OPTIMIZATION IN STOCHASTIC SERVICE SYSTEMS WITH DISTINGUISHABLE SERVERS

by

James Patrick Jarvis

B.S., University of North Carolina
(1971)

S.M., Massachusetts Institute of Technology
(1973)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1975

Signature of Author...
Department of Electrical Engineering, May 14, 1975

Certified by...
Thesis Supervisor

Accepted by...
Chairman, Departmental Committee on Graduate Students

OPTIMIZATION IN STOCHASTIC SERVICE SYSTEMS WITH DISTINGUISHABLE SERVERS

by

James Patrick Jarvis

Submitted to the Department of Electrical Engineering on May 14, 1975 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

The effects of alternative resource allocations in stochastic service systems can be difficult to predict. Diverse, and sometimes conflicting, measures of performance for the operational effectiveness of these systems complicate the search for effective allocations. This report describes a class of models which estimate multiple operating characteristics for systems in which the identities of both the customer and server are important is determining the effectiveness of response. Important applications of these results are found in police, ambulance, and fire services. These public safety systems comprise a class of spatially distributed queuing systems that are discussed at length in various parts of the thesis.

The starting point for the analysis is the continuous time Markov "hypercube" model, an M/M/N queuing model which identifies the busy or idle status of each server in its (2**N)-element state space. A generalization of the hypercube model is given in conjunction with a procedure for determining dynamic allocations of servers to customers which minimize time-averaged costs of assignment. For spatially distributed systems where the cost of assignment is given by response distance, the optimization yields little improvement when compared to the strategy which dispatches the closest available unit to each call for service, but does result in substantial improvements in workload imbalance among the servers. The solution procedure for the optimization problem is a considerable simplification of previous derivations.

For systems in which expected service times are a function of both customer and server, an approximation procedure is developed for estimating steady-state performance. The procedure offers an inexpensive and relatively simple alternative to simulation as a means for analyzing these systems. The approximation is compared with several analytic models and is found to yield estimates within a few percent of the exact values for most measures of system performance. Applications of this procedure to spatially distributed systems for which travel time is a significant component of overall service time are presented.

An iterative procedure which seeks the optimal locations for facilities providing service under conditions of congestion is developed for spatially distributed systems. As opposed to previously developed deterministic location models, the stochastic interaction of the response facilities is explicitly considered in determining locations which minimize either average response distance or more complex geographically derived variables. Computational experience is given.

The use of the approximation procedure and location model is demonstrated in determining the optimal locations of ambulances for an emergency medical system. Special consideration is given to the evaluation of the use of specialized mobile coronary care units as a means for reducing the risk of death following certain coronary emergencies.

THESIS SUPERVISOR: Richard C. Larson
TITLE: Associate Professor of Electrical Engineering
and Urban Studies

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Professor Richard C. Larson for his guidance during the research reported in this thesis.

I am grateful to Keith Stevenson, Tom Willemain, Ken Chelst, and Amedeo Odoni for valuable comments on initial drafts of this document.

Also, recognition is due the MIT Operations Research Center in general and the associated graduate students in particular for providing an interesting environment in which to work.

Special thanks to my wife, Chris, and to Loki; also Mark, Tam, Mac, Fitzy, and CC.

Financial support was provided by the National Science Foundation, through a graduate fellowship and research support through Grant GI38004.

Computations were performed at the MIT Information Processing Center.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. INTRODUCTION

A. Background: Emergency Services

During recent years, there has been an increasing interest in applying the techniques of Operations Research to problems arising in the public sector (Ref. 46 and 16). One important application of this kind of analysis is to the provision of emergency services such as fire, police, and emergency medical systems. This report is an effort to identify some of the problems arising in the spatial design of emergency services and to develop models which can be used to answer questions concerning the appropriate allocation of resources for such systems.

In order to address these issues it is necessary to understand both the objectives of these services and the operational characteristics which will determine whether those objectives can be met. Although there would be little disagreement that these services are intended to provide some degree of protection to the public, it is not easy to relate an objective of this generality to specific, quantifiable performance measures. In order to design a system which will provide acceptable service, there must be a consensus as to "acceptability."

Past research efforts have emphasized such characteristics as response time or queuing delays as measures of system effectiveness. These quantities can be directly related to the operation of these services as _spatially distributed queuing systems_. In part, the same emphasis will be continued here. However, the models described below are designed to incorporate more general measures of system effectiveness as a function of both the spatial and queuing aspects of these services.

This kind of analysis can be very difficult. Particularly in large urban environments, emergency services must be viewed as dynamic probabilistic systems. Demands for service are not deterministic, but require a description which is probabilistic in both time and space. An immediate consequence of this non-determinism is a corresponding uncertainty as to the availability of resources at a random instant. When an immediate response is important, as is the case in emergency services, the response unit which would normally be assigned to provide service might not be available because it is servicing a previous demand. If that service cannot be interrupted, or _preempted_, some alternative response must be made. This type of probabilistic behavior should be considered when answering questions of resource allocation.

In addition to the problem of merely describing the dynamic behaviour of these systems, the relation between alternative allocation schemes and the desired performance must also be considered. Is there an optimal allocation of resources? If performance is defined in terms of multiple objectives, how should conflicts among these objectives be resolved? What are the procedures for determining answers to these questions?

In summary, the general problem that we address here is the provision of services in a stochastic environment. Our empahsis is on the development of models which allow for the simultaneous consideration of multiple performance measures for system effectiveness. This development has two facets. The initial issue is an adequate description of the dynamic operating characteristics of these systems for a given configuration of resources. The second area deals with problems of resource allocation. If we can describe how a system behaves, can we determine those designs which result in improved performance?

B. Objectives

The class of systems which we consider can be characterized in general terms as queuing systems with distinguishable servers and classes of customers. Although much of the development contained in the following chapters will be given in this general framework, it is instructive to analyze each part of this nomenclature as it relates to the provision of emergency services.

"Queuing" implies that there is some contention for resources. For example, if your house catches fire, you would probably prefer that the closest fire house immediately dispatch all of its apparatus to deal with your immediate need. In practical terms, those particular units may be currently engaged elsewhere and some alternative response is required. This could take the form of a response by another unit or a queuing delay if no units were available. The design of a fire system should reflect these kinds of considerations.

This example points up another important feature of these systems; that is, the distinct identity of the responding unit. For spatially distributed systems, the response units might be distinguished solely on the basis of their location. More generally, the individual units could have specialized skills which could make them more or less appropriate for assignment in particular situations. An example presented in Chapter 7 distinguishes between standard ambulances and mobile coronary care units for response to certain coronary emergencies.

These same distinctions of locality and appropriateness of response can be applied to demands for service. When there is an empahsis on response time, a police department is going to choose a car for dispatch at least partially on the basis of proximity. Again, this choice can also be influenced by the nature of the demand for service. The response to a report of a robbery in progress in likely to be very different from that to a complaint of a stereo playing too loudly.

The models which are described below are designed to incorporate the stachastic nature of the arrival of calls for service and a suitable response by the system providing service based on the current availability of resources and the particular demand for service. For the special case of spatially distributed systems, the acceptability of response can be defined in terms of geographic variables, such as response time, or more complex measures for system performance. Instead of only using quantities relating to the process of delivering service, the formulations permit an emphasis on the outcomes of providing that service (Ref. 64); that is, did the patient live? was the fire extinguished? was the thief apprehended? An example in Chapter 7 focuses of some of the effects an ambulance system can have on the risk of death associated with certain medical emergencies.

Because the response of these systems to each alternative configuration is difficult to predict, an attempt is made to do more than just describe their operation. Procedures for optimizing certain measures of performance will be given. In specific terms, the procedures deal with either the allocation of servers to customers, or, for spatially distributed systems, a determination of the optimal location for response units with respect to specified performance measures.

C. An Outline of Contents

The contents of this report can be divided into three areas: a review and summary of models for stochastic services

(Chapters 2 and 8), methodological developments (Chapters 3, 4, 5, and 6), and an example of the application of some of the newly developed models (Chapter 7).

In Chapter 2, we review the literature as it pertains to resource allocation for queuing systems with distinguishable servers in the context of emergency services. Particular attention will be paid to the "hypercube queuing model" developed by Larson (Ref. 41 and 36). The flexibility and philosophy of the hypercube formulation typifies much of the development contained here.

In spatially distributed systems, a central issue in resource allocation is the determination of appropriate locations for facilities. Both deterministic location models and their stochastic variations are examined for their applicability to problems arising in stochastic service systems. Particular attention is paid to models which incorporate some of the queuing aspects of spatially distributed systems.

In Chapter 3, we review the formulation of the hypercube model as used in the analysis of police operations. A generalization of this continuous time Markov model to include other than spatially distributed systems is presented in conjunction with a procedure to determine the assignment of servers to customers which minimizes the expected cost of service. Predictions from simple models for determining average response distance for spatially distributed systems are compared with data generated by the hypercube model for a specific region.

Chapter 4 develops alternative models in response to some limitations of the hypercube model. These models are used to explore the effect of various service time assumptions on the operating characteristics of queuing systems with distinguishable servers. Although these models have limited practical applicability because of associated computational difficulties, they provide counterexamples to two interesting conjectures and suggest a steady state characterization for a much wider class of systems.

A location model for use in spatially distributed queuing systems is developed in Chapter 5. The location model incorporates the interaction among the service units in a manner to improve the location of each unit on the basis of its particular spatial responses. Since the location of the response units can affect the queuing behavior of the system, an iterative procedure for determining the optimal positions is given. The iteration involves successive improvements in unit positions through the alternate use of a descriptive model for analyzing current unit locations and the location model for improving these locations. For applications to the location of emergency medical units, the model is modified to allow constraints on maximum response time. For police operations, similar modifications to include preventive patrol are given.

Using a result derived in Chapter 4, we develop in Chapter 6 an approximation procedure for analyzing systems in which service times depend on both the server and the class of customer. This procedure is similar to that developed by Larson for the hypercube model (Ref. 40). This technique is particularly useful in analyzing spatially distributed systems for which travel time is a significant portion of the overall service time. In these circumstances, the travel time depends on both the initial location of the response unit and the spatial origin of the call for service (customer). We conclude this chapter by comparing the approximation procedure to previously developed analytic models.

Chapter 7 is the culmination of the preceding development. An example of the use of the locational model and approximation procedure is presented in the context of locating emergency ambulances. The flexibility of the models is demonstrated by first locating standard ambulances to minimize average response time and then evaluating the addition of a specialized mobile coronary care unit. The latter example utilizes work recently completed by Cretin (Ref. 13) in modeling the risk of pre-hospital death following certain coronary emergencies. This analysis focuses on a more direct measure of a system's effectiveness than a surrogate such as response time; that is, the risk of pre-hospital death following a myocardial infarction.

Chapter 8 contains a summary and recommendations for further research.

D. Notation

The notation used for both equations and variables corresponds closely to that of a high level programming language such as PL/I or FORTRAN. Insofar as possible, variables are given mnemonic names (such as AWL for average workload). The mnemonic will be indicated in the text by underlining as in the previous example.

Subscript lists and the arguments of functions will follow the associated variable name in brackets, { and }. For example, the i-th component of the vector V will be denoted by V{i}. An arbitrary component of V is sometimes denoted by V{-}. The entry in the i-th row and j-th column of the matrix M is denoted by M{i,j}. M{i,-} denotes the i-th row of M. Similarly, the function F evaluated at t has value F{t}. Parentheses in equations are used only to indicate the order of operations.

Unless altered by the use of parentheses, the precedence of arithmetic operations is exponentiation (**), followed by multiplication (*) and division (/), and then addition (+) and subtraction (-).

In conditioning indices for summations, i:V{i}=j is read "those indices i such that the i-th component of the vector V is equal to j."

A glossary of variable names and mnemonics is given in

Appendix A with an indication as to the location of the first use of the variable in the text.

Chapter 2. LITERATURE REVIEW

A. Introduction

There is a substantial literature on queuing and optimization of stochastic systems. The reader is referred to Ross (Ref. 50), Karlin (Ref. 32), Cox and Smith (Ref. 12) or Feller (Ref. 17 and 18) for a general discussion of these subjects. Our purpose here is to review a specific subset of this literature. In particular, we will look at methodologies and applications which are oriented toward the provision of emergency services, which we view as spatially distributed queuing systems.

The literature which is relevant to the discussion given here can be divided into two categories. The first, which we refer to as the set of predictive models, is largely concerned with the analysis of a service system with a specified level and configuration of resources. The objective of these models is to incorporate the stochastic elements of demands for service and availability of response units into a description of the dynamic operating characteristics of the system. Some of these models are purely descriptive, while others incorporate optimization techniques to determine the allocations of fixed resources which result in the "best" performance of the system.

The second category deals with models for determining the otimal location of facilities in spatially distributed systems. The distinction between these two groupings is not always precise, but it has the advantage of closely paralleling the development given in the subsequent chapters.

B. Predictive Models

The predictive models to be discussed here are designed to estimate the operating characteristics of a system as a function of the spatial and temporal distribution of demands for service, the number and placement of response units, and a service discipline for assigning servers to customers (demands). For spatially distributed systems, an additional imput is given by a description of the local geography. Needless to say, it is very difficult to incorporate all of these variables into a single model. Although simulation techniques have been applied to these systems with some success (Ref. 41 and 43), our emphasis will be on analytic models.

There are many models dealing with particular facets of the operation of emergency services. Perhaps the most comprehensive single work in the field is Urban Police Patrol Analysis by Larson (Ref. 41). The specific issues addressed by Larson include travel time models (for example, the effects of barriers or one-way streets to travel time); preventive patrol (including models for the probability of intercepting crimes in progress); and sector design (with an analysis of fixed position versus mobile units and the effects of overlapping sectors). Although these models are developed in the context of police operations, they have some general applicability to the analysis of other spatially distributed emergency services.

Other examples of models for specific aspects of system performance are the "square root laws" postulated by Larson (Ref. 41) and Blum and Kolesar (Ref. 2) for predicting average response distance in spatially distributed systems. We will examine these models in more detail in Chapter 3.

Chaiken and Larson (Ref. 5) have compiled an excellent survey of techniques for resource allocation in emergency services (through 1972). Their paper notes many of the models for specific aspects of a system's performance and contains an extensive bibliography. Although many of these models are quite useful, there are recently developed models which are more interesting from our viewpoint.

One such model was developed by Hall (Ref. 21) for analyzing a service system consisting of police and emergency medical vehicles. This semi-Markov model partially incorporates the effects of travel time by utilizing different service time distributions for responses of greater than and less than one mile. Although an exponential distribution satisfactorily describes the service time distribution for the particular system examined by Hall, it is not clear whether the model would be analytically tractable for other distributions. In Chapter 6, we develop a procedure which explicitly incorporates travel time as one component of the overall service time.

Insofar as the topics to be discussed here are concerned, one of the most important developments was the Markov "hypercube queuing model." This modeling approach was suggested initially by Larson (Ref. 39) and then detailed by Campbell (Ref. 3). The model explicitly incorporates the probabilistic nature of the arrival of calls and their subsequent service in a framework including the interaction of mobile or fixed units in a spatially distributed system. Initial numerical difficulties encountered by Campbell were largely overcome by an iterative procedure devised by Larson (Ref. 36). Problems related to the sheer size of the model have been solved by an approximation scheme developed by Larson (Ref. 40).

Although the hypercube model has been used mainly in the context of urban police operations (Ref. 38, 30, 14, and 8), Jarvis and Larson (Ref. 29) suggest alternative uses for the model. The generalization of the model given in Chapter 3 is developed in that spirit.

A typical use of models such as the hypercube is the evaluation of alternative system configurations. Since the model produces estimates of many differenct aspects of system performance, the user can base his evaluation of the system on a subjective estimate of its overall effectiveness. If the performance is not deemed satisfactory, then changes in configuration can be made and the system re-evaluated. This kind of iterative improvement can be difficult for several reasons.

In the first place, some level of expertise is required of the user in order to know what kinds of changes in the system are likely to produce the desired results. As noted by Larson and Stevenson (Ref. 42), some quantities, such as travel distance, are largely insensitive to changes in the configuration of a spatially distributed system and thus do not lend themselves to significant improvement in a simple manner. Finally, as noted by Chelst (Ref. 7), some performance measures, for example travel time and workload imbalances, cannot in general be optimized simultaneously.

Implicit in this discussion is the assumption that resources are not unlimited. At least in partial response to considerations such as these, models which determine optimal allocations of the available resources have been developed. For example, Swersey (Ref. 55) has developed a Markov decision model for determining how many fire-fighting units to dispatch to an alarm. Part of the information considered in this decision is the frequency of false alarms and the congestion in the system.

Similar techniques for allocating servers to customers in fire operations have been developed by Ignall (Ref. 25). Specifically, allocation schemes which minimize response distance subject to constraints on workload imbalances are determined. Although this formulation includes queuing phenomena, its general applicability is limited by the assumption that exactly one half of all the response units are assigned to any particular call for service (although not necessarily the same half).

Carter, Chaiken, and Ignall (Ref. 4) developed a procedure for determining response areas to minimize response distance for two fixed position emergency units and noted that the optimization procedure also improved the workload imbalance between the two units. An algorithmic procedure for determining the optimal response areas for more than two units was given by Jarvis (Ref. 28) utilizing Markov decision theory. A generalized version of the algorithm is given in Chapter 3 with a much simplified development of the main results of the solution procedure.

For spatially distributed systems, the question of how to allocate response units to demands for service is conplicated by the additional question of where to locate the response units. Models for solving these allocation-location problems are referred to simply as location models.

C. Location Models

There is a substantial literature on location models. Cooper (Ref. 10) and Revelle, Marks, and Liebman (Ref. 49) have surveyed a large class of deterministic economic models.

Typically, these models determine locations for facilities which minimize average Euclidean or rectilinear distances from sources of supply to points of demand and can include constraints on capacity and feasible allocations. Included in this class of models are the network formulations surveyed by Odoni (Ref. 48). A frequent objective in network problems is the minimization of the maximum distance between any source and demand point. This problem is treated by Handler (Ref. 22) among others.

The difficulty in using these location models for stochastic service systems is due to the deterministic assumptions underlying their formulation. For example, although links between points of supply and demand may have finite capacities, the supply facility is always available to meet demands subject to those fixed capacity constraints.

Deterministic location models have been proposed as a method for locating emergency facilities by several authors. Toregas, Revelle, Swain, and Bergman (Ref. 58) use an integer linear programming model to determine the minimum number of facilities required to meet constraints on maximum distance to a facility. Modifications to determine those locations which minimize weighted response distance are discussed and some computational experience is presented. Formulations of this type lead to the classical set covering problem (Ref. 20) or variations on the p-median problem (Ref. 48). An application of these ideas is given by Keeney (Ref. 33) in determining district boundaries for facilities. The set covering problem is used by Walker (Ref. 60) to determine allocations of fire apparatus to fire houses.

Several different solution procedures have been proposed for solving the p-median problem. These include a modified branch-and-bound procedure suggested by Jarvinen, Rajala, and Sinervo (Ref. 26) and a heuristic technique devised by Shannon and Ignizio (Ref. 54). Comparative computational experience using some of these procedures has been reported by White and Case (Ref. 63).

The applicability of these models and their associated solution procedures to the location of emergency service facilities is limited by the assumption that the facilities are always available to provide service. In general, this condition is satisfied only under circumstances of limited interaction between the response units or very low utilizations of the service. An example of the use of these models when these conditions are satisfied is given by Jarvis, Stevenson, and Willemain (Ref. 31) in determining ambulance locations.

There are some techniques which incorporate time varying demand in determining facility locations. Wesolowsky and Truscott (Ref. 62) and Scott (Ref. 52) have developed dynamic programming models for facility location when demands are known but are not constant in time. Again, these models have limited applicability in stochastic systems.

Mirchandani (Ref. 44) and Chapman and White (Ref. 6) have considered location problems which allow uncertainty in the time or distance between the service facility and the customer. Chapman and White address a crucial issue in applying these models to spatially distributed queuing systems; that is, the availability of a server to a randomly chosen customer. Although they give an algorithm for determining server locations in a queuing environment, no computational experience is given. The authors state that their procedure is too difficult to use except under very special circumstances. The models developed by Chapman and White appear to be more useful in a reliability context where the availability of a service facility is independent of all other facilities.

In summary, there appear to be no location models which have general applicability to spatially distributed queuing systems, although some of the procedures described above are useful in particular situations. In spite of this lack of general models, specific location problems have been analyzed.

Savas used a simulation model to evaluate alternative ambulance locations in New York (Ref. 51) and concludes that dispersion of the fleet is more cost-effective than location at a central facility such as a hospital. The quantities of primary interest in this study were expected response and service times.

Using empirical travel time data, Hogg (Ref. 23) determined allocations of fire units and the location of fire houses to minimize travel time to the scene of an incident. Although the formulation allowed the dispatch of multiple units to a fire and based the allocations on this sort of cooperation, it is assumed that the demand for service is low enough that the units are always available for dispatch.

Fitzsimmons (Ref. 19) and Volz (Ref. 59) have developed models for the allocation and location of ambulances. The model developed by Fitzsimmons employs simulation techniques for the analysis of queuing aspects and a heuristic search routine to find optimal vehicle locations. Although it appears that the queuing analysis of the model could be handled more efficiently by a model such as the hypercube, the idea of successively improving the location of response units employed by Fitzsimmons (as well as Volz and Chapman and White) will be used in Chapter 5.

Optimal locations are determined by Volz under the assumption of instantaneous relocation of all available response units as the number of available units changes (at the receipt of a call for service or the completion of a service). If the system has such a small utilization that the overhead associated with the relocation does not affect performance, then the deterministic location models are probably a more appropriate means for determining locations. The difficulties associated with relocation appear insurmountable for practical applications in systems with representative workloads.

D. Conclusions

Although many of the references cited above contain elements which are useful in certain aspects of resource allocation in stochastic service systems, none incorporate all of the desired features. One of the most comprehensive formulations is given in the hypercube model. In the following pages, models are developed which combine the flexibility of the hypercube model as a descriptive tool with generalizations of the hypercube's service time assumptions. An iterative procedure for response unit location is given which utilizes these descriptive models and includes the interaction among units in spatially distributed queuing systems.

Perhaps one of the most important features of this iterative procedure is exemplified by the spatial design of an emergency response system on the basis of patient outcomes rather than the usual measures of performance such as response time. A major strength of the models developed here is an emphasis on the inclusion of more general measures of system effectiveness than simple geographically derived variables such as travel time.

Chapter 3. THE HYPERCUBE MODEL

A. Introduction: An Application to Police Patrol

As indicated in the introduction and literature review, the problem of describing the response function for emergency services is a difficult one. The recently developed 'hypercube' queuing model represents one of the most comprehensive approaches to this task. In this chapter, we summarize the development of the model in conjunction with an example of its use in modeling police response.

Although much of the material in this chapter is contained in the references, it is included here because it forms the basis for much of the development in the following chapters. We give a brief theoretical description in conjunction with the police example, an optimization procedure for certain aspects of system performance, and a summary of applicable numerical techniques.

A.1 Police Response

Since the hypercube model has been used mainly in describing police operations, that is the example used here. The model focuses on the preventive patrol and emergency response aspects of police units in a geographic setting. Before discussing the exact model formulation, we give a scenario of the police operations which we are trying to model.

We restrict our attention to an autonomous subset of the overall police force and its particular geographic area of responsibility. There are a certain number of mobile or fixed patrol units (servers) in the field which respond to calls for service as assigned by a dispatcher. When a call arrives at the dispatcher's desk, several events may occur.

If the call demands immediate attention, the dispatcher chooses one or more units in the field and directs them to respond to the incident. If the incident is not pressing, it might be held, or queued , for later response. This might be particularly likely to occur if the system is congested. Normally, the unit assigned to service the call would be performing only routine preventive patrol which would be preempted to provide the service. In fact, a unit servicing a routine call might be assigned to another call in an emergency situation, preempting its current service. A unit spends some time providing on-scene service to a call and then returns to preventive patrol.

While the preceding description is realistic for actual police operations, we shall make a few simplifications. When a call for service arrives, every unit is assumed to be either busy (currently providing service) and hence unavailable for other service or free (currently on preventive patrol) and available for service. If any unit is available, the call receives immediate attention. A single unit is dispatched to provide the service. Calls are queued only if the system is saturated; that is, all of the units are busy. In this case, the call is either handled by some means external to the system or held for later service when a unit becomes available; service is never preempted. In the former case, the call is lost from the viewpoint of the system being examined.

It is crucial to note that this scenario focuses on unanticipated demands for service. Scheduled events, such as meal breaks or administrative tasks, although very important in actual operations, are ignored insofar as the model is concerned. In addition, whether the calls represent real emergencies is irrelevant. With respect to the response of the system, they must be treated, at least initially, as emergencies and receive an immediate response if possible.

There are many factors which may be considered in choosing the particular unit which will respond to a call. Since the call arrivals cannot be anticipated but must receive immediate attention, the actual assignments will be a probabilistic variation of an idealized assignment policy. For example, the units in the field might be assigned to patrol disjoint areas in order that each might to develop some familiarity with a particular region. We refer to these regions as sectors and the corresponding unit as the sector car. Under most circumstances, we would like the sector car to respond to all calls from its own sector. In practice, we have to deal with the problem of which unit to assign when the sector car is busy. (Recall that no calls are preempted and all calls receive immediate attention if there is an available unit).

A reasonable solution to this problem is to dispatch the closest available unit. In fact, preserving sector identities may not be very important and we might always try to dispatch the closest available unit. Usually there is some uncertainty as to the exact positions of the available units and the dispatch decision must be based on partial knowledge. (Larson has addressed the implications of different levels of locational information in Ref. 41). Implicit in this discussion is the importance of minimizing the time until the arrival of a unit at the scene of the incident. In some circumstances, such as family disturbances, it may be preferable to assign units with special skills in dealing with a particular problem, even though they are farther from the scene than another unit. In any case, it should be obvious that the dispatch policy will have a great effect on the availability of units, but that this effect is somewhat complicated by the random nature of the arrival of calls.

The time that a unit spends servicing a call is another source of uncertainty in the dynamics of the police system. In general, the time that a unit spends providing service consists of several components, none of which are deterministic. There may be dispatch delays which depend on the work level of the dispatcher. The component of the service at the scene of the incident could easily depend on the exact nature of the call. In addition, for geographically distributed systems, the travel time to and from the scene must be considered. The uncertainty in this component of service time arises from two sources: variations due the travel conditions between two points in the region and uncertainty as to the starting position of the responding unit.

In summary, we are concerned with the response to calls for service. This response depends on the particular dispatch rule being used and on the availability of units. In turn, the availabilities are uncertain due to the random arrival of calls for service and nondeterministic service times. In the next section, we give a more precise formulation of the system dynamics in a framework which allows the use of an analytic model for the system.

A.2 An Analytic Description of Police Patrol

As discussed above, we are focusing on the response function for calls which are presumed to be emergent. For the geographic area of interest, we assume calls arrive according to a time-homogeneous spatial Poisson process (Ref. 35). In particular, this implies that the arrival of calls is independent of the availability of servers and the past history of the system. Also, the time between successive calls is distributed as a negative exponential random variable.

(The validity of the Poisson assumption is examined in Ref. 39 for police operations. In fact, the arrival process is generally not time homogeneous but can be expected to exhibit time independence over non-overlapping intervals. This is not suprising since the pooled output of a large number of sources can be shown to be Poisson over small time intervals (Ref. 11). We use the steady state analysis as an approximation to the actual time inhomogeneity).

An important advantage of the hypercube model is that it preserves the separate identities of the servers, which may be based on the presumed location of the units or on their specialized functions. In either case, every server is assumed to be in one of two states: either busy (unavailable for service) or free (available for service). When a call arrives, a single unit is chosen from those which are free and is immediately assigned to provide service. In the event that all servers are busy, the call is either lost or handled by external means (zero line case) or queued until a unit becomes available (infinite line case).

The service time for each unit is assumed to be distributed as a negative exponential random variable with an expected value which may depend on the particular unit providing the service, but not on the call being serviced. Since calls are at least partially classified by their origin, this assumption is not strictly correct. For geographically distributed systems, one component of the service time is travel to the scene. This, of course, will depend on the initial position of the server and the location of the call. At least for police operations, an analysis by Larson indicated that travel time is usually small compared to overall service time and hence the complication caused by the location of the incident can be ignored (Ref. 39). Even though travel time may be short compared to the total service time, it still has importance as a measure of system performance.

