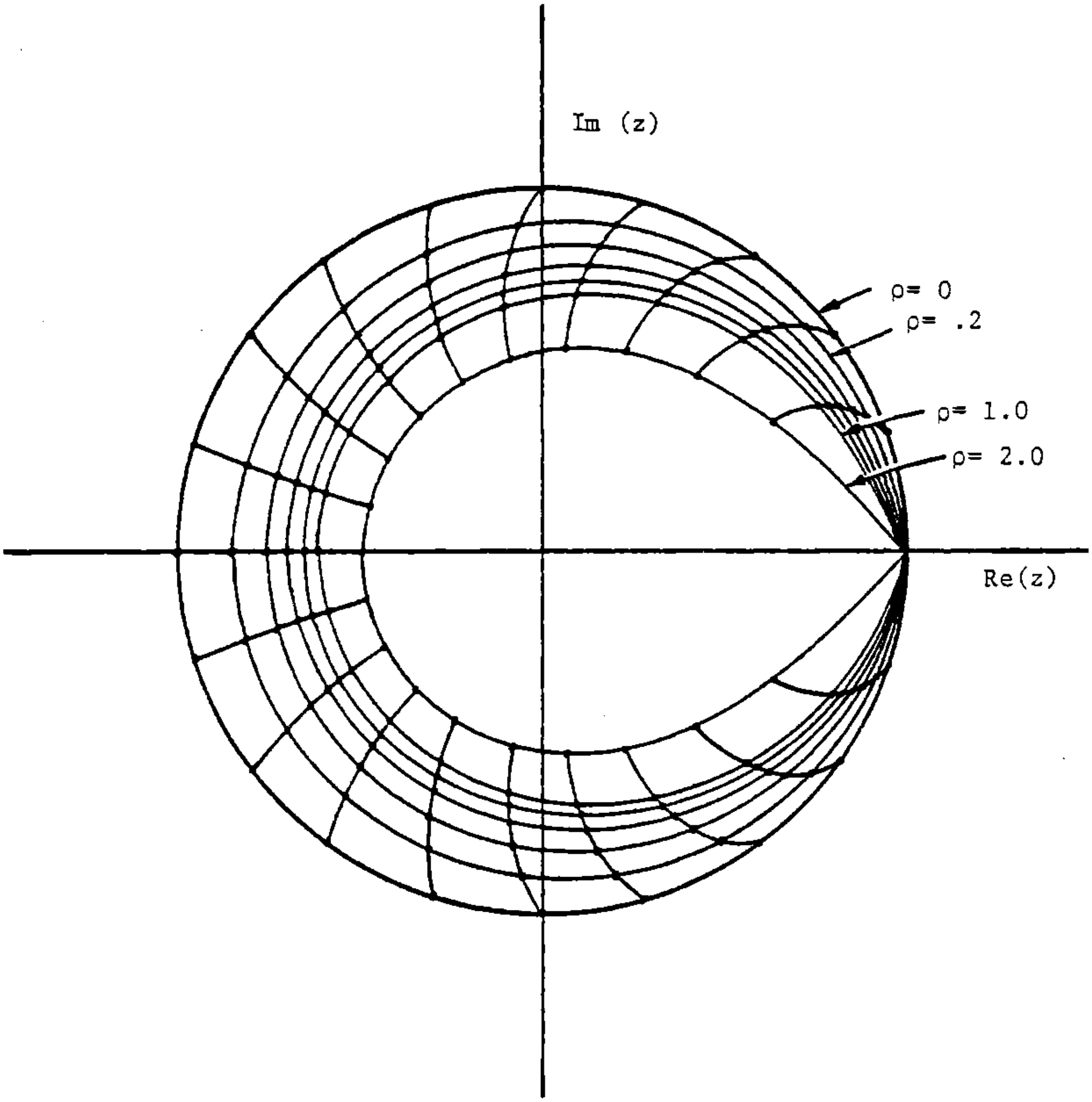
The third and final characteristic is that the c-1 roots (plus the root z=1) fall along a continuous contour. To see this, consider the case where $\rho=0$, implying that $Y(z)=1$, whereby all the roots are located uniformly along the unit circle. Since $Y(z)$ is a continuous function of ρ, its roots must also be continuous functions of ρ and hence one would expect that the contour on which they fall as ρ increases would remain reasonably well-behaved. What we cannot say, however, is how the contour behaves as $\rho \to 1$, since Rouche's theorem does not even guarantee their existence in the unit circle for $\rho > 1$. We can point out that $Y(z)$ is perfectly continuous for $\rho > 1$ and hence we would not expect any major problems, but this is a question that can only be answered numerically.

Figures 4.1 through 4.3 show plots of the zeroes in each case for different values of ρ. Shown are both the contours along which all the zeroes lie for a given value of ρ, as well as the radial paths describing the movement of each zero as ρ is changed. For case III, ρ was changed by increasing the frequency of arrivals from each terminal without changing the size of the loads; the values of ρ are then those that resulted for a given arrival rate and load size. In the other cases, ρ is specified exogenously. In all cases, the contours along which the zeroes lie are extremely smooth and, in view of the differences in the three examples, surprisingly similar. This also made it very easy to provide good starting points for each value of k. In fact, it is these contours that suggested the initialization procedures given in the root finding algorithms in appendix E.1. By using polar coordinates, where the k^{th} root may be expressed as $z_k = r_k \exp(i\theta_k)$, we could take advantage of the fact that as k is increased in equation 4.7, r_{k+1} becomes only slightly smaller than r_k, and $\Delta\theta_{k+1} = \theta_{k+1} - \theta_k$ is only slightly smaller than $\Delta\theta_k$. In most



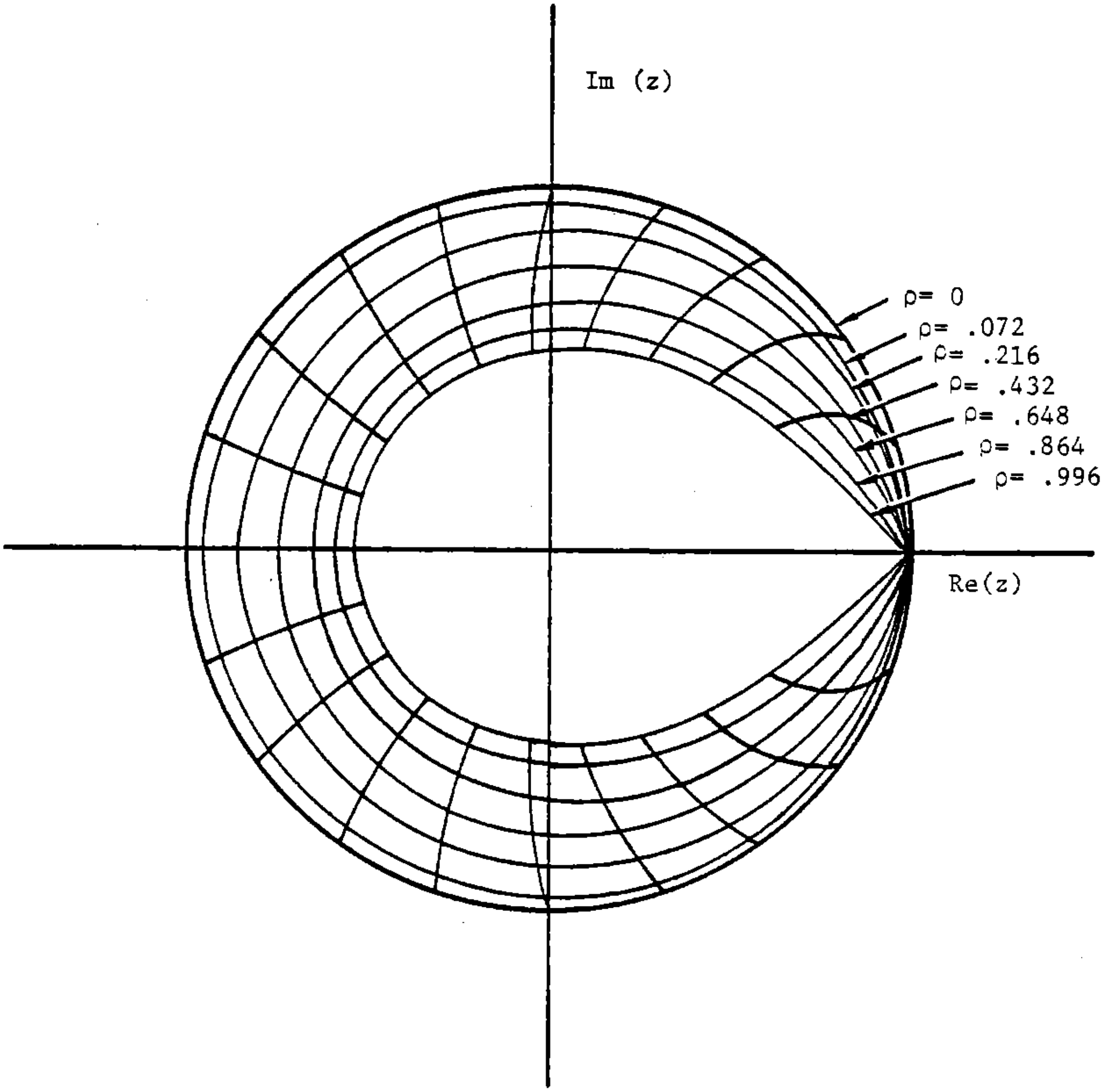
Location of zeroes for case I: simple Poisson arrivals and deterministic headways

Figure 4.1



Location of zeroes for case II: simple Poisson arrivals and Erlang distributed headways

Figure 4.2



Location of zeroes for case III: compound Poisson arrivals and deterministic headways

Figure 4.3

cases, the last few zeroes were each determined in only two iterations of the Newton-Raphson algorithm. What is particularly interesting, however, is that the zeroes are well-defined for values of ρ greater than 1, implying that the root finding procedure will be very stable for values of ρ near 1. The algorithm used for case I proved to be unstable for values of ρ less than .15, although the method used for cases II and III, both of which performed satisfactorily for all values of ρ, could easily be adapted to case I. We should point out that queues with $\rho < .5$ exhibit virtually no queueing (when c=20) and hence can accurately be solved by assuming Q(z) = Y(z) without finding zeroes.

It is significant that case III appeared to pose no special problems. As complex as Y(z) is, its Jacobian is even more complicated suggesting that the contour containing its zeroes might not be quite as "nice" as was found for the simpler case. The experiments undertaken thus far, however, indicate that the problem is still well-behaved and that each root can be found with approximately as many iterations as required for the simpler problems, although with considerably more computational effort per iteration.

We can tentatively conclude from these investigations that the root finding procedure is very stable and should not pose any problems in this respect. We now consider the issue of numerical accuracy.

4.1.2 Accuracy

The factors influencing the accuracy of numerical methods can be divided between those that are controllable and those that are not.

The controllable element in iterative methods is the stopping rule which can guarantee, up to a point, almost any level of accuracy. The stopping criterion must be chosen carefully to ensure that the desired level of accuracy is in fact being obtained, without performing calculations that are unnecessarily accurate. A more important problem is the uncontrollable element introduced by computer roundoff error which in most cases is difficult to detect or correct. This factor puts a limit on the accuracy of iterative search procedures such as root finding algorithms, but is much more troublesome in other areas such as solving systems of simultaneous linear equations. Thus while search procedures tend to be self correcting (if round-off produces an error in one direction, the procedure corrects for this in the next step), other methods such as Gaussian elimination tend to accumulate and compound errors. A similar problem exists in the inversion formulas for the queue length probabilities (eq. 4.26) where q_{i+c} depends on $q_o, \dots, q_i$. Since these errors are basically unpredictable it is necessary to use other methods to determine under what conditions these errors become significant.

An alternative method for finding the steady state vector $\{q\}$ is the method of numerical convolutions mentioned briefly in chapter 3. Here the probability distribution of Q_{n+1} is calculated using the recursion:

$$Q_{n+1} = R_n + Y_n \qquad 4.8$$

where we assume the Y_n are i.i.d.. Thus we have:

$$q_i^{n+1} = \sum_{j=0}^{i} r_j^n \, y_{i-j} \qquad i=0,1,\dots \qquad 4.9$$

Since the vector $\{q_i\}$ is theoretically of infinite length, we truncate it by choosing the smallest constant M such that:

$$1 - \sum_{i=0}^{M} q_i < \varepsilon_1 \qquad 4.10$$

In all the experiments, ε_1=.0001 is assumed. The recursion in 5.8 is carried out repeatedly until:

$$\sum_{i=0}^{M} |q_i^{n+1} - q_i^{n}| < \varepsilon_2 \qquad 4.11$$

The parameter ε_2 is experimented with below, since it can have a considerable impact on both the accuracy of the result and the computational effort required.

At each step in the process, the vector $\{q_i\}$ is normalized to prevent an accumulation of roundoff errors. As is easily seen, the method is self-correcting and hence the only important sources of error are reflected in the choice of ε_1 and ε_2. We point out again the extreme simplicity of this approach compared to the transform method and, while there are some drawbacks, it is felt nonetheless to represent a very powerful tool and one not to be slighted in the face of more elegant techniques.

In appendix E.1, the root finding algorithms are formulated using the stopping parameter ε which we now label ε_3. What we propose is to test the sensitivity of the transform approach to ε_3 and the numerical convolutions method to ε_2 in terms of their effect on the accuracy of the first and second moments and on the accuracy of the probability distributions. The outcome of these analyses are fixed choices of ε_2 and ε_3 which we then use in the following section comparing the two methods in terms of computational speed. All these experiments assume simple Poisson input and deterministic headways, corresponding to case I in the previous section.

The analysis of the error in the moments begins by first computing the moments with each method using $\varepsilon_2 = \varepsilon_3 = 10^{-6}$, and then assuming that these are the exact values. The fact that the estimates produced by each method are slightly different at this level of accuracy may be partly a result of the choice of ε_1 for the numerical convolutions procedure. In any case, ε_2 and ε_3 are then increased, and the mean and standard deviation of the length of the queue are compared to the "true" values for an estimate of the relative error.

The results of these tests are shown together in figure 4.4 where ε is given on a horizontal log scale and the relative error, RE, on the vertical. The relative error is computed using the formula $RE = 100 \cdot (x - \hat{x}) / \hat{x}$, where $\hat{x}$ is the most accurate estimate of the mean or standard deviation and x is the estimate corresponding to a particular value of ε_2 or ε_3. The runs are conducted for $\rho=.7$, where virtually no queueing occurs, and $\rho=.95$, where on the average 6 or 7 units are left over following each departure and where the probability a randomly chosen unit left on the first outbound vehicle is .67. For $\rho=.95$, the relative error produced by the transform approach is under 1 percent for all values of ε_3, the same occurring for $\rho=.7$ when the numerical convolutions technique is used. For this particular

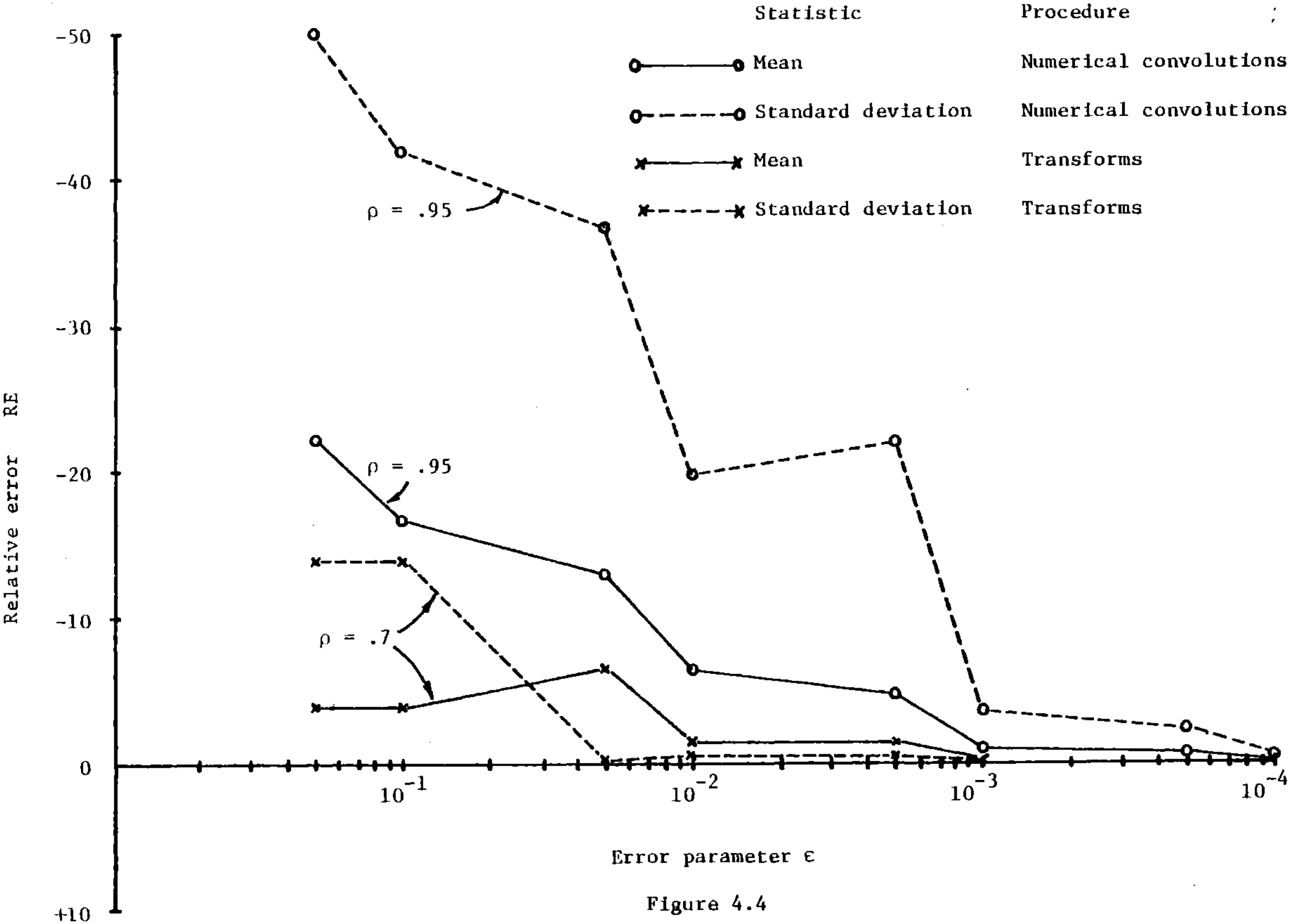


Figure 4.4

Error analysis of numerical convolutions and transforms problem, the error produced by assuming that no queueing occurrs ($Q(z) = Y(z)$) for ρ=.7 is also approximately 1 percent. Based on the figure, it would appear that using ε_3=.001 would be a safe, almost conservative assumption for the transform approach and hence this is used in section 4.1.3. On the whole, the results in terms of predicting moments using transforms are surprisingly, and encouragingly, insensitive to ε_3.

The numerical convolutions technique poses a different problem altogether. Like simulation, it is extremely sensitive to ρ, requiring a large number of successive iterations to reach steady state for even moderate levels of congestion. In addition, for higher values of ρ, it is very sensitive to ε_2, especially when estimating the variance of the steady state distribution. The nature of the algorithm implies there is always some downward bias in the estimates of the mean and variance which, for a given value of ε_2, can produce arbitrarily large (absolute) errors as $\rho \to 1$. In order to choose an appropriate value for ε_2, then, it is necessary to set an upper bound on ρ. For c=20, ρ=.95 produced a 33 percent chance of missing the first departure with an average of 6 units left over following each departure. While this is felt to represent a significant level of congestion, it is not wholly unrealistic and hence is taken to be the upper bound on ρ. Referring to figure 4.4 again, we see that for ε_2=.001 errors in the estimates of both moments are within 2 - 3 percent. While this is higher than that produced by the transform method, it is acceptable for planning purposes and hence will be adopted in the following sections for comparisons on the basis of numerical efficiency.

4.1.3 Computational efficiency

The question of how fast a queue can be solved is of little interest in the analysis of single queues and for this reason has received almost no attention in the queueing literature. When we consider the problem of modelling networks with 1000 links or larger, each with its own queue, then computational efficiency takes on much greater importance. The issue is further magnified if we consider the possibility of incorporating the performance model as part of a search procedure requiring repeated solutions of entire networks under different operating strategies. In this section we compare transforms and numerical convolutions with respect to the time each requires to find the mean and variance of the steady state queue distribution. The comparisons are made assuming simple Poisson input and deterministic or Erlang distributed headways. The execution time for the numerical convolutions method includes the time required to derive the distribution of Y. This poses a negligible burden for the case of deterministic headways but requires the convolution of K geometric distributions for Erlang distributed headways (K=10 is assumed here). If arrivals are in groups, then the distribution of the size of each group would have to be repeatedly convolved, adding substantially to the overhead. The tests being conducted here are therefore case specific and serve only as an indication of the relative efficiency of the two methods.

Each program was executed 50 times and timed using an internal clock which excluded any read and write statements. Both routines were run over a range of values of ρ; the transform approach proved almost

Table 4.1

Comparison of execution times using transforms

and numerical convolutions

	Deterministic headways			Erlang Headways		
ρ	P_m	Numerical convolutions	Transforms	P_m	Numerical Convolutions	Transforms
. 2	.000	.008 secs	.003 secs	.000	.013 secs	.012 secs
. 6	.002	.014 secs	.003 secs	.006	.042 secs	.012 secs
. 8	.034	.036 secs	.003 secs	.205	.097 secs	.012 secs
.85	.067	.050 secs	.003 secs	.301	.148 secs	.012 secs
.90	.134	.089 secs	.003 secs	.431	.241 secs	.012 secs
.95	.318	.301 secs	.003 secs	.601	.572 secs	.012 secs

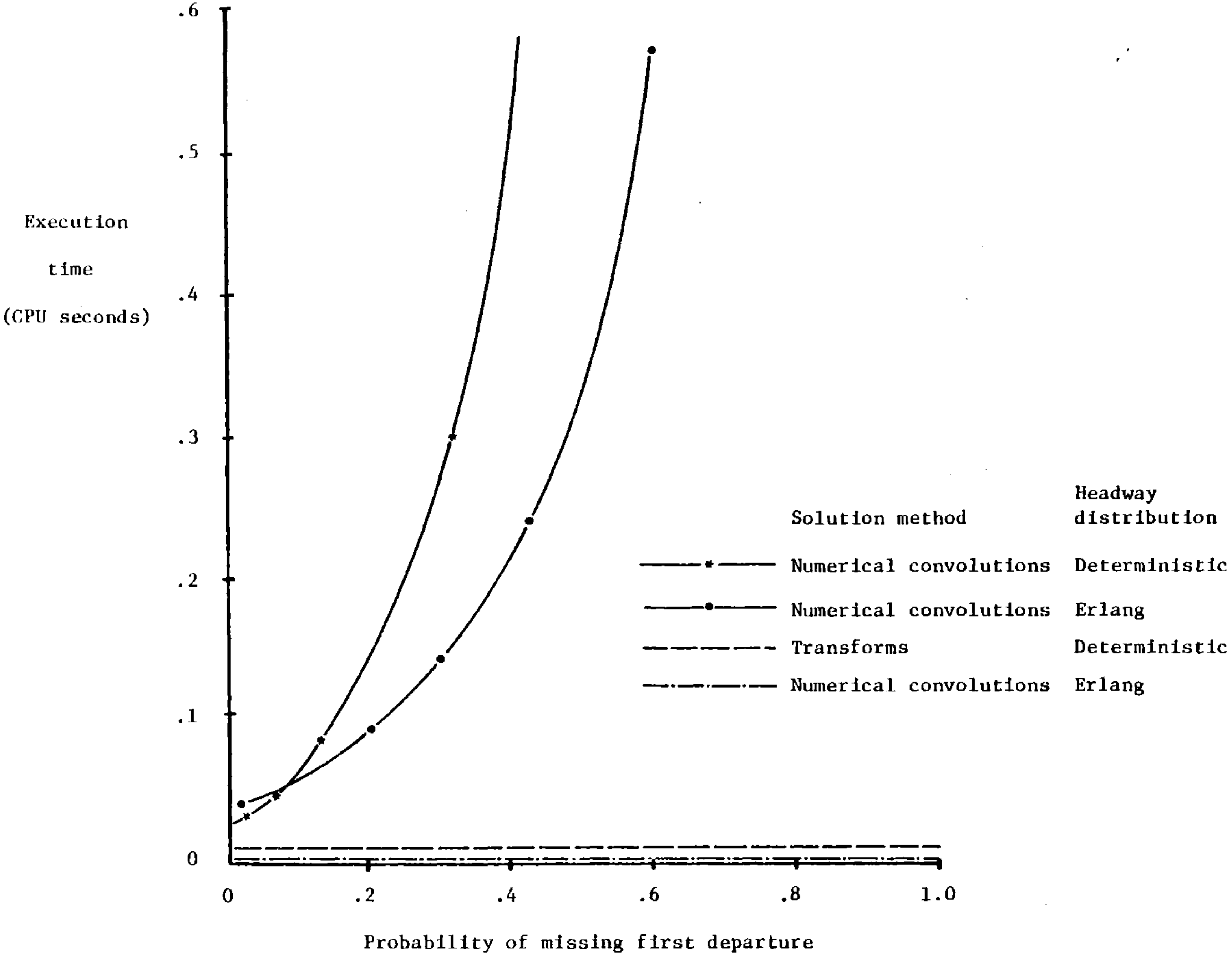

Execution time (CPU seconds)
.6
.5
.4
.3
.2
.1
0
0
.2
.4
.6
.8
1.0
Probability of missing first departure
Solution method
Headway distribution
—*— Numerical convolutions Deterministic
—●— Numerical convolutions Erlang
— — — Transforms Deterministic
—·—·— Numerical convolutions Erlang

Figure 4.5

completely insensitive to ρ while the reverse was true for the numerical convolutions method for values of ρ greater than .8. The results of these runs are summarized in table 4.1 and also plotted in figure 4.5 where, rather than putting execution times versus ρ, they are plotted against the probability P_m a random unit will miss the first departure. This level of service parameter conveys more intuition regarding the degree of congestion than ρ. The plot shows that the transform approach is considerably faster for all but the least congested situations. When P_m is close to 0 (more specifically, less than .05) the efficiency tests become meaningless since virtually no queueing is occurring, and we can accurately assume Q=Y. Since most networks are likely to have a good portion of transfer points where supply is much greater than demand, considerable savings can be realized by using "light traffic" approximations.

What is perhaps more important than the high relative efficiency of transforms is its efficiency in absolute terms. For example, if transforms were 10 times faster than numerical convolutions, but still required 1 minute of CPU time, then it is unlikely that they would be of much use in large scale applications. The fact that run times are on the order of .01 seconds (using transforms) is very encouraging and suggests that the techniques may be efficiently incorporated into network models. If the probabilities $q_0, \ldots, q_{c-1}$ are desired (say for finding P_m), then the polynomial expansion algorithm (described in section 4.2) can be used, requiring only .0012 seconds for c=20 (based on 100 repetitions of the procedure). Of course, as Y(z) increases in complexity so does the execution time. For example, if Y(z) is given by equation 4.4, representing bulk arrival for 10 different terminals, then the execution time jumps to 0.29 seconds. This can probably be reduced first by using equation 3.67 instead of 3.69, where the former expresses the transform of an outbound load size in terms of q_o, ..., q_{c-1} while the latter (used in equation 4.4 and in the root finding routine) uses the zeroes directly. Equation 3.67 is easier to evaluate, and the effort required to find the unknown probabilities is minimal. A much more significant reduction can be obtained if we can replace the individual streams from each terminal, each with its own load size transform, with a single stream with a load size distribution which reflects the combined distribution of the individual streams. Since the execution time for finding roots increases approximately linearly with the number of streams being superimposed, we may realize an order of magnitude improvement if the ten streams in equation 4.4 are replaced by one stream.

4.2 Inverting the transform

A common feature of most papers in queueing theory is that of ignoring the fact that while most transforms cannot be inverted analytically, they can be inverted numerically. The purpose of this section is to indicate how such inversions can be performed and to propose some new results in this area.

