THE UNIVERSITY OF CALGARY

LOCATION-ALLOCATION ON CONGESTED NETWORKS

by

RONALD R. MANDOWSKY

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF BUSINESS ADMINISTRATION

FACULTY OF MANAGEMENT

CALGARY, ALBERTA

MARCH, 1984

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled, "Location-Allocation on Congested Networks" submitted by Ronald R. Mandowsky in partial fulfillment of the requirements for the degree of Master of Business Administration.

Dr. O. Berman, Supervisor
Faculty of Management

Edward A. Silver

Dr. E. Silver,
Faculty of Management.

Dr. C. Van de Panne,
Department of Economics.

MARCH, 1984

ABSTRACT

This thesis deals with location-allocation in networks under conditions of congestion. The problem is to simultaneously find the optimal districting policy which determines how a region should be partitioned into separate service areas, and the optimal locations of facilities to house mobile service units.

In the event of congestion of service demands, i.e. the arrival of one or more calls for service while the server is busy, these calls enter a queue which is depleted according to a first-come-first-served discipline. State-of-the-art research treating service networks under conditions of congestion has provided an analytic solution for the optimal location of a single mobile server and a heuristic solution method for 2-server network districting.

An alternate location and allocation solution improvement procedure is developed to combine the location algorithm of a single mobile service unit with districting heuristics for two or more servers. This heuristic procedure which is a monotone-decreasing convergent process may or may not result in the optimal solution. For two servers the previously developed single server location algorithm and the 2-server districting heuristic are used in the 2-server location allocation heuristic. In order to find location-allocation policies for more than two servers the 2-server districting heuristic is extended to treat networks with three or more servers. The presentation of a solution method for the general case of m servers and n nodes follows the development of heuristic solutions for the 2 and 3 server location-allocation problems.

The significance of the methods presented in this thesis increases with the degree of congestion. While for low demand rates the expected response time is not very sensitive to changes in location-allocation policies, at high rates of demand even slight changes in location or districting policies could be "disasterous".

Solution procedures in the form of flowcharts are presented as well as computational results for up to five servers and twenty-five nodes.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. Oded Berman for his constant encouragement and invaluable assistance throughout our association. His personality and friendship made this task an enjoyable one.

I am grateful to Karen Lehman for her patience and dedication in typing the many drafts of this thesis.

This study would not have been possible without the understanding and moral support of Tsipi, my wife, who bravely endured living with me throughout this period.

This study was supported by National Sciences Engineering Research Council (NSERC), Canada, grant number A4978.

להורי,

לציפי ולאורי

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

This thesis deals with the location-allocation (LA) problem in congested networks. The LA problem may generally be defined as simultaneously finding the optimal districting (allocation, zoning, area designation) policy, which determines how a region should be partitioned into separate service areas, and the optimal locations of facilities to house the service units. LA policy decisions apply to public emergency services such as police, ambulance and fire services as well as to business services such as courier, taxi and road maintenance. The nodes of the network are demand points while service units may be located anywhere on the network (on nodes or on links) and travel along links representing transportation arteries of an urban environment.

Ahituv and Berman[1] consider the hierarchy of policy decisions in service networks, and state that the positioning of these policies in the hierarchy depends mainly on financial and political implications. The LA policy making process which has a long-term impact on expenditures and is highly subject to political considerations, captures the top level of this hierarchy. In addition, the consequences of LA policies prevail over the long-term and the degree of flexibility that is available to modify these policies, once implemented, is very low. Thus, the choice of reliable LA models for

policy decisions is of great importance.

Studies in recent years have dealt with LA both in the context of planar and network problems. However, no attempt has been made to present a method of optimal LA taking congestion into account. Solving this problem with the consideration of congestion, i.e. the arrival of one or more calls for service while the server is busy, is more realistic. In such a case, demands for service may either be rejected or placed in a queue.

The purpose of this thesis is to solve the above problem, allowing the queueing of service demands to take place. The consideration of congestion in location and allocation models makes them highly complex. An analytic solution exists for the optimal location of a single mobile server on a network allowing the queueing of service demands, but the same problem for two servers cannot be solved analytically. A solution of the districting problem with congestion has only been provided for two servers (and two districts) using a heuristic rather than an exact algorithm. Here, the problem is extended to combine the location algorithm with the heuristic solution to the districting problem for two and more servers. The approach is to begin with a solution to the LA problem with two mobile servers. The next step is to find a solution to the districting problem with three mobile service units and to combine this solution with the location algorithm in such a way as to obtain a LA policy for three servers. Finally, an attempt is made to present a heuristic

solution for the general case of n nodes and m servers.

Note should be made to the fact that heuristics are used both for the solution of the districting problem and the LA problem. Therefore, although reference is made throughout this thesis to "optimal" districting and LA policies or solutions, they may or may not be equal to the global optimum.

The next section reviews some of the relevant research done in this area. Section 1.3 summarizes the two papers which provide a basis for the development of this thesis and the last section of this chapter describes the content of the thesis.

1.2 A Brief Literature Review

1.2.1 General Discussion

Research of LA problems covers a wide range of areas and may be subdivided into various categories. In terms of the type of system dealt with, solution methods have been suggested for public emergency and non-emergency as well as business services, manufacturing systems, distribution systems and LA problems related to financial planning. Studies in this area may also be categorized according to the type of representation given to the modeled system, i.e. as a planar or a network problem. Some studies deal with the solution of multicriteria LA problems while most are based on the single objective of average or maximum cost minimization.

By their nature, LA solution methods treat the problem of facility location as well as the districting problem. This literature review first deals with studies on facility location and is followed by a review of some studies related to districting. Finally, research dealing with various LA problems is discussed.

1.2.2 Facility Location Studies

Studies in the field of location theory have dealt both with planar and network problems. When it is assumed that facilities can locate anywhere on a plane, a co-ordinate system is used to specify locations. This assumption is appropriate when the transportation network is highly developed and there is no restriction as to the location of facilities on the plane. Network location problems constrain facility locations to the network only, and for this type of problem the additional data required is presented in form of a distance matrix. As this thesis deals with networks only, the review of location literature is limited to location problems on networks.

A recent paper by Tansel, Francis and Lowe[28] presents the state-of-the-art in network location theory. It includes approximately 100 references, roughly 60 dating from 1978. It divides the studies into various categories and points out relations between them. The authors devote a major portion of their survey to single objective location problems through a discussion on studies related to the p-centre and p-median problems with and without mutual

communication (cooperation) between servers. The paper further treats the distance constraints problem which involves locating new facilities on a network so that they are within specified distances of existing facilities as well as within specified distances of one another, multiple objective location problems and path location problems.

A text by Handler and Mirchandani[11] discusses recent developments in median and centre problems as well as other location problems such as those dealing with multiobjective location, congestion and multistop location (or the traveling salesman location problem).

Related directly to this thesis are publications dealing with optimal location in stochastic service systems. Jarvis[13] discusses spacially distributed queueing systems which are closely related to public safety systems. He presents an iterative procedure which seeks the optimal locations for facilities providing service under conditions of congestion and cooperation among the servers. Berman and Larson[4] treat the case of congestion and extend Hakimi's median to the uncertain availability of servers to a random call. In this paper they treat the case of cooperation between servers but develop results just for the special case in which on-scene service time is much greater than travel time for all servers. Berman, Larson and Chiu[3] present an analytic solution to the case of a single mobile server on a network where stochastic calls for service enter a queue when the server is not available. For the one facility one server case, the problem is solved in its full generality. This paper is discussed in detail in Sub-Section 1.3.1.

1.2.3 Studies on Districting

The treatment of districting has not been as extensive as that of facility location. Some research deals solely with districting while others treat this subject in the wider context of LA policy decisions and are discussed in Sub-Section 1.2.4.

It should be noted that throughout this thesis and in the literature surveyed, the concept of districting is interchangeable with the concepts of zoning, area designation and allocation. All these refer to the partitioning of the whole region into separate service areas based on one or more criteria.

Larson and Stevenson[16] deal with the sensitivity of expected response times to the design of response areas. Keeney[14] presents a method of determining response areas based on the criterion of dispatching the closest service unit. Ignall[12] treats the allocation problem with specific reference to fire services with the criterion of minimizing response distance subject to constraints on workload imbalances. Berman and Wand[6] consider optimal response areas based on two objectives: the minimization of the expected cost of the operating system and the minimization of the percentage of customers that will not be served within some pre-specified time period. Carter, Chaiken and Ignall[7] deal with the issue of cooperation between two urban emergency service units serving separate response areas. They determine district boundaries that minimize average response time and those that equalize workload. Jarvis[13] presents a

procedure for determining dynamic allocations of servers to customers for the general case of m servers and notes similar results to those of Chaiken et al. for the 2-server case. His analysis is based on the continuous time Markov "hypercube" model developed by Larson[17].

A recent study by Berman and Larson[5] focuses on 2-server districting in stochastic networks. The paper contains exact methods for finding the optimal districting policy for low or high (feasible) demand rates, and for a general demand rate, a heuristic is presented. This paper is further discussed in Sub-Section 1.3.2.

1.2.4 Studies on Location-Allocation

Most studies on location-allocation do not restrict the location of facilities to a network and rather deal with locations on a plane. As previously stated, this topic has been dealt with in a variety of contexts some of which are presented here.

LA problems were first formulated by Cooper[8] who later developed heuristic solutions to these problems[9] and an exact solution to the transportation LA problem.[10] Cooper was also the first to formulate the "Alternate Method"[9] for solving LA problems which is used in this thesis in the context of congestion. Beaumont[2] presents a review and discussion of LA models in the plane dividing them into median problems, centre problems, covering problems, hierarchial LA models and dynamic LA models.

Neebe[23] presents a branch and bound algorithm for the p-median

transportation problem. Madsen[19] surveys methods of solving combined location-routing and LA-routing problems. Love[18] describes an algorithm which locates variable facilities in relation to existing facilities situated on one route and performs the allocation simultaneously with the location. Wesolowsky and Truscott[30] formulate a dynamic (multiperiod) LA model with the objective of specifying a plan for facility locations and relocations and for optional allocations of demands which minimize static distribution costs and the costs of relocating facilities.