For services in which travel time is a more significant component of service time, as in emergency ambulance services, these assumptions are unrealistic. We will return to this problem in Chapters 4 and 6.

A.3 Performance Measures

There are several measures of performance which are important in evaluating the effectiveness of police response and patrol. While there has not been a great deal of work concerning the relation of response time (the time from the reception of the call to the arrival of a unit on the scene) to crime prevention or interception, there is a consensus that response time should be relatively low (Ref. 56). For the purposes of developing a local identity or familiarization with a particular area in preventive patrol, it may be important to keep the fraction of calls which take a unit out of its sector as low as possible. In addition, both the absolute unit workload (fraction of time a unit spends servicing calls) and that workload relative to the other units (workload imbalances) can be important administrative and morale factors.

Both response times and workloads may be considered in local and global terms. For instance, the overall response time for a system may be very low but this performance may not be acceptable because of inequities between different areas. On a microscopic level, the fraction of calls which send a unit to any particular part of the region might be of interest.

The hypercube has seen its major application in terms of police sector design. In this context, for a particular number of units assigned to patrol a certain geographic area, the sectors determine the preventive patrol patterns in the absence of calls for service. The hypercube model is used to predict the actual patterns of patrol and response under various assignment alternatives for that sector design. These patterns are given in terms of workloads, response times, the probability of queuing, and intersector dispatches. The reader is referred to References 3 and 40 for a more complete discussion of this application. In what follows, we abstract the general features of the hypercube formulation and give examples of how this formulation may be applied to police and other service systems.

B. The General Hypercube Queuing Model

In this section, we give a general formulation of the hypercube model for a system with distinguishable servers and distinguishable classes of customers. The focus will be on the development of assignment (dispatch) rules to minimize the average cost of providing service when the costs depend on both the class of the customer and the particular unit providing service. This formulation is a generalization of the model developed by Campbell (Ref. 3) for police sector design. Jarvis and Larson indicate other applications of the formulation and summarize the development in Ref. 28. Although we focus on the zero line case here, the modifications required to deal with infinite line capacity are simple and are indicated as appropriate.

B.1 Model Assumptions and Notation

The hypercube model is a continuous time, finite state Markov description of an N server queuing system with distinguishable servers. A state in the Markov process is denoted by a binary vector, VH (state vector, hypercube model), of length N, where the j-th component of the vector is zero if and only if server j is free (available for service) in that state. (Note: the servers are indexed from 0 to N-1, from right to left in the state vector). For convenience, a state is often referred to by the integer associated with the binary vector. Hence there are 2**N states, indexed from 0 to

(2**N)-1 inclusive, where state i has binary representation

$$VH\{i,-\} = (\ VH\{i,N-1\},\ VH\{i,N-2\},\ \ldots\ ,\ VH\{i,0\}\ ) \tag{3.1}$$

and

$$i = \sum_{k=0}^{N-1} VH\{i,k\} * 2^{**}k \quad . \tag{3.2}$$

For example, with three servers, there are 2**3 or 8 states: 000, 001, 010, 011, 100, 101, 110, and 111. State 3, binary 011, corresponds to units 0 and 1 being busy and unit 2, free. The state vector for that state is given by VH{3,0}=1, VH{3,1}=1, and VH{3,2}=0; VH{3,-} = (0 1 1).

There are NC (number of customer classes) distinct classes of customers. Customers of type j arrive for service according to a Poisson process with parameter CR{j} (call rate), j=1,2,...,NC, independent of all other customer types and the state of the system. The total call rate CRT (call rate, total) is given by

$$CRT = \sum_{j=1}^{NC} CR\{j\} \quad . \tag{3.3}$$

For geographically distributed systems, the customer classes are defined in terms of geographic *atoms* or *reporting areas*, which partition the area of interest. In the police patrol example, customers could be classified according to the atom containing the origin of the call for service as well as the actual type of incident reported; the N servers correspond to mobile patrol cars.

B.2 The Steady State Equations

The Markov description of the hypercube model is completed by specifying the transition rates. As is obvious from the state space, transitions involve either an available server becoming busy (an *upward transition*) or a server completing service and becoming available (a *downward transition*). An upward transition corresponds to a customer arrival and the subsequent assignment of exactly one server. Such events occur at rate CR{j} for customer type j. The total rate at which an upward transition occurs depends on the customer types serviced by a particular server in the state of interest. This leads to the notion of an assignment rule.

For every state i (except the saturation state), there is a vector POL{i,-} (policy vector) of length NC where POL{i,m}=n if and only if server n is assigned to customers of type m in state i. (Note: this requires that the server be available, i.e. VH{i,n}=0. Assignments are unique; randomized rules are not allowed). Hence for states i and j where VH{i,r}=VH{j,r} for r not equal to n and VH{i,n}=0, VH{j,n}=1, there is an upward transition from state i to state j with rate R where

$$(3.4) \qquad R = \sum_{m:POL\{i,m\}=n} CR\{m\} \quad .$$

Each server is assumed to have a service time which is exponentially distributed with mean 1/RH{n} (service rate, hypercube model) for server n, independent of the particular customer being serviced or the past history of the system. Therefore, a downward transition occurs from a state j to a state i with rate RH{n} where VH{i,r}=VH{j,r} for r not equal to n and VH{j,n}=1, VH{i,n}=0 (corresponding to server n completing service).

State i is said to be *adjacent* to state j if a transition from i to j is possible. Note that every state is adjacent to exactly N other states corresponding to a change in the status of each server. If PH{i} (steady state probability, hypercube model) is the steady state probability of state i, a set of 2**N simultaneous linear equations can be solved to obtain the PH{i}. Heuristically, the equations may be derived by equating the rate at which a state is entered to the rate at which it is departed (Ref. 15). The resulting *equations of detailed balance* are written

$$(3.5)\qquad PH\{i\} * \left[CRT*DEL + \sum_{j:VH\{i,j\}=1} RH\{j\}\right]$$

$$= \sum_{j:VH\{i,j\}=1} PH\{i-2**j\} * \left[\sum_{m:POL\{i-2**j,m\}=j} CR\{m\}\right]$$

$$+ \sum_{j:VH\{i,j\}=0} PH\{i+2**j\} * RH\{j\}$$

$$\text{for } i=0,1,\ldots,(2**N)-1;$$

where DEL is zero for i=(2**N)-1 (all servers busy) and one otherwise.

Any one of the equations in (3.5) is redundant and may be replaced by a normalization constraint to obtain a unique solution for the PH{-}. This development assumes that calls arriving during saturation (state s, s=(2**N)-1, all servers busy) are irrevocably lost (or handled by resources external to the system).

Infinite line capacity is easily handled in the same framework; one simply adds the infinite tail associated with an M/M/N queue. Equation (3.5) for i=s is modified to

$$(3.6)\qquad PH\{s\}*(CRT + SRT) = \sum_{j=0}^{N-1} CRT*PH\{s-2**j\} + SRT*QP\{1\}$$

where QP{j} is the probability of j customers in queue, j=1,2,... and SRT (service rate, total) is the sum of the RH{-}. The equations for the QP{-} are

$$QP\{j\}*(CRT + SRT) = CRT*QP\{j-1\} + SRT*QP\{j+1\} \tag{3.7}$$

for j=1,2,... and QP{0}=PH{s}.

The addition of queueing just rescales the probabilities obtained for the zero-line case (Ref. 36). State s corresponds to all servers busy but no calls queued. In order for the steady state probabilities to exist when queuing is permitted, it is necessary that CRT be less than SRT. That is, the system must be able to service calls at a greater rate than that at which they arrive.

The hypercube model is parametrized by the state dependent assignment rules. In practice, the choice of assignments would be based on the state of the system and the costs associated with the particular server-customer pair. Explicitly, let C{i,j} (cost of assignment) denote the expected cost associated with assigning server i to a customer of class j. In the context of police patrol analysis, the cost term might be the expected time for unit i to reach the location of atom j (focusing on the travel time component of response time). Larson has considered several "standard" dispatch rules for police patrol which incorporate a probabilistic description of the location of mobile units with expected travel times between points in the geographic area determining the cost of assignment (Ref. 41).

Of course, our formulation is not restricted to the use of travel times as the cost of assignment. For police operations, the costs could be expressed in terms of zero-one variables corresponding to a maximum acceptable travel time. More generally, a utility function for response time, such as developed by Keeney for fire operations (Ref. 34), could be used to incorporate subjective preferences for response times. In the case where the system includes specialized servers, such as bilingual police officers, the cost structure might ignore travel time entirely. Instead, the emphasis could be placed on deriving the maximum benefit from the skills of the specialized servers. This could mean dispatching a bilingual officer across town for calls originating from particular ethnic communities.

In Chapter 7, we will examine a system in which costs are expressed in terms of the risk of death in certain medical emergencies. The formulation relies on work by Cretin (Ref. 13) in which response time is one component in determining the risk of dying after a myocardial infarction.

B.3 Computing Performance Measures

As noted previously, the hypercube model focuses simultaneously on many performance measures for a system.

Given a particular assignment rule, the equations of detailed balance (3.5) can be solved for the steady state probabilities. From the steady state probabilities it is easy to compute the workloads, WL{i}, or fraction of time server i is busy, by

$$(3.8) \qquad WL\{i\} = \sum_{j:VH\{j,i\}=1} PH\{j\} \quad ;$$

the fraction of arrivals of customer class j handled by server i, FSC{i,j} (fraction by server and customer)

$$(3.9) \qquad FSC\{i,j\} = \sum_{m:POL\{m,j\}=i} PH\{m\} \quad ;$$

and the system wide expected cost per customer, EC (expected cost),

$$(3.10) \qquad EC = \sum_{j=1}^{NC} (CR\{j\}/CRT) * \left[\sum_{i=0}^{N-1} FSC\{i,j\} * C\{i,j\} + PH\{s\} * CS\{j\} \right]$$

where CS{j} (cost during saturation) is the expected cost of a call of type j arriving during a period of saturation and

PH{s} is the probability of saturation. Note that one minus the sum of PSC{i,j} over i is the fraction of calls of type j arriving during a period of saturation; that is, PH{s}. Other quantities of interest are detailed in Campbell (Ref. 3) Larson (Ref. 36).

The main drawback of Campbell's work was an inability to deal with a system with more than six servers. By using an iterative technique based on the structure of the model, Larson (Ref. 36) extended the problem size that could be handled by two orders of magnitude (roughly fifteen servers). Even greater improvement in this area has recently been achieved by Larson using an approximation method which will be discussed in more detail in Chapter 5 (Ref. 40).

C. Optimization of Assignment Rules

As noted above, one quantity of interest is EC, the average cost per customer. It is natural to ask if it is feasible or advantageous to minimize this quantity by varying the assignment rule.

This problem has been solved in closed form for the case of a spatially distributed system with two emergency response units by Carter, Chaiken, and Ignall (Ref. 4). For calls arriving according to a spatially distributed Poisson process and serviced by one of two fixed location units, the optimal response area A for unit 0 is given by

$$A = \left\{ x: D\{0,x\} - D\{1,x\} \leq K \right\} \tag{3.11}$$

where K is a constant which depends on the call distribution, system geometry, and utilization (average workload) when RH{0} is equal to RH{1}. The D{i,x} are functions specifying the cost (here travel time or distance) associated with unit i servicing a call at location x. Note that in the model formulation, only in the state with both units available, VH{0,-}=(0 0), is there any choice as to which unit to dispatch.

Small system size was the critical item in developing an analytic solution for the two server case. The Markov process has only four states and may be solved for the PH{i} explicitly. By parametrizing the response areas by the difference in travel time or distance as in equation (3.11), the average costs may be minimized by testing the boundary conditions and extreme points of the objective function as a function of that difference. Such an approach is impractical for larger numbers of servers. Although possibly quite large in number, the alternative assignment rules for a finite number of customer classes are also finite. Since each assignment rule yields an associated set of transition rates and expected costs, standard techniques for minimizing expected costs in Markov processes may be employed.

C.1 Markov Decision Theory

Before dealing with the hypercube optimization problem, we briefly describe the procedure for solving a Markov decision problem. Probably the best known solution procedure is due to Howard (Ref. 24). This technique, referred to as an iteration in "policy space," can be adapted to deal with either continuous or discrete time, finite state processes. We restrict our attention to the discrete time problem (successive call arrivals) as we are interested in the cost of assignment events in the hypercube framework.

For the general M-state problem, define TC{i,j,k} (expected transtion cost) to be the expected cost of a transition from state i to state j under policy k. (Policy choices are made on a state by state basis). Define TP{i,j,k} (transition probability) to be the one-step transition probability of going from state i to state j under policy k in state i. The unconditional expected transition cost from state i under the application of policy k, ETC{i,k}, is given by

$$(3.12) \qquad \mathrm{ETC}\{i,k\} = \sum_{j=1}^{M} \mathrm{TC}\{i,j,k\} * \mathrm{TP}\{i,j,k\} \quad .$$

Finally, define CT to be the expected cost per transition in the steady state operation of the system under a particular policy scheme. The Markov decision problem is to find the policy scheme which minimizes CT.

This problem is solved in an iterative procedure with two phases: value determination and policy improvement. The value determination phase for a particular policy scheme involves solving the set of M simultaneous linear equations in SV{i}, i=1,2,...,M, and CT,

$$\text{(3.13)} \qquad CT + SV\{i\} = ETC\{i,k\} + \sum_{j=1}^{M} TP\{i,j,k\} * SV\{j\}$$

for i=1,2,...,M,

where SV{i} is the state value associated with state i under that particular policy scheme. The set of equations in (3.13) determine the SV variables up to an arbitrary constant. A unique solution may be obtained by setting any one of the SV{i} equal to zero.

The policy improvement routine uses the state values determined from the value determination phase. In particular, for each state i, we choose the policy k which is the minimum in

$$\text{(3.14)} \qquad \min_{k} \left\{ ETC\{i,k\} + \sum_{j=1}^{M} TP\{i,j,k\} * SV\{j\} \right\}$$

If $\underline{k}$ does strictly better in (3.14) than the previous policy k (for any state), the policy scheme obtained by replacing k with $\underline{k}$ will result in a lower cost per transition. The value determination phase is now reapplied to the new policy scheme.

This technique proceeds iteratively until it is not possible to obtain a strict improvement in (3.14). At that point the optimal policy has been obtained. The optimal state values have the property that they satisfy

$$(3.15) \qquad SV\{i\} = \min_{k} \left\{ ETC\{i,k\} + \sum_{j=1}^{M} TP\{i,j,k\} * SV\{j\} - CT \right\}$$

for i=1,2,...,M.

In fact, if there are numbers SV{1}, SV{2}, ... , SV{M} satisfying (3.15), that is a sufficient condition to guarantee that the minimizing policy scheme in (3.15) is the optimal choice (Ref. 50). This result will prove useful in applying Markov decision theory to the hypercube model.

C.2 Application of Markov Decision Theory to the Hypercube

Before expressing the choice of assignment rule as a Markov decision problem, we make a slight digression. Several alternative system descriptions are possible. We can look at the optimization problem in terms of either an infinite or zero line system. In addition, the system may be treated in continuous or discrete time. That is, we can focus on the cost per assignment or the cost per unit time.

In the original formulation of the hypercube model, the service rates for each server were identical. In that case, by focusing on the number of busy servers, it is possible to show that the hypercube model reduces to an M/M/N queuing system (with either zero or infinite line capacity). The important point to note is that this result holds regardless of the assignment rule being used (Ref. 24). In fact, given the steady state probabilities and average cost in any of the system descriptions mentioned above, it is possible to write the same quantities for all other descriptions in terms of linear transformations whose parameters can be expressed as a function of CRT, SRT, and PH{s}, the saturation probability for the continuous time, zero line description.

The fact that PH{s} is independent of the assignment rule implies that if a policy is optimal in one system description, it must be optimal in every other description. An examination of the development in Ref. 24 shows that the relation between the steady state probabilities and average costs also holds for the more general hypercube model in that the saturation probability is the crucial quantity. Unfortunately, if the RH{i} are not all equal, the saturation probability is no longer independent of the assignment rule.

This fact can be demonstrated by a simple two server

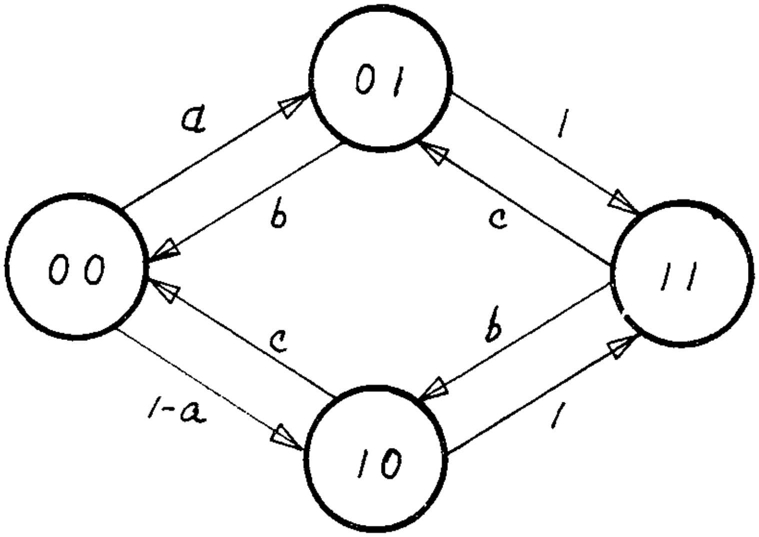


Figure 3.1 State transition diagram for a two server system. CRT=1, RH{0}=b near zero and RH{1}=c much larger than one. Assignment rule one sets a=1; rule two sets a=0. The saturation probability is not independent of the assignment rule.

example (See Figure 3.1). In this instance, unit 1 is assumed to have a very long expected service time (RH(0) near zero) and unit 1 has a very short service time (RH(1) large). For convenience, we set CRT equal to one and consider only two assignment rules. The first (a=1 in Figure 3.1) assigns every call to unit 0 if possible. The second (a=0 in Figure 3.1) never makes an assignment to unit 0 if possible. It is easy to see that under the first policy the system will usually be in state (01); server 1 free and server 0 busy. Alternatively, the second policy will leave the system in state (00) most frequently; both servers free. In this instance, the probability of a particular number of servers being busy is not independent of the assignment policy.

We now focus on the discrete time, zero line capacity formulation of the hypercube model for the application of Markov decision theory. The techniques, of course, are applicable to the other descriptions as well. (It should be noted that it is fairly easy to specify a two server system for which the optimal policies differ between the zero and infinite line case).

In order to apply the techniques of Markov decision theory to the choice of assignment rules we have only to specify the transition probabilities and expected transition costs in the manner of section C.1. For the sake of completeness, we allow self transitions in the saturation state corresponding to the arrival of calls which are not handled by the system but which may contribute to the cost of system operation.

The transition probabilities may be specified for any state i, other than the saturation state s, as follows: Let A be the total rate at which service is completed in state i. That is,

$$(3.16) \qquad A = \sum_{m=0}^{N-1} RH\{m\} * VH\{i,m\} \quad .$$

If server n is free, hence VH{i,n}=0, it is possible to make an upward transition to state j, where j=i+2**n. This event will occur at rate B, where B is given by

$$(3.17) \qquad B = \sum_{m:POL\{i,m\}=n} CR\{m\} \quad .$$

Then the transition probability from state i to state j, calling the current policy k is

$$(3.18) \qquad TP\{i,j,k\} = B \ / \ (A + CRT) \quad .$$

If server n is busy, VH{i,n}=1, then there can be a downward transition to state j where j=i-2**n. The transition probability is given by

(3.19) $$TP\{i,j,k\} = RH\{n\} \;/\; (A + CRT) \quad .$$

Noting that only upward transitions (call arrivals) incur a cost, we can write the expected cost of a transition from state i by conditioning on the type of customer arrival as

(3.20) $$ETC\{i,k\} = \sum_{m=1}^{NC} CR\{m\} * C\{POL\{i,m\},m\} \;/\; (A + CRT)$$

The description is completed by treating the saturation state. For this state, we allow a self transition to occur. This event has rate CRT. Hence

(3.21) $$TP\{s,s,k\} = CRT \;/\; (CRT + SRT)$$

which is independent of the assignment rule. From state s, a downward transition to state j=s-2**n occurs at rate RH{n} and with probability

(3.22) $$TP\{s,j,k\} = RH\{n\} \;/\; (CRT + SRT) \quad ;$$

again, independent of the policy. The cost of a transition in the saturation state also takes a special form.

In this case, instead of using the C{i,j}, the saturation costs, CS{j}, are employed. As before, conditioning on the type of customer arrival and assigning zero cost to downward transitions, we have

$$(3.23) \qquad ETC\{s,k\} = \sum_{m=1}^{NC} CR\{m\} * CS\{m\} \quad / \ (CRT + SRT) \quad ,$$

also independent of the policy. With these definitions, we can directly apply the techniques of Markov decision theory to the problem of choosing assignment rules to minimize the average cost per call.

The difficulty with this approach becomes apparent after examining a typical system. For an N server system, consider the number of assignment alternatives in the zero state alone. There are NC types of calls, each of which may be assigned to any of the N available servers. Hence there are N**NC distinct alternatives. For the moderate sized system of N=6 and NC=25 this number is roughly 10**19. Even under the generous assumption that one of the evaluations required in the policy improvement phase, equation (3.14), could be computed in a tenth of a microsecond, a single iteration for state 0 would take 10**12 seconds; roughly one hundred thousand years. An alternative approach is indicated.

C.3 A Characterization of Optimal Policies

One method for dealing with the large number of possible assignment rules is to find some characterization of the optimal policy. Of course, the more detailed the characterization, the smaller the number of alternatives which must be examined. One approach is to use the functional form of the optimal policy given in equation (3.15).

We focus on a state i in which there is a policy choice. That is, there must be at least two free servers. Denote these servers by indices n and m. Let $\underline{k}$ denote the optimal policy in state i. Let r be a particular class of customer assigned to server n in state i under policy $\underline{k}$. Consider the alternative policy k which is the same as $\underline{k}$ except that a type r customer is assigned to server m instead of server n.

Since $\underline{k}$ achieves the minimum in (3.15), we have

$$(3.24) \qquad ETC\{i,k\} + \sum_{j} TP\{i,j,k\} * SV\{j\} \geq ETC\{i,\underline{k}\} + \sum_{j} TP\{i,j,\underline{k}\} * SV\{j\} \quad .$$

where in both summations, j indexes those states adjacent to state i. Now, since policies k and $\underline{k}$ are the same except for the assignment for type r calls, most of the terms on each side of the inequality (3.24) cancel. After this simplification, we are left with

$$(3.25) \qquad C\{m,r\} + SV\{i+2**m\} \geq C\{n,r\} + SV\{i+2**n\}$$

or

$$(3.26) \qquad C\{m,r\} - C\{n,r\} \geq SV\{i+2**n\} - SV\{i+2**m\}.$$

If for some call type q, we have that C{m,q} minus C{n,q} is strictly greater than C{m,r} minus C{n,r}, then m cannot be the optimal unit to assign to type q in state i.

This last result may be seen by supposing that server m is optimal for calls of type q. Applying the same logic that led to (3.26) we have

$$\begin{aligned}(3.27) \qquad C\{n,q\} - C\{m,q\} &\geq SV\{i+2**m\} - SV\{i+2**n\} \\ &\geq C\{n,r\} - C\{m,r\} \quad .\end{aligned}$$

The last inequality follows from (3.26) and contradicts the assumed relation between the differences in cost of assigning unit n versus unit m for call types q and r.

This result may be stated in a more symmetric form as follows:

> If C{m,q} - C{n,q} > C{m,r} - C{n,r} , then the optimal assignment rule does not assign server m to customer type q and server n to customer type r in any state where both servers m and n are available.

Heuristically, if we were unwilling to incur the extra cost $C\{m,r\} - C\{n,r\}$ to assign server m to customer type r instead of server n, to be consistent we should be unwilling to make the same switch for customer type q and incur the larger difference. The similarity between this result and that given by Carter, Chaiken, and Ignall for two servers should be noted. The proof given above, which is applicable to an arbitrary number of servers, is a considerable simplification of the original approach used by Jarvis (Ref. 24).

The practical significance of this result is largely computational. Equation (3.25) can be used in the policy improvement phase of Howard's algorithm to determine the exact difference in cost of assignment that will minimize the cost per transition at that step of the solution procedure. Units are considered pairwise and only the current "best" transition probabilities and expected transition costs need be retained at any step of the iteration. A more detailed description of the implementation of this solution scheme may be found in Jarvis (Ref. 24).

The solution procedure devised by Jarvis is a variation of a method detailed by Odoni (Ref. 47). Basically, this procedure solves the Markov decision problem by successive approximations in a dynamic programming framework which gives monotonically decreasing bounds on the cost per transition. The computational variation used by Jarvis is intermediate between the dynamic programming procedure and Howard's "iteration in policy space."

Various implementations of the hypercube model have been used over the past three years at the Massachusetts Institute of Technology. Larson (Ref. 37) and Weissberg (Ref. 61) have documented a set of hypercube programs which are designed mainly for use by police planners with a minimum knowledge of the theoretical foundations of the model. The general hypercube model described above has been implemented in a computer program by Jarvis (Ref. 27). This version includes the optimization of dispatch rules. All of the programs are written in PL/I and are in the public domain.

As an example of the computer costs associated with using the model, the optimal dispatching problem for a system with N=6 (servers) and NC=62 (customer classes) can be solved for roughly two dollars using the time sharing option on an IBM 370/168. This amounts to less than 10 seconds of CPU time.

In the next section, we give an example of the use of the hypercube model to find the optimal assignment rule for a three server example. In the following section, the model is used to generate average response distance data for comparison to the "square root law" models which relate average response distances to the average number of units available in a geographic region.

D. A Three Server Example

This example is included to illustrate the sort of problem that may be addressed by the hypercube model. The reader who is more interested in the specific application to police patrol analysis should refer to Larson (Ref. 39, 40, and 37).

D.1 A Description of a "Sample City"

Our example deals with a spatially distributed system in which we are interested in minimizing the average distance traveled in response to a call for service. We consider three fixed units, variously located in the famous vacation spa, Sample City (See Figure 3.2). The area has been partitioned into 16 atoms which serve as a basis for classifying the calls for service.