The task of inverting transforms of bulk service queues must proceed in two stages. First, the probabilities q_o, ..., q_{c-1} must be computed using the zeroes found from the denominator. Second, the remaining probabilities q_c, q_{c+1}, ..., must be calculated from the first c probabilities using a set of recursive formulae. For each stage, two methods for performing the calculations are described. For the first stage, the method commonly referred to in the literature finds the first c probabilities by setting up a system of simultaneous linear equations (see, for example, Ohno (1978). The details of the method are outlined in appendix E.2, but the general logic is as follows. Using the fact that the zeroes of the denominator of 3.9 must coincide with the zeroes of the numerator, we obtain the following set of equations using the zeroes z_j, j=0, ..., c-1:

$$\sum_{i=0}^{c-1} q_i(z_j^c - z_j^i) = 0 \quad j=0, \ldots, c-1 \qquad 4.12$$

The mechanics of setting up these equations is a little more involved than this, but the bottom line remains that we are forced to solve a c x c system of equations. A more efficient scheme, termed here the polynomial expansion technique, uses the zeroes to expand the numerator of equation 3.14 into a polynomial of order c. That is, define $\Psi(z)$ as:

$$\Psi(z) = \sum_{i=0}^{c} \psi_i z^i \qquad 4.13$$

$$= (c - \bar{Y})(z - 1) \prod_{i=1}^{c-1} \left(\frac{z - z_i}{1 - z_i}\right) \qquad 4.14$$

Now, observing that the polynomials in the numerators of 3.9 and 3.14 must be equal, we obtain the desired probabilities by equating coefficients of like powers of z as follows:

$$q_i = -\psi_i \qquad i=0, \ldots, c-1 \qquad 4.15$$

An efficient algorithm for calculating the polynomial $\Psi(z)$ is given in appendix E.3 and is shown to require approximately $c^2/4$ additions and multiplications. On the other hand, solving the system of simultaneous equations requires c^2 multiplications to set up and $c^3/3$ to solve, introducing a much greater risk of round-off error.

With the first stage completed, we now describe two methods for finding the rest of the probability vector, q_c, q_{c+1}, ... The first, while it has not been explicitly described in the bulk queueing literature, is a straightforward extension of a well known method for inverting the queue length transform for the M/G/1 queue (see Gross and Harris (1974), p. 229). Let P denote the one step transition matrix for the Markov chain $\{Q_n\}$, given by:

$$P = \begin{array}{c|cccc|} & 0 & 1 & 2 & \cdots \\ \hline 0 & y_0 & y_1 & y_2 & \cdots \\ 1 & y_0 & y_1 & y_2 & \cdots \\ 2 & y_0 & y_1 & y_2 & \cdots \\ \vdots & & & & \\ c & y_0 & y_1 & y_2 & \cdots \\ c+1 & 0 & y_0 & y_1 & \cdots \\ \vdots & 0 & 0 & y_1 & \cdots \\ & \vdots & \vdots & & \end{array} \qquad 4.16$$

where y_i = prob {i arrivals during the service period}. Observing that the steady state probability vector q must satisfy the following relation:

$$q = q \cdot P, \qquad 4.17$$

we obtain:

$$q_i = y_i \sum_{j=0}^{c-1} q_j + \sum_{j=c}^{c+i} q_j \, y_{c+i-j} \qquad 4.18$$

Solving for q_{c+i} gives:

$$q_{c+i} = \frac{1}{y_0} \left[q_i - y_i \sum_{j=0}^{c-1} q_j - \sum_{j=c}^{c+i-1} q_i \, y_{c+i-j} \right] \qquad 4.19$$

Given the probabilities $q_0, \ldots, q_{c-1}$, equation 4.19 can be used recursively to find $q_c, q_{c+1}, \ldots$.

A second procedure which has not appeared in the literature is obtained directly from the queue length transform. The method begins by writing the queue length transform as the ratio of two functions, $Q_1(z)$ and $Q_2(z)$ (representing the numerator and denominator) as follows:

$$Q(z) \;=\; \frac{Q_1(z)}{Q_2(z)} \qquad 4.20$$

Multiplying both sides by $Q_2(z)$ gives:

$$Q(z)Q_2(z) \;=\; Q_1(z) \qquad 4.21$$

The vector of probabilities q_c, q_{c+1}, ..., can be found by expanding both sides of 4.21 as polynomials and then equating the coefficients of like powers of z. For the case of bulk arrival queues with service in single units, the entire probability vector can be computed in this way. To demonstrate the procedure, consider the case of the $M/D^c/1$ queue, where service intervals are of constant length T. Equation 3.9 is then given by:

$$\sum_{i=0}^{\infty} q_i z^i = \frac{\sum\limits_{i=0}^{c-1} q_i \;(z^c - z^i)}{z^c \; e^{\lambda T(1-z)} \; -1} \qquad 4.22$$

Bringing the denominator on the right over to the left as shown in 4.21 and expanding $e^{-\lambda Tz}$ as a power series gives:

$$z^c e^{\lambda T} \left(\sum_{i=0}^{\infty} q_i z^i \right) \left(\sum_{j=0}^{\infty} \frac{(-\lambda T)^j}{j!} z^j \right) - \sum_{i=0}^{\infty} q_i z^i = \sum_{i=0}^{c-1} q_i (z^c - z^i) \tag{4.23}$$

Finally, changing the indexes in the first set of summations so that terms with z^{i+j} appear as z^m produces:

$$z^c e^{\lambda t} \sum_{m=0}^{\infty} \left(\sum_{\ell=0}^{m} \frac{(-\lambda T)^{\ell}}{\ell!} q_{n-\ell} \right) z^m - \sum_{i=o}^{\infty} q_i z^i = \sum_{i=0}^{c-1} q_i (z^c - z^i) \tag{4.24}$$

Noting that two polynominals are equal if and only if the coefficients of terms with like powers are equal, we obtain the following equations:

$$q_c = q_o e^{\lambda T} - \sum_{i=0}^{c-1} q_i \tag{4.25}$$

$$q_{i+c} = e^{\lambda T} \sum_{\ell=0}^{i} \frac{(-\lambda T)^{\ell}}{\ell!} q_{i-\ell} \qquad i = 1,2,\ldots \tag{4.26}$$

Thus, given q_0, ..., q_{c-1}, we may compute the rest of the probability vector using 4.25 and 4.26. There are, however, potential numerical problems with the use of either equation 4.19 or 4.26. Specifically, both formulas involve subtracting small numbers, the difference of which is then multiplied by a possibly very large number. This type of calculation tends to greatly magnify small roundoff errors and raises serious questions regarding the practical usefulness of either method. To get a sense of the numerical performance of the entire inversion process, several experiments were conducted comparing the probability vector calculated using inversion techniques to that computed using the numerical convolutions procedure.

Beginning with the first stage of the inversion process, the two methods for finding the first c probabilities were applied to an $M/D^c/1$ queue, using c = 20. The results, shown in Table 4.2, exhibit at worst four digit accuracy between the two approaches (for q_{19}) with increasing accuracy as i gets smaller. Similar results were obtained for other values of ρ. When c was increased to 40, however, the method using the simultaneous equations broke down entirely, yielding probabilities that were all negative. The polynomial expansion technique, on the other hand, performed quite well in addition to being computationally much faster.

Unlike the first stage, the recursive formulae needed to complete the second stage of the inversion proved to be so sensitive to roundoff errors as to make them practically useless. Sample results, for c = 4, ρ = .85 and c = 20, ρ = .9, are shown in Table 4.3 alongside the results obtained using the numerical convolutions approach. In both cases, there is reasonably good agreement until the inversion equations appear to suddenly break down, producing unusually high or negative probabilities. In both cases, double precision arithmetic was used with $\varepsilon_3=10^{-12}$ to ensure to maximum possible accuracy in computing the zeroes. Despite these precautions, the problem with equation 4.26, however, appears to be in the errors contained in the first c probabilities. The inversion formulae were applied to the case of c=1 where we know *a priori* that $q_0=1-\rho$. The results, shown in table 4.4, appear quite

Table 4.2

		q_i	
i	Simultaneous equations	Polynomial expansion	Numerical convolutions*
0	0.0000000	0.0000000	0.00000
1	0.0000002	0.0000002	0.00000
2	0.0000014	0.0000014	0.00000
3	0.0000083	0.0000083	0.00001
4	0.0000378	0.0000378	0.00004
5	0.0001375	0.0001375	0.00014
6	0.0004171	0.0004171	0.00042
7	0.0010859	0.0010860	0.00110
8	0.0024771	0.0024772	0.00251
9	0.0050310	0.0050314	0.00510
10	0.0092143	0.0092148	0.00934
11	0.0153765	0.0153772	0.01558
12	0.0235843	0.0235852	0.02389
13	0.0334960	0.0334970	0.03393
14	0.0443399	0.0443411	0.04491
15	0.0550247	0.0550260	0.05572
16	0.0643555	0.0643559	0.06516
17	0.0712893	0.0712914	0.07216
18	0.0751438	0.0751494	0.07604
19	0.0757223	0.0756714	0.07659
20	0.0732025**	0.0734454**	0.07406

* Using $\varepsilon_2 = 10^{-6}$

** Found using equation 4.25

Comparison of first 21 probabilities computed using numerical convolutions and transform inversion

Table 4.3

		q_i			
	$c = 4, \rho = .85$			$c = 20, \rho = .90$	
i	Eq. 4.26	Numerical convolutions	i	Eq. 4.26	Numerical convolutions
4	.14735	.14797	20	.07325	.07340
5	.12987	.13038	21	.06836	.06849
6	.10385	.10420	22	.06180	.06192
7	.07855	.07881	23	.05440	.05448
8	.05783	.05805	24	.04676	.04686
9	.04266	.04232	25	.03960	.03958
10	.03080	.03076	26	.03276	.03296
11	.02050	.02234	27	.02759	.02717
12	.01809	.01621	28	.02115	.02224
13	.02642	.01173	29	.02075	.01812
14	-.03954	.00849	30	.00884	.01473
15	.02705	.00614	31	.02336	.01195
16	.16738	.00442	32	-.00307	.00969
			33	-.02796	.00786
			34	.37999	.00637

Comparison of queue length probabilities found via transform inversion (equation 4.26) and numerical convolutions for bulk service systems.

Table 4.4

$c = 1 \quad \rho = .90$

	q_i	
i	Eq. 4.26	Numerical convolutions
0	.10000	.10496
1	.14596	.15306
2	.13764	.14406
3	.11505	.12001
4	.09380	.09735
5	.07625	.07857
6	.06198	.06327
7	.05039	.05082
8	.04096	.04072
9	.03330	.03253
10	.02707	.02591
	.	.
	.	.
	.	.
20	.00341	.00187
21	.00277	.00131
22	.00225	.00086
23	.00183	.00051
24	.00149	.00023
25	.00121	.00000

Comparison of queue length probabilities found via transform inversion (equation 4.26) and numerical convolutions for single service systems.

good and may easily be more accurate than that produced by numerical convolutions (compare the estimates of q_0 produced by both methods). Thus our inversion equations may be quite useful for single service operations with bulk or single arrivals, but are not recommended for bulk service queues.

We can conclude from these experiments that while we cannot fully invert queue length transforms, we can safely use the polynomial expansion method to find the first c probabilities. It is important to point out, however, that the first c probabilities can, by themselves, be quite useful. For example, it is computationally easier to use the transform as it is expressed in equation 3.9 using $q_0, \ldots, q_{c-1}$ than equation 3.14 using the zeroes. Also, operating statistics such as the probability a random unit will miss the first outbound departure require only the first c probabilities. If, however, the entire probability vector is required, then approximate, fitted distributions must be used, as is discussed in the next section.

4.3 Approximations for queue length distributions

In section 4.1 we show that the task of finding roots posed no numerical problems and could be executed both safely and efficiently. We did find, in section 4.2, that we could not completely invert the distribution, although most quantities of interest could be computed using the portion of the probability vector that could be found. In any case, the nature of the procedure is still relatively complicated and requires some knowledge of both transforms and complex variables. Thus, while the method may present few numerical problems, it is unlikely to attract any attention from many engineers simply as a result of the mathematical background required. In this section, we present a set of approximations that provide not only the first two moments of the steady state length of a queue in closed form, but also a method for estimating the distribution itself.

The approach consists of first approximating the first two moments of the queue distribution and then using this information to determine the parameters of a known distribution. We begin by observing that the moment formulas for the length of the queue (given by eqs. 3.17 and 3.18) can be broken into two parts, the first being a closed form function of the first three moments of Y, and the second being expressed in terms of our now familiar zeroes. Thus we may rewrite 3.17 and 3.18 as:

$$E(Q) = \psi_1(\overline{Y}, \overline{\overline{Y}}) + \phi_I(Y) \qquad 4.27$$

$$\text{Var}(Q) = \Psi_2(\bar{Y},\bar{\bar{Y}},\bar{\bar{\bar{Y}}},) + \phi_2(Y) \qquad 4.28$$

where:

$$\phi_1(Y) = \sum_{i=1}^{c-1} \frac{1}{1-z_1} \qquad 4.29$$

$$\phi_2(Y) = -\sum_{i=1}^{c-1} \frac{z_i}{1-z_i} \qquad 4.30$$

The functions $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are self-evident from 3.17 and 3.18. The functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are expressed as functions of Y since in principal the zeroes reflect the full distribution of Y as opposed to a few of its moments. The problem now is to replace $\phi_1(\cdot)$ and $\phi_2(\cdot)$ with simpler functions of key parameters. To do this, we assume simple Poisson arrivals to a scheduled departure queue with Erlang distributed headways with density function b(t):

$$b(t) = \frac{\Theta^r\,(\Theta t)^{r-1}\,e^{-\Theta t}}{(r-1)!} \qquad 4.31$$

The mean and variance of this distribution is $\frac{r}{\Theta}$ and $\frac{r}{\Theta^2}$ with coefficient of variation $C_b = \frac{1}{\sqrt{r}}$ By varying r, we can obtain distributions ranging from negative exponential headways (r=1) to deterministic headways by letting $r \to \infty$ while holding $\frac{r}{\Theta}$ constant. The transform of b(t) is:

$$B^*(s) = \left(\frac{\Theta}{\Theta+s}\right)^r \qquad 4.32$$

Assuming simple Poisson input, we have:

$$Y(z) = B^{*}(\lambda - \lambda z) = \left(\frac{\theta}{\theta + \lambda - \lambda z}\right)^{r} \qquad 4.33$$

With Y(z) defined thus, we find the utilization ratio ρ given by $\rho = \frac{\lambda \cdot r}{\Theta c}$. The parameters r and ρ can be used to define a family of distributions where ρ controls the mean and r controls the shape. We now assume that $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are functions of r and ρ only (for a fixed value of c). Figure 4.6 shows plots of these functions versus ρ for r = 1, 5, 10 and ∞, from which we can see that both functions are very smooth and well-defined for all values of both ρ and r. Interestingly, whereas the actual moments become infinite as $\rho \to 1$, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ tend to level off as $\rho \to 1$. This implies that ϕ_1 and ϕ_2 become quite small relative to ψ_1 and ψ_2 for ρ close to 1. On the other hand, as $\rho \to 0$, each set of curves converges to a single point independent of r. This behavior can be verified theoretically by observing that both the mean and variance must vanish for $\rho = 0$, implying that:

$$\lim_{\rho \to 0} \psi(\rho, r) + \phi(\rho, r) = 0 \qquad 4.34$$

This gives us:

$$\phi_1(0, r) = \frac{c-1}{2} \qquad 4.35$$

$$\phi_2(0, r) = \frac{c^2-1}{12} \qquad 4.36$$

This result could also have been obtained by observing that for $\rho = 0$, Y(z) = 1, and hence the zeroes must be located uniformly around the unit circle. These could then be substituted into 4.29 and 4.30 to obtain 4.35 and 4.36.

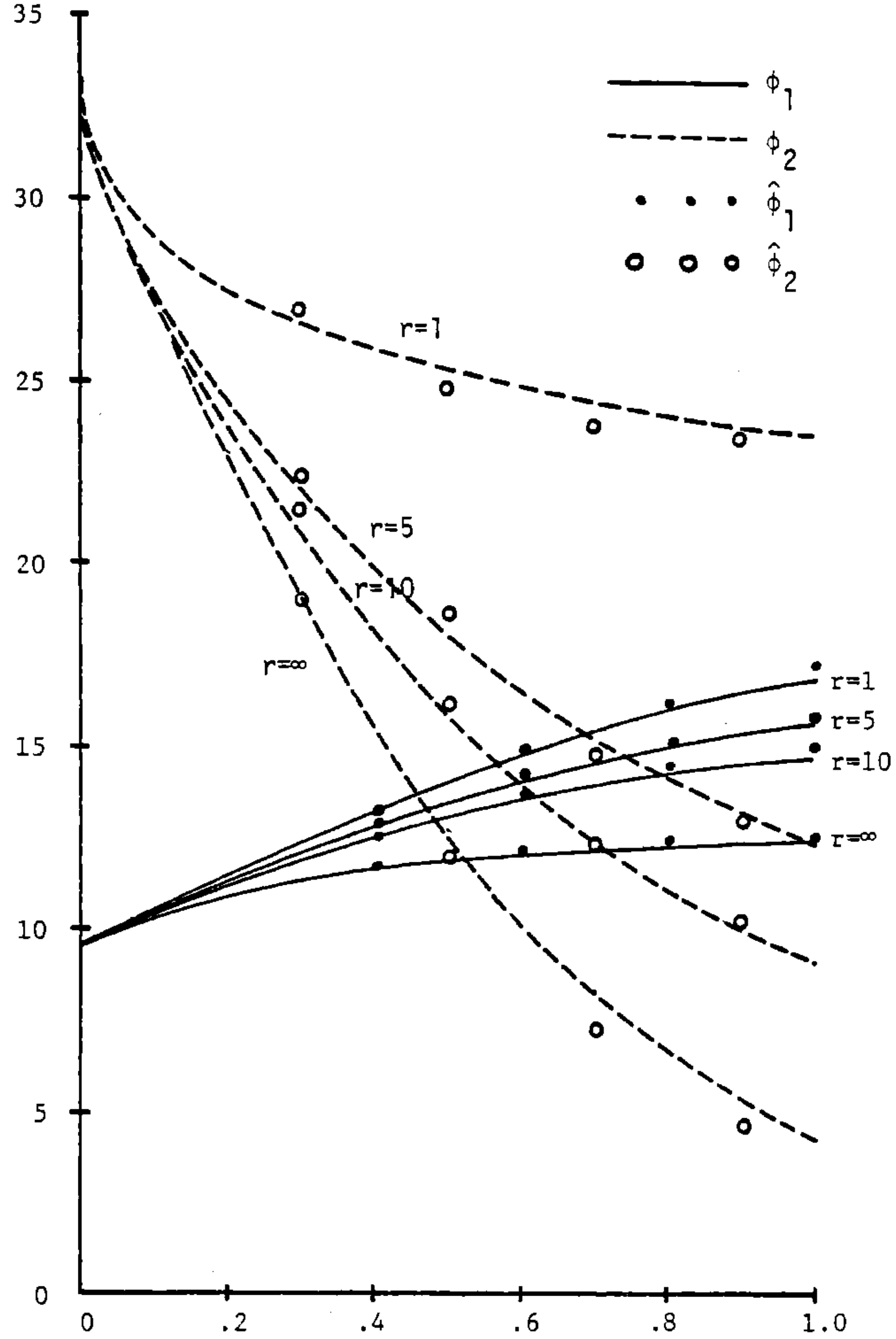


Figure 4.6

Evaluation of approximations for ϕ_1 and ϕ_2

Still assuming c = 20, the following expressions were specified for ϕ_1 and ϕ_2:

$$\phi(r, \rho) = \frac{c-1}{2} + \alpha_1^1 \rho + \alpha_2{}^1\rho^2 + \alpha_3^1 \rho/ \sqrt{r} + \alpha_4^1 \rho^2/ \sqrt{r} \qquad 4.37$$

$$\phi_2(r, \rho) = \frac{c^2-1}{12} + \alpha_1^2 \rho + \alpha_2^2 \rho^2 + \alpha_3^2 \rho/ \sqrt{r} + \alpha_4^2 \rho^2/ \sqrt{r} \qquad 4.38$$

Using a set of 30 points corresponding to ρ = .2, .5, .7, .8, .9 and .99, and r = 1, 5, 10, 20 and ∞, the following estimates were made using ordinary least squares:

$$\hat{\phi}_1(r, \rho) = 9.5 + 11.0468 \rho - 3.3974 \rho^2 - 4.2349 \rho/ \sqrt{r} - .3899 \rho^2/ \sqrt{r} \qquad 4.39$$

$$\hat{\phi}_2 (r, \rho) = 33.25 - 55.2158 \rho + 25.6323 \rho^2 + 30.6362 \rho/ \sqrt{r} - 10.6138 \rho^2/ \sqrt{r} \qquad 4.40$$

The functional specifications of $\hat{\phi}_1(\cdot)$ and $\hat{\phi}_2(\cdot)$ were motivated solely by the shape of the curves and apply only to scheduled departure queues with c = 20. The closeness of the fit is also illustrated in figure 4.6 where both functions have been evaluated at various points and shown as dots ($\hat{\phi}_1$) and circles ($\hat{\phi}_2$). Although we do not explicitly do so, both equations can be easily generalized to other values of c. Figures 4.7 and 4.8 show plots of ϕ_1 and ϕ_2 versus $\xi_1 = \frac{c-1}{2}$ and $\xi_2 = \frac{c^2-1}{12}$ for ρ = .95 and r = 5 and ∞. It is easily seen that ϕ_1 is almost perfectly

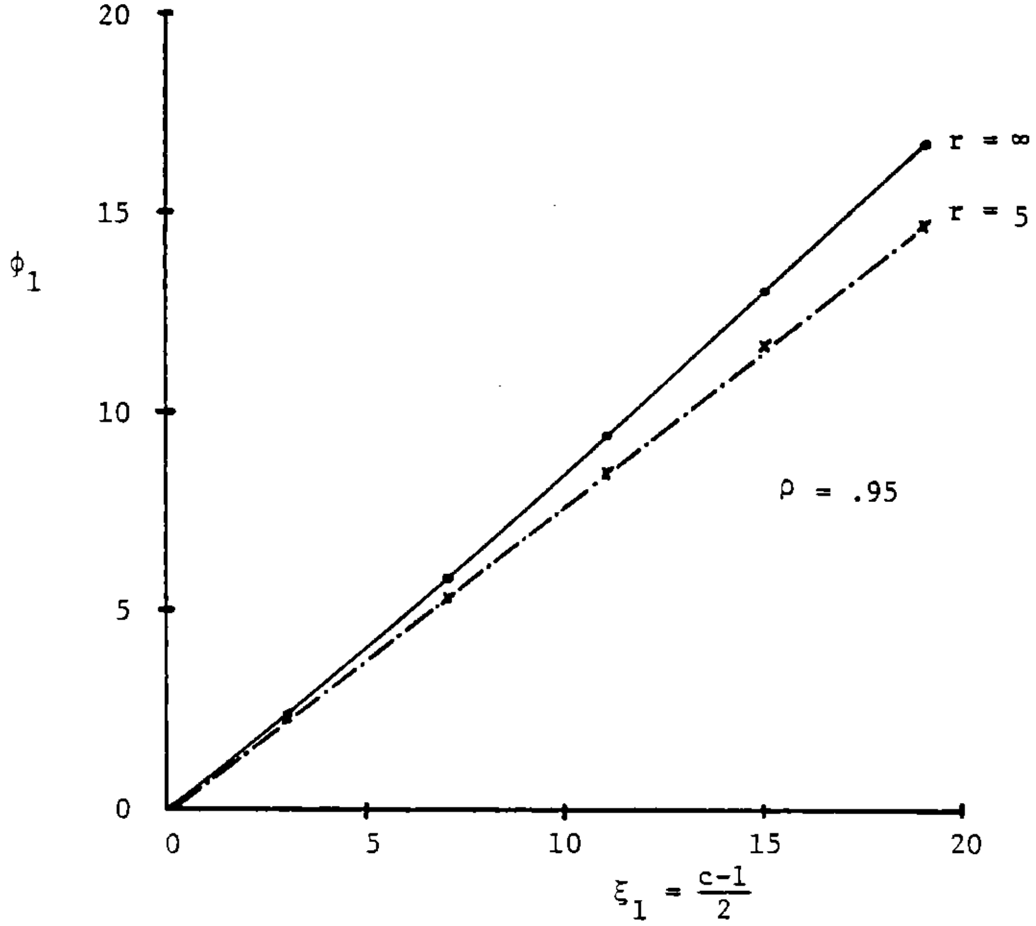


Figure 4.7

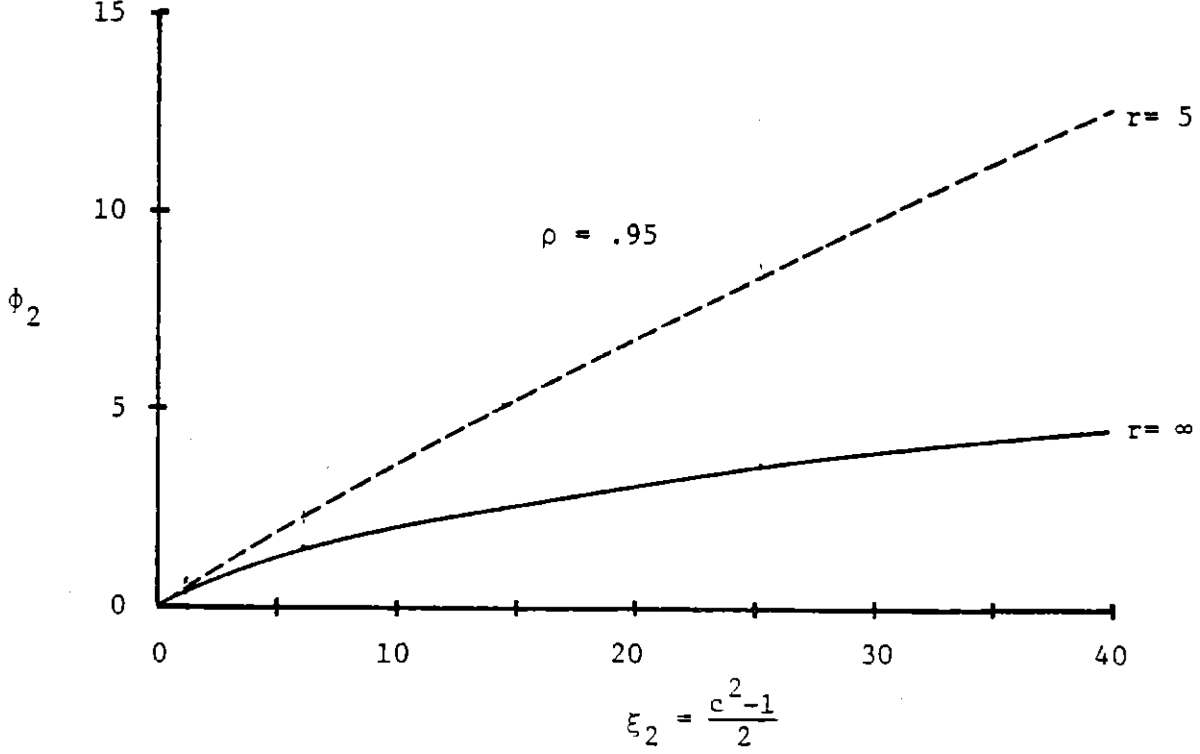


Figure 4.8

linear in ξ_1 and of course becomes even more so for smaller values of ρ. A linear approximation may also be used for ϕ_2 as well, although the fit is not quite as good and a quadratic term may be useful. A possible specification would be:

$$\hat{\phi}_2 = (\xi_2 + \alpha_1 \rho \xi_2^2)(1 + \alpha_2 \rho + \alpha_3 \rho^2 + \alpha_4 \rho / \sqrt{r} + \alpha_5 \rho^2 / \sqrt{r}) \qquad 4.41$$

Naturally 4.41 would have to be expanded, but all the coefficients are identifiable. Note that the specification retains certain desirable properties, namely that it is equal to $\frac{c^2-1}{12}$ when ρ = 0 and it vanishes for c = 1. A similar formulation could also be used to refine 4.39 using ξ_1. It would be desirable to compare the predictive accuracy of the approximations for headway distributions other than those out of the gamma family and for bulk arrival systems as well. It seems reasonable that equations 4.37 and 4.38 could be reformulated to incorporate the coefficient of variation for Y (instead of B), where the latter variable would simultaneously reflect variations in both the headway distribution as well as the sizes of incoming groups.

Once we have determined the moments of the distribution, the next problem is to fit an approximate distribution for the length of the queue. The family of distributions that are used can be looked upon as the discrete analog of the Erlang distribution; whereas the Erlang can be found by convolving negative exponential distributions together, we can convolve geometric distributions together. To add an additional parameter for greater flexibility, we consider as well the shifted geometric with p.m.f.:

$$p_i = (1 - \Theta)\Theta^{i-\gamma} \qquad i = \gamma, \gamma + 1, \ldots \qquad 4.42$$

$$0 \leq \Theta \leq 1,$$

The parameter γ shifts the distribution to the right and of course must be integer. The transform of the shifted geometric is:

$$P(z) = z^{\gamma}\left(\frac{1-\Theta}{1-\Theta z}\right) \qquad 4.43$$

Convolving $\{p_i\}$ K times, we obtain the approximate distribution for Q, denoted by $\hat{Q}$, with transform:

$$\hat{Q}(z;\, \Theta,\, \gamma,\, K) = \left\{z^{\gamma}\left(\frac{1-\Theta}{1-\Theta z}\right)\right\}^{K} \qquad \begin{matrix} K = 1,2\ldots \\ \gamma = 0,1,\ldots \end{matrix} \qquad 4.44$$

The transform of $\hat{Q}$ is used only for notational simplicity since we are interested only in its probability vector $\{\hat{q}_i\}$. The moments of $\hat{Q}$ are easily verified to be:

$$E(\hat{Q}) = K\left[\left(\frac{\Theta}{1-\Theta}\right) + \gamma\right] \qquad 4.45$$

$$\mathrm{Var}(Q) = \frac{K\Theta}{(1-\Theta)}^2 \qquad 4.46$$

The only problem keeping us from using 4.45 and 4.46 is that we have three parameters and only two equations, 4.39 and 4.40. In principle, we could develop a third equation for the third moment in a manner similar to that used to find the approximate equations for the first two moments. As a simplification, we will guess at γ and use 4.45 and 4.46 to find K and θ. $\hat{Q}$, of course, cannot be used to approximate any distribution, being restricted to those with a coefficient of variation less than 1. If necessary, we may generalize this approach to discrete versions of the hyperexponential or hyperstage distribution with transform:

$$Q(z;\ \Theta_i,\ \alpha_i,\ \gamma_i,\ K_i,\ M) = \sum_{i=1}^{M} \alpha_i \left\{ z^{\gamma_i} \left(\frac{1 - \Theta_i}{1 - \Theta_i z} \right) \right\}^{K_i} \qquad 4.47$$

where $\sum_{i=1}^{M} \alpha_i = 1.$

Letting $\gamma_i = 0$ and $K_i = 1$ we obtain the discrete hyperexponential distribution with a coefficient of variation greater than 1. For our problem, however, the simpler distribution ($M = 1$) will suffice.