Solutions to large scale LA problems are treated by Nambiar, Gelders and Van Wassenhove[21] through a heuristic approach to the location of processing factories, the siting of collection stations and the vehicle routing problem of transporting raw material to the central processing factories. The problem described involves 15 processing factories, 300 collection stations and 50 vehicles. Murtagh and Niwattisyawong[20] describe a procedure based on large-scale non-linear programming for solving the multi-depot LA problem. In this paper both the locations of depots and the allocation of customers are allowed to vary simultaneously and numerical experience with 5 and 10 depots and 50 customers is described.

LA in the context of emergency medical services is treated by Or and Pierskalla[24] for regional blood banking. They present algorithms to decide how many blood banks to set up, where to locate them, how to allocate hospitals to the banks and how to route a periodic supply operation. Narula and Ogbu[22] suggest several heuristics to solve a

2-hierarchial LA problem where p1 health centres and p2 hospitals are to be located among n potential locations so as to minimize total weighted travel distance.

The treatment of LA problems under stochastic conditions has been very limited. Sheppard[25] formulates a general conceptual framework for dealing with uncertainty in LA problems and Stidham[26] deals with the application of probabilistic demand to static LA models.

1.3 Summary of Research Leading to this Study

1.3.1 Location of the Stochastic Queue Median

Berman, Larson and Chiu[3] deal with the problem of finding the optimal location for a single mobile server and extend Hakimi's one-median problem to include the treatment of congestion. In the median problem the objective is to minimize the average distance (or the average travel time or the average cost) traveled by the server (or servers) in the case of mobile servers and static demand points. While Hakimi's work took into account only the probabilistic spatial nature of service demands, Berman et al. also consider a probabilistic arrival time process for service demands and probabilistic service times. The incorporation of temporal as well as spatial uncertainties under conditions of congestion may be treated either by service demands being rejected or placed in a queue due to the unavailability of the server. In the case of service demands placed in a queue the

criterion for server location is the minimization of the average time of response which is the sum of mean in-queue delay and mean travel time.

When congestion is not considered, the optimal server location exists on a node of the network, and an analogous result exists for the multi-median problem. In the case of a network incorporating an M/G/1 queuing system (Poisson input, general [independent] service times, and a single server) operating in steady state, it was found that for the case of zero queue capacity the optimal facility location is at the median, and that for the infinite queue capacity case the optimal solution can either be on a node or on a link.

The problem is described as follows: Service demands occur at the nodes of an undirected network $G(N,L)$ with node set $N(|N|=n)$ and link set L. Each node j generates an independent stream of demands with rate λh_j ($\sum_{j=1}^{n} h_j=1$), where $\lambda \geq 0$ is the network-wide demand rate and h_j is the probability that a random service demand originates at node j. Travel distance from point $x \in G$ to node $j \in N$ is $d(x,j)$ and travel distance on link (i,j) is d_{ij}. A single mobile server is located at a facility at point $x \in G$ and can be dispatched for immediate service whenever it is located and free at point x. From the point in time that the server is dispatched to node j there is a travel time of $d(x,j)/v$ where v is the travel speed. The server then incurs an expected on-scene service time of $\overline{R}_j$, a return travel time equal to $(\beta-1)d(x,j)/v$ where $\beta \geq 1$ and an expected off-scene service time of $\overline{W}_j$ (for a given node j,

R_j and W_j are assumed to be independent random variables). The components of the service time are shown in Figure 1.1.

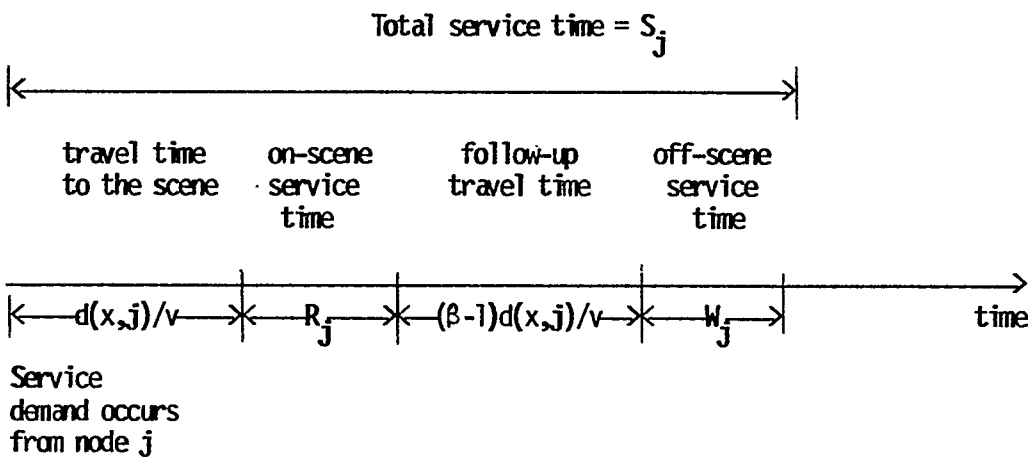


FIGURE 1.1 Time Sequence for a Demand for Service

If additional demand occurs whenever the server is busy, the new demand enters a queue that is depleted according to a first-come-first-served (FCFS) discipline. Therefore, given facility location x, the expected response time $\overline{T}_R(x)$ associated with a random service demand is the sum of the mean in-queue delay $\overline{W}_q(x)$ and the expected travel time $\overline{t}(x)$.

$$\overline{T}_R(x) = \overline{W}_q(x) + \overline{t}(x)$$

If the mobile server is assumed to be located between nodes a and b at a distance x from node a (where ℓ is the length of link (a,b) and d(i,j) is the shortest distance between nodes i, j ε N), then the mean service time is given by:

$$\overline{S}(x) = \overline{R}_j + \overline{W}_j + \frac{\beta}{v} \left[\sum_{j=1}^{n} h_j \min\{x + d(a,j);\ (\ell - x) + d(b,j)\} \right]$$

and the expected travel time to the site of the demand is given by:

$$\overline{t}(x) = \frac{1}{v}[\sum_{j=1}^{n} h_j \min\{x + d(a,j);\ (\ell-x) + d(b,j)\}]$$

For an M/G/1 queueing system the mean in-queue delay is given by:[15]

$$\overline{W}_q(x) = \begin{cases} \dfrac{\lambda \overline{S^2}(x)}{2(1-\lambda\overline{S}(x))} & \text{For } \lambda\overline{S}(x) < 1 \\ +\infty & \text{For } \lambda\overline{S}(x) \geq 1 \end{cases}$$

where: $\overline{S^2}(x) = \sum_{j=1}^{n} h_j E[\frac{\beta}{v} \min\{x + d(a,j);\ (\ell-x) + d(b,j)\} + R_j + W_j]^2$

and E denotes the expectative operator. The objective is to find x^* such that

$$\overline{T}_R(x^*) \leq \overline{T}_R(x)\ \forall\ x \in (a', b'),\ (a', b') \in L$$

x^* is called a stochastic queue median.

Next, the concept of breakpoints is introduced. The node set N may be divided into two disjoint sets A and B as follows:

$$A = \{j;\ x + d(a,j) \leq (\ell-x) + d(b,j)\};\ B = N-A$$

When changing x along the link (a,b) the sets A and B may change. Breakpoints are all the points on G(N,L) at which the sets A and B change.

It was found that $\overline{T}_R(x)$ is a continuous piece-wise convex differentiable function of x for any interval on link (a,b) bounded by two breakpoints when it is finite, and that the only points of

nondifferentiability are at the breakpoints. Therefore, given any interval $[x_1, x_2]$ where x_1 and x_2 are adjacent breakpoints, if the right derivative of $\overline{T}_R(x)$ at $x=x_1$ is negative and the left derivative at $x=x_2$ is positive then $\overline{T}_R(x)$ has a local minimum over (x_1,x_2). This local minimum can be calculated analytically.

The algorithm for finding the stochastic queue median is based on the local convexity of $\overline{T}_R(x)$ and the method for finding breakpoints. The set of all breakpoints is calculated for each link of the network, the minimum value of $\overline{T}_R(x)$ is found for each link and compared to all other links to come up with the optimal location x^* on link $(a,b)^*$ and the corresponding mean travel time $\overline{T}_R(x^*)$. The complete algorithm is presented in the appendix to this chapter.

Examination of the variation in x^* with the total demand rate λ, from 0 to a maximum possible value λ_{max} (λ_{max} is the smallest value of λ for which the queue explodes for all possible server locations) shows that the trajectory of the optimal locations $x^*(\lambda)$ starts at the Hakimi median when $\lambda=0+$ and returns to the median as λ approaches λ_{max}. Mean travel time $\overline{t}(x)$ dominates the solution for low values of λ and the denominator of the expression for the mean in-queue delay dominates for high values of λ. For intermediate values of λ, the numerator of the mean in-queue delay plays an important role in determining x^*.

The paper[3] also includes a heuristic for finding the optimal location. The advantage of this heuristic is that not all links of the network have to be considered but only those links that lie on an

"assumed feasible trajectory" of the optimal solution. Computational experience has shown that the heuristic provides identical results to those derived by the exact algorithm.

1.3.2 2-Server Network Districting in the Presence of Queueing

A second source which serves as a basis for this thesis is a paper by Berman and Larson[5] which deals with the problem of finding an optimal districting policy in a network that is served by two units which act independently as M/G/1 servers. The locations of the two servers are fixed and the assignment criterion of nodes to servers is the minimization of the overall expected response time to a new call for service. The solution provides two independent districts that operate as separate M/G/1 systems.

In this system $X^i \varepsilon G(N,L)$ is the fixed location of unit i,i=1,2 and a districting policy determines the partition of the set N into two sets N^1, N^2 such that $N^1 \cap N^2 = \emptyset$ with N^i being the set of nodes in the district of service unit i. No cooperation exists between the two districts and therefore, if server i is busy when a demand for service in its district occurs, the call enters a queue that is depleted according to a FCFS discipline.