Assume that unit 0, located in atom 1, can service calls at a rate of 1 per hour. Unit 1, in atom 11, can service 1.5 calls per hour; unit 2, in atom 16, 0.75 per hour. Calls arrive at a rate of 1.3 per hour with the spatial distribution shown in Figure 3.2 and Table 3.1. Table 3.1 summarizes the parameters of the system. These include the cost per call in terms of travel distance (one way, arbitrary units), and the call and service rates specifying the various distributions. The cost for a call arriving during a period of saturation is taken to be the same as the travel distance that would be incurred by a unit stationed in atom 9. This cost may be

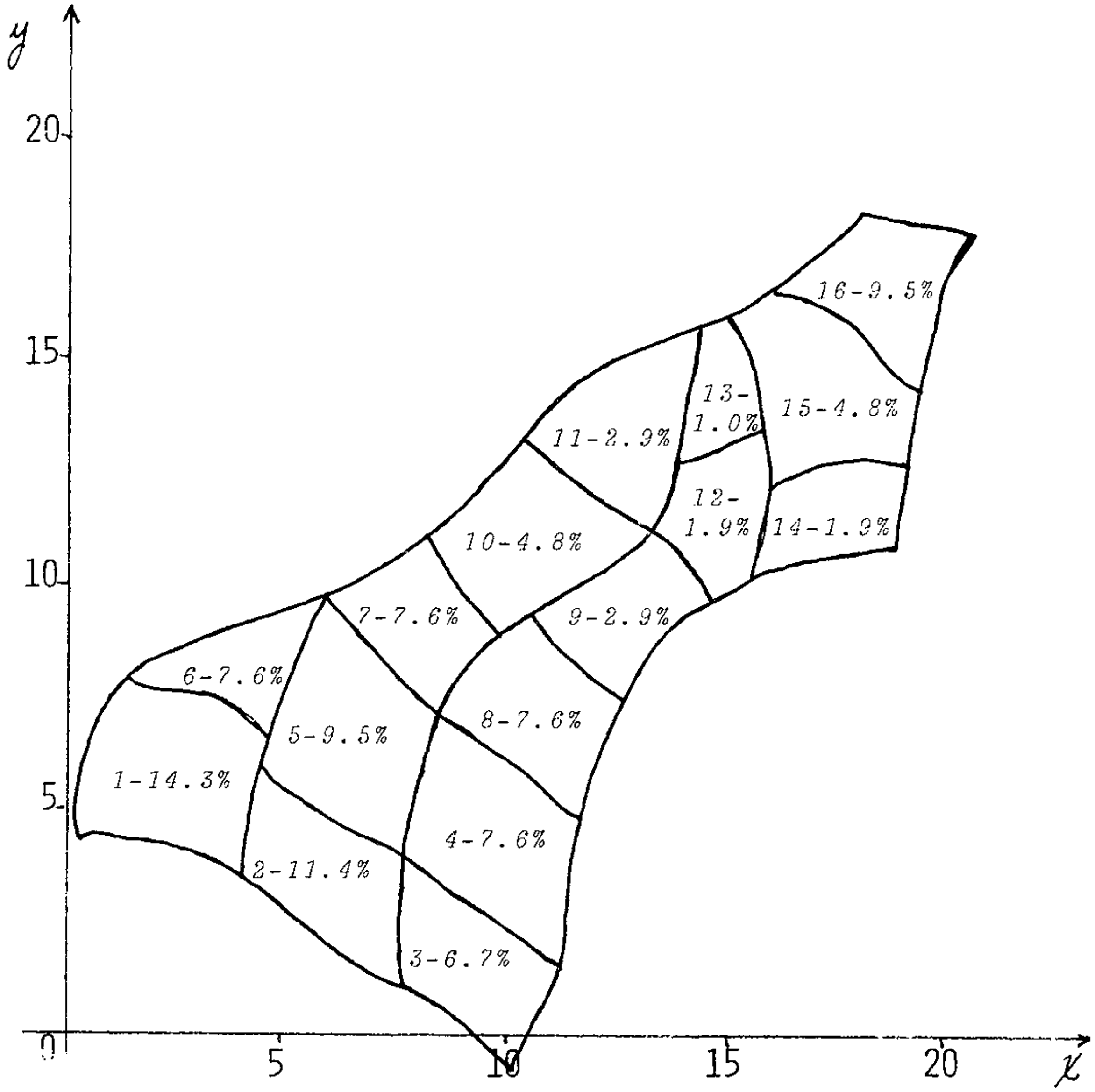


Figure 3.2 A map of Sample City. The area is partitioned in 16 atoms, each shown with its index and the percentage of the calls being generated from that atom.

Table 3.1 System parameters for the three server example in Sample City. Costs of assignment and service and call rates are given below in tabular form.

A. Costs per assignment (C{i,j} and CS{j}) with percentage of calls from each atom (100*CR{j}/CRT):

ATOM j:	C{0,j}	C{1,j}	C{2,j}	CS{j}	%CR{j}
1	2.4	17.2	26.4	13.9	14.3
2	5.9	16.1	25.3	12.8	11.4
3	11.5	14.7	23.9	12.8	6.7
4	8.7	11.3	20.5	8.0	7.6
5	4.8	12.4	21.6	9.1	9.5
6	3.7	13.5	22.7	10.2	7.6
7	8.5	8.7	17.9	5.4	7.6
8	9.2	8.0	17.9	4.7	5.7
9	13.9	4.5	12.5	1.8	2.9
10	13.2	4.0	13.2	4.1	4.8
11	17.2	2.1	9.2	4.5	2.9
12	17.7	4.1	8.7	3.8	1.9
13	20.2	3.0	6.2	6.3	1.0
14	20.4	6.4	6.0	6.5	1.9
15	22.8	5.6	3.6	8.9	4.8
16	26.4	9.2	1.9	12.5	9.5

B. Service rates (RH{i}):

RH{0}=1.00 RH{1}=1.50 RH{2}=0.75

C. Total call rate (CRT) and service rate (SRT):

CRT=1.3 SRT=3.25

thought of as the distance traveled by a centrally located backup unit. The distances are approximately equal to the right angle distance between atom centroids with respect to the coordinate system of Figure 3.2. The atom call rates shown in Table 3.1 are not the absolute call rates but are expressed as a percentage of the total. Units always return to their original location before servicing subsequent calls.

In the zero line case, the last assumption is immaterial; it is crucial if calls are allowed to queue. Without the assumption that units return to their original location before dispatch, there is no simple way to calculate the cost of a call arriving during a period of saturation (Ref. 36).

D.2 The Optimized Strategy

One reasonable assignment procedure is to always assign the closest available unit to each call for service as it arrives. The continuous time state transition diagram for this assignment rule is shown in Figure 3.3. The closest unit is easily determined from Table 3.1. This strategy can be thought of as a myopic (or short-term) optimization. As is often the case in Markov decision theory, the average cost can be reduced by assignment rules which are not optimal in an immediate cost sense but which better anticipate future events. Table 3.2 is a summary of some system performance measures as computed for two assignment rules: the myopic and the optimal. Although not evident from Table 3.2, the only

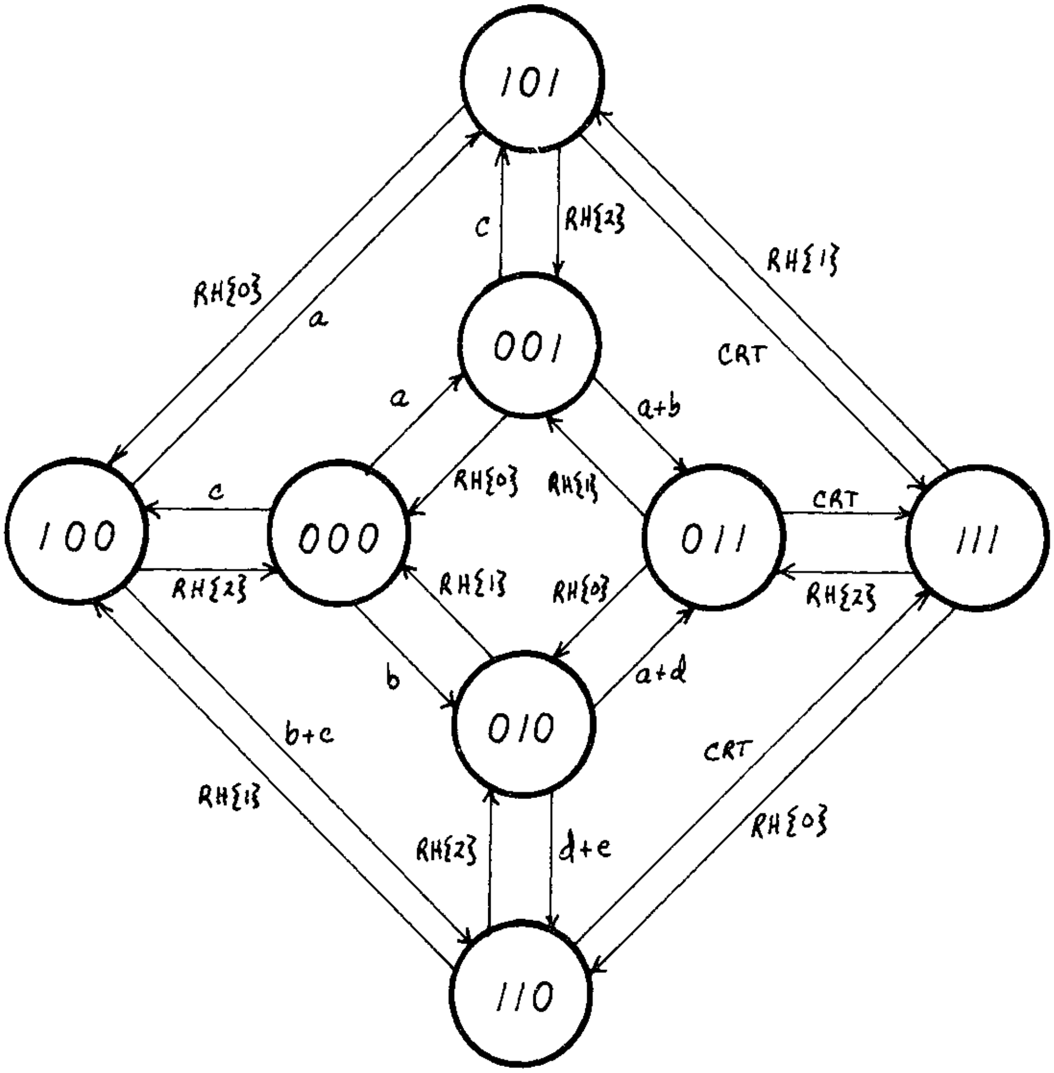


Figure 3.3 State transition diagram for the three server example of section D, Chapter 3. The assignment rule is to dispatch the closest available unit. In the diagram,

a=CR{1}+CR{2}+CR{3}+CR{4}+CR{5}+CR{6}+CR{7},
b=CR{8}+CR{9}+CR{10}+CR{11}+CR{12}+CR{13},
c=CR{14}+CR{15}+CR{16},
d=CR{8}+CR{10},
e=CR{9}+CR{11}+CR{12}+CR{13};

as determined from Table 3.1.

Table 3.2 Summary of system performance for the myopic versus optimal assignment rules.

		MYOPIC	OPTIMAL
A. Steady State Probabilities (PH{i}):			
State:	0-000	0.2982	0.3081
	1-001	0.1981	0.1595
	2-010	0.0720	0.1048
	3-011	0.0922	0.0888
	4-100	0.1087	0.1118
	5-101	0.0886	0.0827
	6-110	0.0476	0.0543
	7-111	0.0914	0.0904
B. Cost per Assignment:			
		8.7335	8.6761
C. Unit Workloads (WL{i}):			
Unit:	0	0.4703	0.4214
	1	0.3031	0.3384
	2	0.3363	0.3392
D. Maximum Workload Imbalance:			
		0.1672	0.0830
E. Fraction of Calls by Unit:			
Unit:	0	0.3630	0.3239
	1	0.3510	0.3901
	2	0.1946	0.1956

diffference in these rules occurs in state 000 and state 100.

In state 000, the optimal rule assigns unit 1 to calls from atoms 3,4 and 7 even though unit 0 is closer. In state 100, atom 7 is assigned to unit 1 instead of unit 0. In every other respect, the rules are identical. It should be noted that the optimal rule decreases the probability of saturation. This is consistent with the earlier remark concerning the variability of the saturation probability for systems in which the unit service rates are not identical.

The global cost per call is roughly the same for the two rules. In fact, the optimal rule results in only a 0.66 percent decrease as compared to the myopic policy. This is consistent with previous experience in using the optimization for spatially distributed systems. Larson and Stevenson (Ref. 42) have investigated this type of problem and reach the conclusion that travel time or distance in spatially distributed systems is largely insensitive to changes in system configuration.

There is one additional item to be noted in Table 3.2. Although the global travel distance does not decrease very much under the optimal policy, the maximum workload imbalance, the difference between the busiest and least busy unit, has been halved by the optimal policy. This effect is characteristic of the optimal policy. In heuristic terms, the improvement in global travel distance is made by avoiding situations in which a unit must be dispatched the relatively long distance into an area normally covered by another unit.

In this particular example, unit 0 has basic responsibility for the left corner of Sample City, an area with a large internally generated workload. By occasionally assigning unit 1 to respond to areas which are slightly closer to unit 0, the workload of unit 0 is decreased. As a result, fewer calls arrive when unit 0 is busy, therefore unit 1 is not so often dispatched deep into the area normally covered by unit 0. By incurring these slightly larger costs, the system avoids the much larger differences. These results are similar to those noted by Carter, Chaiken, and Ignall (Ref. 4) for the two server case.

All other computational experiences to date for geographically distributed systems have shown the same general characteristic: the optimization procedure does not result in a significant improvement in global expected travel distances as compared to dispatching the closest available unit but does decrease workload imbalances. Hence, the minimization of response distances can serve as a surrogate for reducing workload imbalances directly. (It should be noted that the optimizaition can be expected to result in larger imbalances in travel distances. In some situations this increase may not be significant (Ref. 28), but it should not be ignored).

The characterization of optimal policies might be used to further reduce workload imbalances with a minimal increase in average response distance by "over-relaxation." The workload of the least busy unit could be increasel by assigning those calls which have a small difference in cost of assignment as compared to the optimal policy. In certain instances, the optimal policy itself can be perturbed in order to balance workloads. If the optimization results in ties between pairs of units for assignment to a type of call, any resolution of those ties will result in the same average cost of assignment (Ref. 24). This result may be utilized to decrease the workload between those pairs of servers by making the assignments to the server with the smaller workload.

As a final remark concerning the optimization, it should be noted that it supplies very useful negative information for spatially distributed systems. That is, the myopic policy yields average travel distances which are very close to those obtained from the long range optimum. This result will prove very useful in Chapter 5. At the present time, it is not known whether these remarks hold for the alternative cost structures described in Section B.2 of this chapter.

E. The Square Root Law for Response Distance

As we have seen in the previous example, the hypercube model can be used to predict average response distance for spatially distributed queuing systems. There is a class of models, referred to as the "square root laws" for expected response distance, for predicting this one performance measure without using a complex model like the hypercube.

The square root law was first postulated by Larson (Ref. 41 and 39) for estimating travel time and distance in police operations. The hypothesis was that average response distance is proportional to the inverse square root of the density of response units within an area. For a particular dispatch rule (Ref. 41), Larson estimated expected response distance, ED, as

$$(3.27) \qquad ED = (2/3) * (1 + AWL) / (A/N)**0.5 \quad ,$$

where AWL is the average unit workload; A is the area of the region being considered; and N is the number of response units. In (3.27), as congestion in the system increases, so does the average workload and hence the expected response distance. This phenomenon corresponds to the increasing frequency of dispatches of distant units because the usual, closer servers are unavailable. Larson found (3.27) to be a good approximation for AWL less than seven tenths.

A similar model has been proposed by Blum and Kolesar (Ref. 2). Using simple analytic models, such as those proposed by Larson (Ref. 41 and 39) and Larson and Stevenson (Ref. 42), as well as simulation and historical data, Blum and Kolesar estimate ED as inversely proportional to the square root of the expected number of available units in a region. That is,

$$(3.28) \qquad ED = K / (N*(1 - AWL))**0.5 \quad ,$$

where K is a constant of proportionality to be determined from data from the particular region being examined. It is assumed in the development of (3.28) that the probability of saturation is negligible and that the service times of the units are roughly the same. Note that N*(1-AWL) is the average number of available units within the region.

Since the validity of (3.28) has been based largely on comparisons with simulation and historical data, the hypercube model is an interesting alternative means for investigating this particular square root law. Using data collected by Jarvis and McKnew (Ref. 30) for an area of 5.2 square miles (62 atoms), expected travel distances were computed for two situations. In the first, fifteen units were located more or less uniformly over the region. The total call rate (for a fixed service rate) was varied so that the average workload increased from ten percent to almost eighty percent. A second set of response distances was computed by holding the total call rate constant and increasing the number of units from 2 to 13, with a corresponding decrease in the average workload from fifty one percent to eleven percent. The results of these computations are shown in Table 3.3 (3.3.A and 3.3.B respectively).

Table 3.3 also gives the expected response distance as predicted by (3.28). The root mean square difference between these estimates and those given by the hypercube model is 0.105, corresponding to 15 percent of the average response

Table 3.3 Expected travel distance as computed by the hypercube model (HM) and the Blum-Kolesar estimate (BK) from equation (3.28) for various numbers of servers (N) and average workloads (AWL).

N	AWL	HM	BK
3.3.A N=15 constant.			
15	0.100	0.365	0.403
15	0.200	0.402	0.428
15	0.300	0.452	0.457
15	0.400	0.520	0.494
15	0.497	0.608	0.540
15	0.588	0.710	0.596
15	0.667	0.816	0.663
15	0.731	0.914	0.738
15	0.782	0.999	0.819
3.3.B N*CRT/SRT=1.5 constant.			
13	0.115	0.401	0.437
11	0.136	0.456	0.481
9	0.167	0.509	0.541
7	0.214	0.646	0.632
5	0.296	0.817	0.790
3	0.433	1.056	1.136
2	0.517	1.269	1.508

distance for this particular set of data. It should be noted that the use of (3.28) requires that K be determined from available data, whereas a model such as (3.27) predicts average response distance directly from simple geographic variables. For this particular example, a least squares fit gives a value of 1.48 for K. Two-thirds of the square root of the area of the region is 1.52. This one example is an argument for combining (3.27) and (3.28) to estimate average response distance by

```
(3.29)      ED = (2/3) * (A / (N*(1 - AWL)))**0.5  .
```

The response distances from Table 3.3 are plotted in Figure 3.4 as a function of the average workload. The figure shows fairly close agreement between the Blum-Kolesar estimates and the values computed by the hypercube model except when the workloads are large (greater than 0.6) or the number of units is small (less than 4).

Although this one example does not establish the validity of the square root law, it does indicate the usefulness of the hypercube model as an alternative to simulation or historical data for verification of other models. The example also suggests that particular performance measures can often be adequately estimated without resorting to complex models.

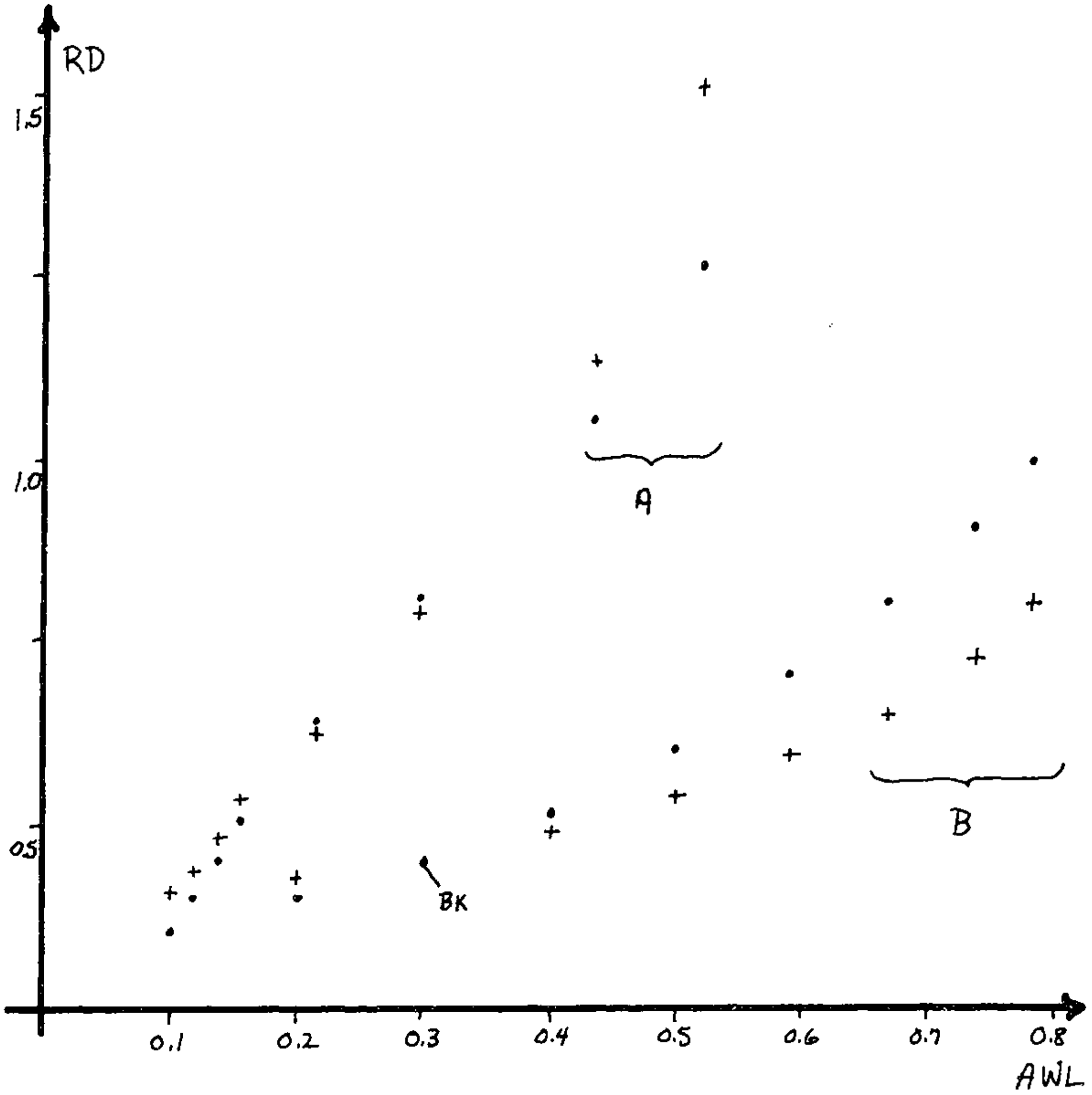


Figure 3.4 Expected response distance as predicted by the hypercube model ( ) and the Blum-Kolesar estimate (+) as a function of the average workload. The regions of largest disagreement are for large workloads (B) or small N (A).

F. Summary

In this chapter, we have introduced the ideas which are basic to modeling emergency response. The emphasis has been on the use of the hypercube model as it is applied to police patrol analysis. At the time of this writing, initial implementation activities in several police departments were underway (Ref. 38 and 14). Many of the references cited here deal with this issue in much greater detail than is appropriate here.

To date, the hypercube model has been used largely as a descriptive tool. In this context, the optimization has seen little practical use. The model is typically used by a planner to analyze various modes of system operation and then to chose the best on the basis of his own priorities.

Such usage involves tradeoffs between workload imbalance and travel time; between local and global performance. Larson (Ref. 40) and Jarvis and McKnew (Ref. 29) deal with this problem at length for large and small scale police operations respectively. Two techniques have recently been developed to facilitate the use of the hypercube model. One of these is an interactive program developed by Weissberg (Ref. 61) which allows the planner to use the model without knowing either the technical details or its particular computer implementation. Chelst (Ref. 7) has developed programs which automatically search for configurations which decrease workload or travel time imbalances for geographically distributed systems.

In practical terms, the optimization example given above does not really address the problem which we would like to solve. Instead of being given the location of units and determining optimal dispatch rules based on those locations, a more realistic approach is to determine the optimal location for a fixed number of units. This problem is addressed in Chapter 5.

An additional area of investigation concerns the travel times associated with geographically distributed systems. In Chapters 6 and 7, we incorporate travel time explicitly into the service times and examine systems in which travel times can be directly related to outcome measures of system performance. This work is motivated by systems such as emergency ambulance services in which travel time is a substantial part of the total service time. In such systems, the service time can depend on both the server and the location of the call.

Chapter 4. GENERAL SERVICE TIME MODELS

A. Introduction

As noted above, the major deficiency of the hypercube model, at least conceptually, is its inability to incorporate more general service time assumptions. This particular difficulty has two facets. In the first place, the service time distribution cannot make the same intuitive appeal to being "memoryless" as the arrival process. This is most evident if the service time consists of two or more distinct components, such as travel to the scene and on-scene service for spatially distributed systems.

In addition, at least on first inspection, the service distributions considered to this point do not allow the flexibility one would like in describing the behavior of more general service systems. While we can specify server dependent service times, in many situations it is more natural to specify times for the type of incident being serviced. For example, for a single volunteer fire department serving a community, one would like to specify service time at least as a function of the severity of the fire or rescue call. In the case of spatially distributed units, the service time could easily depend on both the server and the call type. This chapter introduces that level of generalization.

B. Alternative Continuous-time Markov Models

For all of the objections voiced above, the hypercube model is still attractive in that it is analytically tractable by way of the theory of Markov processes and the other procedures outlined above. For that reason, we first consider models which can be placed in that framework. The first of these arises from an effort to incorporate travel times explicitly; the second allows for server and call type specific service times with exponential distributions. In part, each is motivated by a result due to Sevast'yanov (Ref. 53). That is, the steady state probabilities in an M/M/N:0 queuing system are the same as for an M/G/N:0 system if all servers have identical service times in each system.

The interesting question is whether this result can be extended to problems with distinguishable servers and more general service times. If this is the case, a Markov hypercube-type model would be sufficient for all steady state calculations.

Before addressing the general problem, we consider two situations in which continuous-time Markov models can be used. In the first, which we call the "convolution model," service times are sums of exponentially distributed random variables depending only on the server. In the second, the exponential model, service times are again exponentially distributed but now may depend on both the server and the call type.

B.1 Formulation of the Convolution Model

On a gross level, when we consider spatially distributed systems, there are two major components to service time: travel and on-scene time. If we assume that each of these times can be characterized by possibly server dependent exponential distributions, the system can be modeled as a continuous-time, discrete-state Markov process. In the hypercube model, each state specified information as to the availability of each server. In this more general convolution model, the state space details availability by a '0' if the server is free, a '1' for a server in the first component of service, and a '2' for the second or last component of service.

Specifically, for an N server system, an element of the state space will be denoted by an N-vector, VC (state vector, convolution model), where the i-th element of VC is 0 if server i is free, 1 if server i is in the first component of service, and 2 if server i is in the second component of service. We will consider only the zero line capacity case. For this system, there are obviously 3**N states. As suggested by the hypercube model, if we view a vector in the state space as the ternary expansion of an integer between 0 and (3**N)-1 inclusive, we have a natural ordering of the states. We use either notation as is convenient.

Let RC (service rates, convolution model) denote the N by 2 matrix of service rates for this system. That is, RC(i,j) is the rate at which server i completes the first component of service if j is one; the second for j equal two. Using the same notation as Chapter 3 for the call rates and assignment preferences associated with the call types and making the same independence assumptions, we write the detailed equations of balance for the steady state probabilities PC{-} (steady state probabilities, convolution model) as:

$$
\begin{aligned}
(4.1)\qquad PC\{i\} * \Bigg[& CRT{*}DEL + \sum_{j:VC\{i,j\}=1} RC\{j,1\} + \sum_{j:VC\{i,j\}=2} RC\{j,2\} \Bigg] \\
= & \sum_{j:VC\{i,j\}=0} PC\{i+2*(3**j)\} * RC\{j,2\} \\
+ & \sum_{j:VC\{i,j\}=2} PC\{i-3**j\} * RC\{j,1\} \\
+ & \sum_{j:VC\{i,j\}=1} PC\{i-3**j\} * \Bigg[\sum_{m:POL\{i-3**j,m\}=j} CR\{m\} \Bigg]
\end{aligned}
$$

$$\text{for } i=0,1,\ldots,(3**N)-1.$$

In (4.1), DEL is one except for those states i in which every server is busy (VC{i,n} greater than zero for every n), when it is zero; VC{i,-} is the N-vector ternary expansion of the integer i; CR{i} (call rate) is the call rate for customer type i; CRT (call rate, total) is the sum of the CR{-}; and POL{i,-} is taken to be the assignment vector associated with the availability of servers given by state i.

B.2 Formulation of the Exponential Model

The *exponential* model allows for exponential service times with means being a function of call type and server. Let RE{i,j} (service rates, exponential model) be the service rate for server i and call type j. (RE is an N by NC matrix. NC is the number of call types). Now the state space will specify server status by a '0' for a free server and a 'j', j=1,2,...,NC, for a server busy with a customer of type j. For the zero-line case, this state space has (NC+1)**N elements and we make the usual association between the vector and integer representation.

In a similar fashion, we can write the equations of balance for the steady state probabilities, PE{-} (steady state probabilities, exponential model) as:

$$(4.2) \qquad PE\{i\} * \left[CRT*DEL + \sum_{j:VE\{i,j\} \geq 1} RE\{j,VE\{i,j\}\} \right]$$

$$= \sum_{j:VE\{i,j\}=0} \; \sum_{1 \leq m \leq NC} PE\{i+m*(NC+1)**j\} * RE\{j,m\}$$

$$+ \sum_{j:VE\{i,j\}=m \geq 1} CR\{m\} * PE\{i-m*(NC+1)**j\}$$

for i=0,1,...,((NC+1)**N)-1.

where VE{i,-} is the N-vector (state vector, exponential model) associated with state i and DEL is one except for those states i in which every server is busy (VE{i,n} not equal to zero for every n), when it is zero. CRT and CR{-} are as in Equation (4.1).

It should be noted that these two models are not being presented as particularly useful formulations for real problems. Both require a large amount of storage and time in a computer implementation for other than very small systems. However, computations with these exact models for even small systems can suggest more general hypotheses and provide counter-examples for some postulates.