Approximate distributions were fitted using equations 4.39 and 4.40 to find K and Θ for $r = 1$, 5 and ∞ and for different values of ρ. In each case, several values of γ were tested manually and the value which gave the best fit was used. Most of the time $\gamma = 1$ was the best and in all cases it worked quite well. For low values of ρ, however, $\gamma = 0$ tended to produce a slightly better fit while $\gamma = 2$ worked a little better for higher values of ρ. Typical values of K were on the order of 2 to 10 for $\gamma \geq 1$, but for $\gamma = 0$, K was usually much higher; in one example, the best fit was found using $\gamma = 0$ and $K = 1013$, suggesting the gain in accuracy might be more than offset by the additional computational requirements of convolving over 1000 geometric distributions together.

The results of these experiments are shown in figures 4.9 ($r = 5$) and 4.10 ($r = \infty$), where the solid lines represent the true distributions, computed using numerical convolutions, and the dashed lines the approximation. The excellent fit is self-evident. The distributions for $r = 1$ are not shown since the true distribution is in fact geometric (as shown by Bailey (1954); see also Kleinrock (1975, p. 139)) and hence the approximate distribution is almost exact, using $K = 1$ and $\gamma = 0$. Somewhat ironically, the approximate approach may be more accurate in this case than the "exact" approach using numerical convolutions. The reason is that as a result of the high variability of the headway distribution, there were extremely long queues even for $\rho = .5$, and, as we found earlier, the latter technique produced significantly downward biased estimates of the variance under high levels of congestion. To compare the two approaches under such conditions would therefore be misleading.

On the basis of these results we can conclude that transform analysis of bulk queues may provide useful directions for developing extremely accurate approximations that are easy to use and implement. What transforms do is provide an indication of the true functional form for the mean and variance and leave a much smaller and more manageable component to be estimated empirically. This is especially useful in estimating variances since it is unlikely that we would ever be able to guess at the function in equation 3.18. We should also point out that the approximations developed here for the length of the queue may be used directly to find the moments of the waiting time using equations 3.100 - 3.104.

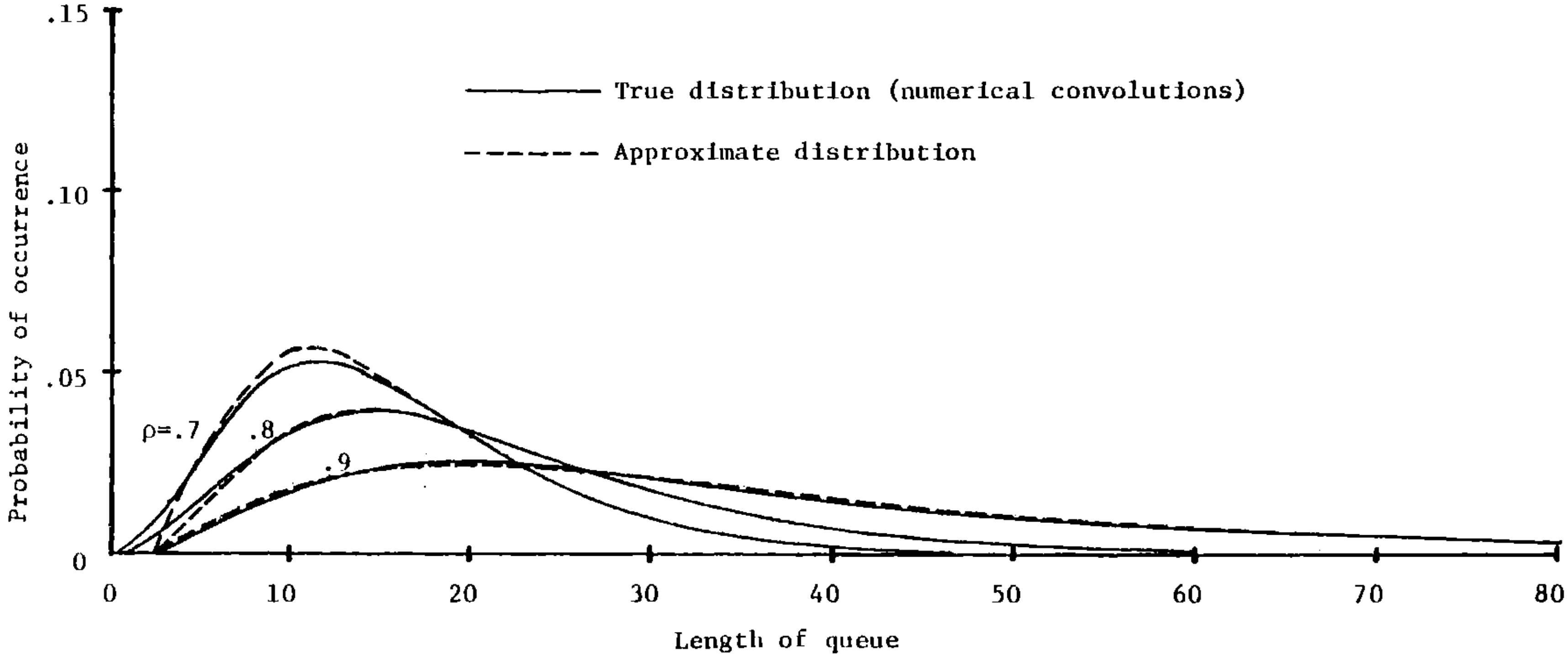


Figure 4.9

Comparison of exact and approximate queue length distributions

Erlang distributed headways - r=5

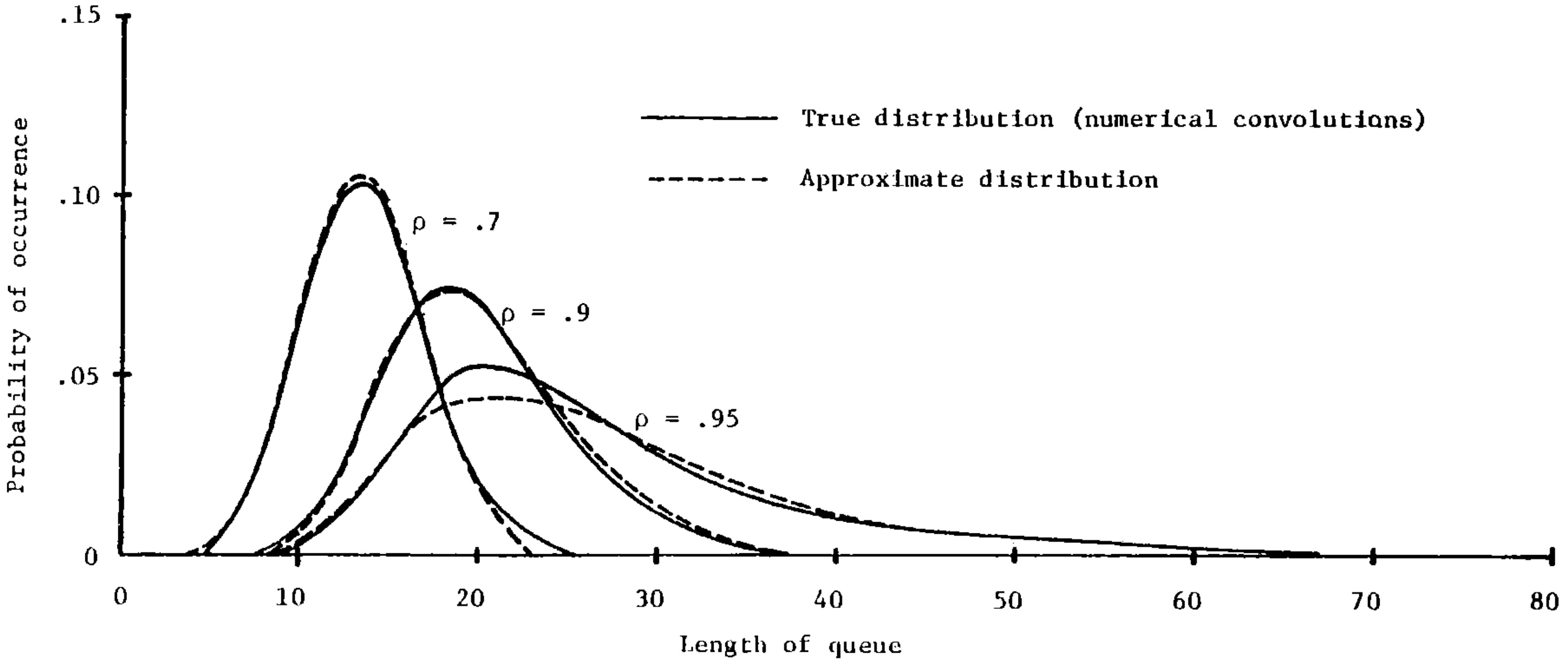


Figure 4.10

Comparison of exact and approximate queue length distributions

Deterministic headways

4.4 Summary

This chapter provides the information needed to obtain numerical results from the transforms derived in chapter 3. Specific algorithms are detailed in appendix E, with research on their numerical performance reported in sections 4.1 and 4.2. The tests indicate that there are no problems with finding the zeroes needed to solve the transforms, an observation that runs contrary to several statements made in the recent literature. Also, given the zeroes, it is possible to safely and efficiently find the first c probabilities which are useful both for computational reasons as well as computing certain operating statistics. It does not appear possible, however, to compute the rest of the probability vector for bulk service queues due to the numerical sensitivity of the recursive formulae used to perform the calculations.

Section 4.3 describes a novel approach for approximating the moment formulas in closed form and demonstrates the accuracy of the method in the case of a specific bulk queueing problem, the $M/E_r^c/1$ queue. The procedure is based on the observation that the term containing the zeroes is extremely smooth and therefore easily approximated. These formulas are then used to fit approximate distributions for the length of the queue which are shown to be very accurate. The ease with which such approximations can be used more than offsets any minor loss in accuracy, and further research expanding the moment formulas to bulk arrivals would be useful.

The use of approximations, of course, can extend beyond simply eliminating the need to calculate zeroes. From a more practical perspective, the most serious errors contained in the moment formulas will arise from violations of the underlying assumptions, most notably, but not exclusively, those regarding steady state and Poisson arrivals.

The purpose of this chapter has been to demonstrate how the transform results presented in chapter 3 can be implemented insofar as obtaining numerical solutions is concerned. All the results reported in this chapter, however, apply only to the case of simple or compound Poisson arrivals. In chapter 5, we turn to the problem of evaluating the Poisson arrival process as an approximation for more general arrival processes.

Chapter 5 Approximating the Arrival Process for $G/G^c/1$ queues

One of the most important assumptions needed for the analytic solution of bulk queueing systems, and which has been used throughout chapters 3 and 4, is that of Poisson arrivals. In this chapter, a series of experiments are presented which test the validity of this assumption and to determine the factors which influence its accuracy. In addition, a methodology is outlined for approximating bulk service queues with general arrival processes.

From a practical perspective, the use of the Poisson arrival process raises two important questions, namely when can an arrival process be approximated by a Poisson and what can be done when this approximation breaks down. Very little work has appeared that addresses either problem adequately, and hence a set of experiments are presented which provide a number of insights regarding the use of Poisson arrival processes. The results of these experiments are divided into two sections. In the first, section 5.1, a range of non-Poisson arrival processes are simulated and analyzed to determine the extent to which the process appears Poisson. The simulations are representative of situations that might arise in a transportation network. The second set of experiments, presented in section 5.2, repeat the simulations in the context of an actual bulk queue. Statistics are gathered on the mean and variance of the queue, which are then compared to the predicted mean and variance if Poisson arrivals are assumed. The results of these tests indicate that the Poisson arrival approximation is not robust, which motivates the discussion in section 5.3. There, a review of recent efforts aimed at approximating arrival processes is presneted, and a methodology for approximating $G^x/G^y/1$ queues is described and illustrated.

5.1 Analyzing general arrival processes

The problem of deciding whether or not an arrival process is Poisson is basically unsolved. Snyder (1975) discusses certain conditions that must be satisfied and suggests that these might be used to characterize the process qualitatively. As he points out, most of the literature on the statistical analysis of arrival processes is concerned with parameter estimation given that the process is Poisson. Gross and Harris (1974) describe several statistical procedures for testing if an observed interarrival time distribution is negative exponential, but ignore the possibility of correlations between successive interarrival times. Notwithstanding the fact that the statistical procedures used are not very powerful, the lack of independence between interarrival times can seriously affect the results.

It is reasonable to conjecture that the question as it is posed is by itself unanswerable. That is, the real question is whether a particular arrival process can be approximated as being Poisson and give good results for a particular problem. The purpose of this section is to present several possible statistics that might be used to characterize an arrival process. Section 5.2 then looks at a particular queue to provide an indication of the relationship between the values of the statistics and the accuracy of the Poisson approximation in estimating means and variances.

Two "views" of an arrival process are used in this section. The first, termed the microscopic view, is based on the usual test for a process to be Poisson which is that the successive interarrival times are i.i.d. with a negative exponential distribution. Thus one approach is to compute the correlation coefficient between successive interarrival times and construct the distribution. The second view, termed the macroscopic approach, looks at the number of arrivals over successive periods of time. This is equivalent to studying the random variables $\{Y_n\}$, where the periods of time represent service intervals. If each time period is of fixed length, then the sequence $\{Y_n\}$ must be i.i.d. with a Poisson distribution if the arrival process is Poisson. Hence another way of testing if a process is Poisson is to compute the correlation coefficient between each $\{Y_n\}$ and construct the distribution for Y_n.

To illustrate the use of these tests, as well as gain a number of insights into non-Poisson processes, a set of experiments were designed which focused on the superposition of independent arrival streams. In transportation networks, vehicles may arrive independently from a number of other terminals. The process describing the arrival of all vehicles, therefore, is a superposition of the arrival streams from each of the individual terminals. The problem is similar to that depicted in figure 5.1. The question, then, is under what conditions a superposition process can be approximated as a Poisson. It is well known that the superposition of two processes is exactly Poisson if and only if the two component processes are also Poisson. In fact, a superposed process is renewal if and only if all the component processes are Poisson. However, it has also been shown by Khintchine (1960) that the superposition of N independent renewal processes tends, as $N \to \infty$, to a Poisson process (see also Cox and Smith (1954) who first considered the case of N identical processes, and the discussion in chapter 6 in Cox (1962)). The problem here is that renewal processes are rare, and are especially unlikely to occur in transportation.

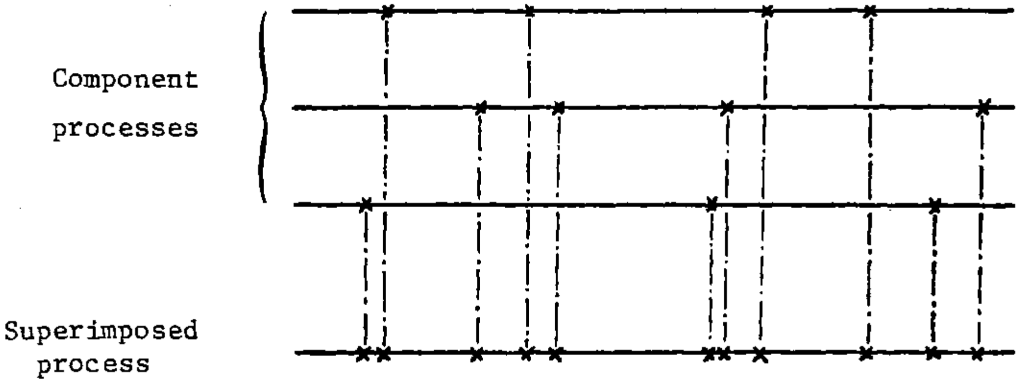


Figure 5.1

Illustration of a superposition process

The simulations that were run look at not only the number of processes being superimposed but also the nature of the component processes. Section 5.1.1 describes the details of the simulation and presents the results of the microscopic view of a process; section 5.1.2 presents the macroscopic view.

5.1.1 The microscopic view

The simulation experiments were conducted by superimposing N independent component process, representing (possibly) the arrival of vehicles from N separate terminals. Different runs were made by changing the characteristics of the component processes which can be divided into two general categories, namely renewal and nonrenewal. Renewal processes were considered as representing the most favorable conditions for producing an approximate Poisson superimposed process (i.e. if, for a given value of N, the superimposed process is not approximately Poisson and the component processes are renewal, then making the component processes non renewal would not improve the approximation). Non-renewal processes were formulated to reflect the presence of a schedule, where actual arrival times were assumed to be uniformly distributed around a scheduled arrival times. This group can be further subdivided on the basis of two criteria which describe the relationship between the component processes. The first criteria is whether the processes have the same frequency of arrivals; if so, the processes are termed phased, and unphased means the frequencies are different. The second criteria is that if the processes are phased, then we may distinguih between those where the scheduled arrival times are the same or not. If so, the system is called coordinated, and uncoordinated implies the scheduled arrival times are independent. Of course, it is impossible to have coordinated service if the process is unphased. An example of phased service would be once an hour service from each terminal. If arrivals are once an hour on the hour, then the service is coordinated.

A hierarchy of the different types of services is provided in Table 5.1. Each of the non-renewal categories can be seen to represent a specific type of scheduled service. In total, four different types of processes can be identified, namely renewal, coordinated, uncoordinated and unphased. In using these names, it should be understood that the last three are all nonrenewal, and the coordinated and uncoordinated process are both phased. We may represent the nonrenewal cases by defining t_n^i to be the arrival time of the n^{th} vehicle from the i^{th} terminal where:

$$t_n^i = t_o^i + n/\lambda_i + \varepsilon_{in} \qquad \begin{array}{l} i=1, \ldots, N \\ n=1, 2, \ldots \end{array} \qquad 5.1$$

and where: t_o^i is the time origin for the i^{th} vehicle

λ_i is the arrival rate of vehicles from terminal i ($1/\lambda_i$ is the headway)

ε_{in} is an error term

For phased service, it was assumed that $\lambda_i = \lambda = 1/(5N)$ for all i, while for coordinated service $t_o^i = 0$ for all i. Letting $\overline{\tau} = 1/\lambda$, we assume for uncoordinated service that t_o^i is uniformly distributed between 0 and $\overline{\tau}$. For unphased service, it was assumed that $\tau_i \sim U(2.5N, 7.5N)$, where $\lambda_i = 1/\tau_i$. In conducting these experiments, the arrival rates of the component processes were always adjusted so that the combined arrival rate remained approximately constant regardless of N, thus allowing us

Hierarchy of component process

I. Renewal - successive interarrival times are i.i.d.

II. Non-renewal - arrivals occur on or around scheduled arrival times which are evenly spaced

   A. Phased - each of the N arrival processes have the same frequency

      1) Coordinated - each of the K arrival processes have the same scheduled arrival times

      2) Uncoordinated - arrival times of the K processes are independent

   B. Unphased - each of the K arrival processes have different frequencies

Table 5.1

to increase N and still compare directly the interarrival time distributions. For the renewal and both phased processes, the combined arrival rate was held constant at .2. For unphased service, since the interarrival times were drawn from a uniform distribution as just described, the arrival rate for each component process is a random variable, and hence so is the combined arrival rate. In this case, the actual average arrival rate must be computed from the simulation. For the renewal process, arrival times for vehicles from terminal i are given by:

$$t^i_{n+1} = t^i_n + \tau^i_n \qquad 5.2$$

where

$$\tau^i_n = 1/\lambda + \varepsilon_{in}$$

and where $\lambda = 1/(5N)$. In all four cases, it was assumed that $\varepsilon_{in} \sim U(-5N\ \delta,\ 5N\ \delta)$. The parameter δ is the relative error in the arrival times and is used to test the effect of variability in the component arrival processes on the convergence to a Poisson.

With the simulation thus described, we turn now to the actual experiments. The basic idea here is to investigate the rate at which the superimposed process converges to a Poisson as N is increased. The rate of convergence is studied with respect to two factors, the first being the type of processes being superimposed, where we have the four cases outlined above, and the second being the parameter δ which determines the degree of variability. In this section, the method by which similarity between a superposition process and a Poisson process is measured is given by what we have referred to as the microscopic view. The distribution of the interarrival times and their first order correlation coefficients are computed and compared to what should occur if the process were Poisson. Thus the interarrival time should have a negative exponential distribution and the correlation coefficient should be zero. Of course, it is impossible to draw any conclusions from these comparisons regarding whether a given process is or is not approximately Poisson. Rather, these experiments are intended to convey how a process looks when viewed using the conventional methods for studying the characteristics of a point process. Then, in sections 5.1.2 and 5.2, we show how the microscopic view can be very misleading, particularly in the context of bulk queues.

A large number of simulations were run for each of the four processes, for different values of both N and δ. It was consistently found that for $\delta \geq .2$, $N \geq 8$, the interarrival time distribution almost exactly fit the predicted negative exponential distribution. For this reason we show only the cases where the fit was not quite as close. These results, given in figures 5.2 - 5.13, show both the histograms for the observed interarrival time distribution as well as the predicted distribution if the process were Poisson. The plots indicate for the most part poor similarity for N = 3 or 5. Thus, on the basis of this information alone, N must be greater than approximately 7 or 8 for the process to appear approximately Poisson.

Figures 5.14 and 5.15 describe the correlation coefficients for each of the four processes as N is increased for two values of δ. These graphs more clearly show the convergence, although at N = 10, the correlation coefficient is still at -.10. It remains to be seen if

Comparison of the observed interarrival time distribution and a negative exponential distribution (N = 3 renewal processes, δ = .2)

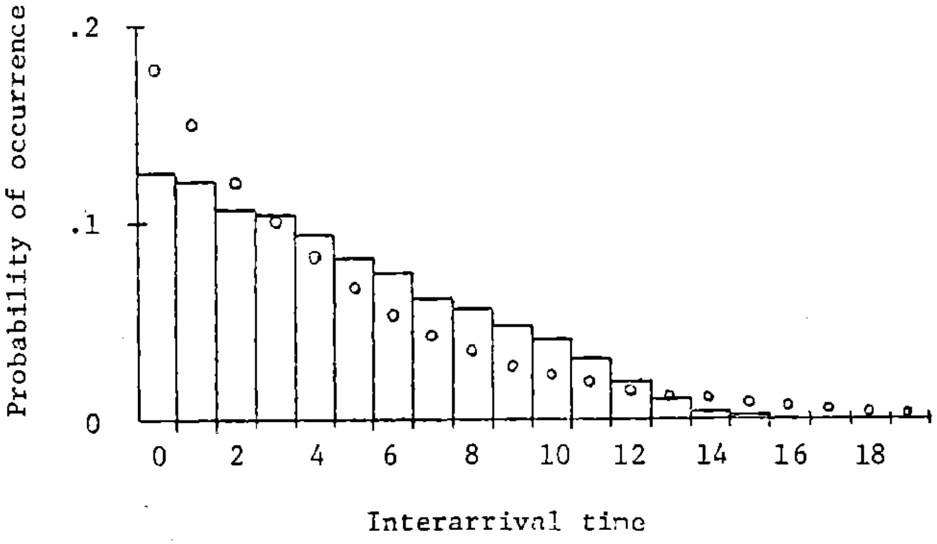


Figure 5.2

(N = 5 renewal processes, δ = .2)

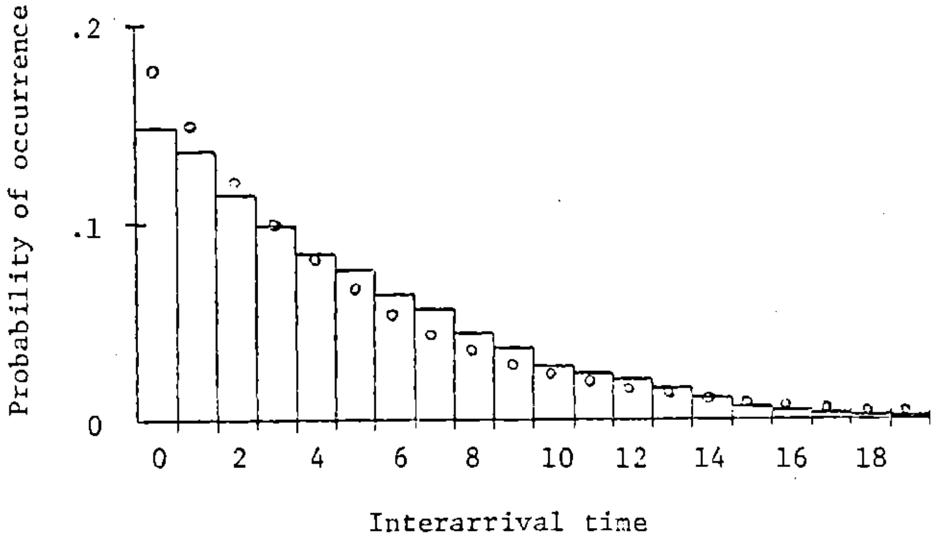


Figure 5.3

(N = 3 processes, coordinated departures, δ = .2)

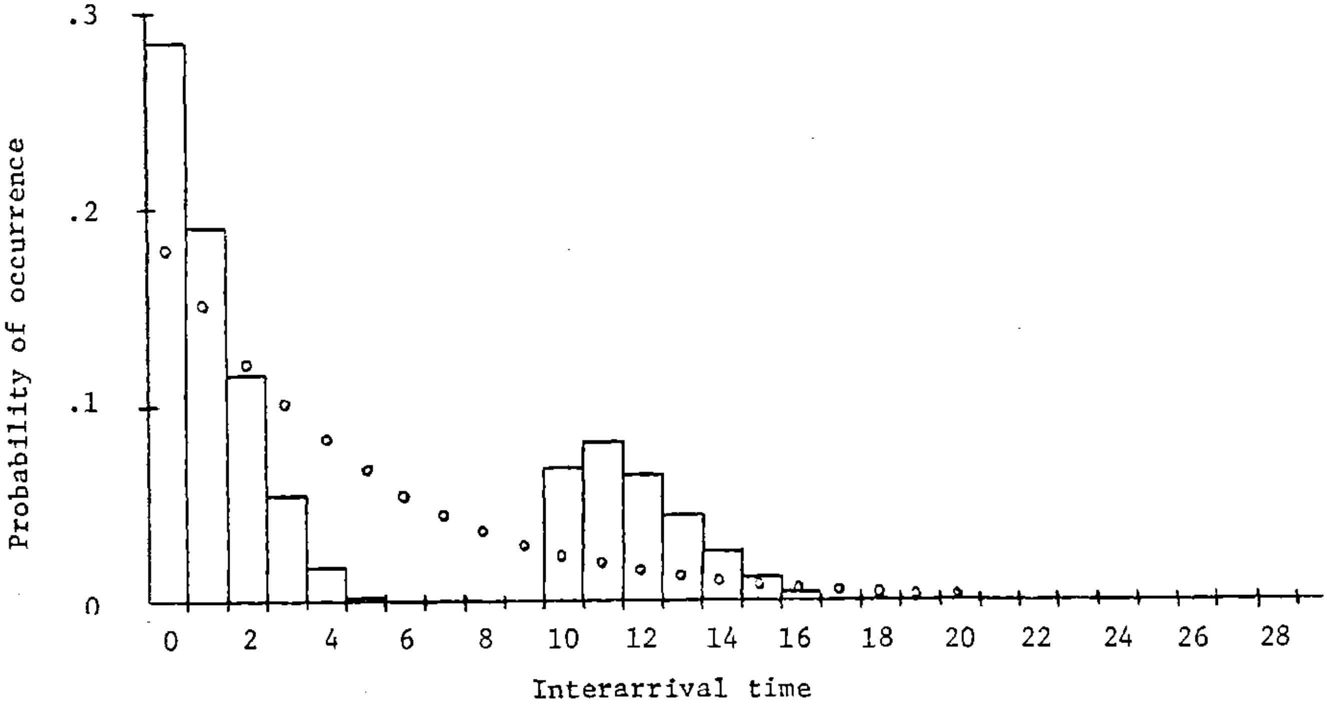


Figure 5.4

(N = 3 processes, coordinated departures, δ = .4)