For each district, the scenario described in Sub-Section 1.3.1 is appropriate. Here, for district i the expected travel time is:

$$\bar{t}^i = \sum_{j \varepsilon N^i} (h_j/h^i) d(X^i,j)/v$$

where $h^i = \sum_{j \varepsilon N^i} h_j$ and the mean in-queue delay is given by:

$$\overline{W_q^i} = \frac{(\lambda h^i)\overline{S^{i^2}}}{2(1-(\lambda h^i)\overline{S^i})}$$

where $\overline{S^i}$ is the expected total service time for server i:

$$\overline{S^i} = \sum_{j \varepsilon N^i} (h_j/h^i)[\beta d(X^i,j)/v + \overline{R_j^i} + \overline{W_j^i}]$$

and $\overline{S^{i^2}}$ is the second moment of the total service time for server i:

$$\overline{S^{i^2}} = \sum_{j \varepsilon N^i} (h_j/h^i)E[\beta d(X^i,j)/v + R_j^i + W_j^i]^2$$

where E is the expectation operator.

For a given policy (N^1,N^2) the overall expected response time to a random call for service, which is the sum of the mean in-queue delay and the expected travel time, is given by:

$$\overline{T_R^{1,2}} = h^1 \overline{T_R^1} + h^2 \overline{T_R^2}$$

The optimal policy $(N^1,N^2)^*$ is the one that minimizes $\overline{T_R^{1,2}}$.

The paper shows that all λ values $\lambda \geq 0$ can be divided into four regions as follows:

Region A includes all λ values, $0 \leq \lambda \leq \lambda_A$ for which the optimal solution is the same one as for $\lambda = 0$, i.e. based on the proximity of each server to the set of demand nodes.

Region D includes all λ values $\lambda \geq \lambda_{max}$ for which a feasible solution does not exist (the system blows up).

Region C includes all λ values $\lambda_C \leq \lambda < \lambda_{max}$ for which there is one optimal policy for all values in the region. This policy is found by solving the following minimax problem:

$$\min_{\substack{X_{ij}=0,1 \\ i=1,2}} \max \left\{ \sum_{j=1}^{n} X_{ij} h_j [\beta d(X^i,j)/v + \overline{R}_j^i + \overline{W}_j^i] \right\}$$

$$\text{s.t.} \quad \sum_{i=1}^{2} X_{ij}=1 \qquad j=1,\ldots,n$$

where:

$$X_{ij} = \begin{cases} 1 \text{ if node } j \text{ is assigned to server } i \\ 0 \text{ otherwise} \end{cases}$$

for $i=1,2 \quad j=1,\ldots\ldots,n$

The above minimax problem is equivalent to the following linear mixed integer programming problem:

$$\min W$$

$$\text{s.t.} \quad \sum_{j=1}^{n} X_{ij} a_{ij} \leq W \qquad i=1,2$$

$$\sum_{i=1}^{2} X_{ij} = 1 \qquad j=1,\ldots,n$$

$$X_{ij} = 0,1 \qquad i=1,2 \qquad j=1,\ldots,n$$

where $a_{ij} = h_j[\beta d(X^i,j)/v + \overline{R}_j^i + \overline{W}_j^i]$.

Region B consists of all the λ values $\lambda_A < \lambda < \lambda_C$ and for this region there may be many optimal policies for different λ values.

A flowchart of the heuristic used to solve this problem is presented in the appendix to this chapter. The heuristic starts with the choice of an initial solution which may either be the optimal policy for Region A or Region C or any other feasible policy. It involves the minimization of a non-linear objective function subject to linear and integrality constraints as follows:

$$\text{minimize} \sum_{i=1}^{2} \left\{ \frac{\left(\sum_{j=1}^{n} X_{ij}h_j\right)\lambda\left(\sum_{j=1}^{n} X_{ij}h_j E[\, d(X^i,j)/v + R_j^i + W_j^i]^2\right)}{2[1-\lambda \sum_{j=1}^{n} X_{ij}h_j(\beta d(X^i,j)/v + \overline{R}_j^i + \overline{W}_j^i]} + \sum_{j=1}^{n} X_{ij}h_j d(X^i,j)/v \right\}$$

s.t.

$$\sum_{i=1}^{2} X_{ij} = 1 \qquad j=1,\ldots,n$$

$$\sum_{j=1}^{n} X_{ij}h_j(\beta d(X^i,j)/v + \overline{R}_j^i + \overline{W}_j^i) < \frac{1}{\lambda} \qquad i=1,2$$

$$X_{ij}=0,1 \qquad i=1,2 \qquad j=1,\ldots.n$$

The "Method of Convex Combinations"[29] is used to solve the non-linear mathematical programming problem by using a linear approximation to

the objective function subject to the same constraints as the original problem, and iteratively improving the solution by finding a direction in which the objective function decreases. The solution obtained by the above method may involve at the most the splitting of one node between two servers. In such a case the solution may be expressed as a convex combination of two integer solutions that are identical except for the split node. A sequence of node switches is performed on each integer solution to search for an improved policy (lower expected response time), and the best solution of the two sequences of node switches is taken as the "optimal" solution. In case of an integer solution, this solution is taken as the best policy and no node switching is required.

1.4 Summary and Outline

1.4.1 Location-Allocation on Congested Networks

In most systems the inter-arrival times of calls for service and service times are probabilistic. Congestion may occur in the sense that all servers of the network or any subset of servers may be busy when new calls for service arrive. The case of possible unavailability of servers has not been taken into account in LA studies to date.

Studies in recent years and specifically those described in Section 1.3 have dealt with urban emergency service networks taking congestion into account. An analytic solution for the optimal

location of one mobile server and a heuristic for the 2-server districting problem in the presence of queueing have been developed, but LA problems have not been treated in this context. This work combines these solutions and extends them in order to provide models for the combined LA problem.

1.4.2 2-Server Location-Allocation

Chapter 2 deals with the LA problem for a network served by two independent mobile service units. The objective is to find the optimal split of demand nodes where the servers operate as two separate M/G/1 sub-systems and to find the optimal location for each server within its service area.

Once the best districting policy is found given a location policy, an improved location (in terms of the expected response time) for the two service units is searched for in each region. If an improved location policy is found, a new districting policy based on the last improved server locations is found, and so on until there is no improvement in the solution. Following a formal presentation of the problem and a heuristic solution method, computational results are given and discussed.

1.4.3 3-Server Location-Allocation

An extension of the study described in Sub-Section 1.3.2 deals with the problem of network districting with three mobile service

units and is presented in Chapter 3. Exact solutions to this problem may be found for low rates of demand (Region A) and high (feasible) rates of demand (Region C). For any (general) demand rate, a heuristic solution is presented which is based on the solution to the 2-server problem. The development of such a heuristic is essential for the solution of the 3-server LA problem. Combining the heuristic for the solution of the 3-server districting problem with the one server location algorithm in a similar manner as described for the two server case, results in an efficient heuristic for the solution of the 3-server LA problem. This heuristic, as well as a summary and discussion of results, are included in the third chapter.

1.4.4 Location-Allocation of m Servers and n Nodes

The final and most valuable product of this study is in the form of a heuristic to solve the LA problem for the general case of any number of servers and demand nodes.

The heuristics developed for networks with two and three servers are extended in order to treat the general case. The m-server districting heuristic iteratively improves the allocation of nodes between a different pair of servers at each step, while keeping all the other districts fixed. Chapter 4 describes the districting and LA heuristics developed for m servers, presents and discusses computational results.

A summary of this thesis and a discussion are included in Chapter 5.

APPENDIX

a. An Algorithm for Finding the Optimal Location[3]

The following algorithm for finding the optimal location x^* is based on the local convexity of $\overline{T}_R(x)$ and the method for finding breakpoints. For any differentiable function f(x) the right derivative of f(x) is defined as:

$$\dot{f}_{\leftarrow}(x) = \lim_{\Delta x \to 0} \frac{f(x) - f(x + |\Delta x|)}{|\Delta x|}$$

and the left derivative of f(x) is defined as:

$$\dot{f}_{\rightarrow}(x) = \lim_{\Delta x \to 0.} \frac{f(x - |\Delta x|) - f(x)}{|\Delta x|}$$

In the following algorithm, $\overline{T_R}^*$ is a running value for minimum mean response time, and $(a,b)^*$ and x^* denote the link and location on the link that yield that value. The algorithm is as follows:

Step 1: Set $\overline{T}_R{}^* = M$ (M very large).

Step 2: Take any link $(a,b) \varepsilon L$ and calculate the set of all breakpoints. Say that the power of this set BP is m+1, so that there are m intervals in which $\overline{T}_R(x)$ is differentiable.

Step 3: Set I = 1.

Step 4: Set $y = I^{th}$ entry in BP

Set $z = I + 1^{st}$ entry in BP

Calculate $\overline{T}_R(y)$, $\overline{T}_R(z)$, $\dot{\overline{T}}_R(y)$, $\dot{\overline{T}}_R(z)$.

If $\overline{T}_R(y) = +\infty$ and $\overline{T}_R(z) = +\infty$, $I = I + 1$ and return to the beginning of Step 4.

If $\overline{T}_R(y) = +\infty$ and $\dot{\overline{T}}_{R\rightarrow}(z) > 0$, go to Step 5.

If $\dot{\overline{T}}_{R\leftarrow}(y) < 0$ and $\overline{T}_R(z) = +\infty$, go to Step 5.

If $\dot{\overline{T}}_{R\leftarrow}(y) < 0$ and $\dot{\overline{T}}_{R\rightarrow}(z) > 0$, go to Step 5.

Otherwise compare $\overline{T}_R(y)$ and $\overline{T}_R(z)$ to $\overline{T}_R^*$. If either $\overline{T}_R(y)$ or $\overline{T}_R(z)$ is less than $\overline{T}_R^*$, update $\overline{T}_R^*$ with a new minimum and set $x^* = y$ or z (whichever yields the lower $\overline{T}_R$) and $(a,b)^* = (a,b)$.