B.3 A Convolution Example

The details of the convolution model are best shown by an example. This example will also be used to test the hypothesis as to whether Sevast'yanov's result can be extended to systems with distinguishable servers having different service distributions. In order to do this, we compare the convolution model to a hypercube formulation with the same expected service time for each server. More precisely, if c and d are the respective rates at which a unit completes the first and second components of service time, the expected service time for that server is just the sum of the reciprocals of c and d, that is 1/c + 1/d. The corresponding rate of service for the hypercube is the reciprocal of this number; with this correspondence each unit has the same average service time in either model.

For this example, set N=2, NC=2, and CRT=3 with CR{1}=a and CR{2}=b. The service rates are given by the 2 by 2 matrix RC with

$$(4.3) \qquad RC = \begin{bmatrix} c & d \\ e & f \end{bmatrix}$$

The state transition diagram for the convolution model is given by Figure 4.1 when server 0 is preferred for call type 1 and server 1 for call type 2. Again we denote the steady state probabilities by PC{-}, but here it is convenient to use the vector notation for the state space. Thus, the state space has eight elements: 00, 01, 02, 10, 11, 12, 20, 21, and 22.

The variables for the hypercube model are denoted by the

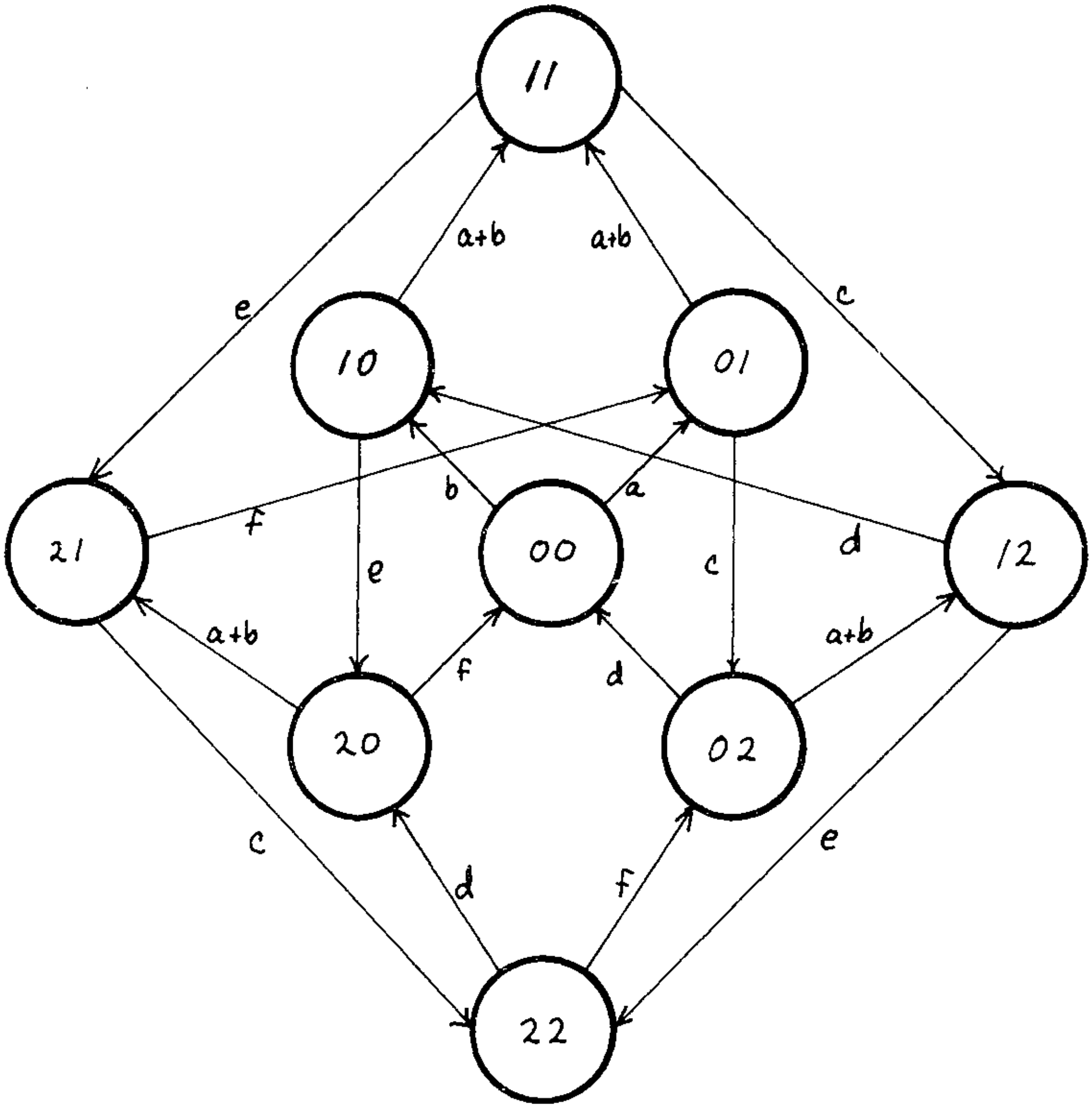


Figure 4.1 State transition diagram for the convolution example of section B.2. Numerical results in Table 4.1 for a=1, b=2, c=1, d=2, e=3, and f=4.

same notation as used in Chapter 3. In particular, the rate of service for the zeroth unit is (1/1 + 1/2)**-1 or 2/3. With PH{-} being the steady state probabilities for the hypercube, the hypothesis is that

$$\begin{aligned} PH\{00\} &= PC\{00\}, \\ PH\{01\} &= PC\{01\} + PC\{02\}, \\ PH\{10\} &= PC\{10\} + PC\{20\}, \\ PH\{11\} &= PC\{11\} + PC\{12\} + PC\{21\} + PC\{22\} \quad . \end{aligned}$$

For this particular example, the steady state probabilities for the hypercube are compared with those for the convolution model as aggregated above (see Table 4.1). The difficulty with this comparison, and several others which were made for more complicated systems, is one which often arises in numerical work. That is, when is the result of a complicated calculation zero? Can the differences shown in Table 4.1 be explained by roundoff errors? For a particular system of linear equations, this difficulty can be overcome by using all integer arithmetic (Ref. 20).

For the values given above, the exact solution procedure gives PH{00} to be 384/2983 or 0.128729 and PC{00} as 10824/84733 or 0.127742. The values are clearly different and hence Sevast'yanov's result cannot be extended to systems in which the unit service times are not identical. It should be noted that some effort was required to establish the counterexample. Other computational experience has indicated

Table 4.1 Numerical results for the example of Section B.3.

PH{00}	= 0.129
PC{00}	= 0.128
PH{01}	= 0.262
PC{01} + PC{02}	= 0.263
PH{10}	= 0.123
PC{10} + PC{20}	= 0.124
PH{11}	= 0.486
PC{11} + PC{12} + PC{21} + PC{22}	= 0.486

that it is difficult to distinguish the steady state values computed by the two models although they are, in general, different. We will make some use of this observation in Chapter 6.

B.4 An Example Using the Exponential Model

As with the convolution model, this model can be more easily understood by way of example. To this end, consider a two-server, two-call type example with CR{1}=a, CR{2}=b. The service rates are given by the 2 by 2 matrix RE with

$$(4.4) \qquad RE = \begin{bmatrix} c & d \\ e & f \end{bmatrix} \quad .$$

When unit 0 is the preferred server for call type 1 and unit 1 is the preferred server for call type 2, the state transition matrix is given by Figure 4.2. For example, state (02) corresponds to unit 0 busy on a call of type 2 and unit 1 free.

The equation of balance for state 20 is given by

$$(4.5) \qquad (a+b+f)*PE\{20\} = b*PE\{00\} + c*PE\{21\} + d*PE\{22\} \ .$$

This particular example can be used to show that the characterization of optimal assignment rules developed in Chapter 3 for the hypercube model does not hold for the

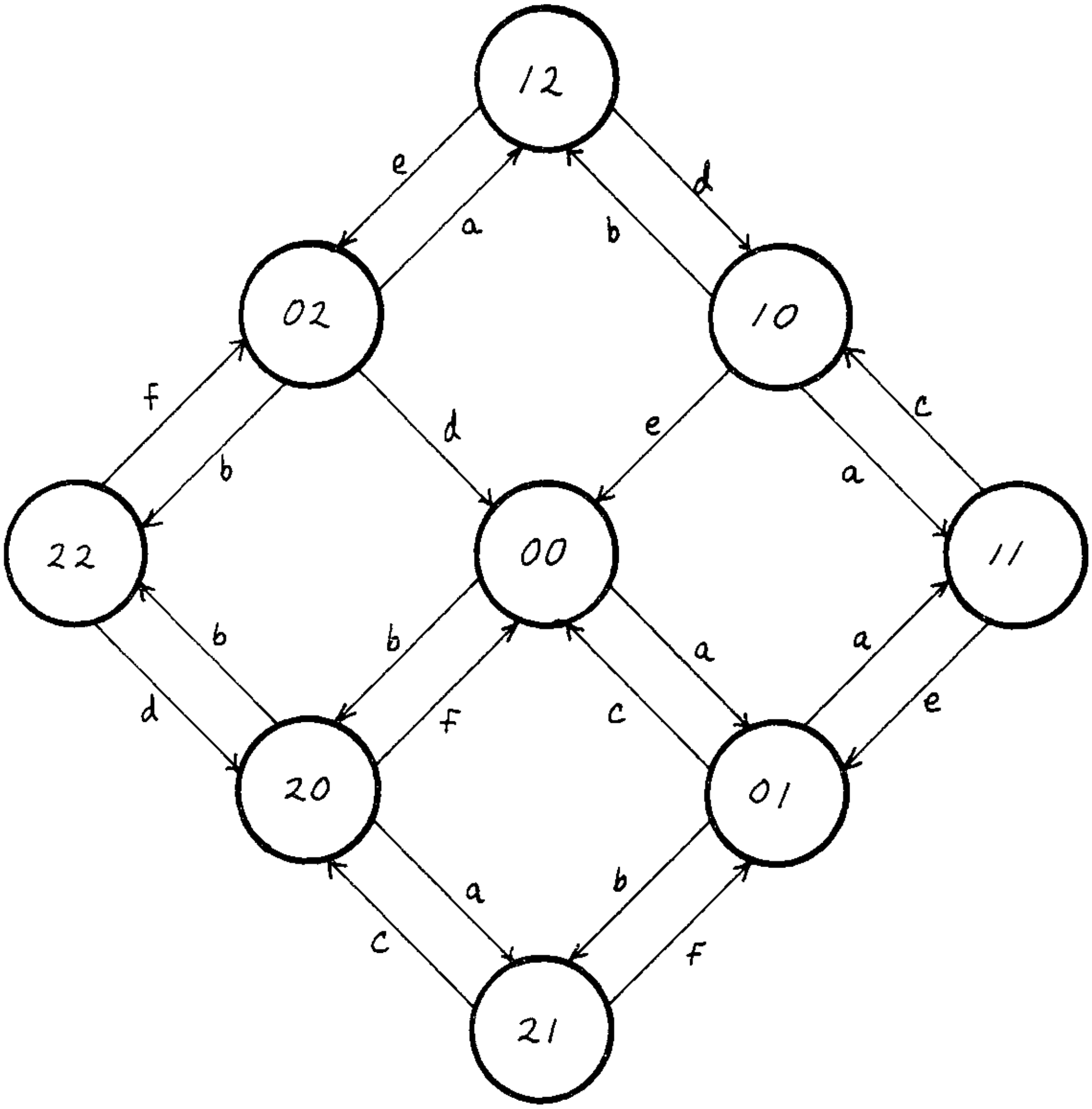


Figure 4.2. State transition diagram for the example of the exponential model in section B.3.

exponential model.

For the two-server case, the theorem states that if C{0,1} - C{1,1} is greater than C{0,2} - C{1,2}, then the optimal policy cannot have type 1 assigned to unit 0 and type 2 assigned to unit 1 in the same state. (Recall that C{i,j} is the cost of assigning server i to a call of type j). To provide the counterexample, consider the present example with a=b=1, d nearly zero, and c,e, and f arbitrarily large. Here, we specify C by

$$C = \begin{bmatrix} 0 & 1 \\ k & 1+2k \end{bmatrix} \tag{4.6}$$

where k is an arbitrary positive constant.

With the service rates as given above, if unit 0 is ever assigned to a call of type 2, then unit 1 will service almost all of the calls. Since CR{1}=CR{2}, the expected cost per call is approximately (k+1+2k)/2. Since the service rate for unit 1 is very high for either type of call, if unit 1 is the preferred unit for both types, we get the same approximate cost per call.

Now, if unit 1 is preferred for call type 2 and unit 0 is preferred for call type 1, since c and e are large, the approximate cost per call is (0+1+2k)/2; a smaller cost than that obtained with any of the other policies if k is greater than zero. The counterexample is completed by noting that

$C\{0,1\} - C\{1,1\} = -k$ is greater than $C\{0,2\} - C\{1,2\} = 1-(1+2k) = -2k$.

Although both the convolution and the exponential models can provide insight into the behavior of systems with distinguishable servers, neither is attractive from a computational point of view or as the general sort of model we would like to develop. In the next section, we derive a steady state result for the exponential model which has an intuitive appeal. We then show that the result holds in a more general setting.

C. The Equilibrium Equation

Recall from the discussion of performance measures in Chapter 3 that the fundamental variables were the fraction of calls of each call type handled by each server. For server i and call type j, this quantity was given by FSC{i,j} (fraction by server and customer), an N by NC matrix. Any performance measure in the hypercube model could be calculated from FSC, the call rates, and the service rates. The exponential model allows us to focus explicitly on these quantities as its state space includes information concerning the type of call being serviced.

C.1 A Derivation for the Exponential Model

Let BUS{i,j} denote the event that server i is busy on a call of type j. The probability of this event is given by

$$(4.7) \qquad \Pr\{\ BUS\{i,j\}\ \} = \sum_{k:VE\{k,i\}=j} PE\{k\} \ .$$

It is a straightforward, if somewhat tedious, problem to calculate the sum in equation (4.7) by summing (4.2) over the indicated states.

In equation (4.2), any CR{-}*PE{-} term will be referred to as an upward transition term and a product of the form RE{-,-}*PE{-} will be a downward transition term. For any upward term on the LHS of (4.2), we can identify the same term on the RHS in exactly one other state equation. Consider the CR{m}*PE{k} term from the LHS. Since we have an upward transition term, at least one element of VE{k,-} is zero. Let n be the preferred unit for call type m is state k. Then equation (4.2) for state k+m*(NC+1)**n is the only state equation which has a CR{m}*PE{k} term on the RHS. Thus, all upward terms from the LHS cancel with the same term on the RHS. (Note that the correspondence is unique).

Similarly, any upward term on the RHS cancels with a LHS term unless it is of the form CR{j}*PE{k} where VE{k,i}=0. No terms of this type appear on the LHS because of the conditioning in the sum of (4.7). Factoring out the CR{j}, we are left with a sum of state probabilities on the RHS which is exactly the probability that server i is assigned to a random call of type j.

The downward terms are treated in a similar fashion. For a term of the form RE{n,m}*PE{k} on the LHS (hence VE{k,n}=m), the same term appears on the RHS of (4.2) for state k-m*(NC+1)**n. The only exception is for terms with a multiplier of RE{i,j}. These terms do not appear on the RHS again because of the conditioning in the sum of (4.7). Recalling that FSC{i,j} is the probability that a call of type j is assigned to server i, we have

(4.8)   RE{i,j} * Pr{ BUS{i,j} } = CR{j} * FSC{i,j} .

In spite of all the manipulations leading to (4.8), the result is almost obvious. The probability that server i is busy on a call of type j is equal to the rate at which the server is assigned such calls, CR{j}*FSC{i,j}, times the expected service time per call, (RE{i,j})**-1. Not too suprisingly, this relation holds for a much wider class of systems than that described by the exponential model.

C.2 The General Service Time Problem

As noted above, equation (4.8) can be derived in a more general setting than the exponential model. We will focus on the event that server i is busy on a call of type j; that is, the event BUS{i,j}. This distinction is made because the results will be used in systems having distinguishable servers and call types. As before, CR{j} is the call rate for type j and FSC{i,j} is the probability that a random call of type j is answered by server i in the steady state.

In addition, define TSC{i,j} (service time by server and customer) as the expected time that server i spends in servicing a call of type j. We make the following assumptions regarding the arrival and service processes:

(i) The arrival process is independent of the state of the system as given by server availability.

(ii) The arrival process is an honest renewal process; that is, the successive times between arrivals of any particular type of customer are mutually independent and identically distributed and the successive interarrival times are finite with probability one. In addition, we assume that the distribution has a finite mean.

(iii) The service times are independent of the arrival times and the state of the system and have finite means and variances. The service time are small enough relative to the arrival rates that a steady state distribution exists for the system.

It is very important to keep in mind that these assumptions refer to an arbitrary server-customer pair.

The probability of the event BUS{i,j} in the steady state is just the expected fraction of time that server i spends on calls of type j over an infinite time horizon. This value can be calculated using the theory of renewal processes (Ref. 50) and the weak law of large numbers (Ref. 17).

Let the random variable X{m} be the time between the (m-1)-st and the m-th arrival of call type j. Define the random variable Y{m} as the time server i spends in servicing the m-th arrival of a call of type j. Note that in the steady state, $Y\{-\}$ is zero with probability $1-FSC\{i,j\}$ (corresponding to unit i not being assigned to service a random arrival of type j) and has mean $TSC\{i,j\}$ when conditioned on the event that unit i is assigned to service the m-th arrival. Finally, let $n\{t\}$ denote the number of arrivals of type j in the time interval from zero to t. With this notation, we can write

$$(4.9) \qquad Pr\{\ BUS\{i,j\}\ \} = E\left\{ \lim_{t\to\infty} \sum_{k=1}^{n\{t\}} Y(k)/t \right\} .$$

If the successive values of $Y\{-\}$ where independently distributed, it would be a simple matter to treat (4.9) using the theory of renewal reward processes (Ref. 50). Unfortunately, the $Y\{-\}$ are not independent. For example, if unit i is always assigned to an arrival of type j if that unit is available, then knowing that $Y\{m\}$ is zero implies that unit i was busy at the m-th arrival (if the service time has no impulse component at zero). For general service times, unit i is more likely to be busy at the (m+1)-st arrival of type j then it would be in the steady state and hence $Y\{m+1\}$ is more likely to be zero than a random arrival of type j.

The limit in (4.9) can be calculated using the weak law of large numbers (Ref. 17). We rewrite the limit as

$$(4.10) \qquad \lim_{t\to\infty} \sum_{k=1}^{n\{t\}} Y\{k\}/t \;=\; \lim_{t\to\infty} \sum_{k=1}^{n\{t\}} (Y\{k\}/n\{t\})*(n\{t\}/t).$$

Writing E{X} as the expected value of X{i} (identical for all i), n{t}/t converges to 1/E{X} for large t with probability 1 since the arrival process is a renewal process (Ref. 50). In addition, n{t} goes to infinity with probability one. Hence, we are left to calculate

$$(4.11) \qquad \lim_{M\to\infty} \sum_{k=1}^{M} Y\{k\} \;/\; M \quad .$$

Even though the Y{k} may not be independent, we can compute (4.11) by using a result from Feller (Ref. 17). That is, if the covariance between Y{k} and Y{m} converges uniformly to zero as the absolute value of k minus m goes to infinity and the Y{-} have finite variances, then the weak law of large numbers still holds for (4.11). Since all service times have finite means and an equilibrium distribution exists, with probability one the system will have all servers free infinitely often (Ref. 50). Since two values of Y{-} are independent if the event that all servers are free intervenes between the two associated arrivals of type j, we have the covariances equal to zero after a sufficiently large number of arrivals.

Applying the weak law of large numbers to (4.11), that quantity converges in measure (Ref. 18) to

$$(4.12)\qquad \lim_{M\to\infty} \sum_{k=1}^{M} E\{\ Y\{k\}\ \}\ /\ M$$

where E{Y{k}} is the expected value of Y{k}. In the limit as M goes to infinity, this expected value is just FSC{i,j} * TSC{i,j}; the probability that unit i services type j in the steady state multiplied by the conditional expected service time.

Collecting these results, we write (4.9) as

$$(4.13)\qquad \Pr\{\ BUS\{i,j\}\ \} = TSC\{i,j\} * FSC\{i,j\}\ /\ E\{X\}\quad .$$

In particular, if the arrival process is Poisson, E{X} is the reciprocal of the call rate and (4.13) becomes

$$(4.14)\qquad \Pr\{\ BUS\{i,j\}\ \} = CR\{j\} * TSC\{i,j\} * FSC\{i,j\}\quad .$$

For the remainder of this work, unless specifically stated otherwise, the arrival process will be Poisson and the conditions leading to (4.14) will be assumed to hold. Equation (4.14) will be used as the cornerstone for further analysis of queuing systems with distinguishable servers.

While this equation does not constitute a very deep theoretical result, its importance should not be underestimated. The conditions leading to its formulation are not stringent. As will be shown in succeeding chapters, this result will allow us to deal with quite general systems.

Chapter 5. OPTIMAL FACILITY LOCATION

A. Introduction: Models for Location

As mentioned at the end of Chapter 3, the optimization technique developed for the hypercube model does not really address the central issue in the allocation of emergency service units. Instead of optimizing dispatch rules for a given set of server locations, we would like to determine the optimal initial location of such units. As noted in Chapter 2, there has been a considerable amount of research in the area of facility location. Although some of the concepts are relevant to the location of emergency facilities, the models are generally inappropriate for our particular problem.

A.1 Economic Models

A good deal of the literature on facility location deals with questions of warehouse location and transportation problems, hence the term "economic models." The survey papers by Cooper (Ref. 10) and Revelle, Marks, and Liebman (Ref. 49) reflect this emphasis. Typically, the problem is to determine the most economical configuration of sources of supply required to meet a specified distribution of demand. Included are Cooper's "generalized Weber problem" and the special techniques for networks discussed by Revelle, Marks, and Liebman. Costs are usually expressed in terms of time or distance and there may be capacity constraints which can force some interaction among the sources.

The difficulty with using these approaches for the location of emergency service units follows from the deterministic assumptions underlying the model formulations. The location and level of the demand is assumed known and the sources are always available to provide services or supplies. These comments notwithstanding, some of the techniques have been applied to the provision of emergency services as noted below.

A.2 Models for Emergency Services

The main problem in applying the deterministic location models to emergency services is an inadequate description of the cooperation between the units. As noted in Chapter 3, simple rules for the assignment of servers to customers can be followed only approximately because of the unavailability of servers. The variability in demand, service times, and unit availability all combine to produce complicated interactions between the units. However, if the arrival rate of calls for service is sufficiently low, there may be little or no interaction between the units. In this case, either the network center or median problems as described by Odoni (Ref. 48) or the set covering formulation used by Jarvis, Stevenson, and Willemain (Ref. 31) may be appropriate for solving a location problem.

When the interaction among the servers is important, descriptive models (such as the hypercube) can be used for facility location on a trial and error basis. In fact, it is exactly this sort of approach which was discussed in Chapter 3. An initial guess is made as to an appropriate system configuration and that configuration is evaluated by the hypercube model. The user then tries certain system changes in an attempt to get the desired levels of performance. With a little practice and some understanding of the assumptions underlying the model, it is not difficult to effect particular changes at least on a semi-quantitative basis. The main difficulty with this procedure is that it requires some level of expertise and could be expected to become increasingly complicated for larger systems.

In recognition of this problem, Chelst (Ref. 7) has developed computer algorithms for balancing workloads and travel distances using the hypercube model. It is this sort of approach that we take here. The objective is to develop an algorithmic procedure for facility location which incorporates the flexibility of the hypercube model into the description of the system.

B. Optimal Location of Response Units

The deterministic location models mentioned above can be useful in the location of emergency units, particularly under conditions of negligible congestion. When the interaction among the servers is important, we would like to incorporate that information into the decision as to the location of the units.

For example, Table 5.1 contains some additional statistics from the three server example of Chapter 3, Section D. Using the FSC(i,j) and the fraction of calls from each atom, we can calculate the fraction of calls which are serviced by other than the first preferred (closest) unit. Overall, slightly more than one third of the calls for service must be handled by the second or third preferred response unit because the closest unit is unavailable. (Under the assumptions of Chapter 4 concerning the arrival of calls for service, if calls of type j are assigned to unit i when it is available, then the workload of unit i, WL(i), is precisely the fraction of type j calls which can be expected to arrive when unit i is busy. These calls must be serviced by another unit. See Ref. 38). As shown in Table 5.1.B, over sixty percent of the responses by unit 1 were to calls which would have been serviced by another unit had that other unit been available.

This kind of information can be included in the decision as to where to locate unit 1. With these response patterns, the average distance traveled by unit 1 in response to a call for service is 9.94 if the unit is stationed in atom 11. If, however, the unit were stationed in atom 8, its travel distance <u>under the same response pattern</u> would fall to 6.78, a

Table 5.1 The extent of inter-unit cooperation in the example from Chapter 3, Section D.

A. Fraction of calls from atom j which are serviced by unit i (FSC{i,j}) and the fraction of calls coming from each atom (CR{j}/CRT):

Atom j:	FSC{0,j}	FSC{1,j}	FSC{2,j}	CR{j}/CRT
1	0.526*	0.287	0.092	0.143
2	0.526*	0.287	0.092	0.114
3	0.526*	0.287	0.092	0.067
4	0.526*	0.287	0.092	0.076
5	0.526*	0.287	0.092	0.095
6	0.526*	0.287	0.092	0.076
7	0.526*	0.287	0.092	0.076
8	0.120	0.694*	0.092	0.057
9	0.048	0.694*	0.164	0.029
10	0.048	0.694*	0.164	0.048
11	0.048	0.694*	0.164	0.029
12	0.048	0.694*	0.164	0.019
13	0.048	0.694*	0.164	0.010
14	0.048	0.197	0.661*	0.019
15	0.048	0.197	0.661*	0.048
16	0.048	0.197	0.661*	0.095

* denotes FSC{i,j} for the unit closest to atom j.

B. Fraction of calls serviced by each unit for which it is not the first preferred unit.

Unit 0:	0.0575
Unit 1:	0.6201
Unit 2:	0.4463
Overall:	0.3581

reduction of roughly thirty percent. It is this sort of improvement that we would like to make in the location of the units. If we are given the level of cooperation between the units; in particular, a description of the frequency with which a unit is dispatched to different atoms; we can locate each unit to best reflect its particular mix of responses.

For the moment, we assume that we are given the FSC{i,j}. These could have been determined by the hypercube model, simulation, or historical data. Whatever the source, we can use these quantities to determine the location of units which best anticipates their overall usage. It is crucial to note that these locations will reflect the same assignment rule which was used to determined the exact extent of the server interaction; that is, the FSC{i,j}.

B.1 A Location Model including Server Cooperation

Rather than formulate the location model in terms of travel time or distance, we use the general cost structure developed for the hypercube model. (Note, however, that we are using the costs in the context of spatially distributed systems). In particular, we assume that the cost of service, formerly denoted by C{-,-}, can also depend on the position of the responding unit prior to its dispatch. (We still assume that the unit returns to its initial position after completion of any service).

For the three server example of Chapter 3, each server could be in any one of 16 distinct positions, each corresponding to an initial location in one of the 16 atoms. In general, we assume there are P possible positions, and let UP{i,p} (unit position) denote the probability that unit i is in position p. By allowing a probabilistic description of the position of each server, we can include such situations as police preventive patrol. For spatially distributed systems, we can model either fixed location or mobile units where the UP{i,p} detail the frequency with which each unit occupies a particular location.

The cost of assigning unit i from position p to a call of type j will be denoted by CP{i,j,p} (cost of assignment with unit positions). The expected cost of assignment can be written in much the same way as equation (3.10) by conditioning on the type of call, the unit providing service, and now the additional information concerning the position of the unit. In a fashion similar to that used in describing the position of mobile patrol units in the police context, the probabilistic description of unit position will be assumed independent of the state of the system or the arrivals of calls for service. However, as will be shown below, the actual mix of locations chosen for a particular server will depend on the call rates and the unit's response distribution.