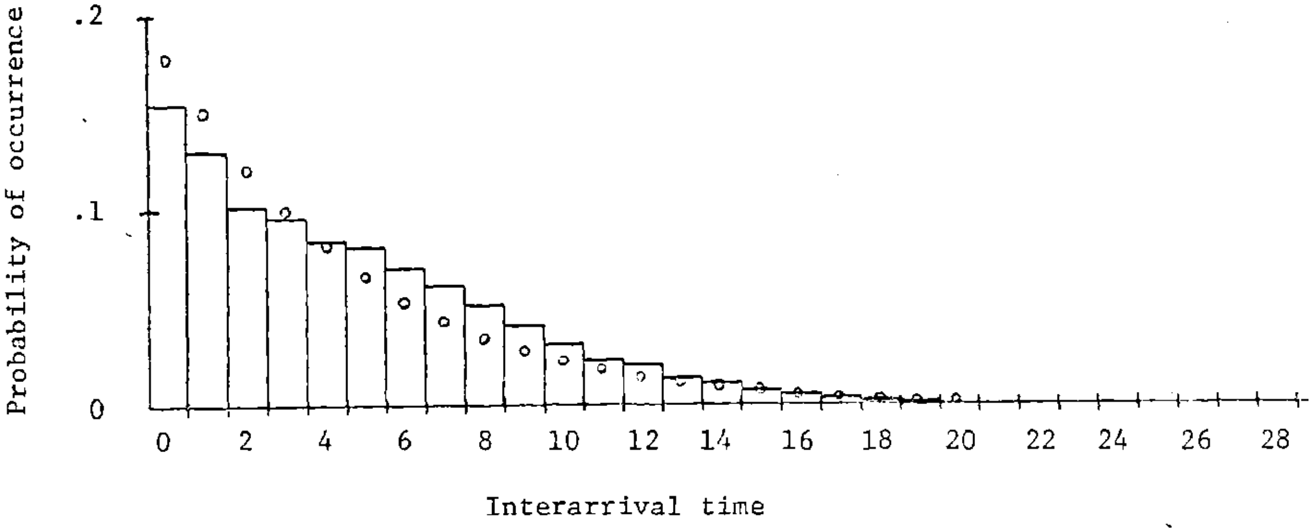


Figure 5.5

(N = 5 component processes, coordinated departures, δ = .2)

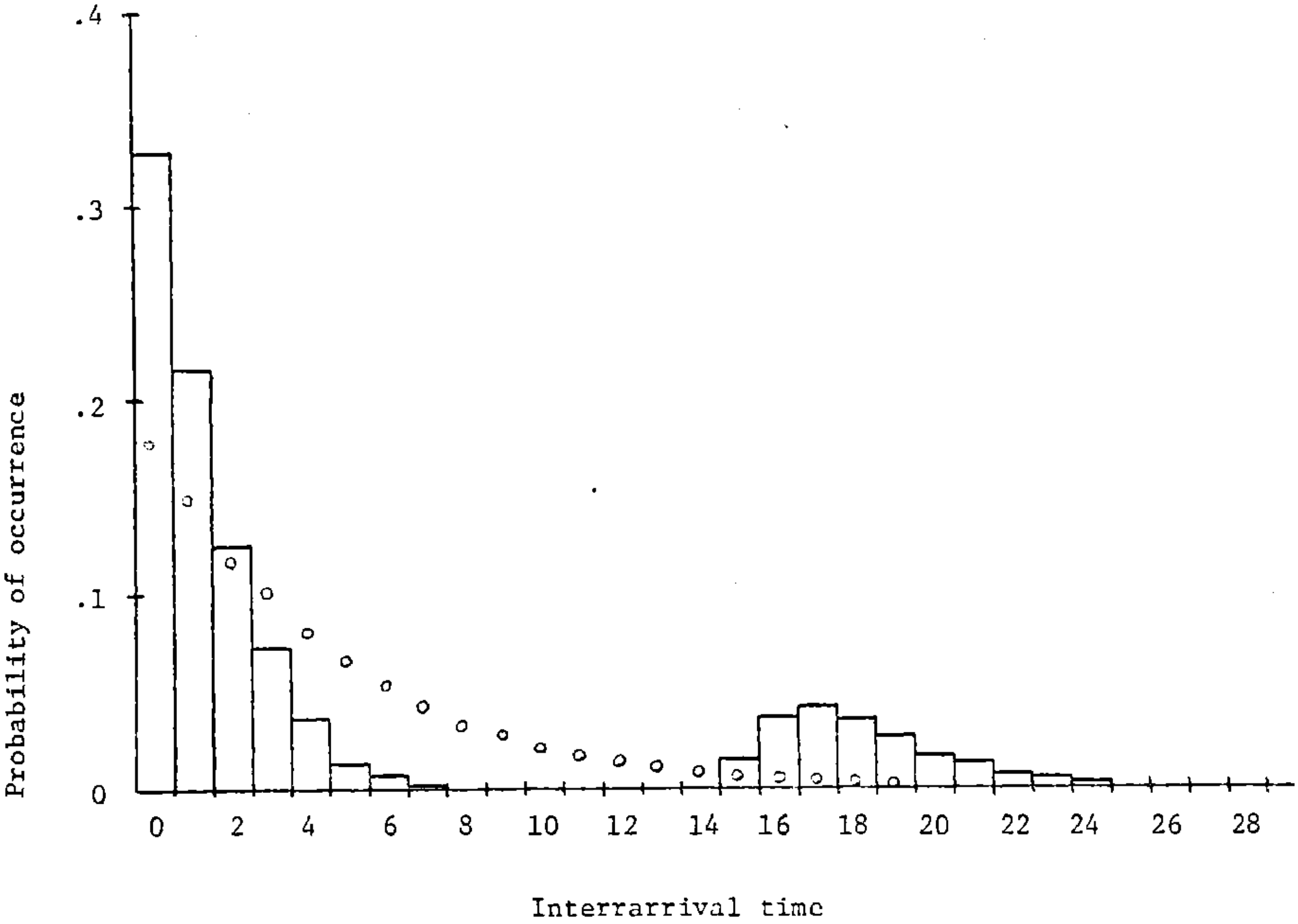


Figure 5.6

(N = 5 component processes, coordinated departures, δ = .4)

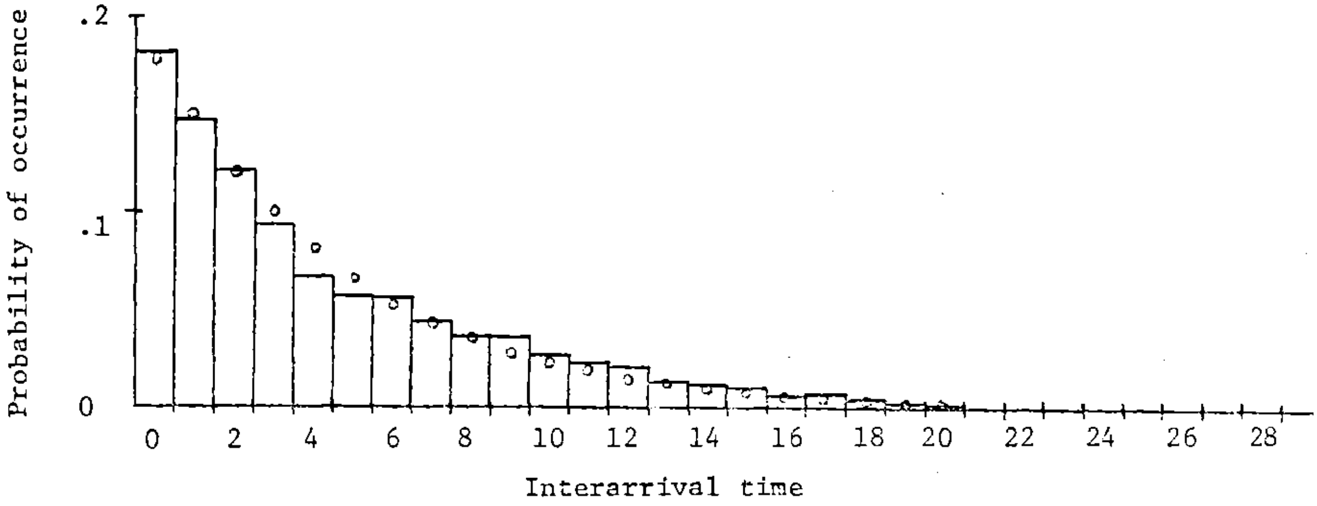


Figure 5.7

(N = 3 component processes, uncoordinated departures, δ = .2)

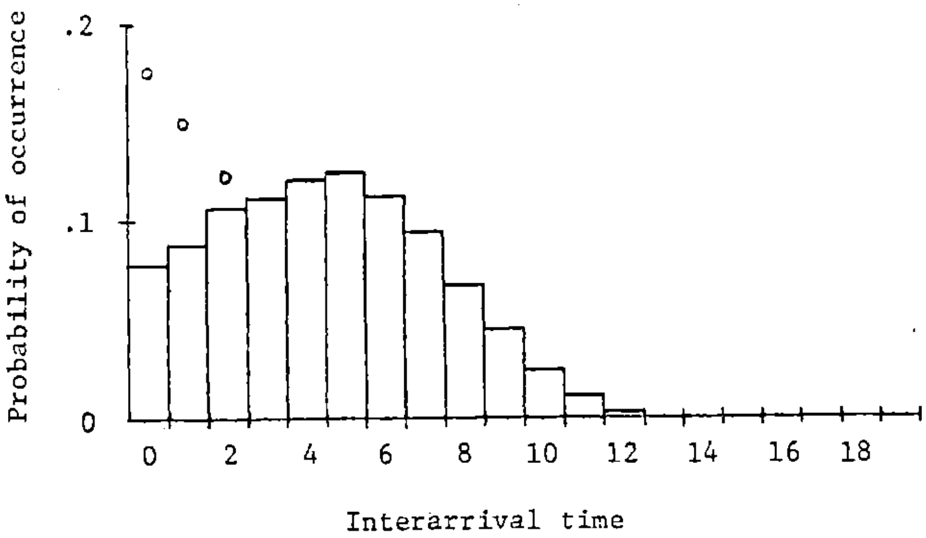


Figure 5.8

(N = 3 component processes, uncoordinated departures, δ = .4)

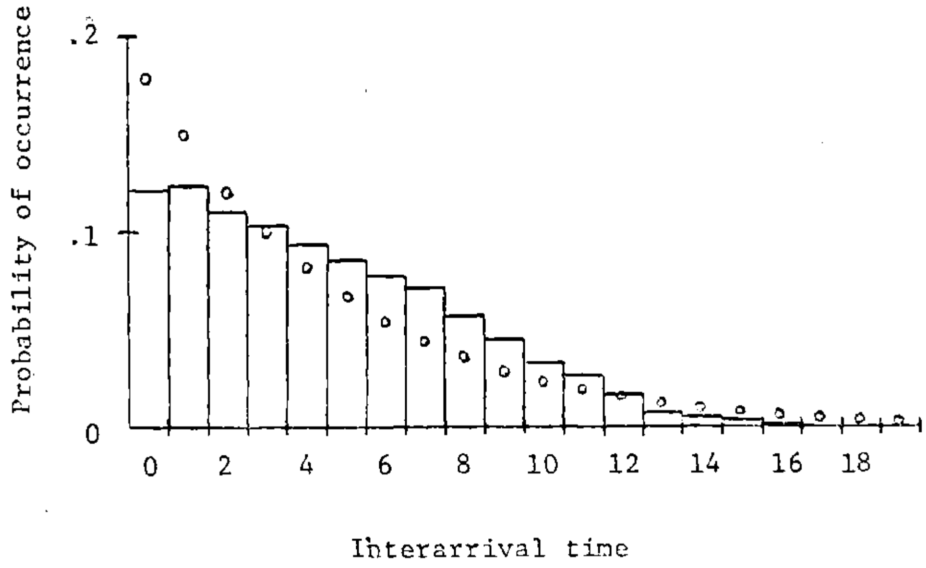


Figure 5.9

(N = 5 component processes, uncoordinated departures, δ = .2)

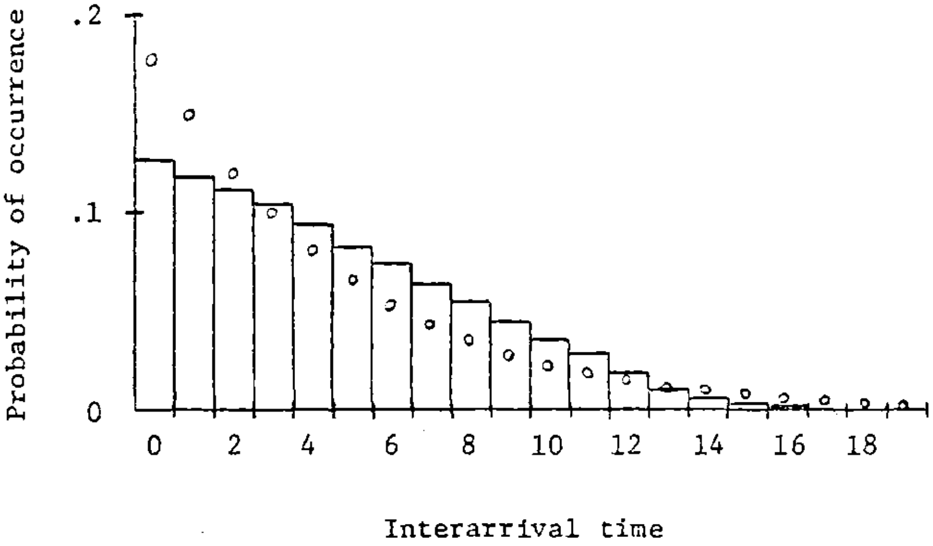


Figure 5.10

(N = 10 component processes, uncoordinated departures, δ = .2)

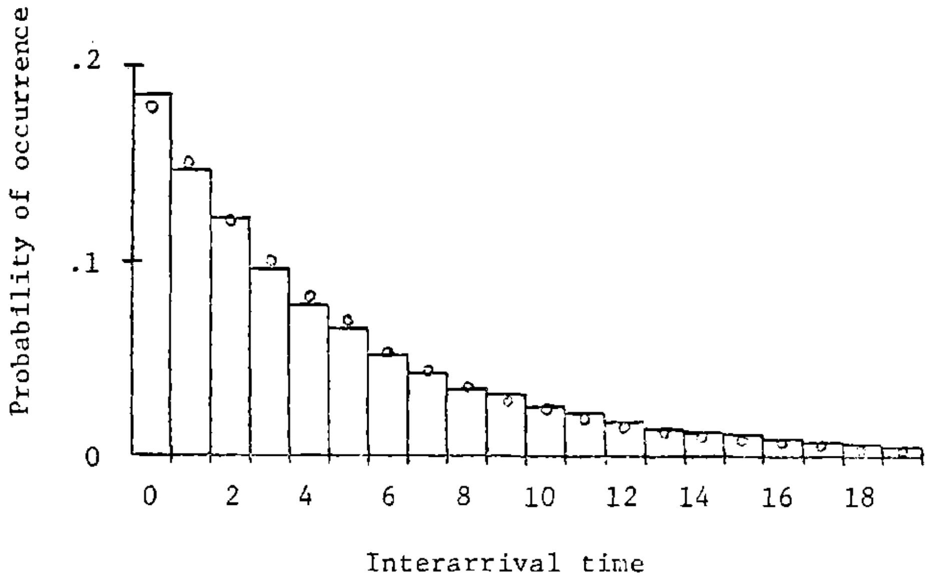


Figure 5.11

(N = 3 component processes, unphased, δ = .2)

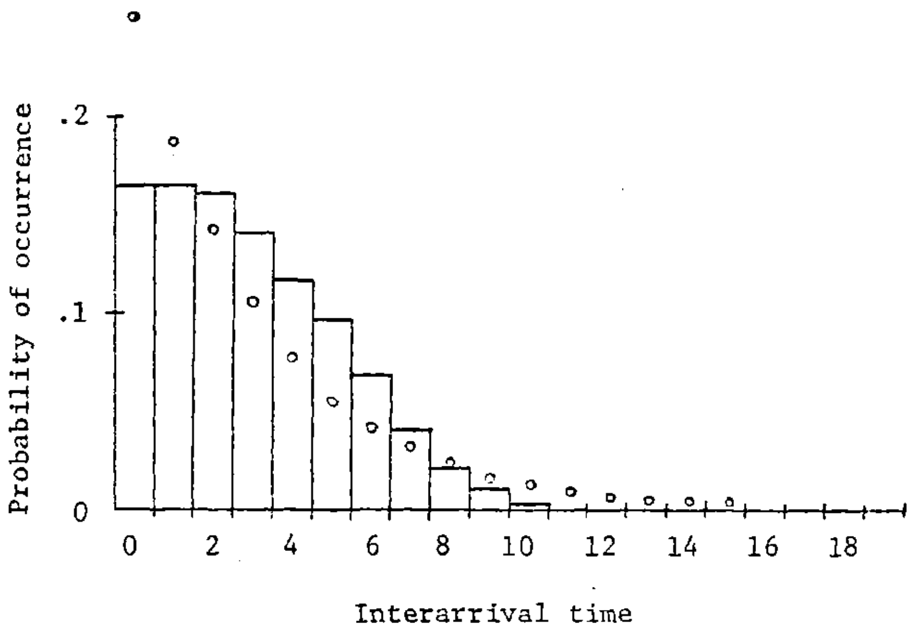


Figure 5.12

(N = 5 component processes, unphased, δ = 0.0)

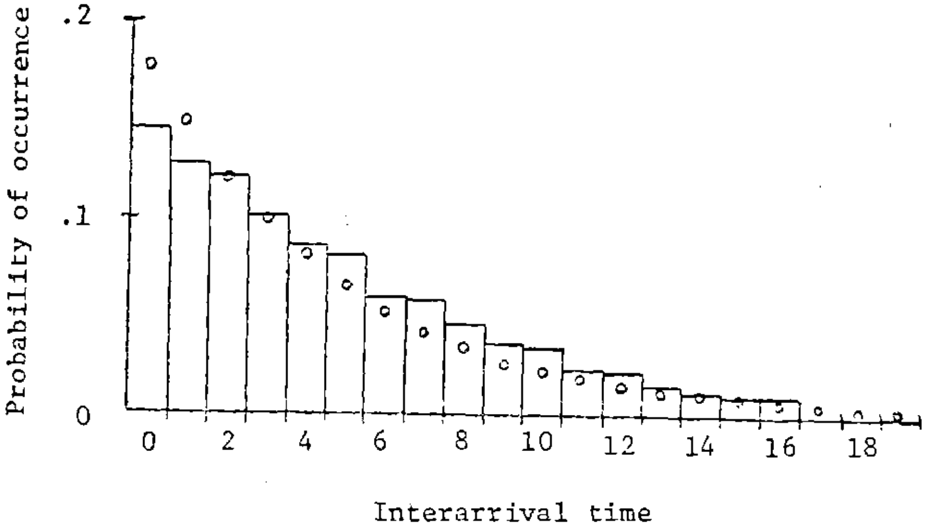


Figure 5.13

Comparison of correlation coefficients for superimposed streams of renewal and unphased processes

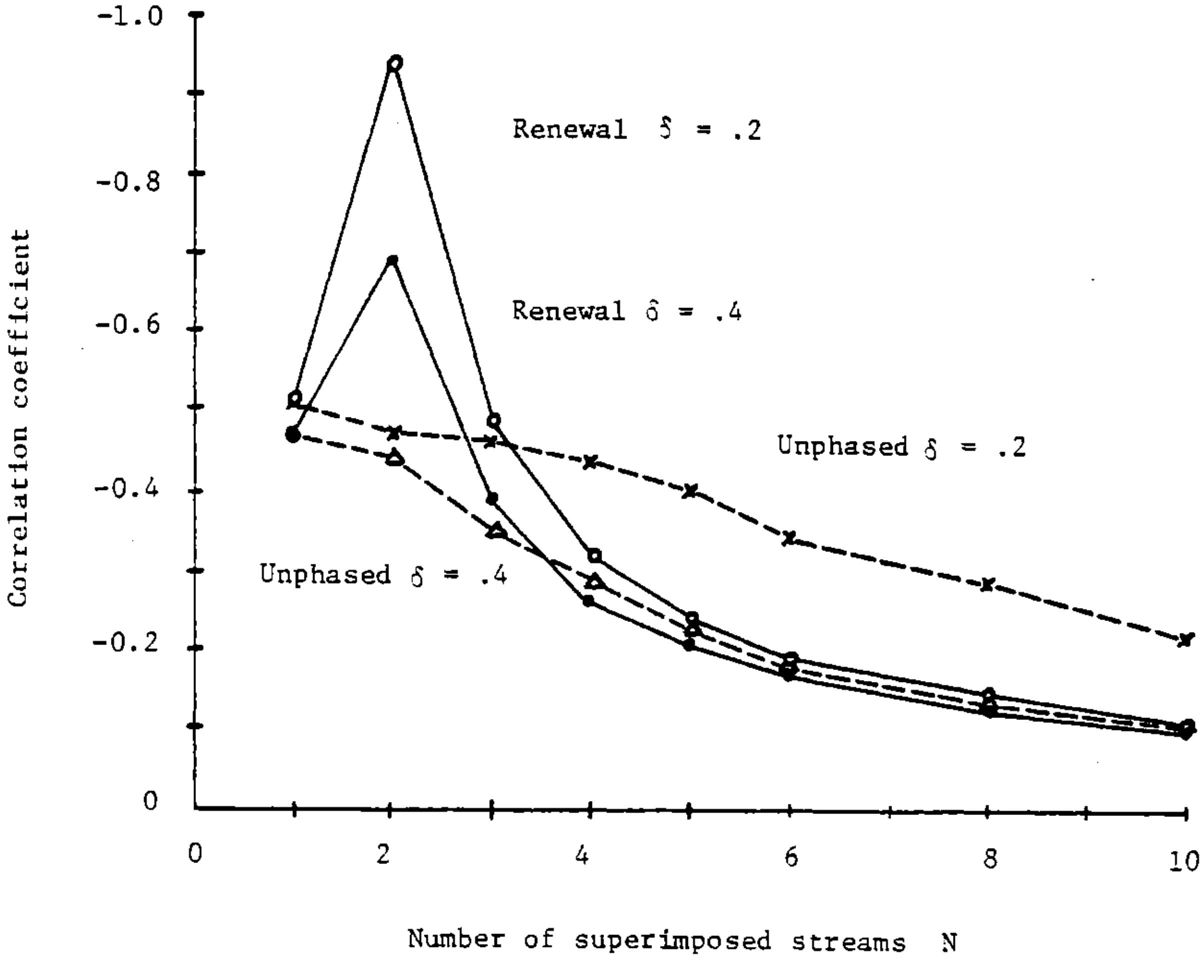
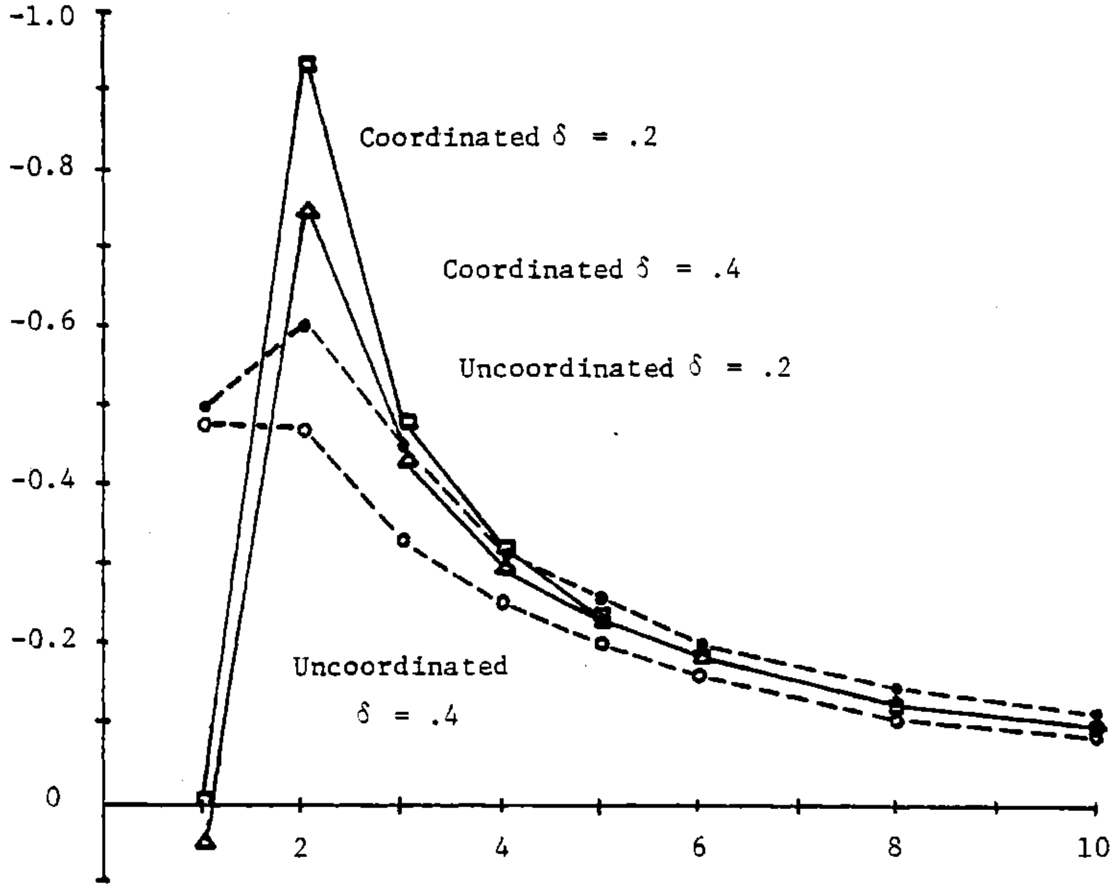


Figure 5.14

Comparison of correlation coefficients for superimposed streams of coordinated and uncoordinated processes



Number of superimposed streams N

Figure 5.15

this is sufficiently low for the resulting process "to be Poisson," but it does seem clear that the rate of convergence will be very slow for N > 10.

Section 5.1.2 repeats some of these experiments, concentrating instead on the number of arrivals in successive periods of time instead of the interarrival times. The results of these tests are then compared to those presented in this section.

5.1.2 The macroscopic view

It is common in the study of random point processes to concentrate on the interarrival time distribution as the most important descriptor. In the analysis of bulk queues, however, the most important quantity is the number of arrivals during a service period, denoted by the variable Y. In this section, then, we study an arrival process by defining a set of points in time, evenly spaced, and analyze the number of arrivals between each set of points. The set of points can be viewed as departure instants for an $M/D^c/1$ queue, but our interest at this point is only studying the arrival process in isolation, and not as part of a queueing process. Denoting the points in time T_n, we assume $T_{n+1} - T_n = T$, where T is a constant equal to the length of each successive interval. As before, Y_n is the number of arrivals between T_n and T_{n+1}.

The study of the variables Y_n is termed here the macroscopic view since the time period of interest, T, is generally much longer than the time period used when studying interarrival times. Also, the microscopic view looks at the time of each arrival, whereas the macroscopic view looks only at the aggregate number of arrivals over a longer time period T. What is significant about the latter approach, however, is that it captures the net effect of correlations between successive interarrival times over a large number of arrivals, That is, if τ_n, τ_{n+1}, . . ., are the sequence of interarrival times, we would find that not only are τ_n and τ_{n+1} correlated, but so are τ_n and τ_{n+2}, τ_n and τ_{n+3}, and so on. The combined effect of such correlations is impossible to capture using the methods described in the previous section.

The simulation experiments were conducted for only two types of arrival processes, renewal and unphased, representing the two cases most likely to produce an approximate Poisson process. The distribution of Y was computed as was the correlation coefficient between Y_n and Y_{n+1}. Similarity to a Poisson arrival process was measured by comparing the distribution of Y to a Poisson distribution with the same mean. Of course, the correlation coefficient should be zero if the process is Poisson. As before, these statistics were computed while increasing N, although here the parameter δ was set to .4. Instead, the effect of the parameter T was studied since this determines the length of the time interval over which arrivals are being counted. Clearly, if T is approximately the length of one interarrival time, then it is unlikely that the two views would differ significantly in their conclusions. As T is increased, the view becomes more macroscopic, and the effect of small correlations in the interarrival times would become more pronounced.

As before, a large number of simulations were run, and in each case the distribution of Y and the correlation coefficient were computed. The results are best described by the plots of the correlation coefficient shown in figures 5.16 and 5.17 for the renewal and unphased processes. In each case, N is increased from 1 to 10, and plots are drawn for T = 5, 10, 15, and 100. Note that along the horizontal axis, we also show the average interarrival time $\overline{T}$ for each of the component processes. Remember that $\overline{T}$ was increased with N in order to keep the average interarrival time for the combined process the same. In figure 5.16, we notice an interesting pattern where the correlation coefficient first increases and then steadily decreases. In each case, the peak occurs when $\overline{T}/T = 2$ or, in other words, when the length of the period of observation, T, is exactly one half the average interarrival time of the component process. In figure 5.17, where the frequencies of the component processes were all different, we notice that the graph remains approximately constant at -.5 until again at $\overline{T}/T = 2$, the curve drops off sharply.