Step 5: Calculate the local minimum x_{min} of $\overline{T}_R$ over (y,z). If $\overline{T}_R(x_{min}) < \overline{T}_R^*$ update $\overline{T}_R^*$ and record new incumbent $x^* = x_{min}$, $(a,b)^* = (a,b)$.

Step 6: If $I < m$, $I \leftarrow I + 1$ and go to Step 4. Otherwise remove (a,b) from L; if there are links remaining in L go to Step 2. Otherwise FINISH. The optimal location is x^* on link $(a,b)^*$, yielding a minimum mean travel time $\overline{T}_R(x^*)$.

b. The 2-Server Districting Heuristic[5]

The following figure depicts the heuristic solution method for the 2-server districting problem:

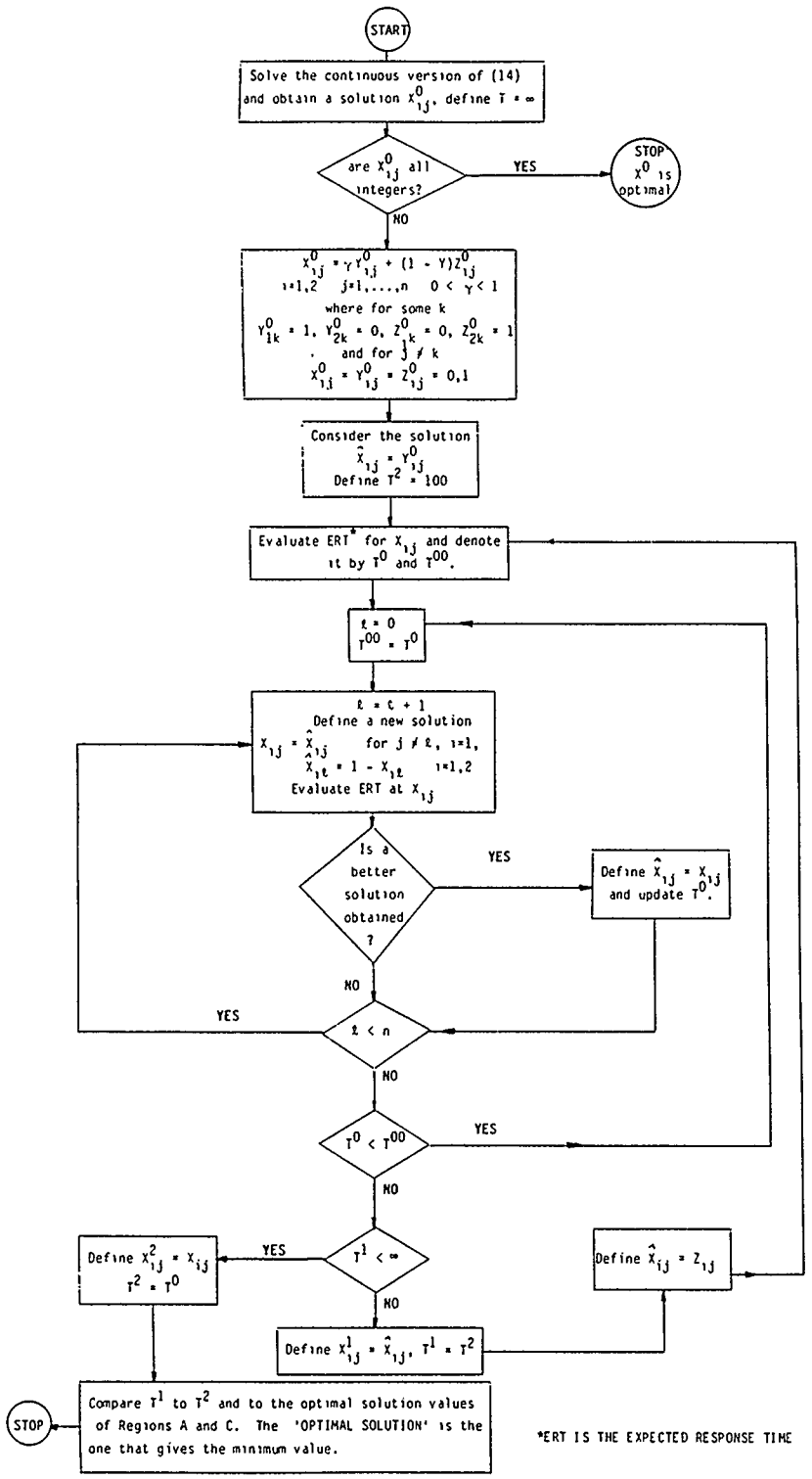


FIGURE 1.A.1 2-Server Districting Flowchart

CHAPTER 2

2-SERVER LOCATION-ALLOCATION ON CONGESTED NETWORKS

2.1 Introduction and Outline

This chapter presents a heuristic to solve the 2-mobile server LA problem allowing queueing of service demands to take place. The heuristic utilizes solution methods discussed in Sub-Sections 1.3.1 and 1.3.2 and involves an alternate location and allocation solution improvement procedure with the criterion of minimizing the overall expected time of response.

The chapter starts with the definition of the problem. Section 2.3 introduces notation and assumptions, explains the method of solution and presents a flowchart describing the heuristic solution method. The subsequent section includes illustrative examples on three different networks. The last section presents a discussion related to the computational results which were derived by a computer code in APL.

2.2 Problem Definition and Procedure

In general, the problem may be defined as simultaneously finding the optimal districting (allocation) policy which determines the

partitioning of the whole region into two independent service areas (districts) and the optimal locations of facilities to house the service units in each area. Here, the queueing of service demands takes place in the event of congestion of demands, i.e. when calls for service arrive while the server designated to serve these calls is busy. As each district is considered independent, no cooperation between servers exists. Therefore, if a call for service in one district arrives while its server is busy, it enters a queue of service demands although the server in the other district may be available at that time.

The criterion of optimality is the minimization of the overall total expected response time to a random call for service. This value is calculated over both regions taking into account the probabilistic spatial nature of the arrival of service calls as well as the temporal nature of the arrival of calls for service and total service times.

The alternate location and allocation solution improvement procedure which is described in this chapter is a logical solution technique for this problem. As the heuristic searches both for the optimal locations of servers (or locations of facilities to house the service units) and the optimal allocation of nodes to servers, once an improved districting policy is found for a given set of server locations, the last set of locations may not be the optimal one for each new district. Therefore, new optimal locations are found treating each district as an independent region. With this new set of server

locations the optimal districting policy is found and this sequence is repeated until there is no improvement in the overall total expected response time.

This is a monotone-decreasing convergent process but the final result of the heuristic may or may not be equal to the global optimum. Results presented later in this chapter indicate that in most cases, using the 2-median as the initial location policy results in the best solution.

2.3 The 2-Server Location-Allocation Heuristic

2.3.1 Notation and Assumptions

Notation and assumptions presented in this section relate not only to the material in this chapter but also to the problems discussed in subsequent chapters. Reference throughout this thesis is made to these notation and assumptions which are therefore presented here in detail.

Let G(N,L) be an undirected network with node set $N(|N| = n)$ and link set L. Service demand can occur only at the nodes of the network with each node j generating an independent Poisson stream with rate λh_j $(\sum_{j=1}^{n} h_j = 1)$. Travel distance from a point $x \in G$ to node $j \in N$ is $d(x,j)$. Travel time is equal to the travel distance divided by travel speed v.

There are two mobile servers on the network which may be located at any point on G(N,L). Let $X^i \in G(N,L)$ be the location of unit i,

$i = 1,2$. Location X^i is presented as a vector of three components: a, xs and b, where server i is located on link (a,b) at a distance of xs from a. Note that if xs^i is equal to zero, server i is located at node a and if xs^i is equal to λ, server i is located at node b.

A districting policy is defined as any partition of the set N into two sets N^1, N^2 such that $N^1 \cap N^2 = \emptyset$, with N^i being the set of nodes in the district of service unit i. Given that server i is free and given a demand from node $j \in N^i$, the server is immediately dispatched to node j. No cooperation is allowed between the two districts and therefore, if server i is not free when a demand at node j occurs, the call enters a queue which operates according to a FCFS discipline.

A LA policy $[(X^1,X^2),(N^1,N^2)]$ is a combination of a location decision (X^1,X^2) and a districting decision (N^1,N^2) as described above, with X^1 and X^2 being the locations of the servers for Districts 1 and 2, respectively. Although a restriction regarding the districting policy exists, i.e. $N^1 \cap N^2 = \emptyset$, there is no restriction as to the location policy. Therefore, X^i, i=1,2, may theoretically be located at any point on the network, even outside its own district, although this would be highly unlikely under an optimal LA policy and could occur only under conditions of extreme congestion.

A 2-server LA policy involves two independent M/G/1 queueing systems. The expected total service time for server i is:

$$\overline{S}^i = \sum_{j \in N^i} (h_j/h^i)[\beta d(X^i,j)/v + \overline{R_j^i} + \overline{W_j^i}$$

where $h^i = \sum_{j \varepsilon N^i} h_j, \quad \beta \geq 1.$

$\overline{R}_j^i$ and $\overline{W}_j^i$ are the expected on-scene and off-scene service times of server i at node j, respectively, and β accounts for the fact that a round trip back to the facility is required for each call. (e.g., if β=2 the travel time from the server's facility to the demand node is equal to the travel time back from the demand node to the facility). The second moment of the total expected service time for server i is:

$$\overline{S^{i^2}} = \sum_{j \varepsilon N^i} (h_j/h^i) E[\beta d(X^i,j)/v + R_j^i + W_j^i]^2.$$

where E is the expectation operator.