Writing EC as the expected cost per call we have

$$\text{(5.1)} \qquad EC = \sum_{i=0}^{N-1} \sum_{j=1}^{NC} \sum_{p=1}^{P} CP\{i,j,p\} * Pr\{i,j,p\}$$

$$+ \; Pr\{S\} * \sum_{j=1}^{NC} CS\{j\}*CR\{j\}/CRT \quad ,$$

where Pr{i,j,p} is the probability that a call chosen at random is of type j and answered by unit i from position p; S is the event that all servers are busy when a call arrives, hence Pr{S} is the saturation probability; and CS{j} is the cost associated with an arrival of a type j call during a period of saturation.

To simplify the notation, define FC{j} (fraction of calls from each class) by

$$\text{(5.2)} \qquad FC\{j\} = CR\{j\} \,/\, CRT \;, \qquad\qquad \text{for } j=1,2,...,NC.$$

(Recall that calls are assumed to arrive according to a Poisson process). Since the position of a server is probabilistically independent of the arrival process, we can write

$$\text{(5.3)} \qquad Pr\{i,j,p\} = UP\{i,p\} * FC\{j\} * FSC\{i,j\} \quad .$$

That is, the probability that a random call is of type j and answered by unit i from position p is the probability that unit i is in position p times the probability that a random call is of type j times the probability that server i responds to a call of type j. It is crucial to note that we are assuming that we know FSC{i,j} for all i and j. That is, the extent of inter-server cooperation has been determined (perhaps by the hypercube model) and we want to choose the positions of the servers in order to reflect that interaction.

Again, to further simplify the equations (at the risk of complicating the notation), define FT{i,j} by

$$FT\{i,j\} = FC\{j\} * FSC\{i,j\} \tag{5.4}$$

$$\text{for } i=0,1,\ldots,N-1;\ j=1,2,\ldots,NC.$$

FT{i,j} (fraction of total service) is the probability that a call chosen at random will be of type j and answered by server i. We now rewrite (5.1) as

$$EC = \sum_{i=0}^{N-1} \sum_{p=1}^{P} UP\{i,p\} * \sum_{j=1}^{NC} CP\{i,j,p\} * FT\{i,j\} + Pr\{S\} * \sum_{j=1}^{NC} FC\{j\} * CS\{j\} \quad . \tag{5.5}$$

It is now evident that the expected cost per call is linear in the UP{i,p}. By adding normalization constraints for each server and defining CO{i,p} as the <u>co</u>efficient of UP{i,p} in (5.5), the choice of the position for each server can be expressed as a linear programming problem which seeks to minimize the expected cost per assignment.

Since the saturation term in (5.5) is constant with respect to changes in the UP{i,p} for a fixed dispatch rule and service times, it can be ignored in the objective function of the linear program. As noted below, if a change in the dispatch strategy is indicated by a change in the location of the units, it is possible for the saturation probability to change. (See Chapter 3). It is possible that the change in saturation probability could result in a larger cost per call than the original configuration. For practical problems, the probability of all units being simultaneously busy can be expected to be of the order of a few percent. Between reasonable dispatch schemes, the saturation probability varies only slightly and system performance would be worsened only if saturation costs were much larger than normal assignment costs. Of course, if a change in positions resulted in worsened performance, the original configuration should be used.

The linear program derived from (5.5) is written

$$(5.6) \qquad \text{Minimize} \quad z = \sum_{i=0}^{N-1} \sum_{p=1}^{P} UP\{i,p\} * CO\{i,p\}$$

$$\text{Subject to} \quad \sum_{p=1}^{P} UP\{i,p\} = 1 \qquad \text{for } i=0,1,\ldots,N-1,$$

$$UP\{i,p\} \geq 0 \quad .$$

Since the constraint equations in (5.6) include no terms of the form UP{i,p} + UP{j,r} for i not equal to j, the linear program can be solved by considering each unit separately. In this case, the optimal solution for unit i is obtained by setting UP{i,p} equal to zero except for UP{i,$\underline{p}$} equal to one, where CO{i,$\underline{p}$} is the minimum of the CO{i,p} over all positions p.

This was exactly the procedure which was applied to the three server example of Chapter 3. For this problem, the CP{i,j,p} terms depend only on j and p: the location of the call for service and the initial position of the responding unit. For the response pattern given by Table 5.1, the travel distance per call for service is minimized by moving unit 0 from atom 1 to atom 5, unit 1 from atom 11 to atom 8, and unit 2 from atom 16 to atom 14. (See Figure 3.2).

With these changes in location, the average travel distance for unit 0 drops from 6.52 to 4.88; for unit 1, from 9.94 to 6.78; and for unit 2, from 10.38 to 8.70. The global average travel distance, incorporating the saturation costs, changes from 8.73 to 6.70; an improvement of over 23 percent. This should be compared with the 0.66 percent improvement resulting from the use of the optimal dispatch rules for the original unit locations when compared to closest available unit dispatching.

To this point, we have assumed that we are given the extent of inter-server cooperation for a particular dispatch rule (in the previous example, through the use of the hypercube model). We have used this information to reposition units in order to best anticipate their usage under that same dispatch rule. The crucial observation is that a change in the location of the units can lead to a change in the dispatch strategy.

For example, from the results of Chapter 3, we know that closest available unit dispatching rule comes very close to achieving the minimum expected travel distance which would be obtained from the optimal assignment strategy. If the application of the location model results in a change in unit positions, then we can expect to further reduce travel distance by determining the closest available unit assignment using the new unit positions. We detail this iterative improvement technique in the next section.

B.2 An Iterative Procedure for Facility Location

The procedure for determining the optimal location of response units is given by the flow chart of Figure 5.1. A reasonable choice is made as to an initial set of locations for a fixed number of units. We determine an appropriate dispatch strategy based on these initial locations and then apply some descriptive model (perhaps the hypercube model) to determine the extent of inter-unit cooperation. This cooperation is expressed in terms of the FSC{i,j}.

Using the FSC{i,j} and the CP{i,j,p} (the costs of assignment conditioned on the type of call, the servicing unit, and the unit's initial location), we relocate the units whenever such a relocation will result in a decreased average cost per call. If the position of the units does change, we use their new locations to determine a new dispatch strategy. If this strategy is different from the one used for the previous iteration, we use the descriptive model to determine new FSC{i,j} and apply the location model again. This procedure is used iteratively until there is no change in either unit positions or dispatch strategy between two

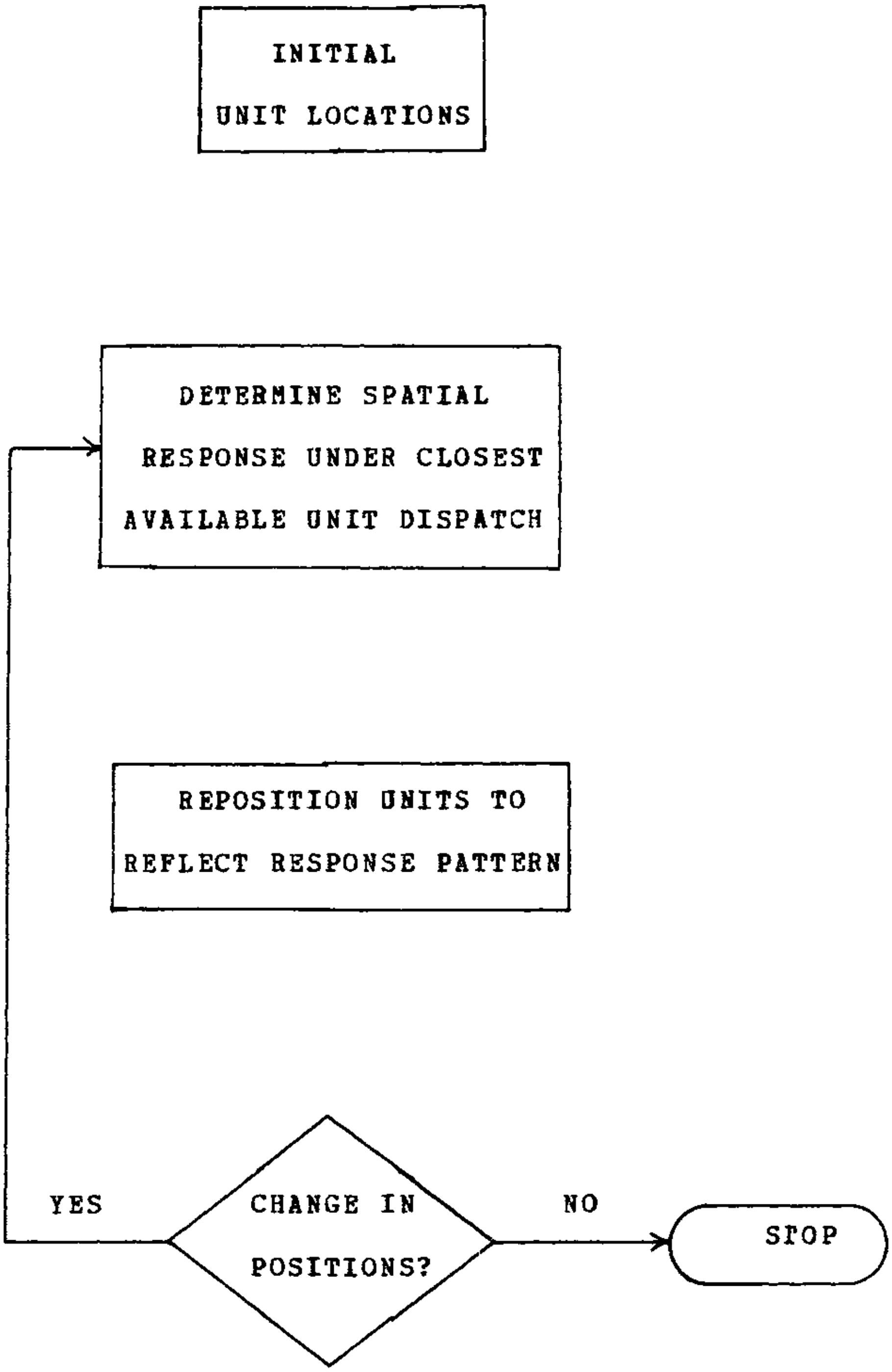


Figure 5.1 Flow chart for an iterative procedure for determining the optimal location of response units under conditions of congestion and cooperation among the servers.

successive iterations.

We illustrate the procedure by determining the optimal location for four response units to be positioned within a grid of 100 atoms. Figure 5.2 shows the atom locations and gives the initial position of the units at the points (0,0), (0,3), (3,0), and (3,3) of the grid.

The hypercube model is used to determine the FSC{i,j}. The cost of assigning a unit at node (x,y) to a call from node (w,z) is given by the absolute value of x minus w plus the absolute value of y minus z; that is, the right angle distance between the two points. (We assume Cartesian coordinates and unit distance between the points of the grid). Finally, the call rates for all nodes (atoms) are assumed identical and very small compared to the service rate for each server (also assumed to be identical). These last assumptions are equivalent to assuming neglible cooperation among the servers.

We determine the assignment rule on the basis of the closest available unit. (We allow ties in the choice of assignment, corresponding to a randomized dispatch rule (Ref. 36), in order to preserve the symmetry of the problem).

Figure 5.3 traces the iteration as it utilizes the hypercube model and then the location model. For each iteration, there are two numbers; the first gives the average travel distance for the system with the indicated unit locations and closest available unit dispatching. The second number is the average response distance when the units are

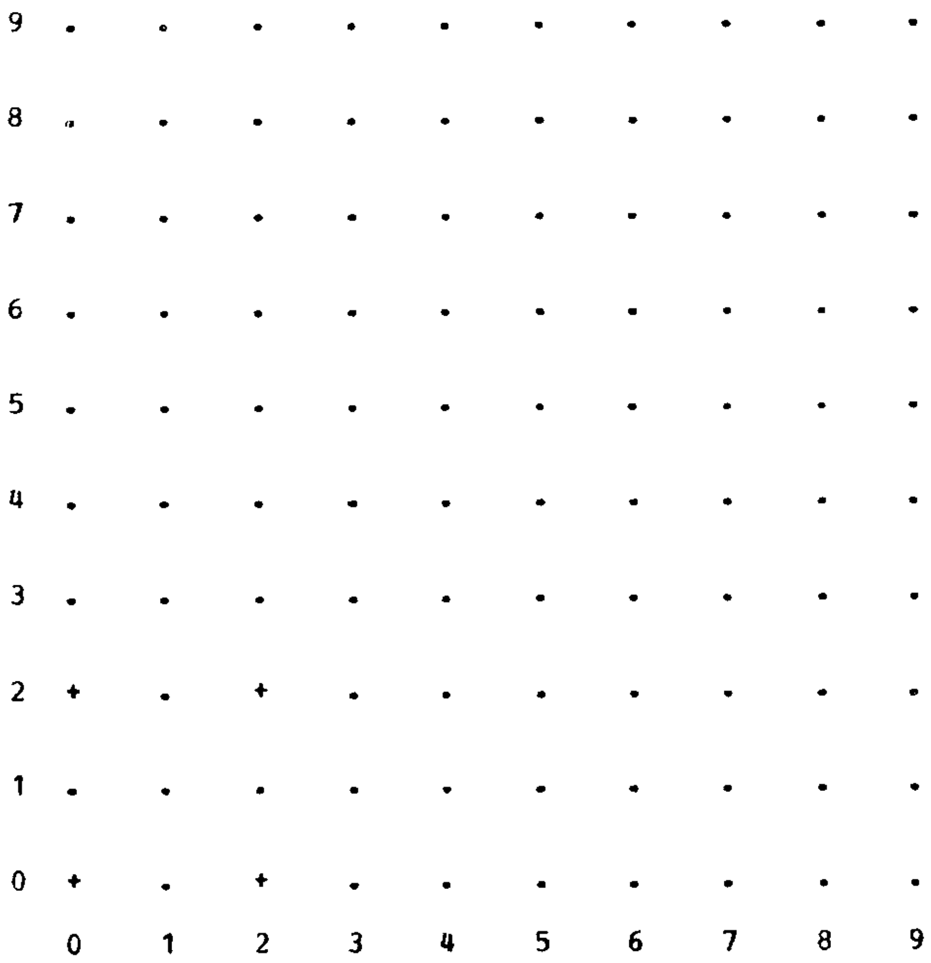
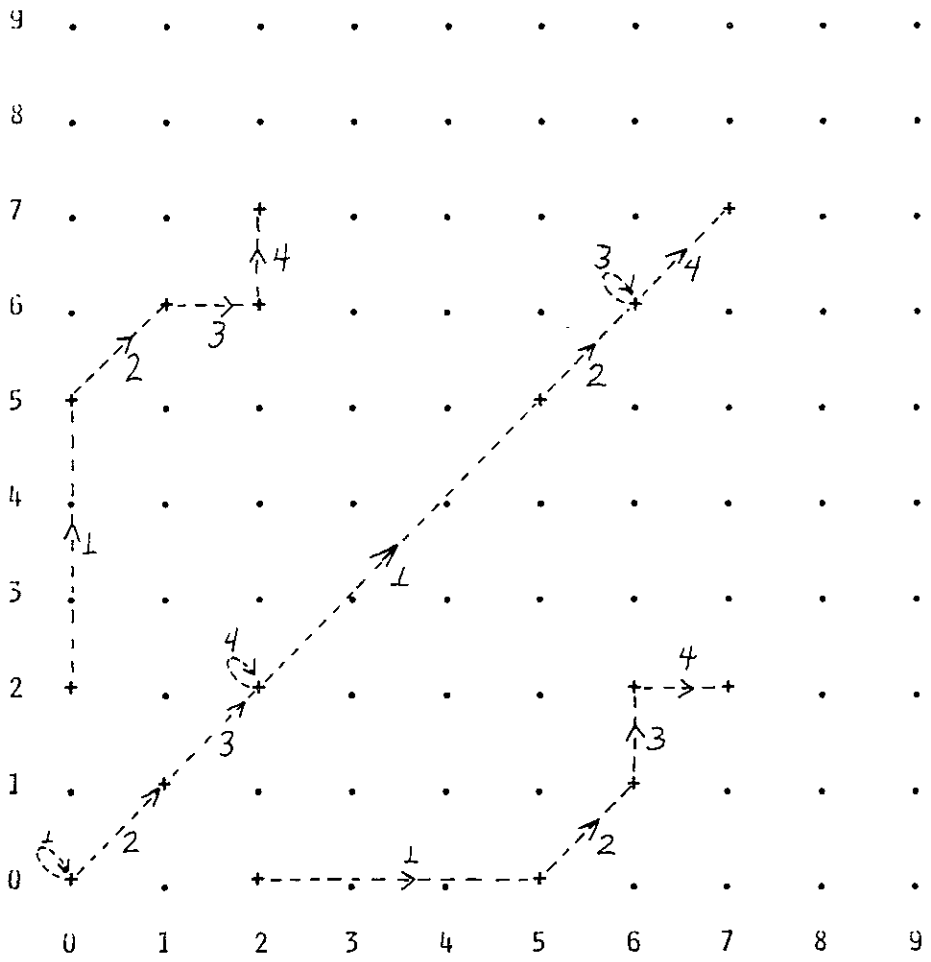


Figure 5.2 Geographic description of a four-server, 100-atom system to illustrate the facility location iteration. Locations are specified in terms of Cartesian coordinates. Initial positions: (0,0), (0,2), (2,0), and (2,2) denoted +.



ITERATION 1:	5.8	-->	3.7
2:	3.2	-->	2.8
3:	2.6	-->	2.6^-
4:	2.6^-	-->	2.5
5:	2.4	-->	2.4

Figure 5.3 Results of successive applications of the location and hypercube model to the system given in Figure 5.2. The diagram shows the successive movements of each unit (+) at each iteration. For each iteration, the average travel distances for the current dispatch policy are shown before and after the units are relocated.

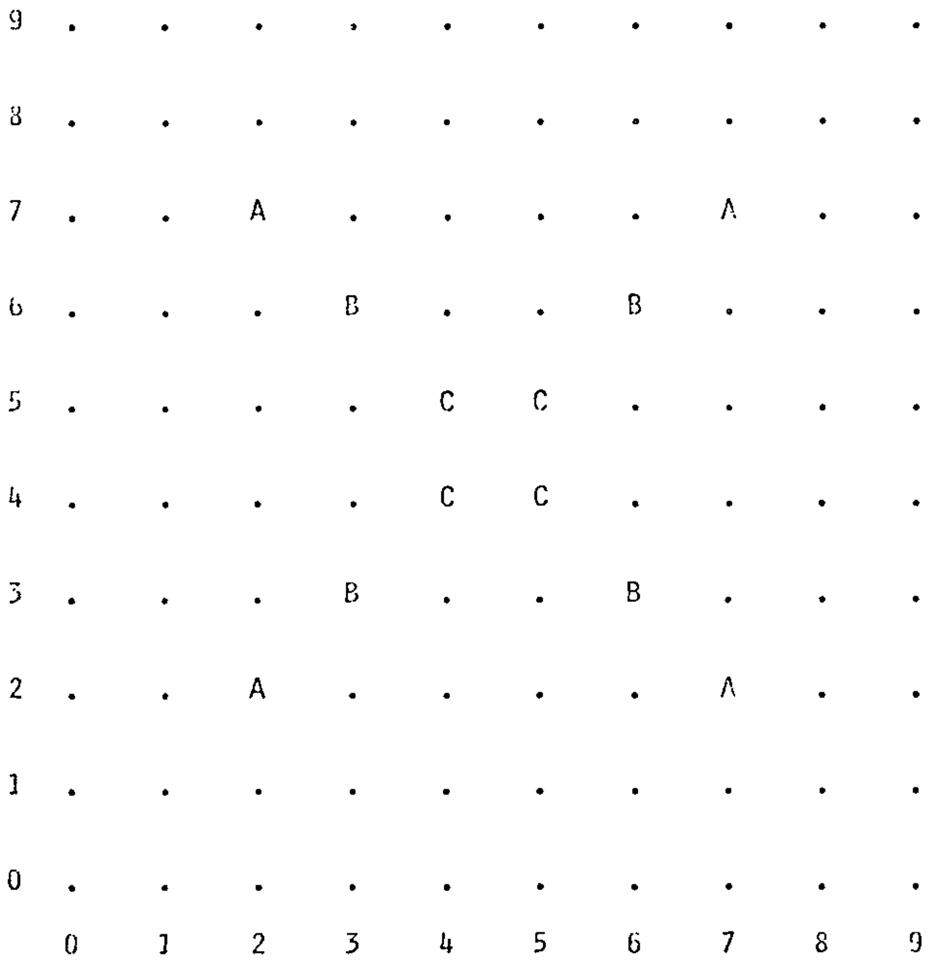
relocated to best reflect their spatial assignments (still under the same dispatch rule).

For example, with the initial locations (iteration 1), the average response distance is 5.80. This can be decreased to 3.70 by respositioning the four units at (0,0), (5,0), (0,5), and (5,5). (Note that one unit does not move). Figure 5.3 indicates that these new unit locations result in a further decrease of response distance to 3.20 when the dispatch policy is changed to reflect the new positions.

After four iterations, the units reach a stable configuration and no further improvement can be made. The optimal locations are exactly.those which would be expected by symmetry arguments for the situation in which the workload of each unit is close to zero.

The same iteration scheme can be used to relocate servers in order to reflect an increased workload. For the same four server system as above, suppose we increase the average workload to 0.5. The average travel distance becomes 4.45 units. The location model moves each unit one step closer to the center of the area with a resultant decrease in the travel distance to 4.18 units. (See Figure 5.4). When the average workload is increased to 0.9, the optimal positions for the units move another step toward the center. The average travel distance decreases from 5.18 to 4.92 with this move.

In this instance, the dispatch strategy remains the same because of the symmetry of the region. The units are relocated



9	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.	.
7	.	.	A	.	.	.	.	A	.	.
6	.	.	.	B	.	.	B	.	.	.
5	.	.	.	.	C	C	.	.	.	.
4	.	.	.	.	C	C	.	.	.	.
3	.	.	.	B	.	.	B	.	.	.
2	.	.	A	.	.	.	.	A	.	.
1	.	.	.	.	.	.	.	.	.	.
0	.	.	.	.	.	.	.	.	.	.
	0	1	2	3	4	5	6	7	8	9

Figure 5.4 Use of the location model to reflect increased unit workloads. The units move toward the center of the region in order to anticipate the increasingly likely event of dispatch uniformly distributed over the region (in the limit as workloads approach unity). Optimal locations are shown for an average workload of 0.0 (A), 0.5 (B), and 0.9 (C).

to reflect the increased probability of a dispatch to a node outside of the unit's own "quadrant." As expected, as the dispatches for each unit become uniform over the entire region, the optimal location of the units shifts toward the center of the area.

The procedure described above focuses only on minimizing the cost of assignment. In some situations, additional factors may be important in the location of facilities. We consider two examples taken from police and emergency medical services.

C. Constraints on Unit Location

The model for improving unit locations given above has two features which may not be desirable in certain circumstances. The objective is the minimization of the expected cost per assignment. As noted in Section A of this chapter, we might like to include some constraints on maximum response time for spatially distributed systems. In addition, the form of the linear program (5.6) results in integer solutions for the decision variables. In the context of police preventive patrol, a fractional solution, corresponding to random mobile patrol, would be preferred. Both of these difficulties can be treated in the same programming framework.

C.1 Limiting Maximum Travel Time

Suppose there is an "acceptable" level of cost associated with each type of call. We want to choose the position of units in order to guarantee at least a minimum level of acceptable responses for each type of call. Define MAC{j} (maximum acceptable cost) as the largest cost of assignment that is acceptable for a customer of type j. Let MAF{j} (minimum acceptable fraction) be the smallest fraction of acceptable responses to type j calls which we will tolerate.

For example, the EMSS Act (PL93-154) specifies (as one criterion for acceptable service in urban areas) that 95 percent of all calls receive a response in ten minutes or less. In our notation, we would set MAC{m} identically equal to 10 minutes and MAF{j} identically equal to 0.95. The costs of assignment are expressed in terms of response time (for the loss system, there are no queuing delays) and we seek to determine locations for ambulances in order to minimize response time subject to the constraints on maximum response time.

We can obtain an acceptable level of response for all types of customers by adding the following constraints to the linear program given by (5.6):

$$(5.7) \qquad \sum_{i=0}^{N-1} \sum_{p=1}^{P} FA\{i,j,p\}*UP\{i,p\} \geq MAF\{j\}$$

for j=1,2,...,NC,

where FA{i,j,p} is the fraction of acceptable responses to a call of type j by unit i in position p. FA{i,j,p} is given by FSC{i,j} if CP{i,j,k} is less than or equal to MAC{j}; otherwise, it is zero.

As noted above, for the particular case of ambulance location, the acceptability of a response might be defined in terms of a maximum travel time. The addition of (5.7) to the linear program forces some equity in service among calls from different locations. It also complicates the solution procedure. The UP{i,p} variables are now coupled and the optimal solution of the linear program can not be expected to be integer. Since non-integer solutions are associated with mobile units, the linear program is modified to an integer linear program by requiring that UP{i,p} be zero or one (corresponding to fixed location units).

Instead of the NC local constraints given by (5.7), we can use a single global constraint by setting a lower bound, K, on the overall fraction of acceptable responses. The constraint is written

$$(5.8) \qquad \sum_{j=1}^{NC} \sum_{i=0}^{N-1} \sum_{p=1}^{P} FC\{j\}*FA\{i,j,p\}*UP\{i,p\} \geq K$$

We will have more to say about the actual use of these constraints in Chapter 7.

In the design of police preventive patrol, fractional solutions to the linear program are actually desirable.

C.2 Police Preventive Patrol

The integer solutions obtained from (5.6) do not reflect the other function of the police service; that is, preventive patrol. We would like to add constraints which provide for preventive patrol while reflecting the response to calls for service. (It is interesting to note that the integer solution to (5.6) implies that response time is minimized by fixed rather than mobile units. See Ref. 41). For the remainder of this section, we focus on mobile units.

Since the workload of unit i, WL{i}, is defined as the fraction of time spent in servicing calls, unit i has the remainder of its time, a fraction 1-WL{i}, for preventive patrol. When UP{i,p} is interpreted as the fraction of preventive patrol time spent in atom p, we would like to specify the fraction of preventive patrol effort to be allocated to atom j. For example, we might want the effort to be roughly proportional to the fraction of calls originating from atom j.

The total preventive patrol effort available from all of the units is the number of units, N, times one minus the average unit workload. Allowing some slack in the preventive patrol required for each atom (for example, seeking a minimum of 100*X percent of the target level), preventive patrol can be specified by the constraints

$$(5.9) \qquad \sum_{i=1}^{N-1} (1-\mathrm{WL}\{i\}) * \mathrm{UP}\{i,p\} \geq X * \mathrm{PC}\{p\} * N * (1-\mathrm{AWL})$$

$$\text{for } p=1,2,...,NC,$$

where AWL is the average workload of the units. (If (5.9) is used with an equality constraint and no flexibility in the level of preventive patrol effort, the patrol allocations are usually unacceptable. A typical solution will divide the preventive patrol of a unit between widely separated atoms or allocate a miniscule amount of its patrol to several disjoint atoms).

This constraint can result in overlapping areas of preventive patrol. This condition can be avoided by adding 0-1 constraints which allow only one unit to patrol any given reporting area. The result is a mixed integer linear program. Needless to say, the degree of difficulty in solving (5.6) increases as further constraints are added.

In addition to constraints on preventive patrol, the equity constraints from the preceding section could also be incorporated. The intent here is not to delineate all possible constraints, but to indicate the degree of flexibility of this model.

D. Summary

The model developed in this chapter allows the position of units to be chosen in a manner which reflects the level of cooperation between the units. This interaction is specified by the FSC(i,j), which were assumed to be given. In that context, these values for a three server example were obtained by applying the hypercube model to an initial set of unit locations.

The optimal location for the units is determined by using the location model alternately with a descriptive model (in these examples, the hypercube model), to choose successively improved unit positions and dispatch strategies. To this point, we have avoided computational difficulties such as local minima or the failure of the iteration to converge to a stable configuration.

Since there are only a finite number of possible unit locations, requiring strict improvement in the average cost per call at each stage guarantees that the iteration will converge. It is possible that convergence will be to a local minimum, as observed in using some deterministic location models (Ref. 6). An obvious procedure for avoiding this difficulty is to solve a particular problem with several different initial configurations. Convergence to local minima has not been a problem in the author's experience except in special circumstances.