These observations can be explained from the nature of the simulation. When $T = \frac{1}{2}\cdot\overline{T}$, one arrival from a particular terminal must occur in either (T_n, T_{n+1}) or (T_{n+1}, T_{n+2}) which produces the high negative correlation between Y_n and Y_{n+1}. For $T<\frac{1}{2}\overline{T}$, if no arrival from terminal i occurs in T_n, it might also be true that none would occur in T_{n+1} either, thereby reducing the dependence between Y_n and Y_{n+1}. On the basis of this analysis, it would appear that the graphs simply reflect the fact that we are correlating only Y_n and Y_{n+1}, and not Y_n and Y_{n+2}, for example. However, the convergence to a Poisson distribution suggested by the reduction in the first order correlation coefficient is supported when we look at the computed distribution of Y itself. Interestingly, the distribution of Y does not even begin to approach a Poisson distribution until $\overline{T}/T$ is at

Correlation coefficient v. number of streams for renewal, phased input

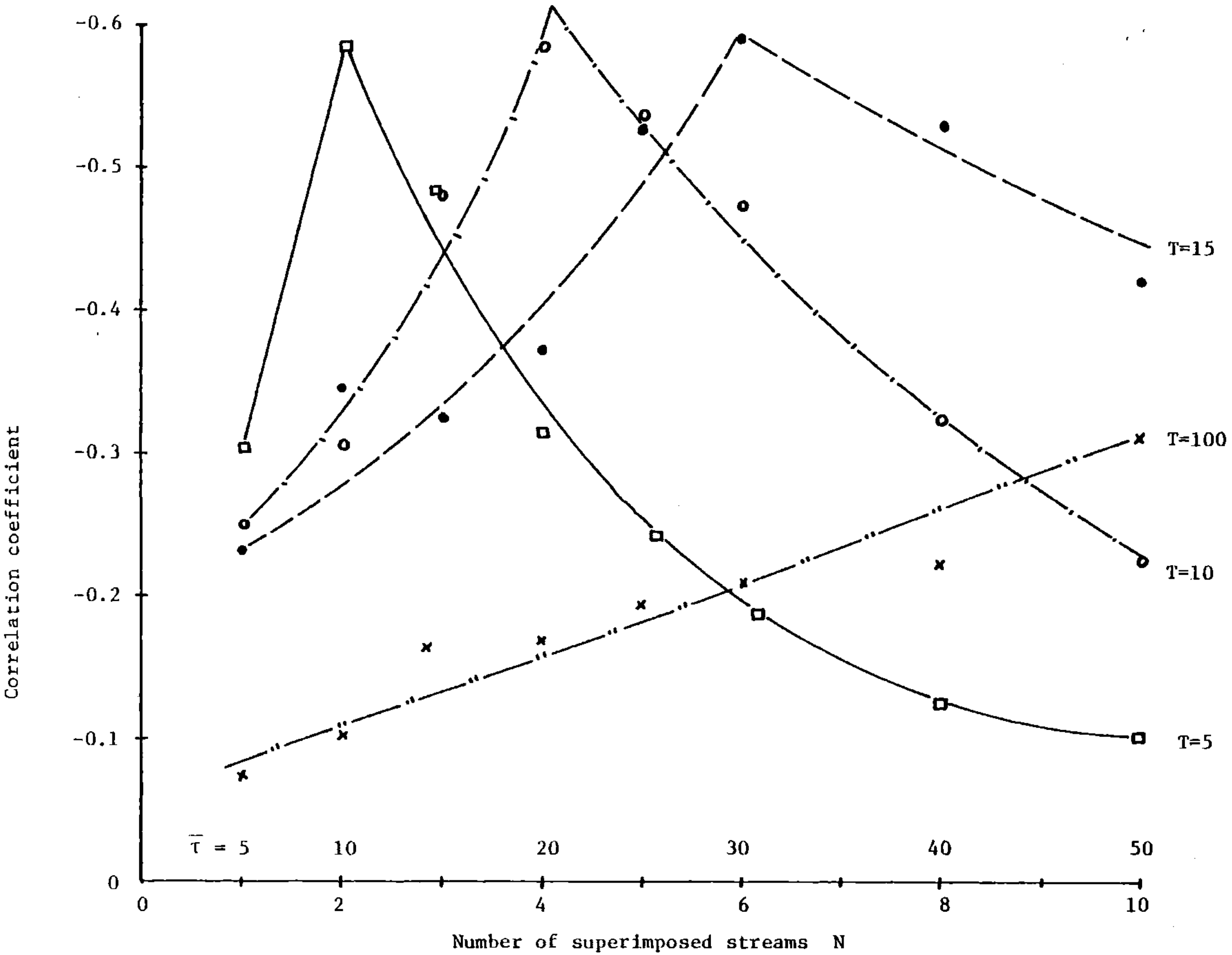


Figure 5.16

Correlation coefficient v. number of streams for nonrenewal, unphased input

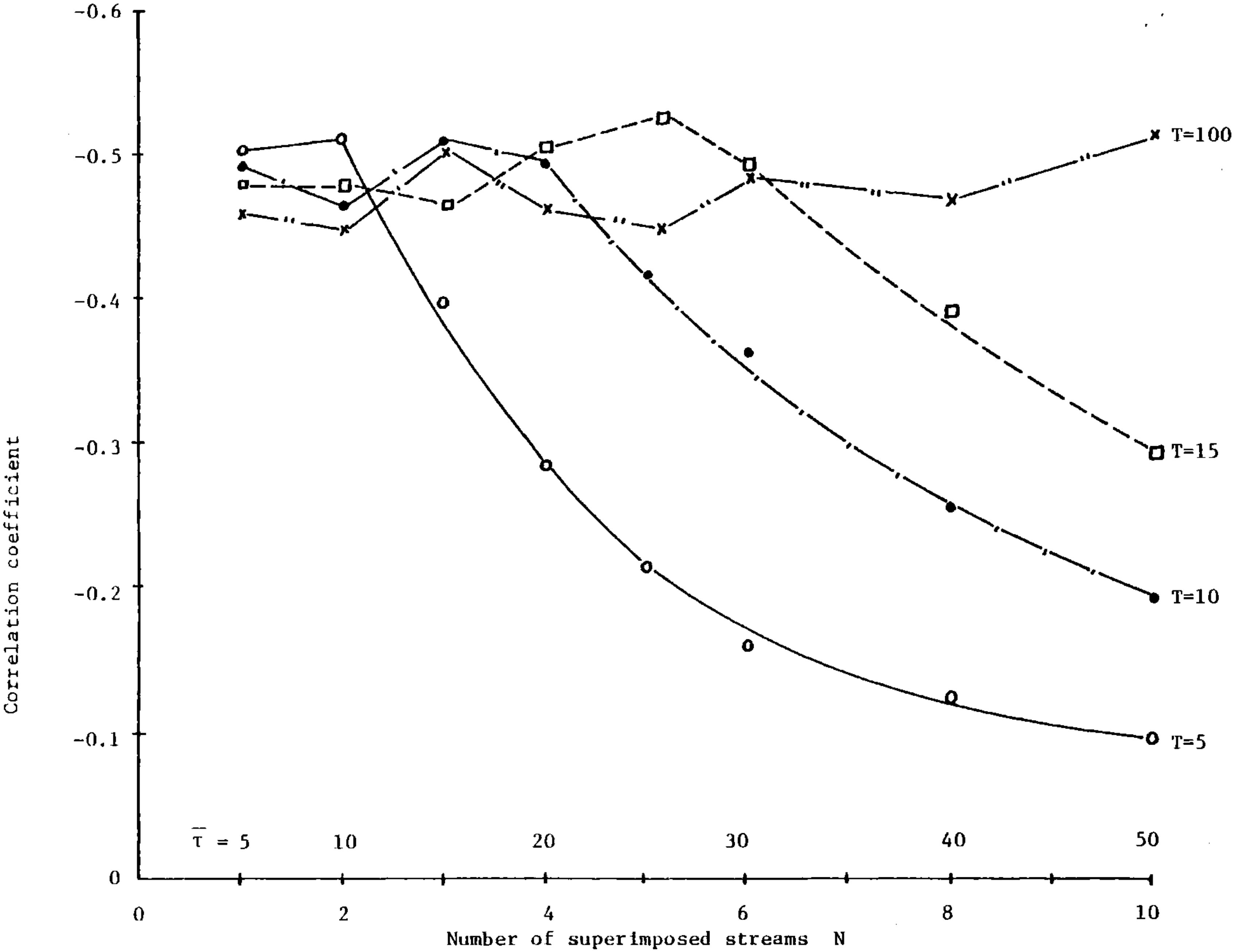


Figure 5.17

least 2. For example, if we are superimposing 10 renewal processes, analysis of the interarrival times (i.e. the microscopic view) suggests an extremely close fit with a Poisson arrival process. On the other hand, the distribution of Y when T = 100 ($\overline{\tau}/T = 50/100 = .5$), shown in figure 5.18, does not even approximately resemble a Poisson distribution with the same mean. As we would expect, the small, negative correlations between successive interarrival times over a long time period produce an actual distribution for Y with a much smaller coefficient of variation than would occur if the process were indeed a Poisson.

On the declining sides of the curves (i.e. when $\overline{\tau}/T > 2$) the agreement between the observed distribution for Y and a Poisson distribution becomes much better. For the cases where N = 10 and T = 5 and 10($\overline{\tau}/T$ = 10 and 5, respectively) the fit is quite good (see figures 5.19 and 5.20). Figures 5.21 and 5.22 show the same information for N = 5 and T = 5 and 10 ($\overline{\tau}/T$ = 5 and 2.5) demonstrating a somewhat poorer fit.

The results of this section suggest that the ratio $\overline{\tau}/T$ is a more important determinant of whether a given arrival process is approximately Poisson than the number of streams being superimposed. On the basis of this observation, one can explain not only why the arrival of passengers at a bus stop <u>would</u> be described by a Poisson process but also why this would <u>not</u> be a good approximation for the arrival of vehicles at a terminal. In the first case, each passenger represents a separate component process, where $\overline{\tau}$ might be equal to 24 hours (i.e. he arrives at a bus stop once a day); since the bus frequency might be departures every .5 hours, we have $\overline{\tau}/T = 48$, which of course is relatively large. On the

Comparison of actual distribution of arrivals between departures for 10 superimposed renewal processes and a Poisson distribution with the same mean

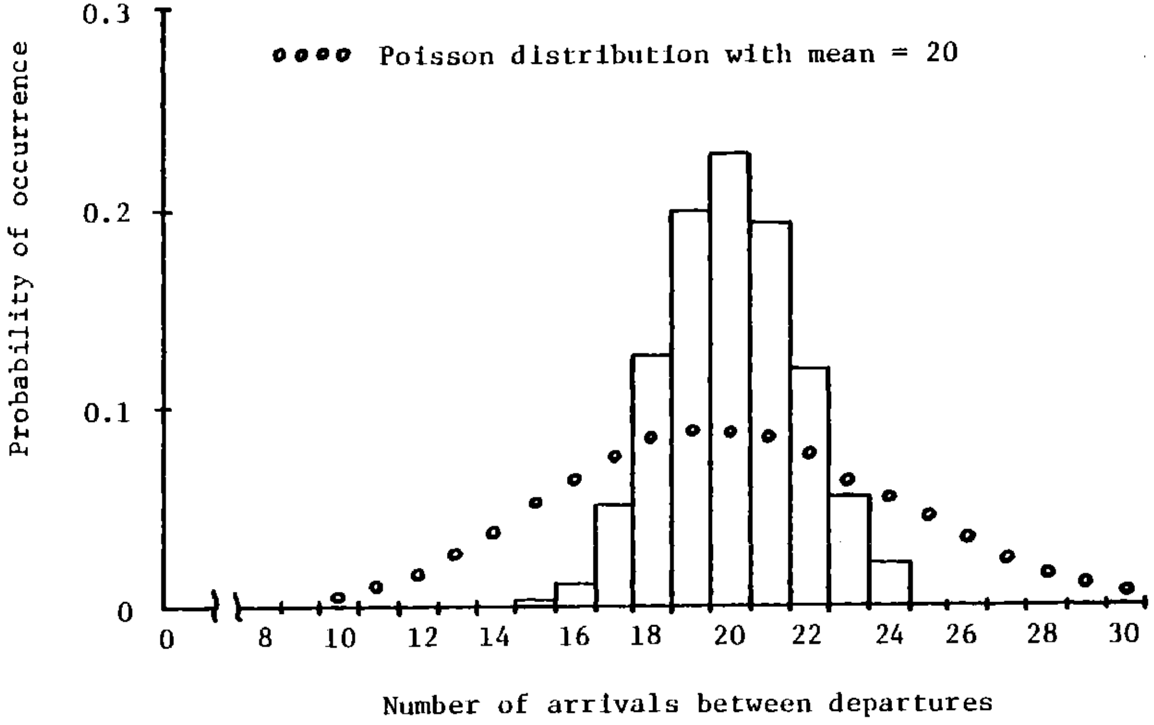


Figure 5.18

Comparison of observed distribution of arrivals and a Poisson distribution with the same mean (T=5; component processes are unphased with $\bar{\tau}$=5)

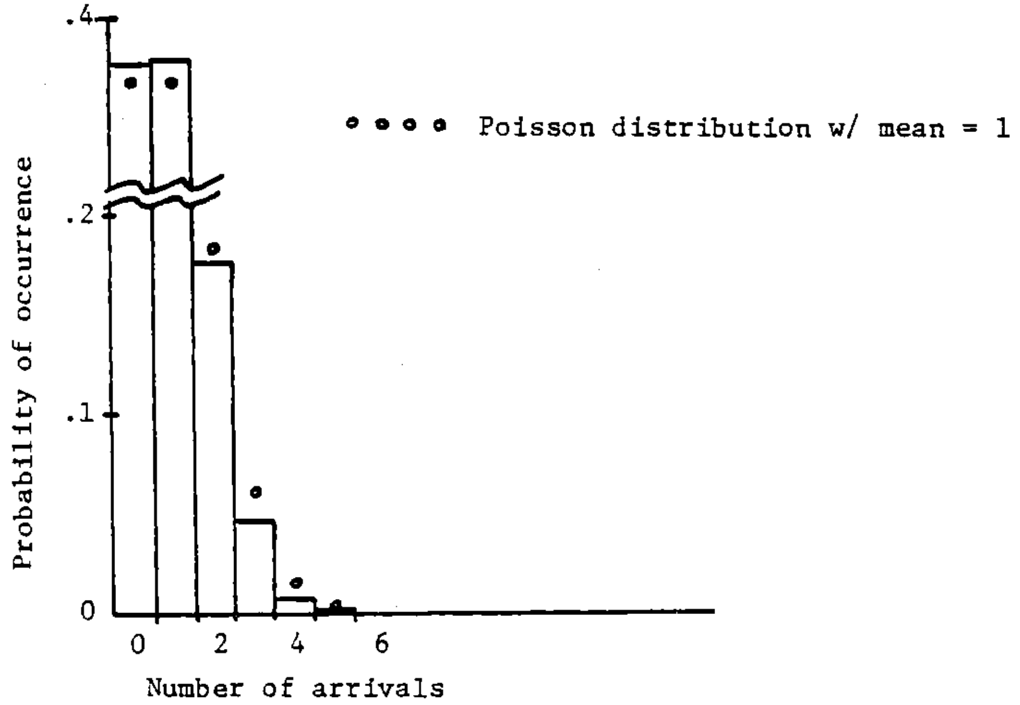


Figure 5.19

Comparison of observed distribution of arrivals and a Poisson distribution with the same mean (T=10; component processes are unphased with $\bar{\tau}$=5)

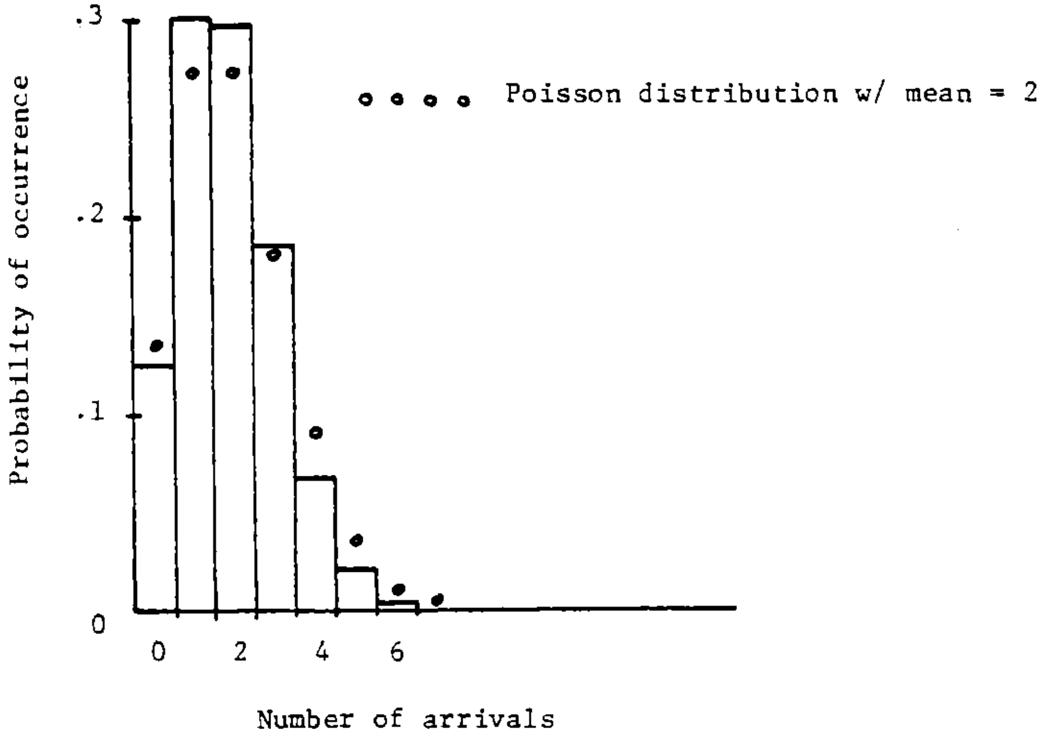


Figure 5.20

Comparison of observed distribution of arrivals and a Poisson distribution with the same mean (T=5; component processes are unphased with T=25)

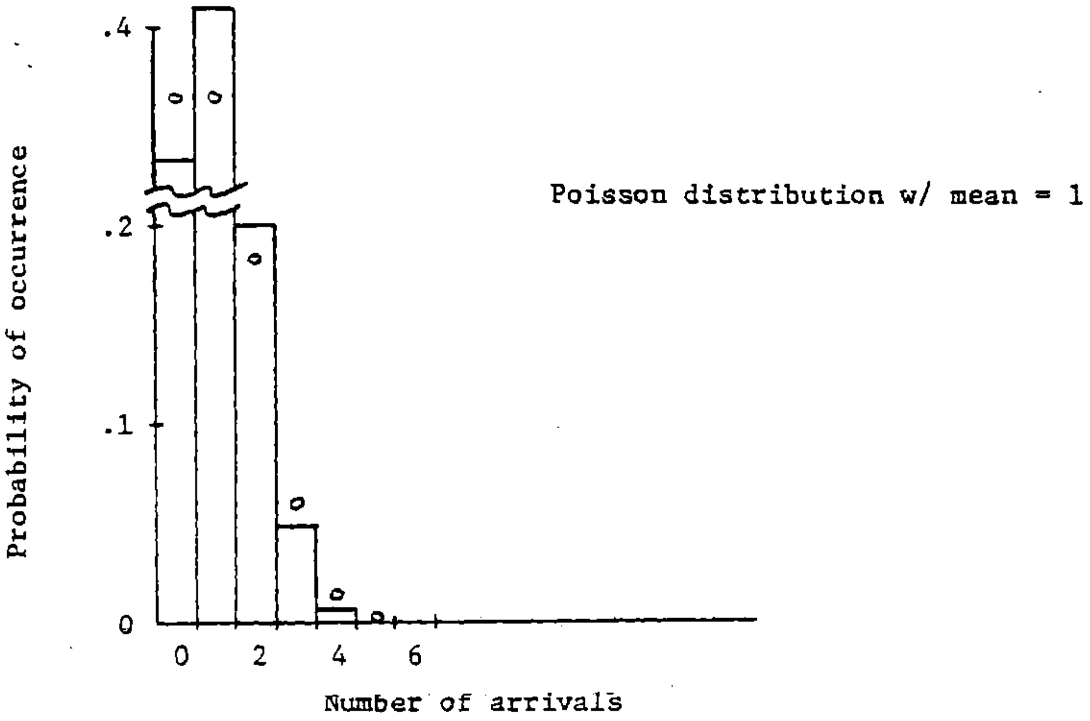


Figure 5.21

Comparison of observed distribution of arrivals and a Poisson distribution with the same mean (T=10; component processes are unphased with $\bar{\tau}$ = 25)

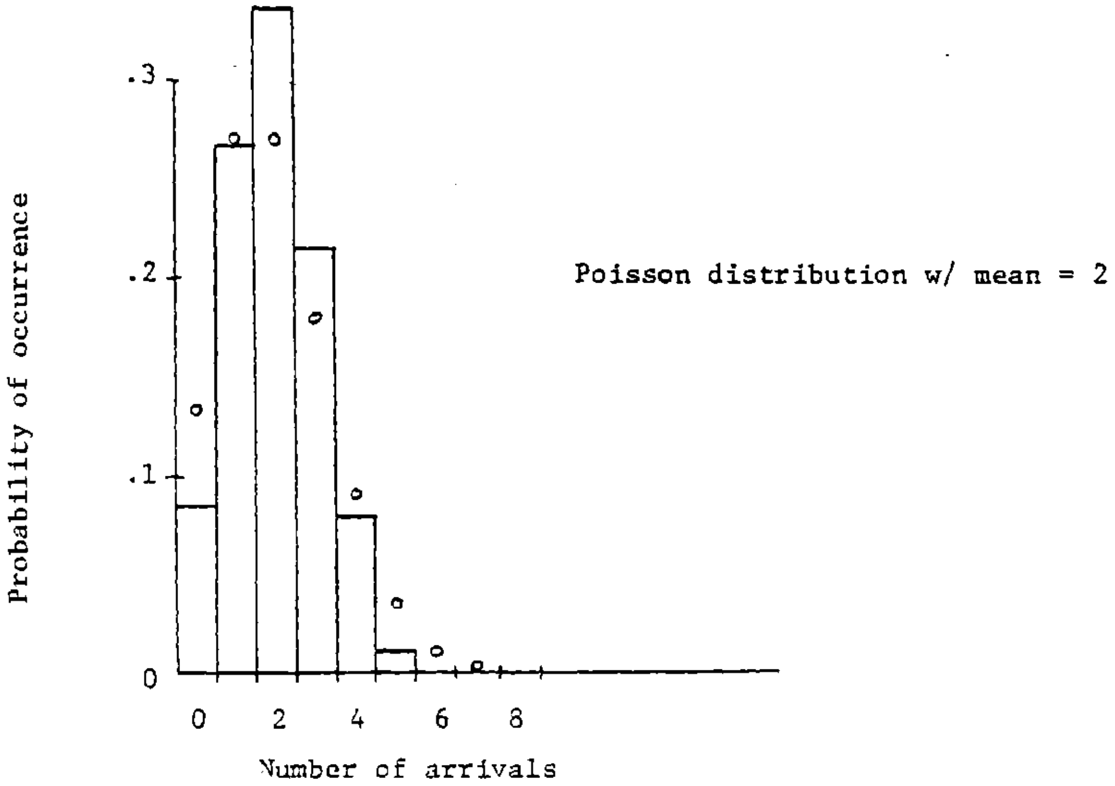


Figure 5.22

other hand, vehicles would arrive at a terminal over each link at approximately the same rate they would depart over each link, in which case $\bar{T}/T \simeq 1$. Thus even though the superposition of arrival streams of vehicles at a particular terminal may appear Poisson based on an examination of interarrival times, it is unlikely that the number of vehicles arriving at a terminal between successive departures over a particular link would be given by a Poisson distribution.

In the next section, we focus on an arrival process in the context of a bulk queue to determine the accuracy of the Poisson arrival process in predicting the first and second moments of the length of the queue.

5.2 A general arrival process in the context of a bulk queue

Section 5.1 presents two approaches for measuring the similarity between a general arrival process and a Poisson, one being the conventional approach based on an analysis of interarrival times, the second looking at successive increments of the process. There it is shown that the latter approach provides a better indication of whether an arrival process is approximately Poisson. Also, the ratio $\overline{\tau}/T$, the mean headway of a component process over the mean departure headway, is a more useful indication of whether a set of superimposed processes is approximately Poisson than N, the number of streams being superimposed. What remains to be tested is how well the Poisson arrival process approximates a general superposition process in terms of estimating the moments of the length of a queue.