For district i the expected waiting time $\overline{W}_q^i$ is:

$$\overline{W}_q^i = \frac{(\lambda h^i)\overline{S^{i^2}}}{2(1-(\lambda h^i)\overline{S}^i)}$$

and the expected travel time to a random call is:

$$\overline{t}^i = \sum_{j \varepsilon N^i} (h_j/h^i) d(X^i,j)/v$$

Therefore, the expected response time for each district is given by:

$$\overline{T}_R^i = \overline{W}_q^i + \overline{t}^i = \frac{(\lambda h^i)\overline{S^{i^2}}}{2(1-\lambda h^i)\overline{S}^i)} + \overline{t}^i$$

For a given LA policy $[(X^1,X^2), (N^1,N^2)]$ the expected response time to

a random demand is given by:

$$\bar{T}_R^{1,2} = h^1\bar{T}_R^1 + h^2\bar{T}_R^2$$

and the objective is to find an optimal LA policy $[(X^1,X^2), (N^1,N^2)]^*$ that minimizes the expected response time, where a call for service may occur in any one of the two districts.

Here, for simplicity of presentation, whenever a server is located at a node, the three components a, xs, and b are replaced by the number of the node at which the facility is located. For example, if one server is located at Node 2 and the second is located between Nodes 3 and 5 at a distance of 2 from Node 3, this location would be represented by:

$$\binom{\text{NODE } 2}{3,2,5}$$

The regular notation for a districting policy is in the form of X_{ij} as described in Sub-Section 1.3.2. Here for easier illustration, the notation used for a districting policy is in form of two rows with the nodes appearing in row i being allocated to server i. Therefore, if in a 5-node network Nodes 1, 2 and 3 are allocated to the first server and Nodes 4 and 5 to the second, this districting policy would appear as follows:

$$\binom{1,2,3}{4,5}$$

A LA policy combining the above location and districting policies

would be represented by the following notation:

$$[(\overset{\text{NODE}\ 2}{3,2,5}),(\overset{1,2,3}{4,5})]$$

2.3.2 The Method of Solution

The alternate location and allocation solution improvement heuristic developed for the LA problem for two mobile servers on a congested network with queues incorporates the two studies described in Section 1.3.

The solution begins with the choice of an initial location policy $(X^1,X^2)_0$. The 2-facility network districting is performed according to the procedure developed by Berman and Larson[5] but with one change. The original heuristic proposes to start the heuristic with an initial feasible solution which could be the solution of Region A or Region C or any other feasible policy. Computational experience has shown that in some cases the heuristic does not provide the "optimal" solution when the better of the solutions of Regions A or C at the given demand rate is taken as the initial policy. Therefore, the procedure used here is to obtain two solutions; one which is obtained based on the solution of Region A as the initial policy and the second based on the solution of Region C. The two final results are compared and the better of the two is taken as the "optimal" districitng policy.

Given the initial location policy, an "optimal" districting policy (N^1,N^2) is found and the network is now split into two independent districts N^1 and N^2. The overall service demand rate for

district i, i=1,2 is given by $\lambda^i = \lambda \sum_{j \varepsilon N^i} h_j$ and the probability that a random call for service in district i originated at node j, $j \varepsilon N^i$ is equal to $h_j^i = h_j / \sum_{j \varepsilon N^i} h_j$. In this heuristic the optimal server location for each district is not restricted to its own district but may be located anywhere on the network. The likelihood of a server being located outside of its own district is nevertheless very small. The best server location in each district is now found independently according to the procedure described in Sub-Section 1.3.1 for locating the stochastic queue median, giving a new location policy $(X^1,X^2)_1$.

The overall expected response time $\overline{T}_R^{1,2}$ given the policy $[X^1,X^2)_1, (N^1,N^2)_1]$ may be equal to or lower than that under the policy $[(X^1,X^2)_0, (N^1,N^2)_1]$. If the expected response times under both policies are equal, no further improvement is possible and the best LA policy (which may be sub-optimal) has been found. If the last policy provides an overall expected response time which is smaller than under the previous policy, the sequence is continued. A new districting policy $(N^1,N^2)_2$ is found with the servers located according to policy $(X^1,X^2)_1$, and the overall expected response time under policy $[(X^1,X^2)_1, (N^1,N^2)_2]$ is compared to that under policy $[(X^1,X^2)_1, (N^1,N^2)_1]$, and so on.

A LA iteration K may consist of a districting policy evaluation or of the evaluation of both districting and location policies as shown in Figure 2.1. Computational experience has shown that the initial location of the servers may affect both the number of iterations

required to minimize the expected response time and the best solution that the heuristic converges onto. As shown in Section 2.4., the most "efficient" initial location policy is that of the 2-median. Intuitively, the superiority of the 2-median as the initial solution is expected. Berman et al[3] showed that for low and high feasible demand rates, the optimal server location (for one server) is at the median. Therefore, the 2-median is a "strong" starting point for low and high λ values for the two server case and in most cases as good as any other starting point for intermediate λ values.

2.3.3 A Flowchart of the Heuristic

The flowchart in Figure 2.1 depicts the heuristic used to solve the LA problem with 2 mobile servers on a congested network with queues.

A solution is found when no improvement occurs in the overall expected response time following the last policy evaluation. This can occur both following the evaluation of a districting policy or following the solution of the optimal location. In each case the expected response time under the updated LA policy is compared to that of the previous policy.

2.4 Illustrative Examples

The results presented in this section were obtained from a computer code programmed in APL.

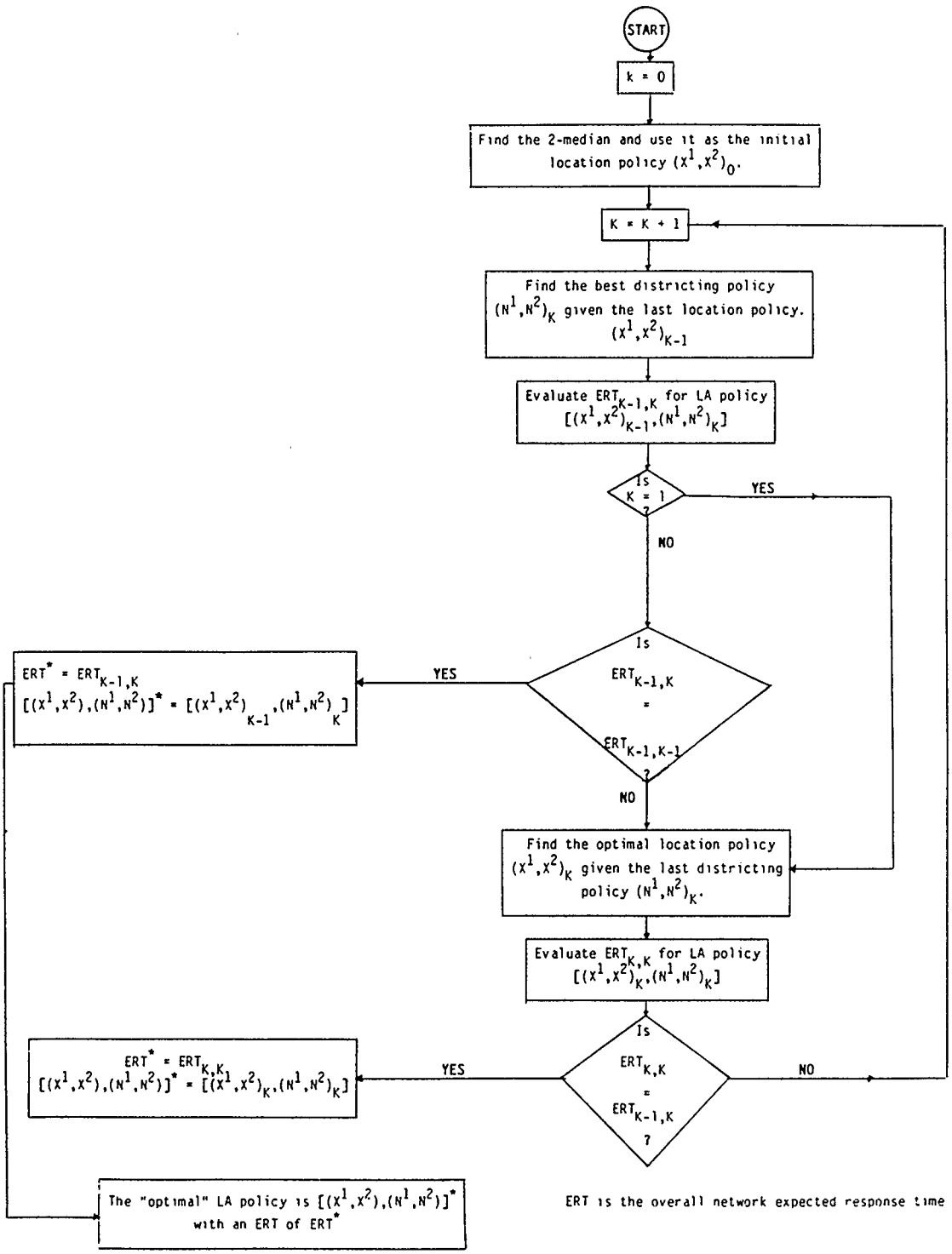


FIGURE 2.1 2-Server Location-Allocation Flowchart

Three networks which consist of 5, 10 and 25 nodes are used as illustrative examples. Initially, solutions were found for a simple 5-node network. The best policies found for the 10 and 25-node networks are more complex, less intuitive and therefore, more interesting to observe. These networks are depicted in Figures 2.2, 2.3 and 2.4, respectively. In these figures, a number adjacent to a node is the fraction of network service requests emanating from that node, and a number adjacent to a link is the link travel distance. For each network "optimal" LA policies were evaluated at various λ values and for some λ values optimal policies were evaluated starting with different initial location policies to examine the sensitivity of this heuristic to changes in the initial policy.

The results presented in Tables 2.1, 2.2 and 2.3 are all based on an initial location of the servers at the 2-median.

For the three networks the following server specifications apply for both servers: $\beta=2$, $v=1$, $R_j^i + W_j^i = 1$. (Here we assume deterministic on-scene and off-scene service times).