For example, if the iteration is applied to the four server example of Section B.2 with a total call rate of zero (WL{i} identically zero, no interaction among the units), a non-optimal local minimum is reached after two iterations. This is due in part to the extreme symmetry of the example. Even in this case, if any unit workload is greater than zero, the local minimum disappears. This symmetry is not typical of real geographies.

For all of the examples above, the hypercube model has been used to incorporate the spatial characteristics of the system and to determine the FSC{i,j}. In the next chapter, we present an approximation procedure for loss systems with distinguishable servers which incorporates the flexibility of the hypercube model in a framework which allows for unit and customer specific service times.

Chapter 6. AN APPROXIMATE ANALYSIS OF THE GENERAL SERVICE TIME MODEL

A. Introduction

Although the hypercube model is quite comprehensive for the analysis of queuing systems with distinguishable servers and classes of customers, there are two important situations which can make its use inappropriate. The first of these concerns the computational effort involved; the second has to do with the service time assumptions of the model.

The computational difficulty becomes apparent after an examination of the "size" of the hypercube model, where size is taken to mean computer storage and time requirements. The earliest implementation of the hypercube model could accomodate at most six servers because of numerical difficulties (Ref. 3). As noted in Chapter 3, Larson increased this number to fifteen by exploiting the structure of the model (Ref. 36). Even so, a system with 10 servers and 100 classes of customers, not unusual for police applications in urban areas, requires roughly 360 thousand bytes of storage in the implementation described by Jarvis (Ref. 27). (Recall that calls can be classified on the basis of location for spatially distributed systems).

The model assumption that is troublesome concerns the service time distribution. As formulated, the hypercube only allows server dependent service times. As noted above, such a description would appear inadequate for systems in which the major component of the service time is travel. For these systems, we would like to be able to specify service time as a function of both the server and the type (location) of customer. In addition, the exponential form of the service time may not be appropriate for some systems.

The first problem has been solved by Larson for the case in which the service times for each unit are identically distributed. The method of solution employs an approximation procedure which avoids the large state space associated with the hypercube and which appears to give the same results (within a few percent error) as the exact model (Ref. 40).

The exponential model described in Chapter 4 allows server and customer dependent service times. Unfortunately, the size of the model is even more restrictive here than with the hypercube. For example, the three server example for Sample City would have roughly sixty-five thousand states (the hypercube only 8). Again there is the additional problem that the service times are exponentially distributed. This distribution was chosen because of its analytic tractability, not because it represented situations often found in real problems.

In this chapter, we give an approximation procedure which is aimed at solving both of the problems mentioned above. The procedure is based in large part on the approximation procedure given by Larson. In fact, the actual approximation is almost exactly that used by Larson. The contribution of this work is to apply that approximation technique to a wider class of systems.

B. An Approximation Procedure

The steady state equation derived in Section C.2 of Chapter 4 forms the basis for the development of the approximation procedure. Before describing the procedure itself, we review the relevant notation and model assumptions.

B.1 Notation and Assumptions

The assumptions and notation used here are consistent with that developed in the previous three chapters. We restrict our attention to the loss system. Calls arriving during periods of saturation are either lost or handled by means external to the system under consideration.

The system consists of N servers, indexed from 0 to N-1, and NC classes of customer, indexed from 1 to NC. Arrivals of customers of type j are distributed according to a Poisson process with rate CR{j}, independent of all other classes of customer and the state of the system. The total rate at which customers arrive is denoted by CRT. If server i is assigned to a call of type j, the service time is independent of the state of the system or the time of the arrival. The expected service time is finite and is given by TSC{i,j}.

If FSC{i,j} is the probability that a random call of type j is serviced by unit i, and BUS{i,j} is the event that server i is busy with a call of type j, then the long term probability of this event is given by equation (4.11), restated here as

$$(6.1) \qquad \Pr\{ \text{BUS}\{i,j\} \} = \text{CR}\{j\} * \text{TSC}\{i,j\} * \text{FSC}\{i,j\} \quad .$$

Since WL{i}, the workload of server i, is just the probability that server i is busy in the steady state operation of the system, we can write an expression for the workload by conditioning on the class of customer being serviced. That is,

$$(6.2) \qquad \text{WL}\{i\} = \sum_{j=1}^{NC} \Pr\{ \text{BUS}\{i,j\} \}$$

$$= \sum_{j=1}^{NC} \text{CR}\{j\} * \text{TSC}\{i,j\} * \text{FSC}\{i,j\} \quad .$$

B.2 Fixed Preference Dispatch Rules

If we know the FSC{i,j}, equation (6.2) can be used to determine the unit workloads. In fact, all of the system performance measures for the loss system can be written in terms of the call rates, the service times, and the FSC terms.

(A detailed list of performance measures can be found in Reference 36. The costs associated with system operation are given in equations (5.2) and (5.4) of Chapter 5. It should be noted that the saturation probability, Pr{S}, is given by one minus the sum of FSC{i,j} over j for any i).

We would like to relate the FSC terms to the workloads of the units and then use equation (6.2) to solve for the workloads and FSC{i,j}. This was the procedure used by Larson for "fixed preference" dispatch or assignment rules.

A fixed preference assignment rule is a state independent strategy. For each type of customer, the units are ordered on the basis of their relative merit or cost with respect to servicing a call of type j. When a call of type j arrives, we dispatch the first preferred server for type j if it is available; otherwise we choose the second preferred unit; if that unit is unavailable, the third preferred; and so on. An example of such a fixed preference rule is given by the three server system used in Chapter 3. The rule used in that case was to dispatch the closest available server.

A fixed preference rule will be specified by an NC by N matrix, DP (dispatch preference), where DP{j,k} is the k-th preferred server for a call of type j. A rule which will often be used is the myopic optimum. In this case, DP{j,1} achieves the minimum in

$$(6.3) \qquad \mathrm{Min}_{i} \sum_{p=1}^{P} UP\{i,p\} * CP\{i,j,p\} \quad ,$$

where UP{i,p} is the probability that server i is assigned from position p and CP{i,j,p} is the associated cost of assignment. DP{j,2} achieves the minimum in (6.3) when i equal to DP{j,1} is excluded. The remainder of the DP{j,-} are determined in a similar fashion.

The FSC{i,j} take a very simple form for fixed preference dispatch rules. Suppose that server i is the k-th preferred server for calls of type j. Then the probability that i is assigned to a random call of type j is the joint probability that the first k-1 preferred servers for type j are all busy and that server i is free. Let B{i} denote the event that server i is busy. F{i} will be the complementary event; that is, server i free. Since the arrival process is independent of the state of the system, the workload of server i is given by

$$(6.4) \qquad WL\{i\} = Pr\{ B\{i\} \} = 1 - Pr\{ F\{i\} \} \quad .$$

Using the dispatch preference matrix, DP, we can write FSC{i,j} as

$$(6.5) \qquad FSC\{i,j\} = Pr\{ B\{DP\{j,1\}\},\ldots,B\{DP\{j,k-1\}\},F\{i\} \}.$$

(Recall that i is the k-th preferred unit for call type j).

If the status of the servers were independent, we could write the probability of the compound event in (6.5) as the product of the probabilities of the separate events. Larson (Ref. 40) has examined this assumption for M/M/N systems (identical servers).

B.3 Systems with Identical Servers

We reproduce the main results of Larson's work as it is crucial to the approximation procedure. Even in the symmetric M/M/N system, the status of the servers is not independent. For the infinite line capacity case, Larson shows that the probability of k servers being busy, given that a randomly selected server is busy, is given by

$$(6.6) \qquad k * \Pr\{k \text{ busy servers}\} / U \quad , \quad k=0,1,\ldots,N,$$

where U is the total call rate multiplied by the average service time (identical for all servers) and divided by the number of servers, N. This quantity will be referred to as the _utilization_ of a system. For the M/M/N queuing model with infinite line capacity, U is equal to the average workload of the servers. Equation (6.6) states in precise terms the qualitative remark that knowing one server is busy biases the distribution of the number of busy servers in the direction of more busy servers. Hence the availability of the servers is not independent.

We return to the case of zero line capacity (the loss system). Consider the problem of sampling the status of randomly selected servers without replacement. Let $A\{k\}$ denote the event that the first k selected servers are busy and that the k+1-st is free. Larson shows that

$$(6.7) \qquad \Pr\{\ A\{k\}\ \} = Q\{N,U,k\} * (AWL**k) * (1-AWL) \quad ,$$

where AWL is the average workload of the units and Q, as a function of the number of servers, utilization factor, and k, is a correction factor which gives the equality in (6.7). Equation (6.7) is a sort of quasi-independence statement regarding the status of the units. $Q\{N,U,k\}$ is given by

$$(6.8) \qquad Q\{N,U,k\} = \sum_{j=k}^{N-1} \frac{(N-j) * (N**j) * (U**(j-k))}{(j-k)!} * \frac{P0 * (N-k-1)!}{((1-PN)**k) * N! * (1-U*(1-PN))}$$

for $k=0,1,\ldots,N-1$,

where P0 is the probability of all servers being free and PN is the probability of all servers being busy. A recursive procedure for the calculation of the correction factors is given in Appendix B. Reference 40 contains tables and graphs of the terms. For general reference, Q{N,U,0} is identically 1. Only for low utilizations (less than two tenths) and large k (greater than 5) is Q{N,U,k} very different from 1.

It should be noted that although these results were derived for the M/M/N loss system, they also hold for the M/G/N system since the steady state probabilities for the two systems are the same (Ref. 53).

B.4 The Approximation Technique

The approximation technique developed by Larson employed the relation suggested by (6.7) above in connection with fixed preference dispatch rules. In particular, although the status of the servers is not independent, the joint probabilities are treated as though they were. The correction term is used as a scaling factor for the product of the workloads to give the correct result for systems with identical servers.

For example, in a six server system, the probability that servers 1 and 3 are busy and server 5 is free at the time of a call for service would be approximated by

$$(6.9) \qquad Pr\{B\{1\},B\{3\},F\{5\}\}=Q\{N,U,2\}*WL\{1\}*WL\{3\}*(1-WL\{5\}).$$

In (6.9), U is given by (CRT*TA)/N where TA is the average service time over all calls not arriving during a period of saturation. approximation procedure developed by Larson assumed that the average service time for all calls was TA, independent of the server or the customer class.

As discussed by Larson (Ref. 40), this approximation to the exact hypercube model yields estimates for the performance measures which are generally within a few percent of those calculated by the hypercube. The advantage of using the approximation instead of the exact model is that the size of the system of equations to be solved grows linearly with the number of servers rather than exponentially. As will be shown below, the approximation requires the solution of a set of nonlinear equations. Solved in an iterative fashion, the computations are actually no more involved than the linear steady state equations derived for the exact model (See (3.5)).

We apply the same sort of approximation technique to the general service time model described in the first part of this section. The notation is slightly simplified by defining an N by NC unit order matrix, UO, where if server i is the k-th preferred unit for calls of type j, then UO{i,j}=k. With this definition, we rewrite (6.5) as

$$(6.10) \qquad FSC\{i,j\}=Pr\{B\{DP\{j,1\}\},\ldots,B\{DP\{j,UO\{i,j\}-1\}\},F\{i\}\} .$$

In much the same manner as Larson, we approximate FSC{i,j} by

$$(6.11) \qquad FSC\{i,j\} \approx Q\{ N,U,UO\{i,j\}-1 \} * (1-WL\{i\}) * \prod_{k=1}^{UO\{i,j\}-1} WL\{DP\{j,k\}\} .$$

Substituting (6.11) into (6.2), we have

$$(6.12) \qquad WL\{i\} = \sum_{j=1}^{NC} CR\{j\} * TSC\{i,j\} * Q\{N,U,UO\{i,j\}-1\} * (1-WL\{i\}) * \prod_{k=1}^{UO\{i,j\}-1} WL\{DP\{j,k\}\}$$

Equation (6.12) can be solved for WL{i} to obtain

$$(6.13) \qquad WL\{i\} = X\{i\} \; / \; (1+X\{i\}), \quad i=0,1,\ldots,N-1;$$

where X{i} is given by

$$(6.14) \qquad X\{i\} = \sum_{j=1}^{NC} CR\{j\} * TSC\{i,j\} * Q\{N,U,UO\{i,j\}-1\} * \prod_{k=1}^{UO\{i,j\}-1} WL\{DP\{j,k\}\} \ .$$

Our intent is to use equations (6.13) and (6.14) in a "linear iteration" technique (Ref. 9) to solve for the workloads of the units. After determining the workloads, (6.12) can be used to determine the FSC{i,j} terms and hence to obtain estimates for the performance measures associated with the system. In order to apply this procedure, we need an initial solution to start the iteration.

One such solution can be derived by assuming that there is no cooperation between the units. (This starting solution corresponds to FSC{i,j} equal to one for i equal to DP{j,1}; otherwise, FSC{i,j} is zero). We first compute an initial value for the average service time TA (used in evaluating the correction factor) by

$$(6.15) \qquad TA = \sum_{j=1}^{NC} FC\{j\} * TSC\{DP\{j,1\},j\} \ .$$

(Recall that FC{j} is the fraction of calls of type j). The initial value of the units' workloads is given by

$$(6.16) \qquad WL\{i\} = \sum_{j:DP\{j,1\}=i} CR\{j\} * TSC\{i,j\} \;.$$

The initial value of U used in the correction factor is given by (CRT*TA)/N.

Each iteration involves using the current estimate of the workloads to compute the X{i} term for each unit i according to (6.14). The new estimate of the workloads is obtained from (6.13). The new estimates of the workloads are then used to compute a new average service time according to

$$(6.17) \qquad TA = \sum_{i=0}^{N-1} \sum_{j=1}^{NC} FC\{j\}*TSC\{i,j\}*FSC\{i,j\}/(1-P\{S\}) \;,$$

the FSC{i,j} being determined from equation (6.11). The utilization for the next iteration is obtained by setting U equal to CRT*TA/N.

The iteration is continued until a convergence criterion is satisfied; for example, the maximum relative change over all the unit workloads being less than one percent.

Of course, there are a great number of questions to be asked concerning the validity of the approximation described here. In the next section some of these issues are explored.

C. Computational Experience

Since the approximation technique described above reduces to that developed by Larson when the servers have identical service times (with the slight difference noted below), a few remarks on the accuracy of Larson's procedure are appropriate. That procedure has been found to yield estimates for the performance measures which are generally within a few percent of the values computed by the exact hypercube model. As the total call rate becomes smaller in relation to the average service time or as the number of units increases, the approximation increases in accuracy. Finally, as the units become more similar in workload, the accuracy increases. At least in part, the last remark might be expected since the approximation is exact for the M/M/N loss system with identical servers. (It should be noted that the approximation technique developed by Larson can also be applied to the infinite line capacity system).

The difference in Larson's approximation and that described here has to do with a normalization constraint employed by Larson. For the M/M/N system, the average workload of the units can be computed independently of the particular dispatch rule used (Ref. 28) and used to scale successive workload estimates. As noted in Chapter 3, with different unit service times, this result does not hold for the hypercube model and consequently cannot be employed in the iteration scheme developed here. For the remainder of this chapter, we deal only with the approximation procedure for general service times as given above.

In trying to give some indication of the accuracy of the approximation technique there is one striking difficulty; there are not very many models with which to compare it. Two obvious candidates for future validation research are simulation models and data collected from the actual operation of a system. In the interim, the approximation can be compared to some of the models developed in Chapters 3 and 4.

Since the approximation closely parallels Larson's work on the hypercube model with identical unit service times, there is no need to pursue that comparison here. There remains the hypercube model with server dependent service times, the convolution model, and the exponential model (See Chapter 4). Since all computational experience to date has indicated that the convolution model is very closely approximated by the comparable hypercube model (Section B.2, Chapter 4), we focus our attention on the hypercube and exponential models.

As noted in Chapter 4, the state space associated with the exponential model is very large; (NC+1)**N elements for a system with N servers and NC customer classes. The computational difficulty in dealing with this model was compounded by the author's use of an interpretative computer language for the solution of the steady state equations. The net result is that systems with more that three servers and three classes of customers became prohibitively expensive (in the present computer implementation) to use for extensive examples.

From Larson's work, we expect to have the most difficulty in describing systems with a small number of servers and widely varying service times. (The latter tends to make the servers "more distinguishable"). In order to test the approximation in such an extreme set of circumstances, computations were performed for a three server, three customer class example with the exponential model. The system parameters are summarized in Table 6.1. Note that the ratio of service times is as large as ten to one and that the dispatch preference matrix used (DP) results in a five to one ratio in the frequency of calls for which units 0 and 1 are first preferred.

For this sort of extreme situation, the average percentage error in the workload estimates was less than one percent at either end of the call rate range, roughly five percent for less extreme values , and then as high as fifteen percent at an average workload of 0.281. It should be noted that the cost of running the approximation technique was roughly one tenth of that for the exact exponential model. This fraction could be expected to decrease as the number of units increases since the problem size in the approximation

Table 6.1 Summary of system parameters for a three server example of the exponential model. Total call rates (CRT) in the range of 0.017 to 1.20 corresponding to an average server workload of 0.014 to 0.928.

$$DP = \begin{matrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 1 & 0 \end{matrix}$$

$$RH = \begin{matrix} 0.416 & 0.058 & 0.379 \\ 0.058 & 0.474 & 0.109 \\ 0.379 & 0.109 & 0.530 \end{matrix}$$

$FC\{1\} = 0.536$
$FC\{2\} = 0.107$
$FC\{3\} = 0.357$

Computations were performed for the following values of CRT (shown with the computed average workload, AWL, for each value of CRT).

CRT	AWL
0.017	0.014
0.086	0.098
0.120	0.162
0.171	0.281
0.286	0.540
0.400	0.700
1.200	0.928

grows linearly rather than exponentially. This example does suggest that some caution be employed when using the approximation for systems with widely varying service times, although the difference in cost and size of the two approaches really leaves no choice for large systems. (For systems with less variance in the service times, say up to five to one, the approximation has yielded estimates for both workloads and the FSC{i,j} with relative errors below two percent).

As noted, the approximation does yield good results as compared to the exact hypercube model with identical service times. This is also the author's experience in examining systems with server dependent service times. To give some idea of the accuracy of the approximation, we will compare it in some detail to the three server example used in Chapter 3. In this problem, the range of average service times is two to one and the range of call rates which are first preferred varies by as much as four to one between two of the servers (0 and 2).

Figure 6.1 is a graph of the maximum percentage difference between the approximation and the exact hypercube model compared over a range of utilizations from 0.05 to 0.95. (Recall that the utilization, U, for the hypercube model, is given by CRT/SRT; SRT is the total service rate). A more detailed comparison of the workload data is presented in Table 6.2 for the same system.

As indicated by Table 6.2 and Figure 6.1, the

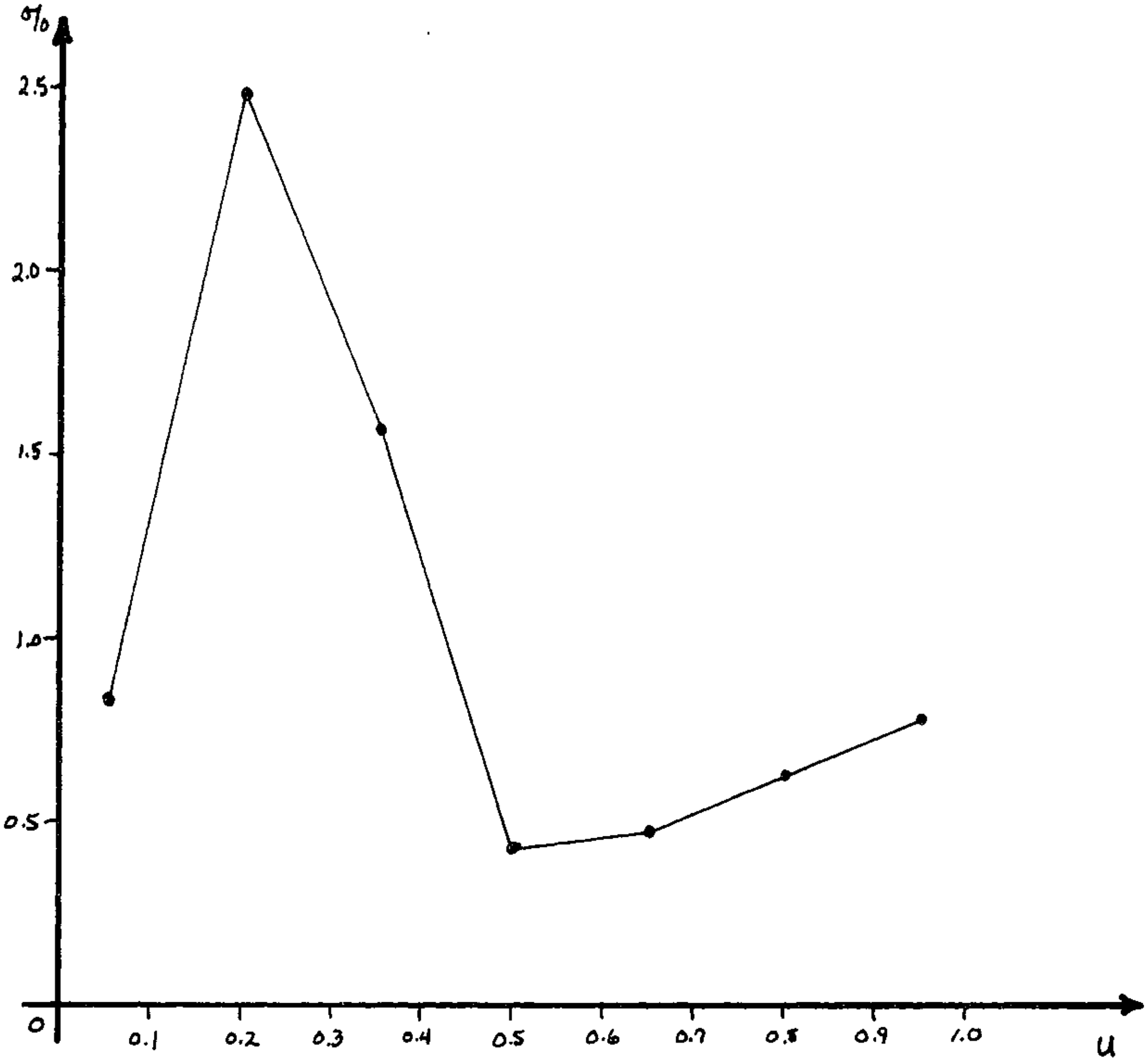


Figure 6.1 Maximum percentage difference in the values of the unit workloads as calculated by the approximation versus the hypercube model. The utilization, U, is CRT/SRT. Example given for three servers in Sample City.

Table 6.2 Unit workloads as calculated by the approximation procedure and the exact hypercube model. For each utilization, the top row is the set of values calculated by the approximation. The bottom row is from the exact hypercube model.

U		WL{0}	WL{1}	WL{2}
0.05	Approx:	0.0955	0.0270	0.0351
	Exact:	0.0955	0.0270	0.0354
0.20	Approx:	0.3009	0.1447	0.1554
	Exact:	0.3006	0.1445	0.1593
0.35	Approx:	0.4369	0.2663	0.2901
	Exact:	0.4362	0.2668	0.2946
0.50	Approx:	0.5339	0.3708	0.4134
	Exact:	0.5327	0.3721	0.4153
0.65	Approx:	0.6058	0.4557	0.5143
	Exact:	0.6042	0.4579	0.5135
0.80	Approx:	0.6603	0.5234	0.5927
	Exact:	0.6587	0.5267	0.5907
0.95	Approx:	0.7026	0.5776	0.6531
	Exact:	0.7013	0.5821	0.6510

approximation gives a very accurate estimate of the unit workloads. The quantities of most interest are the FSC{i,j}, the fraction of calls of type j answered by unit i. Instead of presenting all of these numbers for different utilizations, we concentrate on a single value, U=0.35. It was this utilization which gave the maximum percentage difference between the values of FSC calculated by the approximation and the exact model. Table 6.3 details the values for this example. An examination of the travel distances shown in Table 3.1 will reveal that there are only four distinct preference lists for the particular rule which dispatches the closest available unit. Atoms 1, 8, 11, and 16 are representative of those four preferences.

The largest percentage difference in Table 6.3 occurs at FSC{0,8}. This happens to be for a unit which is the second preferred for that particular atom. The percentage difference is somewhat misleading as the FSC{i,j} terms associated with the first preferred server are clearly the most important in calculating performance measures for the system. It was generally the case for all of the utilizations that the largest percentage differences were associated with dispatches of the second or third preferred unit.

Table 6.4 is an attempt to give some indication of the seriousness of deviations in the FSC terms. The first column gives the maximum percentage difference calculated for FSC terms associated with the dispatch of a first preferred unit.

Table 6.3 Values of FSC{i,j} as calculated for the three server example with a utilization of 0.35. The maximum percentage difference between corresponding values occurs for FSC{0,8} (16.6 percent). This was the largest percentage difference encountered over all the utilizations tested. The values shown below correspond to the four distinct rows of the dispatch preference matrix. The top row is from the approximation; the bottom from the hypercube model.

Atom j:		FSC{0,j}	FSC{1,j}	FSC{2,j}
1	Approx:	0.5631	0.2799	0.0770
	Exact:	0.5638	0.2817	0.0840
8	Approx:	0.1309	0.7337	0.0770
	Exact:	0.1123	0.7332	0.0840
11	Approx:	0.0406	0.7337	0.1651
	Exact:	0.0405	0.7332	0.1557
16	Approx:	0.0406	0.1858	0.7099
	Exact:	0.0405	0.1835	0.7054

Table 6.4 Measures of the relative error in the FSC(i,j) as calculated by the approximation procedure and the hypercube model. For each utilization, Column 1 is the maximum percentage difference in the FSC terms for first preferred dispatches. Column 2 is the maximum absolute difference in the two sets of values divided by 1 minus the saturation probability for that utilization.

U:	Column 1	Column 2
0.05	0.03%	2.7%
0.20	0.46%	1.5%
0.35	0.64%	2.0%
0.50	0.32%	2.0%
0.65	0.41%	1.8%
0.80	0.70%	1.6%
0.95	1.08%	1.3%

The second column is the ratio of the maximum absolute difference in the values of FSC{i,j} as calculated by the approximation and the hypercube to the fraction of calls receiving service (one minus the saturation probability). As is readily seen from the data, only one estimate of FSC for a first preferred dispatch is in error by more than one percent. Although the second column is roughly constant at two percent, it should be remembered that these are maximum relative differences over all the FSC{i,j}.

It is very important to note that the approximation procedure, at least for this last set of examples, yields results which are very close for the types of dispatches which contribute most to the calculation of performance measures for the system; that is, dispatches of the first and second preferred units for each type of customer. It is clear that the approximation requires further validation. Of particular interest would be bounds on the maximum relative errors or some characterization of the type of system which does not lend itself to the use of the approximation.

These comments notwithstanding, the author maintains that the approximation is a good technique for describing these types of loss systems with distinguishable servers and classes of customers. It would appear to be particularly useful in the context of spatially distributed systems in which travel time is a significant portion of the overall service time. As a final note, the approximation is an analytic technique which is easy to implement and inexpensive to run in a computerized version. Some estimates of computer costs will be given for the examples of the next chapter.

D. A Note on Service Times

We would like to examine the differences in workloads which can be attributed to the type of service time information available. For example, the original hypercube formulation included service times which were independent of the customer and server. The hypercube as formulated in Chapter 3 allows the service time to depend on the server but not on the type of customer. Finally, the approximation procedure for the general service time model permits service times to be a function of both the class of customer and the server.

The question raised here is exactly how much of this information is necessary to adequately describe these systems. For the remainder of this section, suppose that the FSC(i,j) are known.

Given the call rates , service times, and response pattern, the workload of unit i is calculated as in equation (6.2) restated here as

$$(6.18) \qquad WL\{i\} = \sum_{j=1}^{NC} CR\{j\} * TSC\{i,j\} * FSC\{i,j\} \quad .$$

Suppose we compute the mean service time for a call answered by unit i for the given response pattern. We define TU{i} (service <u>t</u>ime by <u>u</u>nit), for i=0,1,...,N-1, by

$$(6.19) \qquad TU\{i\} = \sum_{j=1}^{NC} TSC\{i,j\}*FT\{i,j\} \quad / \quad \sum_{j=1}^{NC} FT\{i,j\} \quad ,$$

where FT{i,j} is given by equation (5.5) and is the overall fraction of calls that are of type j and answered by unit i. The denominator of (6.19) divided by (1-P{S}) is equal to the fraction of all calls not arriving during a period of saturation which are serviced by unit i.