To answer this last question, a set of experiments were conducted whereby N renewal processes were superimposed, creating an arrival process to a bulk service queue with departures every T units of time with a vehicle capacity of $c = 5$. The utilization parameter is given by $\rho = \frac{NT}{c\overline{\tau}}$, where $\lambda = 1/\overline{\tau}$ is the arrival rate of customers for each component process. The headway $\overline{\tau}$ was chosen for each value of N such that $\rho = .9$. Given these parameters, the mean and standard deviation of the length of the queue were computed for values of N ranging from 5 to 60 ($\overline{\tau}/T$ increased from .9 to 10.8), as shown in figures 5.23 and 5.24. Also shown are the theoretical mean and standard deviation computed using the methods described in chapter 4. Finally, as a check on the simulation program, the same figures were computed by simulating the superposition of N Poisson processes. It

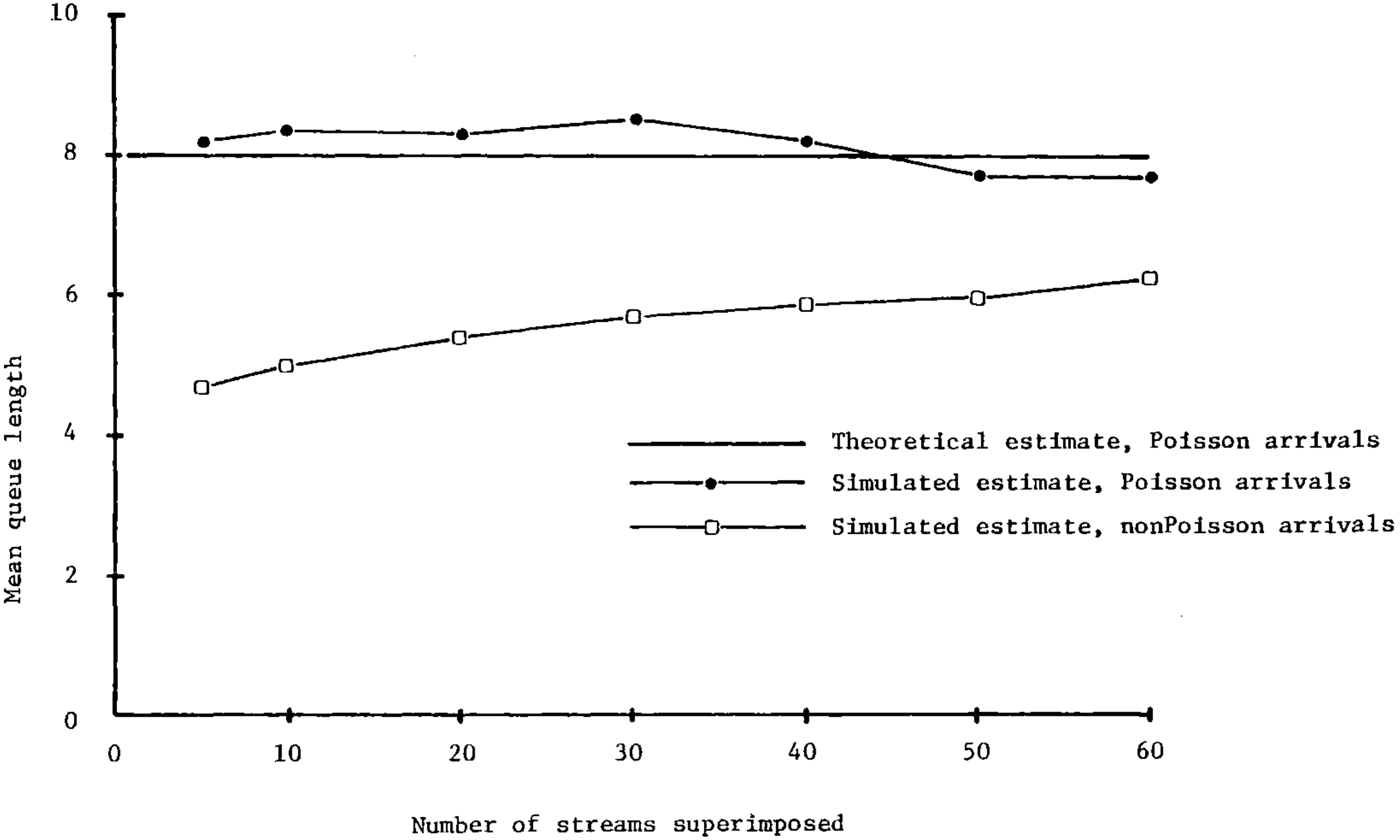

10
8
6
4
2
0
Mean queue length
0
10
20
30
40
50
60
Number of streams superimposed
Theoretical estimate, Poisson arrivals
Simulated estimate, Poisson arrivals
Simulated estimate, nonPoisson arrivals

Figure 5.23

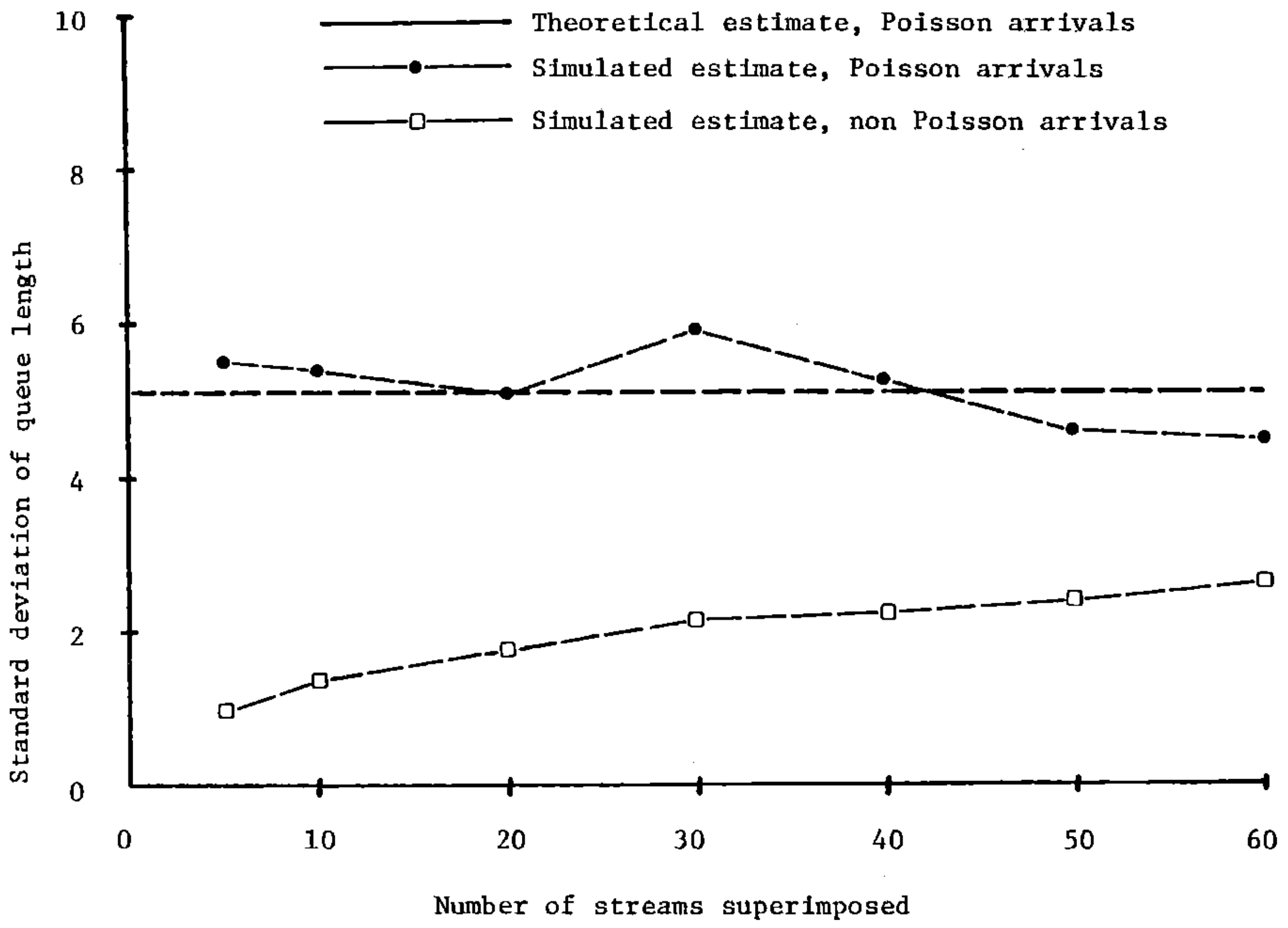

Theoretical estimate, Poisson arrivals
Simulated estimate, Poisson arrivals
Simulated estimate, non Poisson arrivals
Standard deviation of queue length
10
8
6
4
2
0
0
10
20
30
40
50
60
Number of streams superimposed

Figure 5,24

is easily seen that the theoretical values do in fact agree with those obtained when simulating Poisson arrival processes. On the other hand, the simulated general arrival process has consistently lower estimates of both the mean and standard deviation, even for relatively large values of $\overline{\tau}/T$.

This experiment suggests that the Poisson arrival processes is not a particularly robust approximation. More importantly, statistical analyses of an arrival process, using either approach described in section 5.1, may not provide a good indication of whether a given arrival process is "sufficiently Poisson". The problem is that it is unlikely that any process is truly Poisson if viewed over a sufficiently long period of time. For example, arrivals to a bus stop maybe adequately described by a Poisson process, but it is unlikely that the total number of arrivals over 24 hours is given by a Poisson distribution. Hence, the real issue is over what length of time a process must "look" Poisson. For bulk queues, this length must be at least T or 2T, but is more likely to be related to the length of the busy period. Thus as ρ approaches 1, the Poisson approximation is more likely to break down. It would be interesting to compute the ratio of $\overline{\tau}$ over the mean length of a busy period and use this figure as a basis for evaluating the accuracy of the Poisson approximation for a general arrival process.

This discussion brings out the important factors determining the use of the Poisson arrival assumption to solve particular queues. It does not, however, solve the problem of what to do when the arrival process cannot be accurately approximated with a Poisson. Section 5.3 looks at this problem and suggests several methods for developing approximations for $G^y/G^y/1$ queues.

5.3 Approximations for $G/G^c/1$ queues

Since the early sixties, queueing theorists have sought to approximate results for G/G/1 queues which could not be obtained using standard methods based on analysis of the imbedded Markov chain. These include fluid approximations and diffusion approximations as well as a number of bounds on mean waiting time (see Kleinrock (1976) for a review of this material). Only recently, however, have any papers appeared which address the problem of approximating the actual behavior of a G/G/1 queue (i.e. finding the approximate distribution of the queue length or waiting time, instead of simply the average waiting time). Kuehn (1979) and Whitt (1979a, b) have studied the problem of approximating general, non-renewal arrival processes with renewal ones. The problem being studied was the superposition of several message streams in communications networks. Service times were assumed to be negative exponential, and hence the replacement of a G/M/1 queue with a GI/M/1 queue enabled analysis using standard transform methods.

Replacing a general arrival stream with a renewal one does not, of course, simplify the problem of studying bulk queues. The work performed by Whitt does, however, suggest an alternative approach. For this reason, it is appropriate to first describe his research as outlined in Whitt (1979a). Consider the instance of a general arrival stream with interarrival times given by $\tau_1, \tau_2, \ldots$, which are not independent. The problem is to derive a new random variable $\hat{\tau}$ such that an i.i.d. sequence $\hat{\tau}_1, \hat{\tau}_2, \ldots$, produces a renewal arrival stream which produces the same queueing behavior as the original, non-renewal one. The approach used by

Kuehn, which Whitt terms the *stationary interval method*, is to estimate the mean and variance of τ, which are then used to calculate two parameters of an assumed distribution for $\hat{\tau}$. Thus τ and $\hat{\tau}$ have the same mean and variance, the only difference being that $\hat{\tau}_1$, $\hat{\tau}_2$, . . ., are now assumed to be independent.

Whitt then proposes a second approach, which he terms the *asymptotic interval method*, where the means of τ and $\hat{\tau}$ are the same, but the variances differ. Using an important result by Smith (1959), Whitt estimates the limiting distribution of arrivals over a long period of time, and from this obtains a different (lower) estimate for the variance of $\hat{\tau}$. Comparing the two approaches, Whitt found that they produced upper and lower bounds for the actual length of the queue (found using simulation), and that an even sharper estimate could be obtained by taking a convex combination of the two bounds. Furthermore, the appropriate combination was found to depend on ρ, with the stationary method being more accurate for small values of ρ and the asymptotic method improving for large values of ρ.

The approach that is proposed here for approximating bulk queues is to focus on the *increment* of the arrival process. Specifically, let Y(t) denote the associated counting process and let t_1, t_2, . . ., be departure instants. The variables Y_n, then, are given by $Y_n = Y(t_{n+1}) - Y(t_n)$, and hence represent successive increments of the counting process Y(t). All the results described in chapter 3 apply to any process for which the sequence $\{Y_n\}$ is i.i.d.. Of course, this can only occur if the arrival process is compound Poisson, or if the number of *groups* arriving in each interval (t_{n+1}, t_n) is deterministic, and the size of each group forms a sequence of i.i.d. random variables. In most other cases, the variables Y_n will not be independent. However, we may artificially construct a new random variable $\hat{Y}$, where $\hat{Y}_1$, $\hat{Y}_2$, . . ., are i.i.d., and where we specify the moments of $\hat{Y}$ in such a way that the new arrival process behaves in a manner similar to that of the original one. In other words, rather than finding a new arrival process by creating a new interarrival time distribution, we are creating a new distribution for successive increments of a counting process and then assuming that these increments are independent.

The next question is how to estimate the moments of $\hat{Y}$. It seems natural to require that $E(\hat{Y}) = E(Y)$, i.e. the expected number of arrivals in each increment of the new process should equal that of the original. To estimate the variance of $\hat{Y}$, there are two approaches that are suggested by the work of Kuehn and Whitt and are thereby named in an analogous manner. These are:

a) the stationary increment method - simply let var $(\hat{Y})$ = var (Y); in other words, $\hat{Y}$ will have approximately the same distribution as Y, but the successive correlations are ignored;

b) the asymptotic increment method - calculate the variance of the total number of arrivals over a very long period of time, and from this infer the variance of the number of arrivals over a period of time equal to one departure interval.

These approaches can be illustrated by applying them to the same problem considered in section 5.2, where N identical renewal processes were superimposed to form an arrival process to a bulk queue. Interarrival times for each component process are given by:

$$\tau = 1/5 \cdot N + \varepsilon$$

where $\varepsilon \sim U(-2N, 2N)$. The capacity of the outbound vehicle is $c = 5$ and a utilization ratio of $\rho = .9$ is assumed, thus $\overline{Y} = \rho c = 4.5$. The variance of Y, $\overline{\overline{Y}}$, was estimated from the simulation program and is shown in table 5.2 for different values of N, along with the associated correlation coefficients. To apply the stationary increment method to estimate the mean queue length, we would define a random variable $\hat{Y}$ with mean and variance $\overline{Y}$ and $\overline{\overline{Y}}$. Assuming that the number of arrivals between successive departures is independently and identically distributed according to $\hat{Y}$, then the mean queue length is given by equation 3.17 as follows:

$$\overline{Q} = \frac{\overline{\overline{Y}} + c - \overline{Y} - (c - \overline{Y})^2}{2(c-\overline{Y})} + \sum_{i=0}^{c-1} \frac{1}{1-z_i} \qquad 3.17$$

We still have the problem of finding the zeroes for which we would need the actual distribution of $\hat{Y}$, as opposed to just its moments. For simplicity, we will compute the zeroes by assuming $\hat{Y}$ has a Poisson distribution with mean $\overline{Y}$.

To apply the asymptotic increment method, we again assume $\hat{Y}$ has mean $\overline{Y}$, but we must turn to another source to find $\overline{\overline{Y}}$. For this problem we may use a limiting result reported by Smith (1959) for the variance of the number of arrivals in a renewal process over a long period of time. Let $Y_i(t)$ be the number of arrivals in $(0,t)$ from one component process, and let τ_i be a random variable describing the interarrival time. Then, as $t \to \infty$, Smith shows:

Table 5.2

Simulated variance of Y for superimposed process

N	R	$\bar{\bar{Y}}$
5	-.32	1.05
10	-.59	2.54
20	-.29	3.41
30	-.18	3.84
40	-.11	3.93
50	-.07	4.05
60	-.06	4.24

N = number of streams being superimposed

R = correlation coefficient between Y_n and Y_{n+1}

$\bar{\bar{Y}}$ = variance of Y

$$\mathrm{Var}\Big(Y_i(t)\Big) \simeq \frac{\mathrm{Var}(\tau)}{[E(\tau)]^3} \quad \cdot t$$

For the example, $E(\tau) = 5 \cdot N$ and $\mathrm{Var}\ (\tau) = \frac{(10\delta N)^2}{12} = \frac{(4N)^2}{12} = 1.33\ N^2$. Thus:

$$\mathrm{Var}\ Y_i(t) \simeq \frac{.0107}{N}\ t$$

Now let $Y(t) = Y_1(t) + Y_2(t) + \ . \ . \ . + Y_N(t)$ be the counting process for the superposition of all N streams. Since these streams are independent and identically distributed, we find:

$$\begin{aligned}\mathrm{Var}[Y(t)] &= \sum_{i=1}^{N} \mathrm{Var}[Y_i(t)] \\ &= N\ \mathrm{Var}[Y_i(t)] \\ &= 0.107t\end{aligned}$$

which is independent of N. We now want the variance over a length of time equal to the departure interval, which for this problem is $t = 22.5$. Hence $\mathrm{Var}[Y(t = 22.5)] = .24$, which we may use as our estimate of the variance of $\hat{Y}$ (note, as expected, that this estimate of the variance is significantly lower than that reported in table 5.2 for the stationary method). We can substitute $\overline{Y} = 4.5$ and $\overline{\overline{Y}} = .24$ into equation 3.17 (along with the necessary zeroes) to find the estimate of the mean queue length.

Figure 5.25 shows the true estimate of $\overline{Q}$, obtained from simulation and those calculated using the stationary and asymptotic increment methods. As anticipated, these two approximations appear to provide

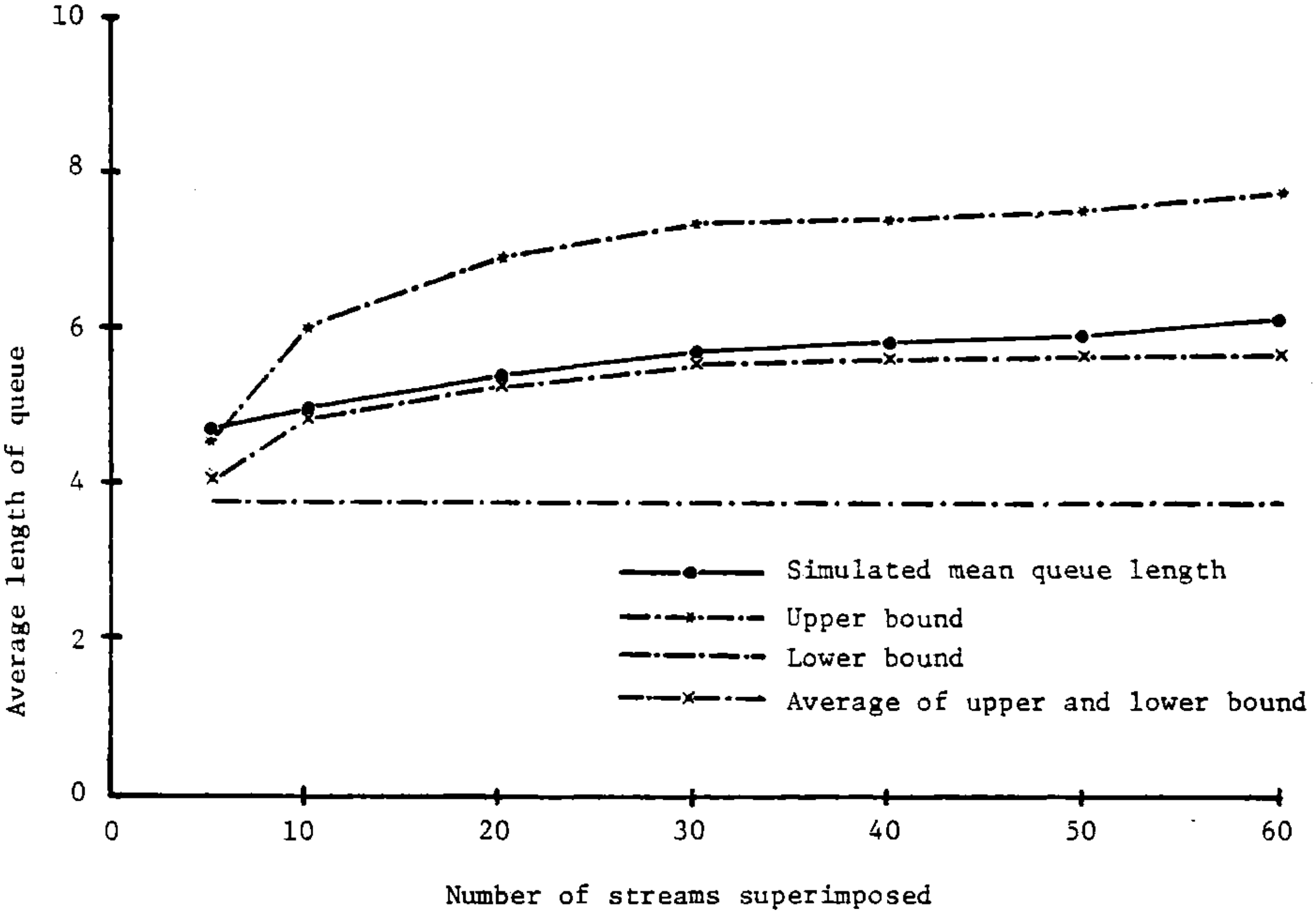


Figure 5.25

Comparison of approximations for estimating mean queue length upper and lower bounds on the mean queue length. To improve the estimate, an average of the two bounds is also shown, which appears to be much more accurate. Of course, it is possible to contemplate more general convex combinations of the two bounds, which would further sharpen the approximation. This, however, represents a significant research problem which extends beyond the scope of this research.

5.4 Summary

The question of when an arrival process is approximately Poisson arises often in queueing problems, and to date, there are no clear criteria for answering it. This chapter develops a number of insights regarding the use of certain criteria for judging whether a process is Poisson, and the accuracy of the Poisson approximation in predicting the moments of queues. Attention is focused on the particular case of examining arrival streams of vehicles from multiple terminals. The results of the simulation experiments suggest the following conclusions:

- Analysis of an arrival process to a bulk queue based on interarrival times can be misleading.
- Analysis of the increments of an arrival process can be useful, but does not necessarily provide a good indication of when a Poisson approximation will provide an accurate estimate of the length of the queue.
- In general, the arrival of vehicles to a queue cannot be accurately approximated using a Poisson arrival assumption.

In light of these observations, section 5.3 proposes a methodology for approximating bulk queues with general arrival processes. This approach is illustrated in the context of a specific problem, but requires considerably more research before it can be applied to more general problems. What is significant about the approximation is that it draws directly off the theoretical work performed in chapter 3, thereby enhancing the applicability of the moment formulas provided there.

Chapter 6 Summary and Directions for Further Research

The motivation for this research is the development of planning tools capable of modeling stochastic delays in transportation terminals. Other tools such as simulation are too slow to be applied to the study of large transportation networks, and in addition, pose problems with respect to model development and the statistical analysis of the outputs. Instead, the theory of modeling bulk queues in steady state is used on the premise that such results strike a better balance between computational efficiency and level of detail. The application of the theory of consolidation terminals, however, posed a number of theoretical and practical problems which must be overcome. The identification, and in some cases, solution of these problems, together with the basic approach used to study stochastic delays, constitutes the contribution of this thesis. In this chapter, these contributions are reviewed, in section 6.1, followed by a summary of important directions for further research in section 6.2.

6.1 Summary of major results

The contributions of this research can be divided into three categories, namely conceptual, theoretical, and numerical. From a conceptual perspective, transportation networks have been described as a network of bulk queues. Stochastic flows over the network are characterized by the flow of groups of loads, with particular interest on the distribution of the number of groups arriving in a given period of time, and the distribution of the size of each group. Outbound links are described as individual queues and the theory of bulk queues are applied to find the delays encountered before departing over a particular link. The observation is made that the most important source of random delays occurred at transfer points in terminals, and hence linehaul delays are ignored. Sources of delay within a terminal are identified as unloading time, connection delay, and loading time, and approaches for modeling each are described. Specifically, the use of steady state bulk queueing theory is identified as a more accurate approach than deterministic models without the computational overhead of a simulation model.

In the second category, theoretical results, the research begins with a thorough review of the literature which identifies potentially useful contributions. At the same time, a set of problems are highlighted which require additional work. In particular, the $M^x/G^y/1$ scheduled departure queue, denoting a system with compound Poisson arrivals, and a general interdeparture time distribution where the outbound vehicle has a random capacity y, is investigated. Also studied is a variant of this system which allows for cancellation of a departure if the queue is less than a specified minimum. New results obtained in this area are:

- transform of the queue length distribution for the $M^x/G^c/1$ queue and formulas for the first two moments;
- transform of the queue length distribution for the $M^x/G^y/1$ queue (extension to random outbound capacities);
- formulas for the first two moments of the waiting time distribution for $M^x/G^y/1$ queues;
- the queue length transform for $M^x/G^y/1$ queue with cancellations;
- a light traffic approximation for queues with cancellations;
- the queue length transform when departure headways form an alternating renewal process;
- the relationship between the distribution between the number of units:

  a) at a dispatch instant,

  b) at a random point in time,

  c) in front of an arriving unit, and

  d) behind a departing unit.

In addition to these results, the derivations used are generally simpler than the usual approach used in the literature.

On the numerical side of the research, several practical issues are raised in connection with implementing the results. The first of these is the immediate problem of solving the transform equations, requiring the calculation of the roots of a certain transcendental equation. Contrary to recent indications in the literature, this task did not present any numerical difficulties and in fact could be performed extremely efficiently. A new approach for performing partial transform inversions is described which is both computationally more efficient and less sensitive to round-off errors than the standard technique. A method for performing complete transform inversions is also presented, but is found to be extremely sensitive to computer round-off. However, it is pointed out that the results of the partial transform inversion, which provides the first c elements of the steady state queue length distribution, could be used to compute several important levels of service statistics. These include the probability a randomly chosen unit leaves on the first outbound vehicle, the probability a vehicle is cancelled due to insufficient demand, and the average load factor on a departing vehicle.

In addition to the work directed at solving the transforms, several approximations are developed which simplify the analysis and enhance the generality of the results. The first of these are simple but accurate approximations for the first two moments of the length of the queue for an $M/E_k^c/1$ system. These formulas are in closed form and eliminate the problem of having to solve for zeroes. Second, a family of discrete distributions is described, and the approximate moment formulas are used to fit an approximate distribution for the length of the queue. When compared to the exact distribution, the approximation is shown to be extremely accurate.

Third and last, a series of experiments were run which test the accuracy of the Poisson arrival process as an approximation for certain non-Poisson arrival processes. It is shown that the Poisson arrival process is not a particularly robust approximation, even when tests based on an examination of the interarrival times suggest that a given arrival process is in fact Poisson. On the basis of these observations, a new approach for approximating the behavior of $G^x/G^y/1$ queues is proposed. The methodology proceeds by replacing a general arrival process with one with independent increments. No attempt is made, however, to actually derive the new process in terms of its interarrival time distribution, since such a process will not in general even exist (in that the Poisson is the only arrival process with independent increments). Rather, we simply estimate a distribution for the increment of the process and then assume that the successive increments are independent. The procedure is illustrated in the context of a specific problem and shown to yield fairly good results.

6.2 Directions for further research

As with most concerted efforts in a new area, a number of questions and problems remain which deserve additional work. A few of these are:

- Approximate moment formulas for the $M^x/G^c/1$ queue - Thus far, formulas have been estimated only for the $M/E_k^c/1$ queue, and only for the case c = 20. It should be possible to incorporate the capacity of the outbound vehicle explicitly, as well as the moments of the size of incoming groups.
- Validation - All the formulas should be tested in specific field experiments. Errors in estimating parameters and other assumptions (e.g. steady state) may overshadow the use of more complex models (such as the use of random outbound capacities versus deterministic ones). The results of an analysis based on stochastic models should be compared with those obtained using simpler, deterministic models.
- Approximating $G^x/G^y/1$ queues - Considerably more work is needed in the area of approximating the performance of bulk queues with general arrival processes. It may be possible to parameterize a range of arrival processes on the basis of the first two or three moments of the number of arrivals during each service period, and the correlation coefficient. Different approximations may prove more accurate for different values of ρ, and this relationship should be explored.

- Approximating queueing networks - Given an approximation for $G^x/G^y/1$ queues, it should be possible to extend the methodology outlined by Kuehn (see section 2.2) to networks of bulk queues. The important problem is describing the output process of a queue, and then using this information to approximate the arrival processes to queues downstream. Stochastic flow over links in the network may be characterized by the first two moments of both the number of groups arriving in a given time interval and the size of each group. Correlations in successive random variables can be incorporated by modifying the variance and higher moments.

References

Arnold, J.H., "Waiting Time Analysis of Container Terminal Queues", S.M. Thesis, Massachusetts Institute of Technology, Department of Civil Engineering (1974).

Arora, K.L., "Two-Server Bulk Queueing Process", *Oper Res* Vol 12 (1964) pp 286-294.

Bagchi, T.P., and Templeton, J.G.C., *Numerical Methods in Markov Chains and Bulk Queues*, Lecture Notes in Economics and Mathematical Systems, (M. Beckmann, G. Goos and H.P. Kunzi, Eds.), Springer-Verlag, New York (1972).

Bahary, E. and Kolesar, P., "Multilevel Bulk Service Queues", *Oper Res* Vol 20, No 2 (1972) pp 406-420.

Bailey, N.T.J., "On Queueing Processes with Bulk Service", *J Royal Stat Soc B* Vol 16 (1954) pp 80-87.

Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F., "Open, Closed and Mixed Networks of Queues with Different Classes of Customers", *J ACM* Vol 22 (1975) pp 248-260.

Bhat, U.N., "On Single Server Bulk Queueing Processes with Binomial Input", *Oper Res* Vol 12, No 4 (1964) pp 527-533.

Bloemena, A.R. "On a Queueing Process With a Certain Type of Bulk Service", *Bull Inst Int Stat*, Vol 37 (1960) pp 219-277.

Borthakur, A., "A Poisson Queue With a General Bulk Service Rule", *J Assam Sc Soc* Vol XIV, No 2 (1971) pp 162-167.

__________, and Medhi, J., "A Queueing System With Arrivals and Departures in Batches of Variable Size", *Cahiers du Centre d'Etudes de Recherche Operationelle* Vol 16 (1974).

Brumelle, S.L., "Some Inequalities for Parallel Service Queues", *Oper Res* Vol 19 (1971), pp 402-413.

Burke, P.J., "The Output of a Queueing System", *Oper Res* Vol 4 (1956) pp 699-704.

__________, "Output Processes and Tandem Queues", *Proc of the Symposium on Computer-Communications Networks and Teletraffic* (1972) pp 419-428.

__________, "Delays in Single-Server Queues with Batch Input" *Oper Res* Vol 23 (1975) pp 830-833.

Chandy, K.M., "The Analysis and Solutions for General Queueing Networks", *Proc Sixth Annual Princeton Conference on Information Sciences and Systems*, Princeton Univ. (1972).

Chandy, K.M., Herzog, U., Woo. L., "Approximate Analysis of General Queueing Networks", *IBM J Res Dev* Vol 19 (1975) pp 43-49.

__________, Howard, J.H., Towsley, D.F., "Product Form and Local Balance in Queueing Networks", *J ACM* Vol 24 (1977) pp 250-263.

__________, and Sauer, C.H., "Approximate Methods for Analyzing Queueing Network Models of Computer Systems", *Computing Surveys* Vol 10 (1978) pp 281-317.

Chaudry, M.L., "The Queueing System $M^x/G/1$ and its Ramifications", *Nav Res Log Quart* Vol 26, No 4 (1979) pp 667-674.

Churchill, R.V., Brown, J.W., and Verhey, R.F., *Complex Variables and Applications*, McGraw-Hill, New York (1974).

Cinlar, E., "Superposition of Point Processes", *Stochastic Point Processes: Statistical Analysis, Theory and Applications* (P.A.W. Lewis, Ed.), John Wiley and Sons, New York (1972) pp 549-606.

Cohen, J.W., *The Single Server Queue*, North Holland, London (1969).

Conolly, B.W., "Queueing at a Single Serving Point with Group Arrival", *J Roy Stat Soc B* Vol 22 (1960) pp 285-298.

Cox, D.R., *Renewal Theory*, John-Wiley and Sons Ltd, New York (1962).

__________, and Smith, W.L., "On the Superposition of Renewal Processes", *Biometrika* Vol 41 (1954) pp 91-99.

Crane, M., "Queues in Transportation Systems I: A Markovian System", *J Appl Probability* Vol 10 (1973) pp 630-643.