The results presented in the tables indicate that for low values of λ the best LA policy is to locate the servers at the 2-median and to allocate demand nodes based on their proximity to the servers. (This policy is named the median-proximity LA policy). Above a certain demand rate for each network (which will be called λ_p for the LA problems) the above LA policy becomes sub-optimal and the LA policy changes for different demand rates.

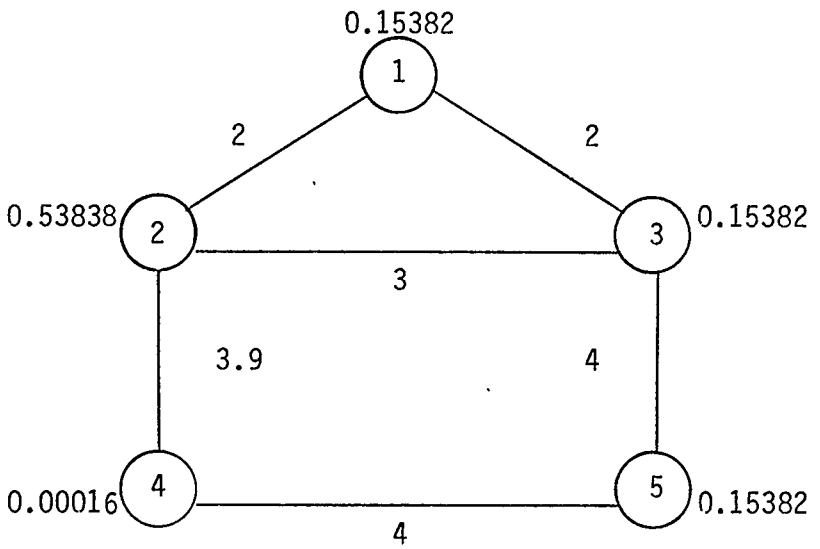


FIGURE 2.2 5-Node Network

TABLE 2.1 2 Server Location-Allocation for the 5-Node Network (2 Median = 2,5)

Demand Rate λ	Initial Location $(X^1,X^2)_0$	"Optimal" Location-Allocation Policy $[(X^1,X^2),(N^1,N^2)]^*$	"Optimal" Expected Response Time ERT*	Number of Iterations K
0.0002	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{\text{NODE 5}},\binom{1,2,3,4}{5}]$	0.770725	2
0.002	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{\text{NODE 5}},\binom{1,2,3,4}{5}]$	0.779882	2
0.01	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{\text{NODE 5}},\binom{1,2,3}{4,5}]$	0.821516	2
0.05	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{\text{NODE 5}},\binom{1,2,3}{4,5}]$	1.056641	2
0.1	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	1.238439	3
0.2	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	1.677555	3
0.3	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	2.334255	3
0.45	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	4.319237	2
0.5	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	5.687277	2
0.60	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	14.387263	2
0.65	$\binom{\text{NODE 2}}{\text{NODE 5}}$	$[\binom{\text{NODE 2}}{3,2,5},\binom{1,2,4}{3,5}]$	5006.6	2

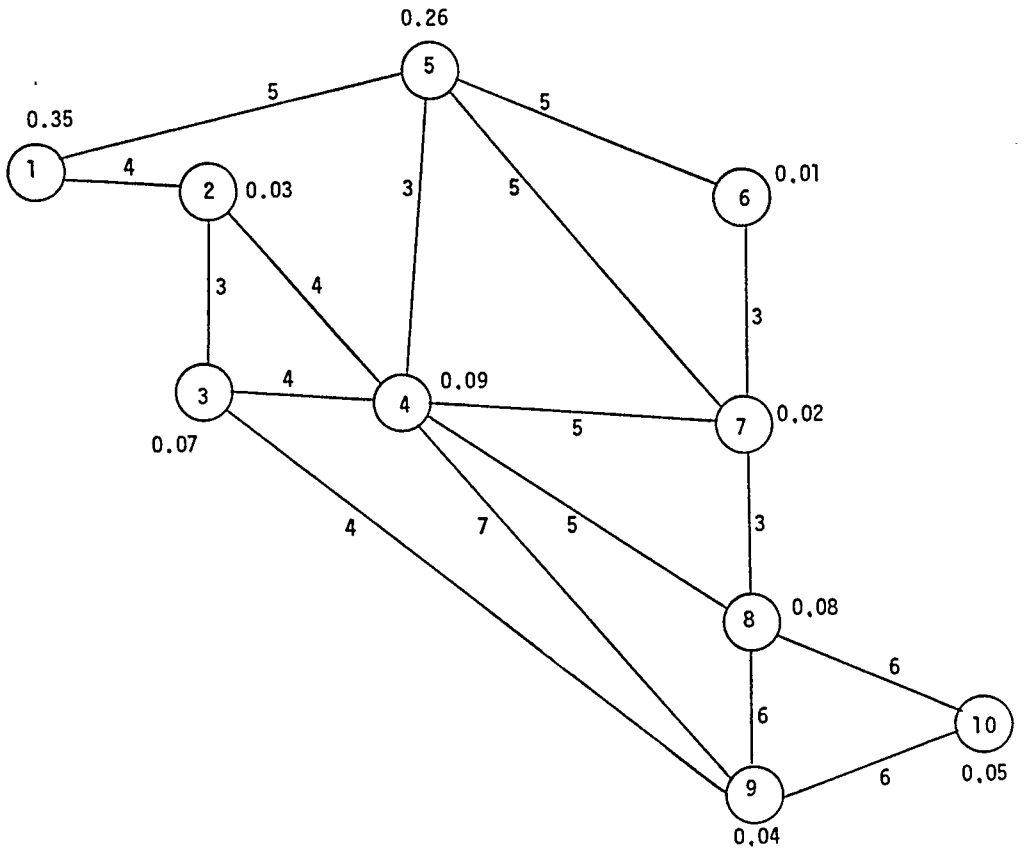


FIGURE 2.3 10-Node Network

TABLE 2.2 2-Server Location-Allocation for the 10-Node Network (2 Median = 1,4)

Demand Rate λ	Initial Location $(x^1,x^2)_0$	"Optimal" Location-Allocation Policy $[(x^1,x^2),(N^1,N^2)]^*$	"Optimal" Expected Response Time ERT*	Number of Iterations K
0.0001	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2, 3,4,5,6,7,8,9,10)]	2.592191	2
0.002	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2, 3,4,5,6,7,8,9,10)]	2.634281	2
0.01	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2, 3,4,5,6,7,8,9,10)]	2.821611	2
0.025	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2, 3,4,5,6,7,8,9,10)]	3.223953	2
0.05	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2,6, 3,4,5,7,8,9,10)]	4.062005	3
0.075	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2,3,6, 4,5,7,8,9,10)]	5.182217	2
0.10	(NODE 1, NODE 4)	[(1,1.5382,5, NODE 4),(1,5, 2,3,4,6,7,8,9,10)]	6.193862	3
0.15	(NODE 1, NODE 4)	[(1,1.0950,5, NODE 4),(1,2,5 3,4,6,7,8,10)]	9.707420	2
0.20	(NODE 1, NODE 4)	[(1,0.6359,5, NODE 4),(1,2,5 3,4,6,7,8,9,10)]	18.136714	3
0.25	(NODE 1, NODE 4)	[(NODE 1, NODE 4),(1,2,5 3,4,6,7,8,9,10)]	63.088462	2
0.274	(NODE 1, NODE 4)	[(NODE 1, NODE 5),(1,2,3,10 4,5,6,7,8,9)]	143.186052	2

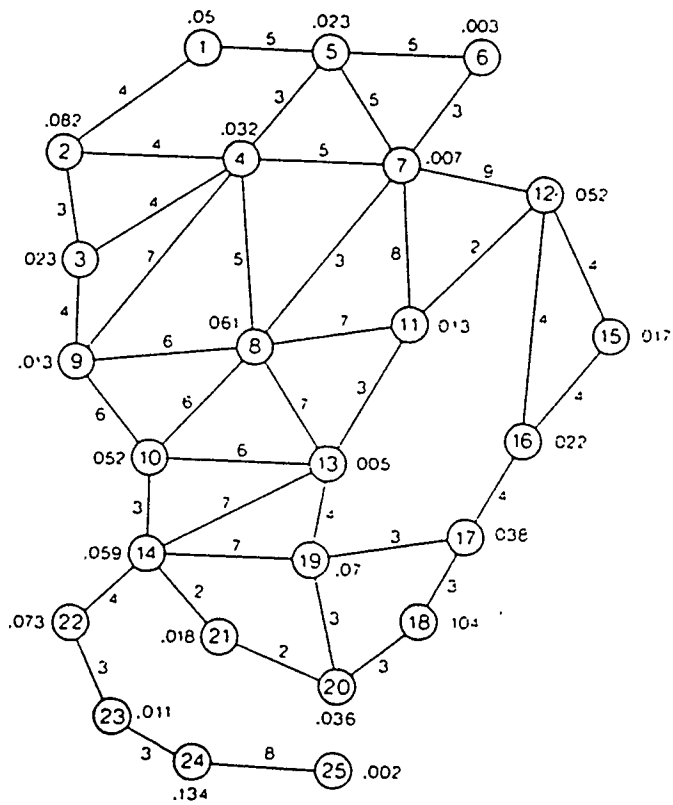


FIGURE 2.4 25-Node Network

TABLE 2.3 2 Server Location-Allocation for the 25-Node Network (2 Median = 2,14)