Now we consider the system with the same response pattern as above except that service times are only server dependent and are given by (6.19). Substituting (6.19) into the RHS of (6.18) we get

$$\sum_{j=1}^{NC} CR\{j\}*TU\{i\}*FSC\{i,j\} \quad = \quad TU\{i\} * \sum_{j=1}^{NC} CR\{j\}*FSC\{i,j\}$$

$$= \sum_{j=1}^{NC} TSC\{i,j\}*FT\{i,j\}*CRT$$

$$= \sum_{j=1}^{NC} TSC\{i,j\}*FSC\{i,j\}*CR\{j\}$$

where the last sum is seen to be WL{i} when compared to (6.18). The result of these manipulations is to say that if the correct average unit service time is used in the general service time model in place of the more detailed server-customer specific service times, we still get the correct unit workloads.

In exactly the same manner we can compute the overall average service time as in (6.17) and substitute that value for the TSC{i,j} in (6.18). Define FU{i}, the fraction of calls serviced by unit i, by

$$(6.20) \qquad FU\{i\} = \sum_{j=1}^{NC} FC\{j\} * FSC\{i,j\} / (1-P\{S\}) .$$

If we substitute TA for TSC{i,j} in the RHS of (6.18), we get (N*AWL*FU{i}). This quantity is not equal to the workload of unit i unless the average service times for each unit are equal. (Note: summing this term and dividing by N does yield the correct value for the average workload, AWL).

The remarks above are interesting in light of a model developed by Bernstein and Thomas (Ref. 57). Their approach is to incorporate travel time information for spatially distributed systems by using Larson's approximation procedure in an iterative manner. An estimate of TA is made for the system, the FSC{i,j} are computed for that service time, and a new estimate of TA is obtained via (6.17). This procedure is repeated until a convergence criterion is satisfied. Since the values of FSC{i,j} depend only on the unit workloads in the approximation scheme, it would appear that such a technique could be used to incorporate travel times for spatially distributed systems if the iteration utilized unit specific service times, TU{-}, rather than the global service time, TA.

E. Sensitivity Analysis

The approximation can be easily adapted to determine first order effects of changes in the system configuration. For example, Chelst (Ref. 7) has developed algorithms for balancing workloads and travel times using Larson's approximation. Such algorithms often require the computation of the change in response patterns of workloads resulting from changes in the dispatch preference matrix. We will focus on one particular kind of change in the dispatch preferences as an example of a simplified computation procedure.

We restrict our attention to calls of type j, where x is the k-th preferred server for type j and y is the k+1-st preferred. That is, DP{j,k}=x and DP{j,k+1}=y. We compute the change in FSC{x,j} and FSC{y,j}, denoted by X and Y respectively, if units x and y are reversed in the dispatch preference list. FSC{x,j} is given by

$$(6.21) \qquad FSC\{x,j\} = Q\{N,U,k-1\}*(1-WL\{x\}) * \prod_{i=1}^{k-1} WL\{DP\{j,i\}\} \ .$$

The values of X and Y are computed on the basis of the first iteration of the approximation starting with the initial workloads, WL{i}, i=0,1,...,N-1. The next iteration yields the following expression for the fraction of calls of type j answered by unit x under the revised dispatch policy.

$$(6.22) \qquad X+FSC\{x,j\} = Q\{N,U,k\} * (1-WL\{x\}) * \prod_{i=1}^{k} WL\{i\} \ .$$

The change for unit x, X, is computed from (6.21) and (6.22) as

$$(6.23) \qquad X = FSC\{x,j\} * (Q\{N,U,k\}*WL\{y\})/Q\{N,U,k-1\} - 1 \ .$$

Y is computed similarly as

$$Y = FSC\{y,j\} * Q\{N,U,k-1\}/(Q\{N,U,k\}*WL\{x\}) - 1 . \tag{6.24}$$

Using (6.18), the first order change in the workload of unit x caused by this change in the dispatch rule will be CR{j} * TSC{x,j} * X. The change in the workload of unit y will be CR{j} * TSC{y,j} * Y. (Note: these computations reflect the changes which will occur in the first iteration of the approximation procedure as applied to the equilibrium workloads derived under the initial dispatch rule).

We apply these concepts to the example begun in Chapter 3. The optimal rule derived there dispatched unit 1 to a call from atom 7 instead of dispatching the closer unit 0. Although the system wide expected travel distance did not change greatly, the workload imbalance was halved. Applying the results developed above to this system, with x=0, y=1, and j=7, we have

Given: CR{7}=0.099, TSC{0,7}=1.0, TSC{1,7}=0.67.

Compute: FSC{0,7}=0.5273, FSC{1,7}=0.2853.
WL{0}=0.4727, WL{1}=0.3033.
X=-0.4019, Y=0.4845.

The workload of unit 0 decreases to 0.4329 and that of unit 1 increases to 0.3353. The maximum workload imbalance can be expected to decrease from 0.169 to roughly 0.10. A more precise computation yields a final imbalance of 0.13. Similar techniques can be applied to the determination of the marginal costs of changes in assignment strategy.

This chapter concludes the methodological developments. Briefly, we have developed several continuous time Markov models for service systems with distinguishable servers and classes of customers. One of these models, the hypercube, was used in conjunction with a location model in order to determine the optimal positions for response units in spatially distributed systems.

This present chapter details the application of an approximation procedure for computing the steady state characteristics of systems in which the service time depends on both the server and the class of customer. In the next chapter, we use the approximation procedure and the location model to treat a spatially distributed system in which travel time is a significant portion of the overall service time. The example is intended to demonstrate the flexibility of these models.

Chapter 7. APPLICATIONS TO THE LOCATION OF EMERGENCY MEDICAL VEHICLES

A. Introduction

The purpose of this chapter is to demonstrate the use of some of the models developed in the preceding chapters. In particular, we would like to use the location model to optimize the positions of response units in an environment in which inter-unit cooperation is important. An emergency medical system (EMS) was chosen for several reasons.

We have focused on emergency service systems and we continue that emphasis here. The descriptive models, including the hypercube model and the approximation procedure, incorporate the emergency aspects of the service (immediate response if a unit is available and unpredictable arrivals of calls for service) with a comprehensive treatment of local geography (including impediments to travel, placement of units, and the spatial distribution of calls for service). Since the hypercube model has seen extensive use in police applications (Refs. 38, 30, 14, and 8), we prefer to demonstrate the use of the approximation procedure developed in Chapter 6 in an EMS application.

In addition, the approximation procedure incorporates a more accurate description of service time for ambulance systems than the hypercube model. In spatially distributed systems, at least one element in the classification of customers is the location of the incident. As a result, the expected service time for an arbitrary call has a component which depends on both the server (the location from which it is dispatched) and the customer (locational origin of the call). That component is the travel time to and from the scene of the indident. Recall that the hypercube formulation allows server but not customer dependent service times. (Of course, any system with spatially distributed responses will exhibit the same server and customer dependence in its service times. For urban police forces, this component is usually small compared to the overall service time. See Ref. 39).

The examples given here are intended to demonstrate the use of the models, not to arrive at general conclusions concerning the optimal location of ambulances. The models are specifically designed to include local information such as peculiarities of geography and the spatial distribution of calls for service as well as the particular placement of response units (ambulances) and facilities (hospitals). For these reasons, the models can be expected to produce results which reflect the characteristics of the specific problem being examined.

We address two problems dealing with the location of emergency medical vehicles. The first is a straightforward location problem. How do we locate ambulances in order to minimize the global expected response time for a region when there are constraints on maximum response time?

The second problem is somewhat more complicated. How is the performance of an emergency medical response system affected by the addition of specialized units such as mobile coronary care units? Before dealing with either of these questions, we spend some examining the components of service time time for ambulance response and relating these components to the particular geography of Sample City.

B. Relating Service Time to Local Geography

Both examples of this chapter are formulated in the context of the geography of Sample City. Of course, the techniques have general applicability. In order to use the approximation procedure and the location model, we have to specify service times and assignment costs for the system.

Ambulance service times consist of several distinct components (Ref. 31). These are travel to the scene, on-scene service time, travel to a hospital, time for the transfer of the patient at the hospital, and return to base (Figure 7.1). We assume that the total service time includes the travel time in returning to the base; that is, no vehicles are dispatched from the hospital. Initially, we seek to minimize expected response time; the time from the reception of a call until a unit arrives at the scene of the incident. The two components of response time are the dispatch delay time and travel time to the scene (Figure 7.1).

We assume that calls arriving when all units are busy

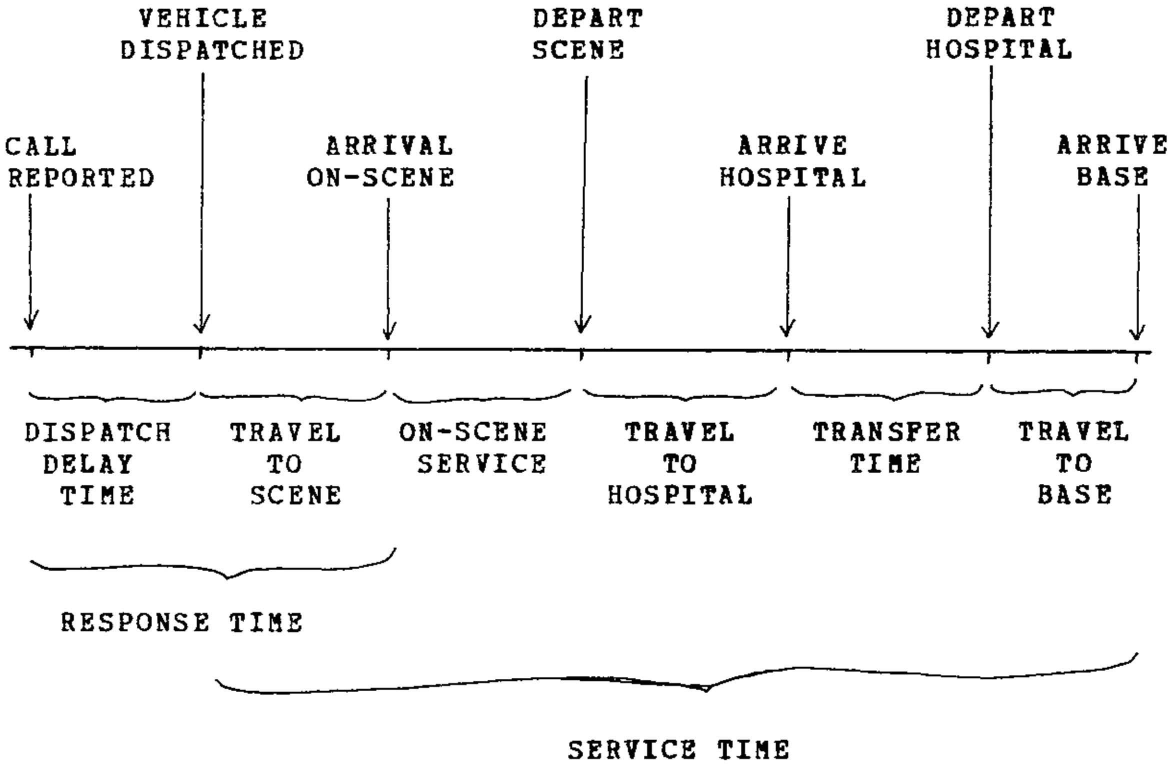


Figure 7.1 Components of service and response time for emergencey medical response vehicles in the case of no queueing delays.

(saturation) are serviced by means external to the system. Hence the dispatch delay time has no component related to queuing delays but consists entirely of the time from the receipt of a call until a response unit is notified and that unit starts to travel to the scene of the incident. During periods of congestion, radio equipped units could be dispatched directly from the hospital. For this example, we assume that all dispatches begin at the response units' home base.

We include local geography in the service time description by classifying customers according to the origin of the call for service. If a call arrives from a point x and server i is assigned to that call, the expected service time is determined by adding the dispatch delay, on-scene, and hospital transfer times to the various travel time components. If unit i is located at point y and the hospital at z, these components are travel time from y to x (travel to the scene), x to z (travel to the hospital), and z to y (return to base).

We partition Sample City into sixteen geographical atoms, each representing a different class of customer (NC=16). In general, the number and size of the atoms for an area are determined by the requisite detail in locational information. (See, for example, Ref. 30). We define an NC by NC matrix TT, with elements TT(i,j) equal to the travel time from atom i to atom j. In practice, this matrix would be determined either from empirical data or by calculating the distance from atom i to atom j and dividing by an effective average travel speed for that pair of atoms (or for the whole region).

The former approach is more realistic (and also more expensive) in that it allows specific consideration of local travel characteristics. For the sake of simplicity, we adopt the latter approach here. Table 7.1 summarizes the geographic characteristics of Sample City (see also Figure 3.2). We assume that the distance from atom i to atom j is given by the right angle (rectilinear) distance from the center of atom i to the center of atom j. For i equal to j, we take the average intra-atom travel distance to be one-half the square root of the area of the atom (Ref. 41).

The specification of the expected service times is completed by giving the non-travel time components of service time and an average travel speed. We use miles and minutes as the units of distance and time. For these examples, we assume a delay time, DT, of 2 minutes; an on-scene service time, OSS, of 10 minutes; a hospital transfer time, TRN, of 5 minutes; and a travel speed, TS, of 0.5 miles per minute (30 miles per hour). Finally, we assume the region contains a single hospital, located in atom 5. (The values of the service time components given here were chosen to be representative of EMS systems. No particular significance should be attached to any of these values. See Ref. 31 and 13).

As an example of a response and service time calculation, suppose that unit 0 is located in atom 1. We compute

Table 7.1 Atom centroids and areas for Sample City with the fraction of calls generated by each atom (FC{j}=CR{j}/CRT). All distances expressed in miles. See Figure 3.2.

ATOM	X-COORD	Y-COORD	AREA	FC*100
1	2.5	6.0	13	14.3
2	6.0	3.6	10	11.4
3	9.5	1.5	9	6.7
4	9.8	4.6	11	7.6
5	6.5	6.8	13	9.5
6	4.0	8.2	6	7.6
7	8.0	9.0	7	7.6
8	10.6	7.1	10	7.6
9	12.9	9.5	7	2.9
10	10.5	11.2	12	4.8
11	12.3	13.4	10	2.9
12	14.6	11.6	7	1.9
13	14.6	14.1	4	1.0
14	17.1	11.8	6	1.9
15	17.0	14.3	12	4.8
16	18.1	16.8	8	9.5

TSC{0,8}, the expected service time for unit 0 responding to a call from atom 8, as the sum of

(i) Travel time to the scene: 18.4 minutes. (The distance from atom 1 to atom 8 is equal to (10.6-2.5) plus (7.1-6.0) or 9.2 miles. Divide by TS=1/2).

(ii) On-scene service time: 10 minutes.

(iii) Travel to the hospital: 8.8 minutes. (The distance from atom 8 to atom 5 is equal to (10.6-6.5) plus (7.1-6.8) or 4.4 miles. Divide by TS=1/2).

(iv) Transfer time at the hospital: 5 minutes.

(v) Travel in returning to base: 9.6 minutes. (The distance from atom 5 to atom 1 is equal to (6.5-2.5) plus (6.8-6.0) or 4.8 miles. Divid by TS=1/2).

The total service time in this case is 51.8 minutes. The response time for this call is the dispatch delay time (2 minutes) plus the travel time to the scene (18.4 minutes); or 20.4 minutes.

In general, if unit i is located in atom p, and a single hospital is located in atom h, we can write TSC{i,j} as

$$\text{(7.1)} \qquad TSC\{i,j\} = TT\{p,j\} + OSS + TT\{j,h\} + TRN + TT\{h,p\}.$$

Since we seek to minimize average response time, we identify the costs per call with the response time and write CP{i,j,p}, the "cost" of assigning unit i from atom p to a call from atom j, as

(7.2) $CP\{i,j,p\} = DT + TT\{p,j\}$.

Note that the cost of assigning unit i to a call from atom j depends only on the location of unit i.

The specification of the system is completed by giving the call rates. The spatial distribution was given in Figure 3.2 and repeated in Table 7.1 (FC{j}). We set CRT equal to 0.8 calls per hour or (0.8/60) calls per minute.

With this description, we can use the approximation procedure to determine the operating characteristics of the system by specifying the number and location of the response units.

C. Locations Minimizing Response Time

In order to apply the location model, we have to determine the response pattern associated with an initial set of unit locations. A reasonable choice is to position three units at atoms 1, 10, and 15. The characteristics of this configuration are shown in Table 7.2. With these locations, the average workload (AWL) is 0.267, and the average response time is 12.5 minutes.

One additional quantity shown in Table 7.2 is the fraction of calls receiving an "acceptable" response. A response is considered acceptable if the response time is less than thirty minutes. (Although the choice of thirty minutes is somewhat arbitrary, that is the figure mentioned by the

Table 7.2 Response characteristics for three units as initially located. Section 7.2.A contains statistics computed for the region as a whole; section 7.2.B contains statistics particular to each response unit. Unit 0 is positioned in atom 1; unit 1, in atom 10; and unit 2, in atom 15.

7.2.A Region-wide statistics.

Average unit workload (AWL): 0.267
Average response time (minutes): 12.6
Average service time (minutes): 62.7
Saturation probability (P{S}): 0.041
Fraction of calls with acceptable response: 0.954

7.2.B Workload (WL{i}), fraction of calls (FU{i}), and average response and service times for each unit.

UNIT	WL{i}	FU{i}	RESPONSE	SERVICE
0	0.25	0.41	10.8	46.3
1	0.31	0.38	14.0	62.4
2	0.25	0.20	13.7	97.2

EMSS Act (PL93-154)).

Table 7.3 summarizes the results of the optimal location procedure as applied to the initial unit positions given above. The optimal configuration is reached after one iteration. Unit 0 is repositioned at atom 5; unit 1, at atom 8; and unit 2, at atom 15. The average response time is reduced to 10.3 minutes, a reduction of 18.5 percent as compared to the initial locations.

The optimization of response times also resulted in improved performance in several other areas. By decreasing response times, the average service time is reduced and the average workload decreases almost 12 percent. In addition, the fraction of calls which must receive backup service (those arriving during periods of saturation) decreases slightly from four to three percent.

The optimization procedure was applied with the additional constraint that at least 95 percent of the calls have a response time of less than thirty minutes. For this particular example, this was not a binding constraint. Tables 7.2 and 7.3 also give some performance characteristics particular to each unit. For this example, the optimization also resulted in some workload smoothing.

The iterative location procedure has the following general effects. The relocation of units is toward positions which are centrally located within the region. If the hospital is also centrally located, the decrease in response

Table 7.3 Response characteristics for three units optimally located. Unit 0, at atom 5; unit 1, at atom 8; unit 2, at atom 15. Region-wide and unit-specific statistics are given.

7.3.A Region-wide statistics.

Average unit workload (AWL): 0.235
Average response time (minutes): 10.3
Average service time (minutes): 54.7
Saturation probability (P{S}): 0.030
Fraction of calls with acceptable response: 0.951

7.3.B Unit-specific statistics.

UNIT	WL{i}	FU{i}	RESPONSE	SERVICE
0	0.22	0.46	8.6	37.8
1	0.23	0.34	11.8	52.4
2	0.25	0.20	11.6	96.4

time associated with the repositioning of the units is accompanied by a corresponding decrease in the time required for the units to return to base from the hospital. These two effects combine to reduce the overall service times of the units with a resultant decrease in their workloads. As the workloads are lessened, the probability of all units being busy simultaneously is also reduced. The net effect is an overall improvement in the performance of the system.

It is important to recall the distinctive characteristic of the location model being used here. The units are not positioned in order to minimize the average distance from a call source to the closest unit. The locations incorporate the information regarding the frequency with which other than the closest unit provides service (due to the unavailability of that unit).

For this particular example, we sought to minimize response time. In the next section, we illustrate the use of the same models in evaluating specialized units; in particular, mobile coronary care units.

D. Location of Specialized Response Units

Thus far, we have used response time or travel distance as a proxy for measuring the effectiveness of an emergency response. In this section, we focus on a more interesting measure of effectiveness for certain medical emergencies; namely, the risk of pre-hospital death from ventricular fibrillation following an acute myocardial infarction. We will not go into detail concerning the specific medical characteristics of this type of emergency, but utilize the work of Cretin (Ref. 13) in estimating the risk of death following a myocardial infarction (MI).

We summarize a much simplified version of Cretin's work here. Again, our intention is to demonstrate the use of the models, not to include all of the detail which might be required in an actual case study. All references in this section to the model of the risk of pre-hospital death are taken from Cretin (Ref. 13) unless specifically noted otherwise.

In simple terms, we wish to calculate the probability that an individual suffering a heart attack (specifically an MI) dies of ventricular fibrillation before reaching a hospital. The basic result, as developed by Cretin from clinical data, is that the risk of death, RD(T), in time less than or equal to T minutes following an MI, if no medical intervention occurs, is given by

$$(7.3) \qquad RD(T) = 1 - EXP\left\{-(0.222)*\left[1-EXP(-(0.015)*T)\right]\right\},$$

where EXP is the exponential function. Taking the limit of (7.3) as T goes to infinity, we have the probability of death following an MI as approximately 0.199 if no action is taken.

Suppose we have the following options in redesigning the ambulance response system of the previous section. We can obtain a relatively expensive mobile coronary care unit (MCCU) to replace one of the three present vehicles or we can use the same amount of money to obtain an additional standard ambulance. The question is whether better performance would be obtained from two standard vehicles and one MCCU or four standard vehicles. We make the following assumptions concerning the operation of the system.

We assume that ten percent of all calls represent MI's. (Again, this is taken as a representative value. See Ref. 45). The difference in treatment of these emergencies between an MCCU and a standard ambulance, insofar as the pre-hospital response of the emergency medical system is concerned, is entirely accounted for by the difference in treatment during the period from the arrival of a vehicle at the scene until the arrival of the patient at the hospital. We overstate the difference between an MCCU and a standard vehicle by assuming that an individual is at risk from the time of the MI until an MCCU arrives at the scene. If the response is by a standard vehicle, the patient is at risk for the additional period of on-scene service time plus the time for the trip to the hospital. This is equivalent to assuming that a standard vehicle has no treatment capabilities; it only provides transport. An MCCU is assumed to provide perfect treatment; there is no risk of death after it arrives on the scene.

The two different periods of risk are shown in Figure

7.2. Note that in either case the patient is at risk before the emergency is reported and during the response time of the unit assigned to the call.

The system with four standard vehicles is assumed to operate as in the previous section. We locate the four vehicles in order to minimize the region wide response time. With these locations, we calculate the risk of pre-hospital death for the coronary emergencies. The crux of the evaluation insofar as coronary emergencies are concerned is whether the addition of an extra standard vehicle can reduce the average response time sufficiently to offset the advantage of the treatment offered by the MCCU.

If the system has an MCCU, we alter the optimal location model in the following manner. Any coronary emergency will be attended by the MCCU if it is available. The only time the MCCU is dispatched to another type of emergency is when all standard vehicles are unavailable. Except for these two situations, all calls are serviced by the closest available standard vehicle. (We assume that it is possible to distinguish all coronary emergencies when the incident is reported).

Under this dispatch policy, we use the decoupled linear program given by (5.7) to choose the optimal positions of the units. The standard vehicles are located with the objective of minimal response time to all calls. The MCCU is located in order to minimize the risk of pre-hospital death for those

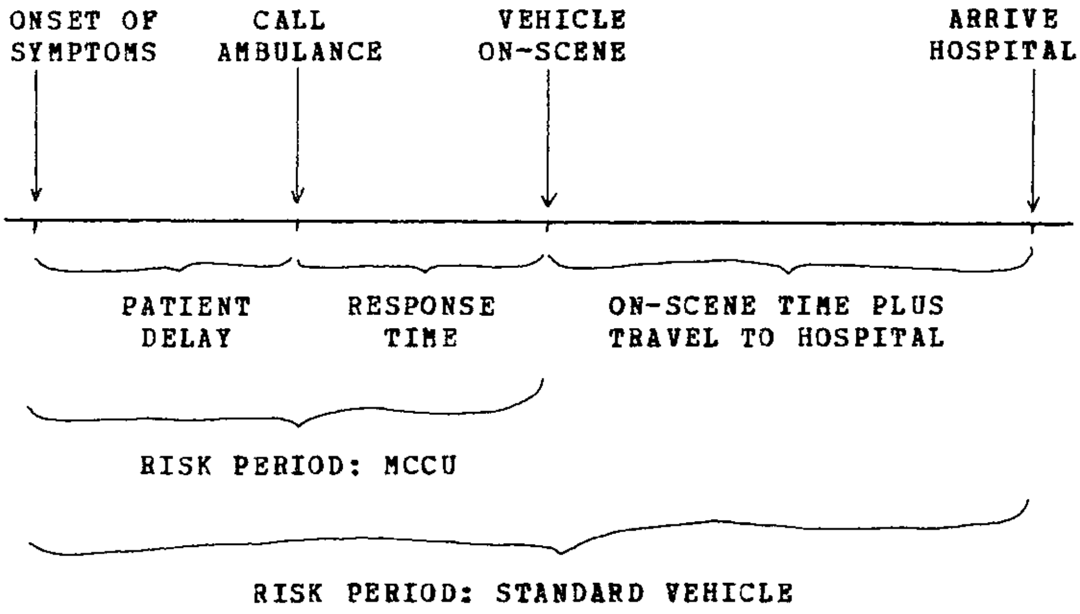


Figure 7.2 Period of risk for an MI victim when attended by an mobile coronary care unit (MCCU) or a standard ambulance having no treatment capabilities.

coronary emergencies serviced by the MCCU. (Note: we are not locating units to minimize the overall risk of pre-hospital death for MI victims). This risk depends on both the patient delay and the response time of the MCCU. Patient delay refers to the time between the onset of symptoms and the time at which the emergency is reported.

Let PD{t} denote the probability of a patient delay of t minutes. Labeling the MCCU as response unit 0, the optimal location for the unit under the response pattern FT{0,-} is found by minimizing

$$(7.4) \qquad \sum_{j} FT\{0,j\} * \left[\sum_{t=0} PD\{t\} * RD\left\{ t + DT + TT\{p,A\{j\}\} \right\} \right]$$

over the position of the unit, p. The sum over j is for those calls which are MI's; the A{j} are the associated atom locations. (The patient delay distribution is assumed to be discretized to an integral number of minutes). The expression given in (7.4) is the risk of pre-hospital death conditioned on the location of the MCCU (p), the patient delay (t), and the event that the MCCU responds to the coronary emergency (from atom A{j}).

With the exception of the dispatch policy and the slight change in the location model to accomodate the minimization of response time for standard vehicles and the risk of pre-hospital death for coronary calls serviced by the MCCU, all other parameters of the system as described in Section B remain the same, save two. Since the MCCU provides more extensive service, we assume it has an on-scene service time of 15 minutes versus 10 minutes for a standard vehicle. In addition, we now have 32 classes of customers (NC=32); the classification being based on location (16 alternatives) and the nature of the emergency (coronary or other).

Before evaluating the alternative systems, we have to specify patient delay; the time from the onset of symptoms until an ambulance is called. As noted by Mogielnicki, Stevenson, and Willemain for a fire rescue squad in Cambridge, Massachusetts, patient delay is often considerably larger than typical ambulance response times (Ref. 45). Hence, we evaluate our two alternatives first with zero patient delay and then with the mean patient delay of slightly more than four hours as described by Cretin (Ref. 13). The distribution for non-zero patient delay is shown in Figure 7.3.