__________, "Queues in Transportation Systems II: An Independently Dispatched System", *J Appl Probability* Vol 11 (1974) pp 145-158.

Dahlquist, G., Bjork, A., and Anderson, N., *Numerical Methods*, Prentice-Hall Inc, Englewood Cliffs, New Jersey (1974).

Daley, D.J., "Notes on Queueing Output Processes", Lecture Notes in Economics and Mathematical Systems (M. Beckmann and H. P. Kunzi, Eds), *Mathematical Methods in Queueing Theory*, Springer-Verlag, New York (1973) pp 351-358.

Disney, R.L., and Cherry, W.P., "Some Topics in Queueing Network Theory", Lecture Notes in Economics and Mathematical Systems (M. Beckman and H.P. Kunzi, Eds.), *Mathematical Methods in Queueing Theory*, Springer-Verlag, New York (1973) pp 23-44.

Downton, F., "Waiting Time in Bulk Service Queues", *J Royal Stat Soc B* Vol 17 (1955) pp 256-261.

__________, "On Limiting Distributions Arising in Bulk Service Queues", *J Royal Stat Soc B* Vol 18 (1956) pp 265-274.

Foster, F.G., "Queues With Batch Arrivals - I", Acta Math Acad Hung Vol 12 (1961) pp 1-10.

__________, Queues With Batch Arrivals - II", Acta Math Acad Hung Vol 16 (1965) pp 275-287.

Foster, F.G., "Batched Queueing Processes", Oper Res Vol 12, No 3 (1964) pp 441-449.

Gaver, D.P., "Imbedded Markov Chain Analysis of a Waiting Line Process in Continuous Time", Ann Math Stat Vol 30 (1959) pp 698-720.

Ghare, P.M., "Multichannel Queueing System with Bulk Service", Oper Res Vol 16 (1968) pp 189-192.

Gordon, W.J., and Newell, G.F., "Closed Queueing Systems with Exponential Servers", Oper Res Vol 15 (1967) pp 254-265.

Gougenheim-Creange, D. "Analysis of Queueing Networks", S.M. Thesis, Massachusetts Institute of Technology, Dept. of Civil Engineering (1976).

Groninger, K.L., "The Relationship Between Carrier Capacity and Mean Passenger Waiting Time", S.M. Thesis, Massachusetts Institute of Technology, Dept. of Civil Engineering (1966).

Gupta, S.K., "Queues with Batch Poisson Arrivals and a General Class of Service Time Distributions", J Industrial Eng Vol 15 (1964) pp 319-320.

__________, and Goyal, J.V., "Queues with Batch Poisson Arrivals and Hyperexperrential Service", Nav Res Log Quart Vol 12 (1965) pp 323-329.

Harris, C.M., "Some Results for Bulk Arrival Queues with State Dependent Service Times", Management Sci Vol 16 (1970) pp 313-326.

Hirasawa, K. "Numerical Solutions of Bulk Queues via Imbedded Markov Chain", Electrical Engineering in Japan Vol 91, No 1 (1971) pp 127- 136.

Iglehart, D.L., and Whitt, W., "Multiple Channel Queues in Heavy Traffic II: Sequences, Networks and Batches", Adv in Appl Prob Vol 2 (1970) pp 355-369.

Jackson, J.R., "Networks of Waiting Lines", Oper Res Vol 5 (1957) pp 518-521.

__________, "Jobshop-Like Queueing Systems", Management Sci Vol 10 (1963) pp 131-142.

Jaiswal, N.K., "Bulk Service Queueing Problem", Oper Res Vol 8 (1960a) pp 139-143.

Jaiswal, N.K., "Time Dependent Solution of the Bulk Service Queueing Problem", Oper Res Vol 8 (1960b) pp 773-781.

Jensen, G.L., and Paulson, A.S., "Explicit Steady State Solutions for a Particular $M^{(K)}/M/1$ Queueing System", Nav Res Log Quart Vol 24, No 4 (1978) pp 651-659.

Kabak, I., "Blocking and Delays in $M^{(x)}/M/c$ Bulk Arrival Queueing Systems", Management Sciences Vol 17, No 1 (1970) pp 112-115.

Kashyap, B.R.K., "The Double Ended Quene with Bulk Service and Limited Waiting Space", Oper Res Vol 14 (1966) pp 822-834.

Keilson, J., "The General Bulk Queue as a Hilbert Problem", J Roy Stat Soc Sci B Vol 24 (1962) pp 344-358.

Kelly, F.P., "Networks of Queues", Adv. Appl. Prob., Vol 8 (1976), pp. 416-432.

__________, Reversibility and Stochastic Networks, John Wiley and Sons, New York (1979)

Khintchine, A.J., Mathematical Methods in the Theory of Queueing, Grinnin, Dondon (1960).

Kleinrock, L., Communication Nets: Stochastic Message Flow and Delay, McGraw-Hill, New York (1964); out of print; reprinted by Dover Publications, New York (1972).

__________, Queueing Systems Vol 1: Theory, John Wiley and Sons, New York (1975).

__________, Queueing Systems Vol 2: Computer Applications, John Wiley and Sons, New York (1976).

Kobayashi, H., "Application of the Diffusion Approximation to Queueing Networks, Parts I and II", J ACM Vol 21 (1974) pp 316-328 and pp 459-469.

Kuehn, Paul J., "Approximate Analysis of General Queueing Networks by Decomposition", I.E.E.E. Transactions on Communications Vol COMM-27 No 1 (1979) pp 113-126.

Lemoine, A.J., "Networks of Queues - A Survey of Equilibrium Analysis", Mgmt Sci Vol 24, No 4 (1977) pp 464-481.

__________, "Networks of Queues - A Survey of Weak Convergence Results" Mgmt Sci Vol 24 No 11 (1978) pp 1175-1193.

Luchak, G., "The Continuous Solution of the Equations of the Single Channel Queue with a General Class of Service Time Distributions by the Method of Generating Functions", J Roy Stat Soc B Vol 20 (1958) pp 176-181.

Medhi, J., "Waiting Time Distribution in a Poisson Queue With a General Bulk Service Rule", Management Sci Vol 21 No 7 (1975) pp 777-782.

Medhi, J., "Further Results on Waiting Time Distribution in a Poisson Queue Under a General Bulk Service Rule",*Cahiers du Centre d'Etudes de Recherche Operationelle* Vol 21, No 2 (1979) pp 183-189.

__________, and Borthakur,A., "On a Two Server Markovian Queue With a General Bulk Service Rule", *Cahiers du Centre d'Etudes de Recherche Operationelle* Vol 14 (1972) pp 151-158.

Miller, R.G., "A Contribution to the Theory of Bulk Queues", *J Royal Stat Soc B* Vol 21 (1959) pp 320-337.

Neuts, M.F., "The Busy Period of a Queue With Batch Service", *Oper Res* Vol 13 (1965) pp 815-819.

__________, "Semi-Markov Analysis of a Bulk Queue", *Bull Soc Math Belgique* Vol 18 (1966) pp 28-42.

__________, "A General Class of Bulk Queues With Poisson Input" *Ann Math Stat* Vol 38 (1967) pp 759-770.

__________, "Moment Formulas for the Markov Renewal Branching Process", *Adv Appl Prob* Vol 8 (1976) pp 690-711.

__________, "Some Explicit Formulas for the Steady State Behavior of the Queue With Semi-Markovian Service Times", *Adv Appl Prob* Vol 9 (1977) pp 141-157.

__________, "Queues Solvable Without Rouche's Theorem", *Oper Res* Vol 27, No 4 (1979) pp 767-781.

Novaes, A.G.N., "Operational Optimization of a Short-Distance Ferryboat System", S.M. Thesis, Massachusetts Institute of Technology, Dept. of Naval Architecture (1963).

Ohno, K., "Computational Algorithm for a Fixed Cycle Traffic Light and New Approximate Expressions for Average Delay", *Trans Sci* Vol 12, No 2 (1978) pp 29-47.

Pack, C.D., "The Output of an M/D/1 Queue", *Oper Res* Vol 25, No 4 (1975) pp 750-760.

Page, E.S., "Computers and Congestion Problems", *Proc Symp on Congestion Theory* (W.L. Smith and W.E. Wilkinson, Eds.), Univ. of North Carolina Press (1965).

Peterson, E.R., Fullerton, H.V. (Eds.), *The Railcar Network Model*, Canadian Institute of Guided Ground Transport, Queen's University, Kingston, Ontario, Report No 75-11 (1975).

__________, "Railroad Modeling: Part I. Prediction of Put Through Times", *Trans Sci* Vol 11, No 1 (1977a) pp 37-49.

Peterson, E.R., "Railyard Modeling: Part II. The Effect of Yard Facilities on Congestion", *Trans Sci* Vol 11, No 1 (1977b) pp 50-59.

__________, "Bulk Queues with Random Batch Size: With Application to Railroad Modeling", Work paper series No 71-3, Canadian Institute of Guided Ground Transport, Queen's University, Ontario (1971)

Ponstein, J., "Theory and Numerical Solution of a Discrete Time Queueing Problem", *Statistica Neerlandica* Vol 20 (1974) pp 139-152.

Restrepo, R.A., "A Queue With Simultaneous Arrivals and Erlang Service Distributions", *Oper Res* Vol 13 (1965) pp 375-381.

Richardson, F.W., "Analysis of Stochastic Demand on Motor Carrier Operating Strategies", S.M. Thesis, Massachusetts Institute of Technology, Dept. of Civil Engineering (1979).

Roes, P.B.M., "A Many Server Bulk Queue", *Oper Res* Vol 14 (1966) pp 1037-1044.

Saaty, T.L., *Elements of Queueing Theory*, McGraw-Hill Book Co., New York (1961).

Sahbazov, A.A., "A Problem of Service With Non-Ordinary Demand Flow", *Soviet Mathematics Doklady* Vol 3 (1962) pp 1000-1003.

Sirbu, M.A., "Waiting Times in a Class of Bulk Queues", S.M. Thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering (1968).

Smith, W.L., "On the Cumulants of Renewal Processes", *Biometrika* Vol 46 (1959) pp 1-29.

Snyder, D.L., *Random Point Processes*, John Wiley and Sons, New York (1975).

Takacs, L., *Introduction to the Theory of Queues*, Oxford Univ. Press, New York (1962).

Teghem, J., Loris-Teghem, J., and Lambotte, J.P., *Modeles d'attente M/G/1 et GI/M/1 a arrivees et services en groups*, Lecture Notes in Operations Research and Mathematical Economics, 8, Springer-Verlag, New York (1969).

Terziev, M., Richardson, F., and Roberts, P., "An Approach for the Evaluation of Freight Network Operations: A Stochastic Supply Model for the Regular Route Motor Carrier Industry", *Transportation Research Forum* (1978), pp 260-269.

Turnquist, M.A., and Bowman, L.A., "Control of Service Reliability in Transit Networks", The Transportation Center, Northwestern University, January (1979).

Whitt, W., "Approximating a Point Process by a Renewal Process I: A General Framework", unpublished working paper, Bell Laboratories, November (1979).

__________ , "Approximating a Point Process by a Renewal Process: The View Through a Queue, an Indirect Approach", unpublished working paper, Bell Laboratories, November (1979).

Wijngaard, J., "A Direct Numerical Method for a Class of Queueing Problems", Management Science Vol 24, No 13 (1978) pp 1441-1447.

Appendix A Neuts' algorithm

The numerical procedure developed by Neuts was motivated by the perceived computational problems associated with finding the roots of a given function. The algorithm is of interest here since it applies to a very wide class of bulk queueing problems. It was not given any further attention, however, since the calculations required suggested that the method would simply be too slow to be of any practical use in the problems being considered. In this section, the major steps of this algorithm will be outlined, the purpose being only to communicate the computational requirements. No attempt will be made to motivate any of the steps, as derivations are often quite involved.

Neuts assumes that the imbedded Markov process can be represented by the following general transition matrix:

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & \cdots \\ A_0 & A_1 & A_2 & \cdots \\ 0 & A_0 & A_1 & \cdots \\ 0 & 0 & A_0 & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix}$$

where the elements of P are c x c matrices with general form:

$$B_j = \begin{bmatrix} b_{o,cj} & \cdots & b_{o,cj + c - 1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ b_{c-1,cj} & \cdots & b_{c-1,cj+c-1} \end{bmatrix} \qquad j=0,1,\ldots$$

$$A_o = \begin{bmatrix} a_0 & a_1 & \cdots & a_{c-1} \\ 0 & a_0 & \cdots & a_{c-2} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & a_0 \end{bmatrix}$$

$$A_j = \begin{bmatrix} a_{cj} & \cdots & a_{cj+c-1} \\ \vdots & & \vdots \\ a_{cj-c+1} & \cdots & a_{cj} \end{bmatrix}$$

To incorporate the special structure of P in the analysis, denote a state by the pair (r,j), $r \geq 0$, $1 \leq j \leq c$, where r will be referred to as the level and j as the state within level r. The vector of probabilities $q_o, \ldots, q_{c-1}$ will now be found by finding the mean recurrence time for each state (o,j), $1 \leq j \leq c$.

Let $G = \{G_{ij}\}$be a cxc matrix of probabilities that given the system is in state (r+1, i), the first state visited in level r will be j. From the structure of P, it is easily seen that this matrix is independent of r. By considering the sequence of possible events following a transition from (r+1, j), it can be seen that G must satisfy the relation:

$$G = \sum_{i=0}^{\infty} A_i \cdot G^i$$

Rearranging terms slightly, Neuts obtains the following algorithm for finding G given an initial starting point $G_o = \{0\}$:

$$G_{n+1} = (I-A_1)^{-1}(A_o + \sum_{j=2}^{\infty} A_j G_n^j)$$

Given the matrix G, the next problem is to find the stationary probability vector g for the (stochastic) matrix G, where g = g G. Neuts does not give an explicit procedure in his last paper for doing this, but presumably the relatively small dimensions of G simplify this problem.

Before describing the next step, some new variables are needed. Let

$$A = \sum_{j=0}^{\infty} A_j,$$

$$e = (1, \ldots, 1),$$

$$\alpha^o = (\alpha_1^o \ \ldots, \alpha_c^o) = \sum_{j=o}^{\infty} jA_j e,$$

and let $\Delta(\alpha^o)$ be the diagonal matrix whose elements are $(\alpha_1^o \ldots, \alpha_c^o)$. Finally let $\tilde{G}$ be the matrix whose rows are g. Now compute the following vector μ:

$$\mu = (I-G+\tilde{G}) \ (I-A+\tilde{G}-\Delta \ (\alpha^o)\tilde{G})^{-1} e \qquad (17)$$

With G, g and μ thus computed, it is now possible to address the problem somewhat more directly. Let

$$K = \sum_{j=o}^{\infty} B_j G^j$$

and find the vector γ such that $\gamma = \gamma K$. Let

$$\beta^{o}=(\beta_{1}^{o},\ldots,\beta_{c}^{o})=\sum_{j=1}^{\infty} jB_{j}e,$$

and

$$B=\sum_{j=o}^{\infty} B_{j}$$

and, as usual, let ρ be the utilization parameter. Now compute the vector θ where:

$$\theta = e + (1-\rho)^{-1}\beta^{o} + (B-K)\ (I-G+\tilde{G})^{-1}\mu$$

The vector $(q_{o},\ldots, q_{c-1})$ is now given by $(\gamma\theta)^{-1}\gamma$. The remaining probabilities q_{m}, q_{m+1}, ... , are easily found from a series of recursive formulae.

In view of the considerable matrix manipulations required, it is difficult to understand how the algorithm can possibly compete with the method of convolutions which is far simpler as well as being more flexible with regard to the variety of dispatching strategies which may be represented. Neuts has tested the algorithm extensively but makes no comparisons against other numerical procedures. For c=10, typical run times were on the order of 5 seconds on a CDC 6500.

Appendix B Proof that the zeroes lie within the unit circle

As described in the text, Rouche's theorem tells us that there are c zeroes for the function $f(z) = z^c - Y(z)$ on or within the unit circle, if it can be shown that $|z^c|>|Y(z)|$ along the contour defined by $|z| = 1 + \delta$, for $\delta > 0$ but arbitrarily small. We will now demonstrate that this is true for any Y(z) that is a legitimate probability mass function of a non-negative, integer valued random variable, as long as a) $Y < c$, and b) the transform Y(z) is defined on the unit circle $|z| = 1 + \delta$, for δ arbitrarily small but positive. The latter of these conditions is satisfied if all the moments of Y are finite, a condition that is likely to be satisfied for any distributions encountered in practice.

First, we let $z = (1 + \delta)e^{i\theta}$, $0 \leq \theta \leq 2$, and note that:

$$\begin{aligned} |z^c| &= |(1+\delta)^c\ e^{ci\theta}| \\ &= (1+\delta)^c \\ &= 1+C\delta + 0(\delta^2) \end{aligned} \qquad \text{B.1}$$

We also have:

$$Y(z) = \sum_{k=0}^{\infty} y_k z^k \qquad \text{B.2}$$

Or:

$$Y(z)\Big|_{z=(1+\delta)e^{i\theta}} = \sum_{k=0}^{\infty} y_k (1+\delta)^k\ e^{ki\theta} \qquad \text{B.3}$$

Applying the triangle inequality we observe:

$$Y\ (z)\ \Big|_{z=(1+\delta)e^{i\theta}} = \left| \sum_{k=0}^{\infty} y_k (1+\delta)^k\ e^{ki\theta} \right|$$

$$\leq \sum_{k=0}^{\infty} \left| y_k (1+\delta)^k e^{ki\theta} \right|$$

$$\leq \sum_{k=0}^{\infty} y_k (1+\delta)^k \qquad \text{B.4}$$

Rewriting the right hand side of B.4, we obtain:

$$\sum_{k=0}^{\infty} y_k (1+\delta)^k = \sum_{k=0}^{\infty} y_k \left(1 + k\delta + \frac{k(k-1)}{2}\delta^2 + \frac{k(k-1)(k-2)}{3\ 2}\delta^3 + \ldots\right)$$

$$= 1 + \overline{Y}\delta + \sum_{k=2}^{\infty} \frac{\gamma_k}{k!}\delta^k \qquad \text{B.5}$$

where γ_k is the k^{th} factorial moment of Y, which is equivalent to $\left.\frac{d^kY(z)}{dz^k}\right|_{z=1}$. Since the right hand side of B.5 is bounded by assumption, we may write:

$$\sum_{k=0}^{\infty} y_k (1+\delta)^k = 1 + \overline{Y}\delta + 0(\delta^2) \qquad \text{B.6}$$

Since B.6 is an upper bound for Y(z), it is sufficient to show that the right hand side of B.6 is less than B.1. Noting that the $0(\delta^2)$ terms become negligible as $\delta \to 0$, we have:

$$1 + c\delta > 1 + \overline{Y}\delta$$

which is true as long as $c > \overline{Y}$. We observe that the condition $\rho < 1$ is sufficient but not necessary for the zeroes to be within the unit circle, since they may also be there (as was in fact found to be the case) for $\rho < 1$. What may happen for $\rho > 1$ is that the number of zeroes inside the unit circle becomes greater than c.

Appendix C Moment formulas for the length of the queue

In this section we will outline the derivation for the mean and variance of the length of the queue. The purpose of this appendix is to assist others in verifying the formulas in 3.17 and 3.18, since the algebra is quite lengthy and mistakes are easily made.

Proceeding to find the mean, we have from equation 3.14:

$$Q(z) = \frac{(c-\bar{Y})(z-1) \prod_{i=0}^{c-1} (\frac{z-z_i}{1-z_i})}{\frac{z^c}{Y(z)} - 1} \qquad \text{C.1}$$

To find the mean, we observe that:

$$\left.\frac{dQ(z)}{dz}\right|_{z=1} = \left.\frac{d}{dz} \sum_{j=0}^{\infty} q_j z^j \right|_{z=1}$$

$$= \sum_{j=0}^{\infty} jq_j$$

$$= \overline{Q} \qquad \text{C.2}$$

Thus we have but to find Q'(1). To simplify this, we rewrite C.1 as:

$$Q(z) = A(z) \prod_{i=0}^{c-1} B_i(z) \qquad \text{C.3}$$

where:

$$A(z) = \frac{(c-\bar{y})(z-1)}{\frac{z^c}{Y(z)} - 1}$$

$$B_i(z) = \frac{z - z_i}{1 - z_i}$$

It is easily verified that (using the fact that A(1) = B(1) = 1):

$$Q'(1) = A'(1) + \sum_{i=0}^{c-1} B_i'(1) \qquad \text{C.4}$$

We will now try to find A'(1) in the simplest manner possible. Let:

$$A(z) = \frac{A_1(z)}{A_2(z)} \qquad \text{C.5}$$

where $A_1(z) = (c-\bar{Y})(z-1)$ and $A_2(z) = \frac{z^c}{Y(z)} -1$. Differentiating once yields:

$$A'(z) = \frac{A_1'(z)A_2(z) - A_1(z)A_2'(z)}{A_2(z)^2} \qquad \text{C.6}$$

Noting that both the numerator and denominator vanish at z=1, apply l'Hopital's rule and reducing slightly yields:

$$A'(z) = \frac{A_1''(z)\ A_2 - A_1(z)\ A_2''(z)}{2A_2(z)A_2'(z)} \qquad \text{C.7}$$

Clearly $A_1''(z) \neq 0$. Again we have 0 /0, so we use l'Hopital's rule once more:

$$A'(z) = \frac{-A_1'(z)A_2''(z) - A_1(z)\ A_2''(z)}{2A_2(z)A_2''(z) + 2A_2'(z)^2} \qquad \text{C.8}$$

Letting z=1 and observing that $A_1(1) = A_2(1) = 0$, we have:

$$A'(1) = \frac{-A_1'(1)A_2''(1)}{2A_2'(1)^2} \qquad \text{C.9}$$

Solving, we note:

$$A_1'(1) = c-\overline{Y} \qquad \text{C.10}$$

$$A_2'(1) = \left.\frac{cz^{c-1}}{Y(z)} - \frac{z^c Y'(z)}{Y(z)^2}\right|_{z=1}$$

$$= c - \overline{Y} \qquad \text{C.11}$$

$$A_2''(1) = \frac{c(c-1)z^{c-z}}{Y(z)} - \frac{2cz^{c-1}\,Y'(z)}{Y(z)^2}$$

$$\left. - z^c \; \frac{Y''(z)}{Y(z)^2} - \frac{2Y'(z)^2}{Y(z)^3}\right|_{z=1}$$

$$= c(c-1) - 2c\overline{Y} - Y''(1) - 2\overline{Y}^2 \qquad \text{C.12}$$

We note that $Y'(1) = \overline{\overline{Y}}(1) + \overline{Y}^2 - \overline{Y}$, giving:

$$A_2''(1) = c\,(c-1) - 2c\overline{Y} - \overline{\overline{Y}} + \overline{Y} + \overline{Y}^2 \qquad \text{C.13}$$

Substituting C.13, C.11 and C.10 into C.9 gives:

$$A'(1) = \frac{\overline{\overline{Y}} + c - \overline{Y} - c^2 + 2c\overline{Y} - \overline{Y}^2}{2(c-\overline{Y})}$$

$$= \frac{\overline{\overline{Y}} + c - \overline{Y} - (c-\overline{Y})^2}{2(c-\overline{Y})} \qquad \text{C.14}$$

Also:

$$B'(1) = \frac{1}{1-z_i} \qquad \text{C.15}$$

Combining C.4, C.14 and C.15 gives us 3.17:

$$\overline{Q} = \frac{\overline{\overline{Y}}+c-\overline{Y} - (c-\overline{Y})^2}{2(c-\overline{Y})} + \sum_{i=0}^{c-1} \frac{1}{1-z_i} \qquad \text{C.16}$$

Turning to the variance, we note that:

$$\overline{\overline{Q}} = Q''(1) + Q'(1) - Q'(1)^2 = Q''(1) + \overline{Q} - \overline{Q}^2 \qquad \text{C.17}$$

To find Q''(1), take logs of both sides of C.3 and define:

$$Q_\ell(z) = \log Q(z) = \log A(z) + \sum_{i=0}^{c-1} \log B_i(z) \qquad \text{C.18}$$

Taking derivatives yields:

$$Q_\ell'(z) = \frac{Q'(z)}{Q(z)} = \frac{A'(z)}{A(z)} + \sum_{i=0}^{c-1} \frac{B'(z)}{B(z)} \qquad \text{C.19}$$

$$Q_\ell''(z) = \frac{Q''(z)}{Q(z)} - \frac{Q'(z)^2}{Q(z)^2} = \frac{A''(z)}{A(z)} - \frac{A'(z)^2}{A(z)^2} + \sum_{i=0}^{c-1} \frac{B_i''(z)}{B(z)} - \frac{B_i'(z)^2}{B(z)^2} \qquad \text{C.20}$$

Solving for Q''(z) and letting z=1 gives:

$$Q''(1) = \overline{Q}^2 + A''(1) - A'(1)^2 + \sum_{i=0}^{c-1} [B_i''(1) - B_i'(1)^2] \qquad \text{C.21}$$

Proceeding to find A''(1), we return to C.6 and differentiate again:

$$A''(z) = \frac{[A_1''(z)\ A_2(z) - A_1(z)A_2''(z)]\ \ A_2(z)}{A_2(z)^3}$$

$$- \frac{2[A_1'(z)\ A_2(z) - A_1(z)A_2'(z)]\ A_2'(z)}{A_2(z)^3} \qquad \text{C.22}$$

Again noting that $A_1''(z) = 0$, we see that we will have to apply l'Hopital's rule three times. Proceeding:

$$A''(z) = \frac{\begin{array}{l}[-A_1'(z)\ A_2(z)\ A_2''(z) - A_1(z)\ A_2'(z)\ A_2''(z) - A_1(z)\ A_2(z)\ A_2'''(z) \\ - 2A_1''(z)\ A_2(z)\ A_2'(z) - 2A_1'(z)\ A_2'(z)^2 - 2A_1'(z)\ A_2(z)\ A_2''(z) \\ + 2A_1'(z)\ A_2'(z)^2 + 4A_1(z)\ A_2'(z)\ A_2''(z)]\end{array}}{3A_2(z)^2\ A_2'(z)} \qquad \text{C.23}$$

Reducing:

$$A''(z) = \frac{\begin{array}{l}[-3A_1'(z)\ A_2(z)\ A_2''(z) + 3A_1(z)\ A_2'(z)\ A_2''(z) \\ - A_1(z)\ A_2(z)\ A_2'''(z)\ ]\end{array}}{3A_2(z)^2\ A_2'(z)} \qquad \text{C.24}$$

Applying l'Hopital's rule a second time, but avoiding any terms with $A_1''(z)$:

$$A''(z) = \frac{\begin{array}{l}[-3A_1'(z)\ A_2'(z)\ A_2''(z) - 3A_1'(z)\ A_2(z)\ A_2'''(z) \\ +3A_1'(z)\ A_2'(z)\ A_2''(z) + 3A_1(z)\ A_2''(z)^2 \\ +3A_1(z)\ A_2'(z)\ A_2'''(z) - A_1'(z)\ A_2(z)\ A_2'''(z) \\ -A_1(z)\ A_2'(z)\ A_2'''(z) - A_1(z)\ A_2(z)\ A_2''''(z)]\end{array}}{6A_2(z)\ A_2'(z)^2 + 3A_2(z)^2\ A_2''(z)} \qquad \text{C.25}$$

Reducing:

$$A''(z) = \frac{[\ -4A_1'(z)\ A_2(z)\ A_2'''(z) + 3A_1(z)\ A_2''(z)^2 + 2A_1(z)\ A_2'(z)\ A_2'''(z) - A_1(z)\ A_2(z)\ A_2''''(z)\ ]}{6A_2(z)\ A_2'(z)^2 + 3A_2(z)^2\ A_2''(z)} \qquad \text{C.26}$$

We now apply l'Hopital's rule one last time. In anticipation of letting $z = 1$, any terms containing $A_1(z)$ or $A_2(z)$ will be omitted, since these will vanish. Thus:

$$A''(z) = \frac{[-4A_1'(z)\ A_2'(z)\ A_2'''(z) + 3A_1'(z)\ A_2''(z)^2 + 2A_1'(z)\ A_2'(z)\ A_2'''(z)\ ]}{6A_2'(z)^3} \qquad \text{C.27}$$

Letting $z = 1$ and reducing, using $A_1'(1) = A_2'(1)$, gives:

$$A''(1) = \frac{-2A_2'(1)\ A_2'''(1) + 3A_2''(1)^2}{6A_2'(1)^2} \qquad \text{C.28}$$

Deriving $A_2'''(1)$, we return to C.12:

$$\begin{aligned} A_2'''(z)\Big|_{z=1} &= c(c-1)(c-2) - 3c(c-1)\ \overline{Y} \\ &\quad -3c\ [Y''(1) - 2\overline{Y}^2\ ] \\ &\quad -[Y'''(1) - 6Y''(1)\ \overline{Y} + 6\ \overline{Y}^3\ ] \end{aligned} \qquad \text{C.29}$$

It is easily shown that $Y'''(1) = E(Y^3) - 3E(Y^2) + 2\overline{Y}$.