Demand Rate	Initial Location $(x^1,x^2)_0$	"Optimal" Location-Allocation Policy $[(x^1,x^2),(N^1,N^2)]^*$	"Optimal" Expected Response Time ERT*	Number of Iterations K
0.002	(NODE 2, NODE 14)	[(NODE 2, NODE 14),(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)]	6.507963	2
0.006	(NODE 2, NODE 14)	[(NODE 2, NODE 14),(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)]	6.855182	2
0.01	(NODE 2, NODE 14)	[(NODE 2, NODE 14),(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)]	7.234175	2
0.02	(NODE 2, NODE 14)	[(NODE 4, NODE 14),(1,2,3,4,5,6,7,8,9,12,15,10,11,13,14,16,17,18,19,20,21,22,23,24,25)]	8.168542	2
0.03	(NODE 2, NODE 14)	[(NODE 4, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,15,10,13,14,16,17,18,19,20,21,22,23,24,25)]	9.225280	2
0.04	(NODE 2, NODE 14)	[(NODE 4, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,15,16,10,13,14,17,18,19,20,21,22,23,24,25)]	10.508229	2
0.05	(NODE 2, NODE 14)	[(NODE 7, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,15,16,10,13,14,17,18,19,20,21,22,23,24,25)]	12.036448	2
0.075	(NODE 2, NODE 14)	[(NODE 7, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,13,15,16,10,14,17,18,19,20,21,22,23,24,25)]	18.056227	2
0.090	(NODE 2, NODE 14)	[(NODE 7, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,13,15,16,25,10,14,17,18,19,20,21,22,23,24)]	25.180175	2
0.10	(NODE 2, NODE 14)	[(NODE 7, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,13,15,16,25,10,14,17,18,19,20,21,22,23,24,)]	33.680345	2
0.128	(NODE 2, NODE 14)	[(NODE 4, NODE 14),(1,2,3,4,5,6,7,8,9,11,12,13,16,17,25,10,14,15,18,19,20,21,22,23,24)]	319.770684	2

The results shown in Table 2.1 for the 5-node network are interesting to observe. For low values of λ the best LA policy is the median-proximity LA policy. Above λ_p which is between 0.002 and 0.01 there is a range of λ values within which the best location policy remains at the 2-median while the districting changes. Although Node 4 is closer to the server at Node 2, within this range of λ values it is allocated to Node 5. The best policy changes again at a demand rate between 0.1 and 0.2, where the best location of servers is at Node 2 and half way between Nodes 3 and 5. The best districting policy allocates Nodes 1, 2 and 4 to the server at Node 2, and Nodes 3 and 5 to the server located half way between them. This LA policy remains "optimal" up to λ_{max} = 0.650100, where the system explodes (i.e., the average demand rate exceeds the average service rate).

It is also interesting to observe the progression of the solution sequence to the point of convergence. As an example, the alternate procedure for the 5-node network at λ = 0.1 is presented in Table 2.4. Starting at the 2-median as the initial location, Nodes 1 and 2 are allocated to the server at Node 2 and Nodes 3, 4 and 5 are allocated to the server at Node 5. For the first district the optimal server location is at Node 2 while for the second the optimal location is between Nodes 3 and 5 at a distance of 2.0068 from Node 3. Based on this new location policy Node 4 switches to the server at Node 2 which also serves Nodes 1 and 2, while Nodes 3 and 5 continue to be served by the second server. The optimal location of the server serving Nodes 1, 2 and 4 is found to be at Node 2 and for Nodes 3 and 5 the

RATE OF DEMAND λ	X_0	$(N^1,N^2)_1$	$ERT_{0,1}$	X_1	$ERT_{1,1}$	$(N^1,N^2)_2$	$ERT_{1,2}$	X_2	$ERT_{2,2}$	$(N^1,N^2)_3$	$ERT_{2,3}$
0.1	(NODE 2, NODE 5)	(1,2, 3,4,5)	1.327774	(NODE 2, 3,2.0068,5)	1.238797	(1,2,4, 3,5)	1.238440	(NODE 2, 3,2,5)	1.238439	(1,2,4, 3,5)	1.238439

TABLE 2.4 A Location-Allocation Solution Sequence for the 5-Node Network at λ=0.1

server is located half way between them as they have the same fraction of demand. No further improvement is possible and the best LA policy is therefore:

$$[(\overset{\text{NODE } 2}{3,2,5}),(\overset{1,2,4}{3,5})]$$

As shown in Table 2.2 for the 10-node network, λ_p is between 0.025 and 0.05, and above that value the best location alternates between the 2-median and locations that are close to it. The best allocation policy changes gradually for demand rates above λ_p, and λ_{max} is equal to 0.274725. Table 2.3 includes the best policies for the 25-node network. Here the results are similar to those for the 10-node network in the sense that while the median-proximity policy is optimal for $\lambda \leq \lambda_p$, there is a gradual change in the best LA policy as the demand rate increases towards λ_{max} = 0.128370.

2.5 Discussion of Results

It was expected that for low demand rates the optimal LA policy would be the median-proximity policy and that is indeed the result for demand rates below λ_p for all three networks. For high demand rates, in all three cases the best location policy is not at the 2-median and for the 25-node network the best districting policy is not the same as the solution for Region C (refer to Sub-Section 1.3.2) when the servers are located at the 2-median. In this case, had the best location been at the median, the optimal districting policy would have been equal to

that of Region C, but due to the shift of the servers away from the 2-median, the corresponding districting is equal to the optimal policy of Region C given that final server location.

The servers are not located at the 2-median for high demand rates although it was shown[3] that for one server the optimal location returns to the median as λ approaches λ_{max}. This is explained by noting that each server is located as expected at the median of its own district, but due to the particular districting at high demand rates the medians of the two districts may not be equal to the 2-median.

An additional example with three nodes is presented in order to illustrate the changes in "optimal" LA policies with increasing λ values. The 2-median of this network which is shown in Figure 2.5 is at Nodes 2 and 3.

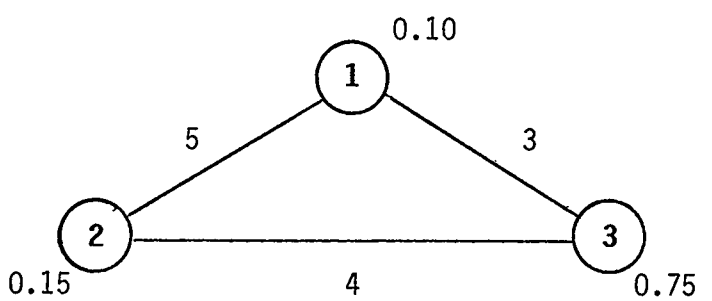


FIGURE 2.5 3-Node Network

The "optimal" LA policies are shown in Figure 2.6 along a line representing λ values from 0 to λ_{max} = 0.80.

Here for $\lambda \leq 0.197$ the best LA policy is the median-proximity policy. Above that value the best districting changes from that based on proximity and the location of the server of Nodes 1 and 2

initially moves towards Node 1 and then changes direction and moves towards Node 2 as the demand rate increases. At approximately $\lambda = 0.64$ the optimal location of the server of this district reaches Node 2 and it remains there until the system explodes. At all λ values above λ_p the best districting is the same as the solution for Region C, i.e. the server for Node 3 is located at that node and Nodes 1 and 2 are served by the server located at Node 2.

In their paper[5], Berman et al. presented this 3-node network with the servers located at Nodes 1 and 3. They showed that given this location policy the optimal districting for Region C is to allocate Node 2 to the server located at Node 1 and to allocate Nodes 1 and 3 to the server located at Node 3. This result is most non-intuitive because the server located at Node 1 does not serve Node 1 itself but is assigned to serve Node 2 which is the farthest node from the server. For this example with the servers located at Nodes 1 and 3 the system explodes at $\lambda_{max} = 0.689$. The LA solution which provides a better location policy at high λ values also permits a greater feasible range of λ values ($\lambda_{max} = 0.80$). Due to this location policy the best districting at high demand rates is more intuitive, with Nodes 1 and 2 allocated to the server at Node 2, and Node 3 allocated to the server at Node 3.

All LA policies derived based on the 2-median as the initial location were found to be best. When an initial location policy other than the 2-median was taken, the solution was either optimal or sub-optimal. Therefore, these results indicate that the 2-median is

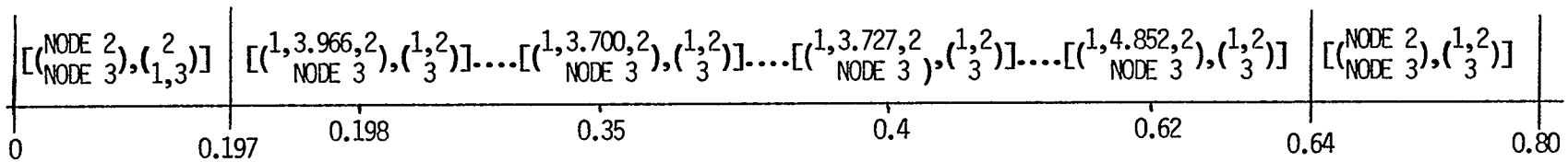


FIGURE 2.6 "Optimal" Location-Allocation Policies at Various λ Values for the 3-Node Network

the preferred initial policy.

A modified procedure to the original districting heuristic was proposed in Sub-Section 2.3.2. The following example illustrates the significance of using both the solutions of Regions A and C as initial policies in this heuristic. For the 10-node network, with a demand rate of 0.15 and the servers located at Nodes 1 and 4 the solution of Region C is found to be $\binom{1,2,6,9,10}{3,4,5,7,8}$ with an expected response time of 12.489371, and that of Region A is $\binom{1,2,}{3,4,5,6,7,8,9,10}$ with an expected response time of 21.988860. The districting heuristic using the solution of Region C as the initial policy provides the following solution: $\binom{1,2,3,6,9}{4,5,7,8,10}$ with an expected response time of 10.955628, while with the solution of Region A the final result is $\binom{1,2,5,}{3,4,6,7,8,9,10}$ with an expected response time of 10.004818. Therefore, a choice of an initial policy based on expected response times would have provided a sub-optimal solution and as a "better" initial policy cannot be determined in advance, both should be used separately and the final results compared in order to determine the "optimal" solution.