Tables 7.4 and 7.5 summarize the operating characteristic of the response system with and without the MCCU respectively. It happens that the optimal location of the MCCU does not change when patient delay is considered so these results hold independently of the two particular distributions of patient delay used as examples here. With two standard vehicles and one MCCU operating under the dispatch policy as described above, the optimal locations for the standard vehicles are in

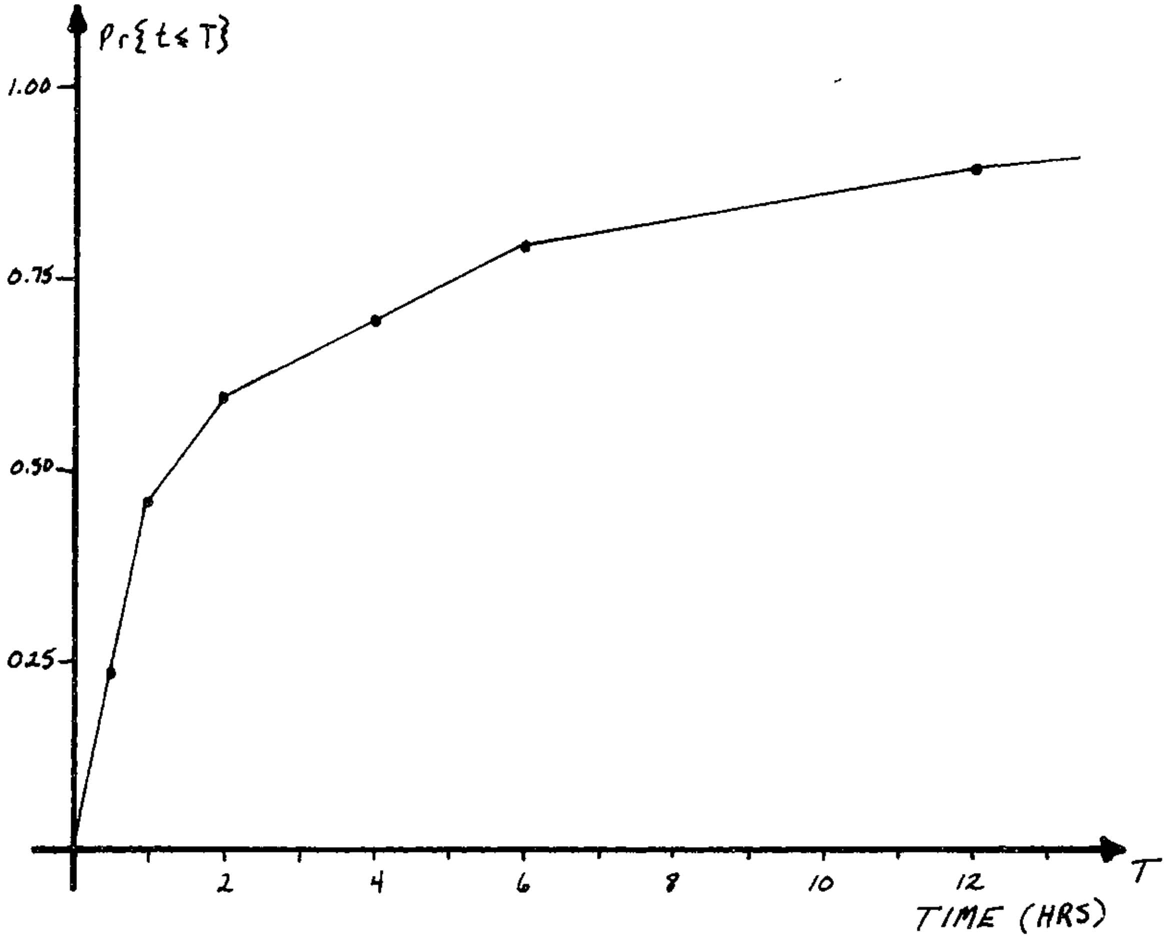


Figure 7.3 Cumulative distribution function for patient delay; the time between suffering an MI until calling for an ambulance. (Ref. 13).

atoms 5 and 10. The MCCU should also be located in atom 5. Although the MCCU is positioned only on the basis of the coronary calls to which it responds, it should be noted that it answers 90 percent of the coronary calls.

We note that the expected response time for the three vehicle system is 13.6 minutes as compared with 10.3 minutes for three standard vehicles optimally located and 8.8 minutes for four standard vehicles. The addition of the MCCU has clearly worsened overall response. The reduction in response time which can be obtained by dispersing three response units over the region has been negated by the addition of the MCCU. The responses of standard vehicles are most frequently within their own local vicinity. Since the MCCU responds to coronary emergencies on a region-wide basis, it is centrally located and the positions of the remaining standard units are adjusted to respond to non-coronary calls throughout the region.

Insofar as response to MI's is concerned, the MCCU offers a definite improvement. With zero patient delay, the risk of pre-hospital death for MI's drops from 0.086 with four standard vehicles to 0.052 when the MCCU replaces two of those vehicles. These figures should be compared with the risk associated with zero response time (0.0) and infinite response time (0.199). If we express the risk of death in terms of the expected number of deaths per year, a risk of 0.086 implies 60.3 deaths per year; a risk of 0.052, 36.4 deaths; a difference of 23.8 lives saved per year.

Table 7.4 Response characteristic for two standard vehicles and one MCCU. The MCCU, unit 0, is located in atom 5; the standard vehicles, in atoms 5 (unit 1) and 10 (unit 2). In addition to the region-wide and unit specific statistics, the response times and fraction of appropriate responses is given for coronary and all other calls (APPROP in 7.4.C). An appropriate response is defined by the dispatch of the MCCU to a coronary call and a standard vehicle to all other calls.

7.4.A Region-wide statistics.

Average unit workload (AWL): 0.24
Average response time (minutes): 13.6
Average service time (minutes): 54.8
Saturation probability (P{S}): 0.029
Fraction of calls with acceptable response: 0.940

7.4.B Unit-specific statistics.

UNIT	WL{i}	PU{i}	RESPONSE	SERVICE
0	0.12	0.16	15.4	56.4
1	0.28	0.50	11.2	43.1
2	0.31	0.33	16.2	71.7

7.4.C Statistics for response to MI and all other emergencies.

CALL	RESPONSE	APPROP
MI	13.4	0.90
OTHER	15.2	0.92

Table 7.5 Response characteristics for four standard vehicles located to minimize region-wide response time. Unit 0 is positioned at atom 1; unit 1, at atom 5; unit 2, at atom 10; and unit 3, at atom 15.

7.5.A Region-wide statistics.

Average unit workload (AWL): 0.18
Average response time (minutes): 8.8
Average service time (minutes): 54.1
Saturation probability (P{S}): 0.006
Fraction of calls with acceptable response: 0.99

7.5.B Unit-specific statistics.

UNIT	WL(i)	FU(i)	RESPONSE	SERVICE
0	0.13	0.24	6.8	42.3
1	0.20	0.39	9.3	39.2
2	0.18	0.22	10.8	62.8
3	0.20	0.15	7.4	97.4

The same computations were made using the empirical distribution for patient delay given by Figure 7.3. Considering patient delay alone, the risk of pre-hospital death in these circumstances is 0.133 with zero response time and, as before, 0.199 with infinite response time. There is clearly much less room for improvement in this situation. With four standard ambulances, the risk of pre-hospital death is computed to be 0.161; with the MCCU, 0.150. These risks convert to 112.8 and 105.1 deaths per year respectively, a difference of 7.7 lives saved per year.

The results of these computations indicate the kind of conflicts which must be resolved in determining the allocation of resources in an emergency service. If an MCCU is added to the EMS system, performance improves with respect to a subset of the population being serviced, but the overall response characteristics of the system are worsened. There are several questions which must be answered in order to choose between the alternative configurations.

In the first place, is the decrease in the risk of pre-hospital death really significant? On a percentage basis, there is clearly less improvement when the effects of patient delay are included in the analysis. Since the difference in the service provided by a standard ambulance and an MCCU was overstated, it would be hard to argue that the difference in risk is significant in this case. Even if patient delay is zero, there is some question as to whether a decrease in pre-hospital mortality contributes substantially to an increase in subsequent life-time expectancies. This issue is addressed in some detail by Cretin (Ref. 13).

The differences in response to other than coronary emergencies can be even more difficult to evaluate. At least for MI's, we have some measure of the effectiveness of response; for other emergencies, it is not clear what benefit is to be obtained from an average response time of 8.8 minutes (four standard vehicles) versus 13.6 minutes (two standard vehicles and the MCCU). An additional consideration is the fraction of demands for service which have a response time of less than thirty minutes; these are 94 percent and 99 percent respectively. In practice, these conflicts must be resolved on the basis of subjective preferences.

Although this evaluation of these systems does not yield purely objective answers for all of the questions which may be raised, it is hoped that this type of analysis will help to remove some of the uncertainty associated with the effects of various alternatives. For this particular example, if the patient delay is distributed as in Figure 7.3, the most cost effective alternative might be a system with three standard vehicles; the funds for an MCCU or additional vehicle being used in an education program designed to reduce patient delay.

E. Summary

As noted above, the purpose of the preceding examples is not to reach specific conclusions but to indicate the use of the models developed in the preceding pages to evaluate the consequences of alternative system designs. One striking ability of the formulation given here is the incorporation of very different measures of a system's effectiveness. For example, the same programs were used to locate vehicles in order to minimize response time and to determine the optimal location of the MCCU under a very different cost criterion.

As a final note, all of the results of this chapter were obtained at an expense of approximately twenty dollars. It is difficult to relate this quantity to CPU time since the computations were performed using an interpretive language, APL (Ref. 1), under the Time Sharing Option on an IBM 370/168. Based on other experience, one-minute of CPU time on that machine using a language such as PL/I or FORTRAN is almost certainly an extreme overestimate of the computation time required.

Chapter 8. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

In the preceding chapters, models for queuing systems with distinguishable servers have been developed and presented in connection with applications to spatially distributed queuing systems. The major developments include a procedure for determining optimal assignment rules in the Markov hypercube queuing model; an approximation procedure for the steady-state analysis of loss systems in which expected service times are a function of both the server and the class of customer; and an iterative procedure for determining the optimal locations for response units in spatially distributed queuing systems.

To date, the hypercube model has been applied only to spatially distributed systems in which the cost of assignment is given by the expected travel time to the scene of the incident. As noted in Chapter 3, the optimization procedure applied to this cost structure yields very little improvement in average travel distance as compared to the "dispatch the closest available unit" strategy. At least for spatially distributed systems, the disadvantage associated with the complexity of the optimal rule would appear to offset its benefit in the reduction of travel times.

However, that result does provide useful information.

That is, the simple, fixed preference rules determined on the basis of proximity yield average response times which are near the optimum for that system configuration. The most useful result of the hypercube analysis could very well be the application of similar techniques to other systems. Applications in communications and medicine are suggested by Jarvis and Larson (Ref. 29). Since the optimization was formulated in terms of a general cost structure, it could be applied to these new applications with, perhaps, more significant results than those obtained for spatially distributed systems.

The location model and approximation procedure offer much promise but also require a substantial amount of further investigation. Insofar as the approximation is concerned, an issue of immediate concern is its robustness or applicability in describing widely varying systems. One approach for resolving this question might be the use of a set of simulation experiments. At the present time, there appear to be no analytic models substantially different from those presented here which might serve as a basis for validation. As an alternative to simulation, the approximation could be compared to historical data if that data was sufficiently detailed and reliable.

An interesting alternative to the above would be the development of analytic bounds on the error associated with the approximation. A possible approach to this problem might include some restrictions on the service time distributions. As formulated, only the error in the estimates of the FSC terms need be bounded. These comments notwithstanding, the computational experience with the approximation has been good. In addition, that procedure is a simple, inexpensive alternative to simulation as a means for explicitly incorporating travel time into a queuing analysis of spatially distributed systems.

There are no analytic problems with the location model per se. In its simplest form, the location model reduces to choosing the minimum of a finite set of numbers. When constraints such as those developed for police preventive patrol or maximum travel time are included, the location model must be solved as a linear or integer linear program respectively. Neither of these problems presents serious computational difficulties.

The main question relating to the location model deals with its use in conjunction with various descriptive models to determine optimal locations for response units in spatially distributed systems. There is no guarantee that the iterative procedure will not converge to a local minimum. In practice, it would appear advisable to use the algorithm with several different initial conditions in order to have some confidence that a global minimum has been reached. Local minima have not been a problem in the author's experience except in situations in which the distance between contiguous atoms is comparable to the overall dimensions of the region being examined or in problems exhibiting a great deal of symmetry.

In summary, the location model and the approximation procedure offer solution procedures to problems which could be treated only in part previously. Although these procedures require further investigation, initial experience has been encouraging. The approximation procedure produces estimates within a few percent of those derived from exact analytic models except where mean service times are very dissimilar. The use of the iterative location model to determine optimal locations for response units has not been complicated by convergence to local minima. Finally, both procedures are very inexpensive to use. An application of the algorithms to problems arising in large urban systems is not expected to require more than a minute or two of CPU time on large computing machines.

BIBLIOGRAPHY

(1) Berry, P., "APL/360 Primer," IBM Program Product GH20-0689-2, August 1971.

(2) Blum, E.H. and P. Kolesar, "Square Root Laws for Fire Engine Response Distance," Management Science, Vol. 19, No. 12, August 1973, pp. 1368-1378.

(3) Campbell, G.L., "A Spatially Distributed Queuing Model for Police Patrol Sector Design," Technical Report No. 75, MIT Operations Research Center, Cambridge, Massachusetts, 1972.

(4) Carter, G.M., J.M. Chaiken, and E. Ignall, "Response Areas for Two Emergency Units," Operations Research, Vol. 20, No. 3, May-June 1972, pp. 571-594.

(5) Chaiken, J.M. and R.C. Larson, "Methods for Allocating Urban Emergency Units: A Survey," Management Science, Vol. 19, No. 4, December 1972, pp. P110-P130.

(6) Chapman, S.C. and J.A. White, "Probabilistic Formulations of Emergency Service Facilities Location Problems," paper presented at the 1974 ORSA/TIMS Congerence, San Juan, Puerto Rico.

(7) Chelst, K.R., "An Interactive Approach to Police Sector Design," IRP-WP-03-74, MIT Operations Research Center, Cambridge, Massachusetts, March 1974.

(8) Chelst, K.R., "Transferring the Hypercube Queuing Model to the New Haven Police Department: A Case Study in Technology Transfer," The New York City Rand Institute, WN-9034-HUD, March 1975.

(9) Conte, S.D., Elementary Numerical Analysis, New York: McGraw-Hill Book Co., 1965.

(10) Cooper, L., "Solutions of Generalized Locational Equilibrium Models," Journal of Regional Science, Vol. 7, No. 1, Summer 1967, pp. 1-18.

(11) Cox, D.R., Renewal Theory, London: Methuen & Co., 1967.

(12) Cox, D.R. and W.L. Smith, Queues, London: Chapman and Hall, 1961.

(13) Cretin, S., "A Model of the Risk of Death from Myocardial

Infarction," IRP-TR-09-74, MIT Operations Research Center, Cambridge, Massachusetts, November 1974.

(14) "Documentation: Application of the Hypercube Model, Sector Design Analysis," Quincy Police Department, Quincy, Massachusetts, April 1975.

(15) Drake, A.W., Fundamentals of Applied Probability Theory, New York: McGraw-Hill Book Co., 1967.

(16) Drake, A.W., R.L. Keeney, and P.M. Morse (ed.), Analysis of Public Systems, Cambridge, Massachusetts: The MIT Press, 1972.

(17) Feller, W., An Introduction to Probability Theory and its Applications, Vol. 1 (3rd edition), New York: John Wiley & Sons, 1968.

(18) Feller, W., An Introduction to Probability Theory and its Applications, Vol. 2 (2nd edition), New York: John Wiley & Sons, 1971.

(19) Fitzsimmons, J.M., "A Methodology for Ambulance Deployment," Emergency Medical Systems Project Working Paper EMS-71-9-W, Graduate School of Business Administration, UCLA, August 1971.

(20) Garfinkel, R.S. and G.L. Nemhauser, Integer Programming, New York: John Wiley & Sons, 1972.

(21) Hall, W.K., "The Application of Multifunction Stochastic Service Systems to Allocating Ambulances to an Urban Area," Operations Research, Vol. 20, No. 3, May-June 1972, pp. 558-570.

(22) Handler, G.Y. "Minimax Network Location Theory and Algorithms," Technical Report No. 107, MIT Operations Research Center, Cambridge, Massachusetts, November 1974.

(23) Hogg, J.M., "The Siting of Fire Stations," Operational Research Quarterly, Vol. 19, 1968, pp. 275-287.

(24) Howard, R.A., Dynamic Programming and Markov Processes, Cambridge, Massachusetts: The MIT Press, 1960.

(25) Ignall, E., "Response Groups of Fire-Fighting Units I - Theory," The New York City Rand Institute, WN-7561-NYC, October 1971.

(26) Jarvinen, P., J. Rajala, and H. Sinervo, "A Branch and

Bound Algorithm for Seeking the p-Median," Operations Research, Vol. 20, No. 1, January-February 1972, pp. 173-178.

(27) Jarvis, J.P., "A Procedure for Determining Optimal Assignments in the Hypercube Model: User's Guide," MIT Operations Research Center, Cambridge, Massacusetts (to appear).

(28) Jarvis, J.P., "Optimal Dispatch Policies for Urban Server Systems," IRP-TR-02-73, MIT Operations Research Center, Cambridge, Massachusetts, September 1973.

(29) Jarvis, J.P. and R.C. Larson, "Optimal Server Assignment Policies in M/M/N:0 Queuing Systems with Distinguishable Servers and Customer Classes," IRP-WP-06-74, MIT Operations Research Center, Cambridge, Massachusetts, April 1974.

(30) Jarvis, J.P, and M.A. McKnew, "Data Collection and Computer Analysis for Police Manpower Allocation," MIT Operations Research Center, Cambridge, Massachusetts, (to appear).

(31) Jarvis, J.P, K.A. Stevenson, and T.R. Willemain, "A Simple Procedure for the Allocation of Ambulances in Semi-Rural Areas," IRP-TR-13-75, MIT Operations Research Center, Cambridge, Massachusetts, March 1975.

(32) Karlin, S., A First Course in Stochastic Processes, New York: The Academic Press, 1969.

(33) Keeney, R.L., "A Method for Districting among Facilities," Operations Research, Vol. 20, No. 3, May-June 1972, pp. 613-618.

(34) Keeney, R.L., "Preferences for the Response Times of Engines and Ladders to Fires," The New York City Rand Institute, R-509-NYC, June 1970.

(35) Kendall, M.G. and P.A.P. Moran, Geometrical Probability, London: Charles Griffen and Company, 1963.

(36) Larson, R.C., "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services," Computers and Operations Research, Vol. 1, No. 1, March 1974, pp. 67-95.

(37) Larson, R.C., "Computer Program for Calculating the Performance of Urban Emergency Service Systems: User's Manual (Batch Processing) Program Version 75-001

(Batch)," IRP-TR-14-75, MIT Operations Research Center, Cambridge, Massachusetts, March 1975.

(38) Larson, R.C., "Illustrative Police Sector Redesign in District 4 in Boston," Urban Analysis, Vol. 2, 1974, pp. 51-91.

(39) Larson, R.C., "Models for the Allocation of Urban Police Patrol Forces," Technical Report No. 44, MIT Operations Research Center, Cambridge, Massachusetts, 1969.

(40) Larson, R.C., "Urban Emergency Services: An Approximate Method for Computing Operational Performance Measures," to appear in Operations Research.

(41) Larson, R.C., Urban Police Patrol Analysis, Cambridge, Massachusetts: The MIT Press, 1972.

(42) Larson, R.C. and K.A. Stevenson, "On Insensitivities in Urban Redistricting and Facility Location," Operations Research, Vol. 20, No. 3, May-June 1972, pp. 595-612.

(43) Lipsett, F.R. and J.G. Arnold, "Computer Simulation of Patrol Operations of a Semi-Rural Police Force," Journal of Police Science and Administration, Vol. 2, No. 2, 1974, pp. 190-207.

(44) Mirchandani, P., "Location of Facilities on a Stochastic Network," IRP-WP-10-74, MIT Operations Research Center, Cambridge, Massachusetts, September 1974.

(45) Mogielnicki, R.P., K.A. Stevenson, and T.R. Willemain, "Patient and Bystander Response to Medical Emergencies," IRP-TR-05-74, MIT Operations Research Center, Cambridge, Massachusetts, July 1974.

(46) Morse, P.M. and L.W. Bacon (ed.), Operations Research for Public Systems, Cambridge, Massachusetts: The MIT Press, 1967.

(47) Odoni, A.R., "Alternative Schemes for Investigating Markov Decision Processes," Technical Report No. 28, MIT Operations Research Center, Cambridge, Massachusetts, 1967.

(48) Odoni, A.R., "Location of Facilities on a Network: A Survey of Results," IRP-TR-03-74, MIT Operations Research Center, Cambridge, Massachusetts, April 1974.

(49) Revelle, C., D. Marks, and J.C. Liebman, "An Analysis of Private and Public Sector Location Models," Management

Science, Vol. 16, No. 11, July 1970, pp. 692-707.

(50) Ross, S.M., Aplied Probability Models with Optimization Applications, San Francisco: Holden-Day, 1970.

(51) Savas, E.S., "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service," Management Science, Vol. 15, No. 12, August 1969, pp. B608-B627.

(52) Scott, A.J., "Dynamic Location-Allocation Systems: Some Basic Planning Strategies," Department of Geography, University of Toronto, Discussion Paper No. 3, November 1969.

(53) Sevast'yanov, B.A., "An Ergodic Theorem for Markov Processes," Theory of Probability and its Applications, Vol. 2, No. 1, 1957, pp. 104-112.

(54) Shannon, R.E. and J.P. Ignizio, "A Heuristic Programming Algorithm for Warehouse Location," AIIE Transactions, Vol. 2, No. 4, 1970, pp. 334.

(55) Swersey, A.J., "A Markov Decision Model for Deciding How Many Units to Dispatch," The New York City Rand Institute, WN-7743-NYC, June 1972.

(56) "Task Force Report: Science and Technology, A Report to the President's Commission on Law Enforcement and Administration of Justice," U.S. Government Printing Office, Washington, D.C., 1967.

(57) Thomas, E.M.P. and S.F. Bernstein, "A Model for Evaluating the Operation of Regional Ambulance Systems," MIT Operations Research Center (IRP), Cambridge, Massachusetts, unpublished working paper, June 1975.

(58) Toregas, C., C. Revelle, R. Swain, and L. Bergman, "The Location of Emergency Service Facilities," Operations Research, Vol. 22, No. 6, November-December 1974, pp. 1363-1373.

(59) Volz, R.A., "Optimal Ambulance Location in Semi-Rural Areas," paper presented at the 1970 ORSA/TIMS Conference, Detroit.

(60) Walker, W., "Using the Set Covering Problem to Assign Fire Companies to Fire Houses," Operations Research, Vol. 22, No. 2, March-April 1974, pp. 275-277.

(61) Weissberg, R.W., "Using the Interactive Hypercube Model,"

MIT Operations Research Center (IRP), Cambridge, Massachusetts, to appear 1975.

(62) Wesolowsky, G.O. and W.J. Truscott, "The Multiperiod Location-Allocation Problem with Relocation of Facilities," McMaster University, Hamilton, Ontario, unpublished paper.

(63) White, J.A. and K.E. Case, "On Covering Problems and the Central Facilities Location Problem," Geographical Analysis, Vol. 6, No. 3, 1974, pp. 281-293.

(64) Willemain, T.R., "The Status of Performance Measures for Emergency Medical Services," IRP-TR-06-74, MIT Operations Research Center, Cambridge, Massachusetts, July 1974.

APPENDIX A

An alphabetical list of variable names and mnemonics with the section (S) or equation (E) where the variable is first used.

AWL: average unit workload; S3.E

B: event that a server is busy; E6.4

BUS: server busy with specified customer; S4.C.1

C: cost of assignment; S3.B.2

CO: objective coefficient in location model; E5.6

CP: cost of assignment given unit position; S5.B.1

CR: call rate for each customer class; S3.B.1

CRT: call rate, total; E3.3

CS: cost for a customer arriving during saturation; E3.10

CT: expected cost per transition; E3.13

DP: dispatch preference matrix; S6.B.2

DT: dispatch delay time; S7.B

E: expectation of a random variable

EC: expected cost per customer; E3.10

ED: expected response distance; S3.E

ETC: expected transition cost; E3.12

F: event that a server is free; E6.4

FA: fraction of acceptable responses; S5.C.1

FC: fraction of calls from each atom; E5.2

FSC: fraction of responses by server for each class; E3.9

FT: fraction of total services by server and class; E5.4

FU: fraction of calls serviced by each unit; E6.20

LHS: left hand side

MAC: maximum acceptable cost; S5.C.1

MAF: minimum acceptable fraction; S5.C.1

N: number of servers; S3.B.1

NC: number of customer classes; S3.B.1

OSS: on-scene service time; S7.B

PC: steady state probabilities, convolution model; E4.1

PD: probability of patient delay; S7.D

PE: steady state probabilities, exponential model; E4.2

PH: steady state probabilities, hypercube model; E3.5

POL: state policy vector, hypercube model; S3.B.2

Pr: probability of an event

Q: correction factor; E6.8

RC: service rates, convolution model; S4.B.1

RD: risk of death following an MI; E7.3

RE: service rates, exponential model; S4.B.2

RH: service rates, hypercube model; S3.B.2

RHS: right hand side

S: saturation event, all units busy; S5.B.1

s: saturation state, hypercube model, (2**N)-1; S3.B.2

SRT: service rate, total; E3.6

SV: state value; E3.13

TA: average service time; E6.17

TC: transition cost; S3.C.1

TP: transition probability; S3.C.1

TRN: hospital transfer time; S7.B

TS: average travel speed; S7.B

TSC: expected service time by server and customer; S4.C.2

TT: inter-atom travel time matrix; S7.B

TU: average service time by unit; E6.19

U: utilization; S6.B.3

UO: unit preference order matrix; S6.B.4

UP: probability of unit positions; S5.B.1

VC: state vector, convolution model; S4.B.1

VE: state vector, exponential model; S4.B.2

VH: state vector, hypercube model; E3.1

WL: server workload; E3.8

APPENDIX B

A recursive procedure for the calculation of the correction factors, Q{N,U,k}, where

$$(6.8) \qquad Q\{N,U,k\} = \sum_{j=k}^{N-1} \frac{(N-j) * (N**j) * (U**(j-k))}{(j-k)!}$$

$$* \frac{P0 * (N-k-1)!}{((1-PN)**k) * N! * (1-U*(1-PN))}$$

Define

$$SM = \sum_{j=0}^{N} ((U*N)**j) \ / \ j!$$

$$PR = ((U*N)**N) \ / \ N!$$

$$PN = PR \ / \ SM$$

$$P0 = 1 \ / \ SM \ .$$

Define H, G, and F recursively by

$$H\{N,U,0\} = PR \ / \ (N*U)$$
$$H\{N,U,k+1\} = H\{N,U,k\} \ / \ U$$

$$G\{N,U,0\} = (SM - PR) \ / \ N$$
$$G\{N,U,k+1\} = N * (G\{N,U,k\} - H\{N,U,k\}) \ / \ (N-k-1)$$

$$\begin{aligned} F\{N,U,0\} &= SM * (1-U*(1-PN)) \\ F\{N,U,k+1\} &= N * (F\{N,U,k\} - G\{N,U,k\}) / (N-k-1) \end{aligned}$$

for k=0,1,...,N-2.

Then

$$Q\{N,U,k\} = \frac{P0 * F\{N,U,k\}}{(1-U*(1-PN)) * ((1-PN)**k)}$$

for k=0,1,...,N-1.

BIOGRAPHICAL NOTE

James P. Jarvis is a research assistant in the "Innovative Resource Planning in Urban Public Safety Systems" project at the Massachusetts Institute of Technology and has taught in the Department of Electrical Engineering at that institution. Mr. Jarvis received his B.S. in Mathematics from the University of North Carolina at Chapel Hill in 1971 and his S.M. from the Department of Electrical Engineering at the Massachusetts Institute of Technology in 1973. He is a member of ORSA and SIGMA XI. His research interests include the development of stochastic models and associated techniques for computer analysis as applied to the provision of emergency services. His publications include "Optimal Dispatch Policies for Urban Server Systems," "Optimal Server Assignment Policies in M/M/N:0 Queuing Systems with Distinguishable Servers and Customer Classes" (with R.C. Larson), "Data Collection and Computer Analysis for Police Manpower Allocation" (with M.A. McKnew), and "A Simple Procedure for the Allocation of Ambulances in Semi-Rural Areas" (with K.A. Stevenson and T.R. Willemain), all available from the MIT Operations Research Center.