Also, $\overset{\equiv}{\overline{Y}} = E(Y-\overline{Y})^3 = E(Y^3) - 3\overline{Y}E(Y^2) + 2\overline{Y}^3$. Expressing $Y'''(1)$ in terms of $\overline{Y}$, $\overline{\overline{Y}}$ and $\overline{\overline{\overline{Y}}}$ gives:

$$Y'''(1) = \overline{\overline{\overline{Y}}} + 3(\overline{\overline{Y}} + \overline{Y}^2)(\overline{Y} - 1) - 2\overline{Y}^3 + 2\overline{Y}$$

$$= \overline{\overline{\overline{Y}}} + 3\overline{\overline{Y}}(\overline{Y}-1) + \overline{Y}^3 - 3\overline{Y}^2 + 2\overline{Y} \qquad \text{C.30}$$

Expressing $A_2'''(1)$ now in terms of $\overline{Y}$, $\overline{\overline{Y}}$ and $\overline{\overline{\overline{Y}}}$ gives:

$$A_2'''(1) = c(c-1)(c-2) -3c(c-1)\overline{Y} -3c(\overline{\overline{Y}} - \overline{Y}^2 - \overline{Y})$$

$$-\overline{\overline{\overline{Y}}} -3\overline{\overline{Y}}(\overline{Y}-1) - \overline{Y}^3 +3\overline{Y}^2 -2\overline{Y} +6\overline{Y}\ (\overline{\overline{Y}} + \overline{Y}^2 - \overline{Y}) - 6\overline{Y}^3$$

$$= c(c-1)(c-2) - 3c^2\overline{Y} + 6c\overline{Y} - 3c\overline{\overline{Y}} + 3c\overline{Y}^2 - \overline{\overline{\overline{Y}}}$$

$$+ 3\overline{Y}\ \overline{\overline{Y}} + 3\overline{\overline{Y}} - \overline{Y}^3 - 3\overline{Y}^2 - 2\overline{Y} \qquad \text{C.31}$$

Substituting C.11, C.13 and C.31 into C.28 produces:

$$A''(1) = \frac{\begin{array}{l} -2(c-\overline{y})\ [c(c-1)(c-2) -3c^2\ \overline{Y} + 6c\overline{Y} - 3c\overline{\overline{Y}} + 3c\overline{Y}^2 \\ \quad - \overline{\overline{\overline{Y}}} + 3\overline{Y}\ \overline{\overline{Y}} + 3\ \overline{\overline{Y}} - \overline{Y}^3 - 3\overline{Y}^2 - 2\overline{Y}\ ] \\ \quad + 3\ [c(c-1) - 2c\overline{Y} - \overline{\overline{Y}} + \overline{Y} + \overline{Y}^2\ ]^2 \end{array}}{6(c-\overline{Y})^2} \qquad \text{C.32}$$

Expanding:

$$A''(1) = \frac{\begin{aligned}&-2c^4 + 6c^3 - 4c^2 + 2c^3\bar{Y} - 6c^2\bar{Y} + 4c\bar{Y} + 6c^3\bar{Y}\\ &- 6c^2\bar{Y}^2 - 12c^2\bar{Y} + 12c\bar{Y}^2 + 6c^2\bar{\bar{Y}} - 6c\bar{Y}\,\bar{\bar{Y}} - 6c^2\bar{Y}^2\\ &+ 6c\bar{Y}^3 + 2c\bar{\bar{\bar{Y}}} - 2\bar{Y}\,\bar{\bar{\bar{Y}}} - 6c\bar{Y}\,\bar{\bar{Y}} + 6\bar{Y}^2\,\bar{\bar{Y}} - 6c\bar{\bar{Y}} + 6\bar{Y}\,\bar{\bar{Y}}\\ &+2c\bar{Y}^3 - 2\bar{Y}^4 + 6c\bar{Y}^2 - 6\bar{Y}^3 +4c\bar{Y} - 4\bar{Y}^2\\ &+ 3c^4 - 6c^3 + 3c^2 - 12c^2(c-1)\bar{Y} - 6c(c-1)\bar{\bar{Y}} + 6c(c-1)\ \bar{Y}\\ &+ 6c(c-1)\bar{Y}^2 + 12c^2\bar{Y}^2 + 12c\bar{Y}\,\bar{\bar{Y}} - 12c\bar{Y}^2 - 12c\bar{Y}^3\\ &+ 3\bar{\bar{Y}}^2 - 6\bar{Y}\,\bar{\bar{Y}} - 6\bar{Y}^2\bar{\bar{Y}} +3\bar{Y}^2 + 6\bar{Y}^3 + 3\bar{Y}^4\end{aligned}}{6(c-\bar{Y})^2} \qquad \text{C.33}$$

Reducing:

$$A''(1) = \frac{\begin{aligned}&c^4 - c^2 - 4c^3\bar{Y} + 2c\bar{Y} + 6c^2\bar{Y}^2 - 4c\bar{Y}^3\\ &+ 2c\bar{\bar{\bar{Y}}} - 2\bar{Y}\,\bar{\bar{\bar{Y}}} + \bar{Y}^4 - \bar{Y}^2 + 3\bar{\bar{Y}}^2\end{aligned}}{6(c-\bar{Y})^2} \qquad \text{C.34}$$

Rearranging terms yields:

$$A''(1) = \frac{(c-\bar{Y})^4 - (c-\bar{Y})^2 + 2c\bar{\bar{\bar{Y}}} - 2\bar{Y}\,\bar{\bar{\bar{Y}}} + 3\,\bar{\bar{Y}}^2}{6(c-\bar{Y})^2} \qquad \text{C.35}$$

Having reduced C.33 to a much more manageable form, substitute C.21 into C.17 to give:

$$\bar{\bar{Q}} = A''(1) - A'(1)^2 + A'(1) + \sum_{i=0}^{c-1}[B_i'(1) - B_i'(1)^2] \qquad \text{C.36}$$

where we have used $B''_i(1) = 0$. Starting with the first three terms we have, combining C.35 and C.14:

$$A''(1) - A'(1)^2 + A'(1) = \frac{\begin{array}{l} 2(c-\overline{Y})^4 - 2(c-\overline{Y})^2 + 4c\overline{\overline{\overline{Y}}} - 4\overline{Y}\,\overline{\overline{\overline{Y}}} + 6\overline{\overline{Y}}^2 \\ - 3\overline{\overline{Y}}^2 - 6\overline{\overline{Y}}(c-\overline{Y}) + 6\overline{\overline{Y}}(c-\overline{Y})^2 - 3(c-\overline{Y})^2 \\ + 6(c-\overline{Y})^3 - 3(c-\overline{Y})^4 + 6\overline{\overline{Y}}(c-Y) + 6(c-Y)^2 \\ - 6(c-\overline{Y})^3 \end{array}}{12(c-\overline{Y})^2}$$

$$= \frac{(1+6\overline{\overline{Y}})(c-\overline{Y})^2 + 4(c-\overline{Y})\ \overline{\overline{\overline{Y}}} + 3\overline{\overline{Y}}^2 - (c-\overline{Y})^4}{12(c-\overline{Y})^2} \qquad \text{C.37}$$

Solving and reducing the rest of C.36 gives the desired result, shown in equation 3.18:

$$\overline{\overline{Q}} = \frac{(1+6\overline{\overline{Y}})(c-\overline{Y})^2 + 4(c-\overline{Y})\ \overline{\overline{\overline{Y}}} + 3\overline{\overline{Y}}^2 - (c-\overline{Y})^4}{12(c-\overline{Y})^2} - \sum_{i=0}^{c-1} \frac{z_i}{(1-z_i)^2} \qquad \text{C.38}$$

APPENDIX D

Moment Formulas for Waiting Times

In the text we found:

$$V(z) = \frac{A(z)}{B(z)} \cdot R(z) \tag{D.1}$$

where:

$$A(z) = \frac{1-Y(z)}{\overline{Y}(1-z)} \tag{D.2}$$

$$B(z) = \frac{1-G(z)}{\overline{G}(1-z)} \tag{D.3}$$

$$R(z) = \frac{Q(z)}{Y(z)} \tag{D.4}$$

We now wish to find $\overline{V}$ and $\overline{\overline{V}}$ in terms of the moments of G, Y, and Q.

Proceeding:

$$\overline{V} = V'(1) = \left(\frac{A'(z)}{B(z)} - \frac{A(z)B'(z)}{B(z)^2} \right) R(z) + \frac{A(z)}{B(z)} \cdot R'(z) \Bigg|_{z=1}$$

$$= A'(1) - B'(1) + R'(1) \tag{D.5}$$

Differentiating A(z), we obtain:

$$A'(z) = \frac{-Y'(z)(1-z) + 1-Y(z)}{\overline{Y}(1-z)^2} \tag{D.6}$$

Now let z ⟶ 1. This gives us 0/0; hence we must apply l'Hopital's rule:

$$\lim_{z\to 1} A'(z) = \frac{-Y''(z)(1-z) + Y'(z) - Y'(z)}{-2\overline{Y}(1-z)} = \frac{Y''(z)}{2\overline{Y}} \tag{D.7}$$

Or:

$$A'(1) = \frac{\overline{\overline{Y}} + \overline{Y}^2 - \overline{Y}}{2\overline{Y}} = \frac{1}{2}\left(\frac{\overline{\overline{Y}}}{\overline{Y}} + \overline{Y} - 1\right) \tag{D.8}$$

To find A'(1), we need merely note that A(z) and B(z) have the same functional form, and hence:

$$B'(1) = \frac{1}{2}\left(\frac{\overline{\overline{G}}}{\overline{G}} + \overline{G} - 1\right) \tag{D.9}$$

Finally:

$$R'(1) = \left.\frac{Q'(z)}{Y(z)} - \frac{Q(z)Y'(z)}{Y(z)^2}\right|_{z=1}$$

$$= \overline{Q} - \overline{Y} \tag{D.10}$$

Substituting D.8, D.9 and D.10 into D.5 and reducing gives:

$$\overline{V} = \frac{1}{2}\left[\frac{\overline{\overline{Y}}}{\overline{Y}} + \overline{Y} - \frac{\overline{\overline{G}}}{\overline{G}} - \overline{G}\right] + \overline{Q} - \overline{Y} \tag{D.11}$$

To find $\overline{\overline{V}}$, we use:

$$\overline{\overline{V}} = V''(1) + V'(1) - V'(1)^2 \tag{D.12}$$

As a shortcut, multiply both sides of D.1 by B(z) to give:

$$V(z) \cdot B(z) = A(z) \cdot R(z) \tag{D.13}$$

Noting that A(z), B(z), R(z), and V(z) are all transfers of random variables, and since the product of two transforms represents the sum of two independent random variables, we must have:

$$\overline{\overline{V}} + \overline{\overline{B}} = \overline{\overline{A}} + \overline{\overline{R}}$$

Or:

$$\overline{\overline{V}} = \overline{\overline{A}} - \overline{\overline{B}} + \overline{\overline{R}} \tag{D.14}$$

To find $\overline{\overline{A}}$, we must first find A"(1). Differentiating D.6 once again gives:

$$A''(z) = \frac{-Y''(z)(1-z)}{\overline{Y}(1-z)^2} + 2\frac{[-Y'(z)(1-z) + 1-Y(z)]}{\overline{Y}(1-z)^3}$$

$$= \frac{-Y''(z)(1-z)^2 - 2Y'(z)(1-z) + 2-2Y(z)}{\overline{Y}(1-z)^3} \qquad \text{(D.15)}$$

Applying l'Hopital's rule:

$$\lim_{z\ 1} A''(z) = \frac{-Y'''(z)(1-z)^2}{-3\overline{Y}(1-z)^2} = \frac{Y'''(1)}{3\overline{Y}} \qquad \text{(D.16)}$$

Again we note that:

$$Y'''(1) = \overline{\overline{\overline{Y}}} + 3\overline{Y}\,\overline{\overline{Y}} + \overline{Y}^3 - 3\overline{\overline{Y}} - 3\overline{Y}^2 + 2\overline{Y}$$

$\overline{\overline{A}}$ is given by:

$$\overline{\overline{A}} = A''(1) + A'(1) - A'(1)^2$$

$$= \frac{\begin{aligned}[4\overline{Y}\,\overline{\overline{\overline{Y}}} + 12\overline{Y}^2\overline{\overline{Y}} + 4\overline{Y}^3 - 12\overline{Y}\,\overline{\overline{Y}} - 12\overline{Y}^3 + 8\overline{Y}^2 \\ + 6\overline{Y}\,\overline{\overline{Y}} + 6\overline{Y}^3 - 6\overline{Y}^2 - 3\overline{\overline{Y}}^2 - 6\overline{Y}^2\overline{\overline{Y}} + 6\overline{Y}\,\overline{\overline{Y}} \\ - 3\overline{Y}^4 + 6\overline{Y}^3 - 3\overline{Y}^2]\end{aligned}}{12\overline{Y}^2}$$

$$= \frac{[4\overline{Y}\,\overline{\overline{\overline{Y}}} + 6\overline{Y}^2\overline{\overline{Y}} + 4\overline{Y}^3 - \overline{Y}^2 - 3\overline{\overline{Y}}^2 - 3\overline{Y}^4]}{12\overline{Y}^2}$$

$$= \frac{[4\overline{Y}\,\overline{\overline{\overline{Y}}} + \overline{Y}^2(4\overline{Y}-1) - 3(\overline{\overline{Y}}-\overline{Y}^2)^2]}{12\overline{Y}^2} \qquad \text{(D.17)}$$

$\overline{\overline{B}}$, of course, is the same as $\overline{\overline{A}}$ using the moments of G instead of the moments of Y. Finally, we note that:

$$\overline{\overline{R}} = \overline{\overline{Q}} - \overline{\overline{Y}} \qquad (D.18)$$

Substituting into D.14 now gives:

$$\overline{\overline{V}} = \frac{[4\overline{Y}\,\overline{\overline{\overline{Y}}} + \overline{Y}^2(4\overline{Y}-1) - 3(\overline{\overline{Y}}-\overline{Y}^2)^2]}{12\overline{Y}^2} - \frac{[4\overline{G}\,\overline{\overline{\overline{G}}} + \overline{G}^2(4\overline{G}-1) - 3(\overline{\overline{G}}-\overline{G}^2)^2]}{12\overline{G}^2}$$

$$+ \overline{\overline{Q}} - \overline{\overline{Y}} \qquad (D.19)$$

APPENDIX E

Solving the Transforms

In this appendix we present the mechanics of actually solving the transforms, first, in E.1 by finding the necessary zeroes, and second, in E.2 and E.3 by then finding the first c probabilities where we will present two alternative ways with which they can be found.

E.1 Finding the Roots

Here we will present two methods for finding roots, the first applying only to the special case of simple Poisson arrivals to a queue with deterministic headways, the second covering all cases. The same algorithm will be used in both cases, namely the Newton-Raphson root-finding procedure, and the only additions that will be made are in the choice of starting points when finding each zero.

The problem is to find the c zeroes within the unit circle of the function f(z) defined by:

$$f(z) = \frac{z^c}{Y(z)} - 1 = 0 \qquad \text{E.1}$$

Or:

$$\frac{z^c}{Y(z)} = 1 = \exp(2\pi i) \qquad \text{E.2}$$

where $i = \sqrt{-1}$. Taking the $1/c^{th}$ power of both sides gives the following c equations for the c roots:

$$zY(z)^{-1/c} = \exp(\frac{2\pi ki}{c}) \quad k = 0,\ldots, c-1 \qquad \text{E.3}$$

We now assume arrivals are simple Poisson and headways are deterministic of length T, hence $Y(z) = \exp(1-\lambda T(1-z))$. We also have that $\rho = \frac{\lambda T}{c}$, thus $\lambda T = \rho c$. Substituting back into E.3 gives:

$$z \exp\ (\ \rho(1-z)) = \exp(\frac{2\pi k i}{c}) \qquad \text{E.4}$$

Setting $z = r \exp\ (i\theta) = \exp\ (\ell n r + i\theta)$ and taking logs of both sides yields:

$$\ell n r + i\theta + \rho - \rho r e^{i\theta} = \frac{2\pi k i}{c}$$

Or:

$$\ell n r + i\theta + \rho - \rho r(\cos\theta + i \sin\theta) = \frac{2\pi k i}{c} \qquad \text{E.5}$$

Taking real and imaginary parts, we obtain two equations for the two unknowns, r and θ:

$$\ell n r + \rho - \rho\ r \cos\theta = 0 \qquad \text{E.6}$$

$$\theta - \rho r \sin\theta = \frac{2\pi k}{c} \qquad \text{E.7}$$

For this special case only, we may solve for r using E.7 and substitute back into E.6, where we now define $g(\theta)$ as:

$$g(\theta) = \ell n \left[\frac{\theta - \frac{2\pi k}{c}}{\rho \sin\theta}\right] + \rho - (\theta - \frac{2\pi k}{c}) \cot\theta = 0 \qquad \text{E.8}$$

We now have the problem of finding, for each k, the single root θ_k of the function $g(\theta)$. For this purpose, the following algorithm is used:

```
Step 1:  $k = 0$
Step 2:  $k = k + 1$
         Initialization:
         If $k = 1$ or $2$, set $\theta_k^o = (k+.5)2\pi/c$
```

```
          If $k \geq 3$,  $\theta_k^0 = \theta_{k-1} + (\theta_{k-1} - \theta_{k-2})$
          Set $n = 0$

Step 3:   $n = n + 1$
          $\theta_k^n = \theta_k^{n-1} - g(\theta_k^{n-1}) / \frac{d\, g(\theta_k^{n-1})}{d\,\theta}$

Step 4:   If $| \theta_k^n - \theta_k^{n-1} | < \varepsilon$ , go to step 2;
          otherwise go to step 3.
```

The derivative in step 3 is:

$$\frac{d\, g(\theta)}{d\,\theta} = \left[\frac{\rho \sin \theta}{\theta - \frac{2\pi k}{c}}\right] \cdot \left[\frac{1}{\rho \sin\theta} - \frac{\theta - \frac{2\pi k}{c}}{\rho \sin^2\theta} \cos \theta\right]$$

$$-\cot\theta + (\theta - \frac{2\pi k}{c})(1 + \cot^2\theta)$$

$$= \left(\frac{1}{\theta - \frac{2\pi k}{c}}\right) - 2 \cot \theta + (\theta - \frac{2\pi k}{c}) (1 + \cot^2 \theta) \qquad \text{E.9}$$

Since complex roots must appear as conjugate pairs, we need only solve for half the roots. More precisely, we must find $[\frac{c-1}{2}]$ complex roots, where [x] is the largest integer not exceeding x. If c is even, then there is one more root on the negative real axis that cannot be found using the above algorithm (since we already know $\theta = \pi$ for this root). In this case we define $\hat{g}(r)$ using E.6:

$$\hat{g}(r) = \ln r + \rho + \rho r = 0 \qquad \text{E.10}$$

This problem is now solved using the same algorithm, where now we are looking for r, using $\hat{g}(r)$ and its derivative instead of $g(\theta)$. For a starting point, use $r^{o}_{c/2} = r_{\frac{c}{2}-1}$ = value of r corresponding to $\theta_{\frac{c}{2}-1}$ found from E.7. The reason for this is that $r_k - r_{k-1}$ becomes smaller as k increases, and hence $r_{\frac{c}{2}} \simeq r_{\frac{c}{2}-1}$.

The reason for presenting this special case is that when it applies, it can be solved extremely efficiently. To illustrate how other cases are solved, assume we have Poisson arrivals and gamma distributed headways with parameters N and $N \cdot \mu$, thus:

$$Y(z) = {\frac{N\cdot\mu}{N\cdot\mu+\lambda-\lambda z}}^{N}$$

$$= (1+\gamma-\gamma z)^{-N} \qquad \text{E.11}$$

where $1/\mu$ = average headway and $\gamma = \frac{\lambda}{N\cdot\mu}$. Returning to equation E.3, we have:

$$z \cdot (1+\gamma-\gamma z)^{N/c} = \exp\left(\frac{2\pi k i}{c}\right) \qquad \text{E.12}$$

Taking logs again gives:

$$\ell n r + i\theta + \frac{N}{c}\ \ell n\ (1+\gamma-\gamma z) = \frac{2\pi k i}{c} \qquad \text{E.13}$$

Taking real and imaginary parts, define $g(r,\theta)$, $g_1(r,\theta)$ and $g_2(r,\theta)$ as follows:

$$g(r,\theta) = \begin{vmatrix} g_i\ (r,\theta) \\ g_2\ (r,\theta) \end{vmatrix} = \begin{vmatrix} \ell n\ r + \frac{N}{c}\ \mathrm{Re}\ \{\ell n\ (1+\gamma-\gamma z)\} \\ \theta + \frac{N}{c}\ \mathrm{Im}\ \{\ell n\ (1+\gamma-\gamma z)\} - \frac{2\pi k}{c} \end{vmatrix} \qquad \text{E.14}$$

where Re(·) and Im (·) refer to the real and imaginary parts of the argument. We must now solve E.14 using the multidimensional Newton-Raphson algorithm. Again, we are looking only for the first $[\frac{c-1}{2}]$ roots, and a different procedure is required if c is even to find the last root on the negative real axis. The algorithm is as follows:

```
Step 1:  $k = 0$
Step 2:  $k = k + 1$ ;  $n = 0$
         Initialize $\theta$:
         If $k = 1$ or 2:
             $\theta_k^o = (k + .5) \frac{2\pi}{c}$
         If $k > 2$:
             $\theta_k^o = \theta_{k-1} + (\theta_{k-1} - \theta_{k-2})$
         Initialize r:
         If $k = 1$:  $r_1^o = .8$
         If $k > 1$   $r_k^o = r_{k-1}$
Step 3:  $n = n + 1$
         $x_k^n = x_k^{n-1} - \nabla_x g^{-1} (x) \cdot g(x)$
         where:
             $x_k^n = \begin{vmatrix} r_k^n \\ \theta_k^n \end{vmatrix}$

             $\nabla_x g(x) = \begin{vmatrix} \frac{dg_1}{dr} & \frac{dg_1}{d\theta} \\ \frac{dg_2}{dr} & \frac{dg_2}{d\theta} \end{vmatrix}$
```

Step 4: If $\left| r_k^n - r_k^{n-1} \right| < \varepsilon$ and $\left| \theta_k^n - \theta_k^{n-1} \right| < \varepsilon$, go to step 2; otherwise go to step 3.

The Jacobian is found to be:

$$\nabla_x g = \begin{vmatrix} \frac{1}{r} - \frac{N}{c} \; \mathrm{Re} \; (\frac{\gamma e^{i\theta}}{1+\gamma-\gamma z} & - \frac{N}{c} \; \mathrm{Im} \; (\frac{\gamma \; z}{1+\gamma-\gamma z}) \\ & \\ - \frac{N}{c} \; \mathrm{Im} \; (\frac{\gamma e^{i\theta}}{1+\gamma-\gamma z}) & 1 - \frac{N}{c} \; \mathrm{Re} \; (\frac{\gamma z}{1+\gamma-\gamma z}) \end{vmatrix} \qquad \text{E.15}$$

Naturally, this must be inverted, but this is trivial for a 2 × 2 matrix. If c is even, the last zero on the negative real axis must be found using g_1 alone, that is, by finding the single zero $r = -z$ of:

$$g_1(r) = \ell n \; r + \frac{N}{c} \; \ell n \; (1 + \gamma + \gamma r) \qquad \text{E.16}$$

This is solved in an analogous manner, using as a starting point $r^o_{\frac{c}{2}} = r_{\frac{c}{2} - 1}$.

E. 2 Solving the simultaneous linear equations

Here we will solve for the unknowns $q_o, \ldots, q_{c-1}$ by solving the set of simultaneous linear equations using the roots $z_o, \ldots, z_{c-1}$. The intent here is simply to demonstrate how these equations must be set up; as an illustration, we will use the equations arising for the scheduled departure queue with no minimum load constraint. Thus we have:

$$\sum_{i=0}^{c-1} q_i \; (z_k^c - z_k^i) = 0 \qquad k = 0, \ldots, c - 1 \qquad \text{E.17}$$

We cannot, however, use all c roots to determine the unknowns. First, the root $z_o = 1$ gives us no information. In its place, we use the fact that $\lim_{z\to 1} Q(z) = 1$, or:

$$\lim_{z\to 1} Q(z) = 1 = \lim_{z\to 1} \frac{\sum_{i=0}^{c-1} q_i \; (z^c - z^i)}{\frac{z^c}{Y(z)} - 1} \qquad \text{E.18}$$

Applying l'Hopital's rule gives:

$$1 = \frac{\sum_{i=0}^{c-1} q_i (c-i)}{c - \lambda}$$

Or:

$$\sum_{i=0}^{c-1} q_i \; (c-i) \;\; = c - \lambda \qquad \text{E.19}$$

Equation E.19 is the first equation. For the rest, we now note that since the complex roots appear as conjugates, only one of the conjugates gives us any information. The remaining equations are made up by using the real and imaginary parts, as follows:

$$\text{Re} \; \{ \sum_{i=0}^{c-1} q_i \; (z_k^c - z_k^i \,)\} = 0 \qquad \text{E.20}$$

$$k = 1, \; ..., \; [\tfrac{c-1}{2}]$$

$$\text{Im} \; \{ \sum_{i=0}^{c-1} q_i \; (z_k^c - z_k^i \,)\} = 0 \qquad \text{E.21}$$

If c is odd, then equations E.19, E.20 and E.21 are all we need to find the missing unknowns. If c is even, then the root $z_{\frac{c}{2}}$ on the negative real axis must be included as well. Once set up, the equations are easily solved using Gaussian elimination.

The computational requirements of this approach are as follows. There are approximately $c/2$ equations to set up, each requiring c complex multiplications or 4c simple multiplications, giving $2c^2$ multiplications altogether. Gaussian elimination then requires $c^3/3$ multiplications and additions.

E.3 Expanding the polynomial

As an alternative to solving the system of simultaneous linear equations, it was observed that we may determine the unknown probabilities by matching the coefficients of the following polynomials in z:

$$\sum_{i=0}^{c-1} q_i \; (z^c - z^i) = (c - \overline{Y}\,)(z-1) \prod_{i=0}^{c-1} \left(\frac{z-z_i}{i-z_i} \right) \qquad \text{E.22}$$

Define P(z) as:

$$P(z) = \sum_{i=0}^{c} p_i z^i = \sum_{i=0}^{c-1} q_i (z^c - z^i) \qquad \text{E.23}$$

Thus:

$$\begin{aligned} q_i &= - p_i \qquad i = 0, \ldots, c-1 \\ \sum_{i=0}^{c-1} q_i &= p_c \end{aligned} \qquad \text{E.24}$$

We now wish to expand the polynomial on the right hand side of E.22, remembering that the roots appear as complex conjugates. The following algorithm is now presented:

Step 1: Initialization

Let $\beta = (c - \overline{Y}) \prod_{i=0}^{c-1} \frac{1}{1-z_i}$

Define $M = \left| \frac{c-1}{2} \right|$

To compute β, set $\beta = c - \overline{Y}$ and do, $i = 1, \ldots, M$:

$$\beta = \beta \cdot / (1-2 \operatorname{Re} \{z_i\} + |z_i|^2)$$

If c is even, then $\beta = \beta/(1 - z_{\frac{c}{2}})$

If c is even, define the vector $A = (a_1, a_2, \ldots)$ as:

$$a_o = -z_{c/2}$$

$$a_1 = -1 + z_{c/2}$$

$$a_2 = 1$$

$$a_i = 0 \qquad i = 3, \ldots$$

If c is odd, then:

$$a_o = -1$$

$$a_1 = 1$$

$$a_i = 0 \qquad i = 2, \ldots$$

Step 2: Compute $B = (b_o, b_1, b_2, \ldots, b_c)$ where:

$$B(z) = \sum_{i=0}^{c} b_i z^i = (z - 1) \prod_{i=0}^{c-1} (z-z_i) ;$$

To do this:

```
Do: $i = 1, \ldots, M$;
    $x_o = |z_i|^2$ ;
    $x_1 = 2\mathrm{Re}\ (z_i)$;
    Compute $b_o, b_1, \ldots$ as follows:
        $b_o = a_o \cdot x_o$ ;
        $b_1 = a_o x_1 + a_1 z_o$ ;

        Do: $j = 2, 2*i + 1$ ;
            $b_j = a_j \cdot x_o + a_{j-1} x_1 + a_{j-2}$ ;
        End;
    Now set $a_j = b_j$, $j = 0, \ldots, 2*i + 1$ ;
End;
```

Step 3: Compute $P = (p_o, p_1, \ldots, p_c)$

$$p_i = \beta \cdot b_i \quad i = 0, \ldots, c$$

Step 4: Find probabilities:

$$q_i = - p_i \qquad i = 0, \ldots, c - 1$$

The primary computational requirements are in step 3. The innermost loop repeats two additions and multiplications $2 \cdot i$ times, which is repeated for $i = 1, \ldots, c/2$ (approximately), or $2 \sum_{i=1}^{c/2} i = 2 \cdot \frac{(c/2)(c/2 + 1)}{2} \cong \frac{c^2}{4}$ additions and multiplications.