Figure 2.7 includes a graph of the expected response time for all three networks as a function of the demand rate. The relationship between the ERT and the demand rate for low λ values is close to linear. At high λ values and especially as λ approaches λ_{max} (which is different for each network) the ERT increases very rapidly. The shape of the three graphs suggests, convexity of the expected response time as a function of λ.

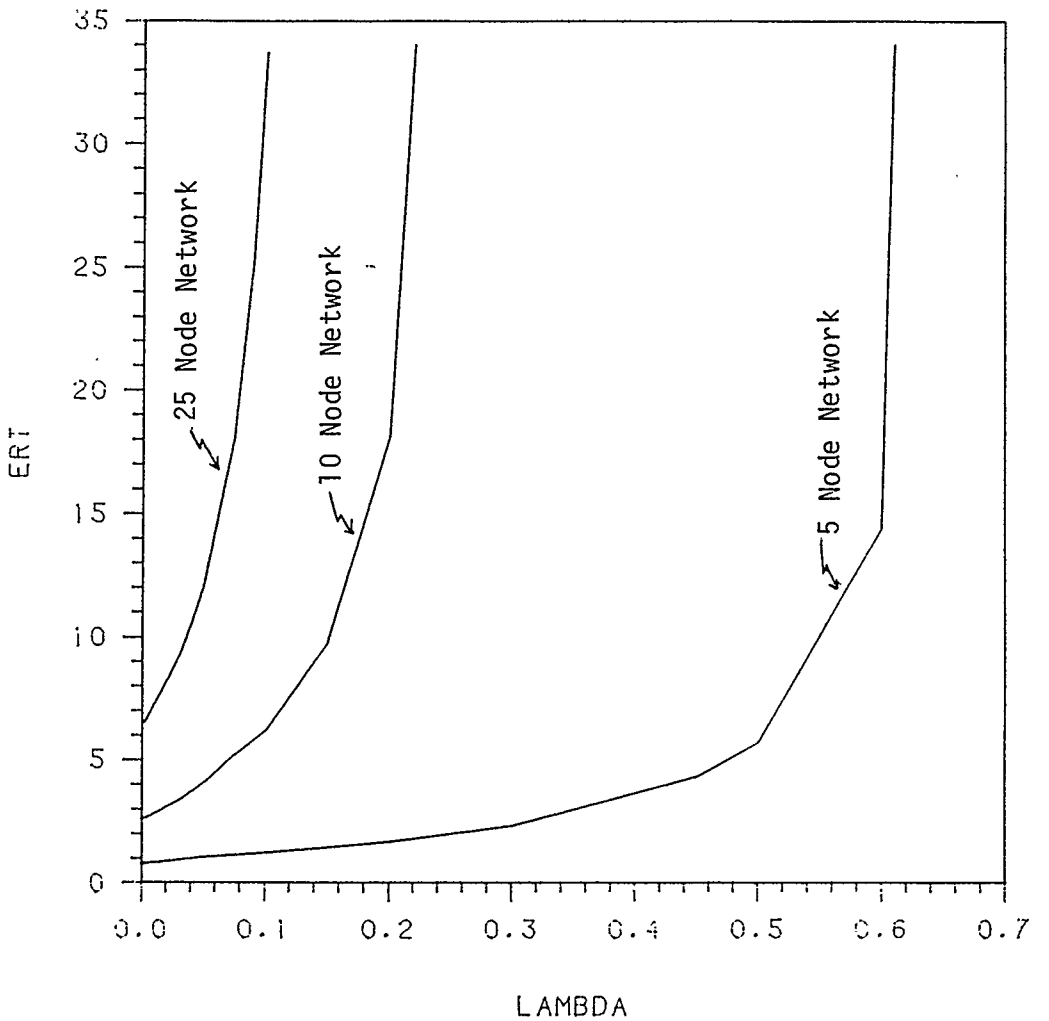


FIGURE 2.7 Expected Response Times at "Optimal" LA Policies for the 5, 10 and 25 Node Networks with Two Servers

CHAPTER 3

3 - SERVER LOCATION-ALLOCATION ON CONGESTED NETWORKS

3.1 Introduction and Outline

In this chapter the LA problem in a 3-server network is considered. The state-of-the-art in network districting under conditions of congestion includes the treatment of two servers and is described in Sub-Section 1.3.2. In order to solve the 3-server LA problem a heuristic solution to the 3-server districting problem was developed and is presented in this chapter.

The solution to 3-server districting is based on the 2-server case. It extends the previous work by introducing an iteration procedure that improves the districting between two servers at each iteration, while keeping the allocation of nodes to the third server constant. At each iteration the 2-server districting is found for two different servers and this sequence is performed until a whole cycle does not improve the solution.

Aside from the 3-server districting heuristic which is an extension of the heuristic solution method for two servers, the LA solution technique follows the same steps outlined in Chapter 2. Due to this similarity, reference should be made to notation and

definitions presented earlier.

The chapter begins with a definition of the problem in Section 3.2. Section 3.3 treats the 3-server districting problem and includes a discussion of the solution method, a flowchart of the heuristic and a summary of experimental results. The LA problem with three servers is treated in Section 3.4 which also includes a summary of results and discussion. All results presented in this chapter were derived through APL computer codes.

3.2 Problem Definition

As the formal definition of this problem is similar to that for two servers, only a brief description is provided and the reader may refer to Sub-Section 1.3.2 for further detail.

Three mobile service units are located when available at fixed locations on the network G(N,L). Let $X^i \varepsilon G(N,L)$ be the location of unit i, i=1,2,3. A districting policy is defined as any partition of the set N into three mutually exclusive sets N^1, N^2 and N^3, with N^i being the set of nodes in the district of service unit i. A LA policy for three servers is a combination of a location decision (X^1,X^2,X^3) and a districting decision (N^1,N^2,N^3) to form a combined policy $[(X^1,X^2,X^3), (N^1,N^2,N^3)]$. Given that server i is free and given a demand from node $j \varepsilon N^i$, the server is immediately dispatched to node j. No cooperation is allowed between the three districts and therefore, if server i is busy when a demand at Node j occurs, the

call enters a queue which is depleted according to a FCFS discipline. The expected total service time for server i, the second moment of the total service time, the expected waiting time and the expected travel time to a random demand are as shown in Sub-Section 1.3.2. For a given location policy (X^1,X^2,X^3) and a districting policy (N^1,N^2,N^3), the expected response time to a random demand is given by:

$$\bar{T}_R^{1,2,3} = h^1\bar{T}_R^1 + h^2\bar{T}_R^2 + h^3\bar{T}_R^3$$

and the problem is to find the optimal LA policy $[(X^1,X^2,X^3),(N^1,N^2,N^3)]^*$ that minimizes $\bar{T}_R^{1,2,3}$ when a call for service may occur in any one of the three districts.

The notation for location and districting policies for three servers follows the same format as for two servers but consists of three rows, each row related to a different server.

3.3 3-Server Network Districting in the Presence of Queueing

3.3.1 The Method of Solution

As in the two server case, the interval of all λ values $\lambda \geq 0$ can be divided into the four regions described earlier. Therefore, for λ values in Regions A and C there is one optimal solution for all λ values in each region, there is no feasible solution for λ values in Region D (where $\lambda \geq \lambda_{max}$) and there may be many possible optimal

policies for different λ values in Region B.

The heuristic begins with the determination of optimal districting policies for Regions A and C. Each policy at a time is taken as the initial solution, the heuristic is solved twice and the better solution of the two is taken as the overall "optimal" solution. Following the determination of an initial solution, one district is kept constant and the best 2-facility districting policy is found for the two remaining servers. The new solution, therefore, consists of the fixed district and the best districting for the two other servers discluding the demand nodes that were fixed. Now, a different district is kept fixed and the best districting policy is found for the two remaining servers in the same manner. In this heuristic a cycle is defined as a sequence of three iterations. At each iteration the allocation of nodes to a different server is kept the same as in the initial policy and the best districting policy is found for the other two servers. For example, cycle L may be composed of the iterations shown in Table 3.1, starting with policy $(N_0^1, N_0^2, N_0^3)^L$.

ITERATION NUMBER	FIXED POLICY	BEST 2-SERVER DISTRICTING	3-SERVER DISTRICTING POLICY
1	$N_1^1=N_0^1$	(N_1^2,N_1^3)	$(N_1^1,N_1^2,N_1^3)^L$
2	$N_2^2=N_0^2$	(N_2^1,N_2^3)	$(N_2^1,N_2^2,N_2^3)^L$
3	$N_3^3=N_0^3$	(N_3^1,N_3^2)	$(N_3^1,N_3^2,N_3^3)^L$

TABLE 3.1 Iterations of One Cycle for a 3-Server Network

In the above table N_Z^i indicates an allocation of nodes to server i in iteration Z. If for a whole cycle there is no improvement of the overall expected response time compared to the expected response time of $(N_0^1, N_0^2, N_0^3)^L$, then this policy is taken as the best 3-facility districting policy.

Special attention should be given to the restructuring of the input variables for each iteration Z. As one district is kept fixed, (say N^1) only the remaining nodes are to be allocated among the two servers located at X^2 and X^3. For the 2-facility districting calculations the fixed nodes are assigned zero demand and the new fraction of demand at node j, $j \in N^2 + N^3$ becomes $h_j^Z = h_j / \sum_{j \in N^2+N^3} h_j$. The total demand rate used for the same calculations reduces due to the fixing of N^1 and it becomes $\lambda^Z = \lambda \cdot \sum_{j \in N^2+N^3} h_j$. For each iteration these values are updated to take into account the new fixed nodes.

The flowchart in Figure 3.1 depicts the heuristic for 3-facility districting in the presence of queueing. Computational experience has justified the intuitive "cycle" criterion of reaching the final solution of the heuristic. The reason that an improvement over a whole cycle is searched rather than a comparison only with the previous iteration, is that a better policy may be obtained at any stage of the cycle. Some examples show no improvement for the first two iterations and a better policy obtained at the third iteration, but once a whole cycle has not provided an improvement, all the possibilities (according to this heuristic) have been exhausted and no better solution can be obtained.

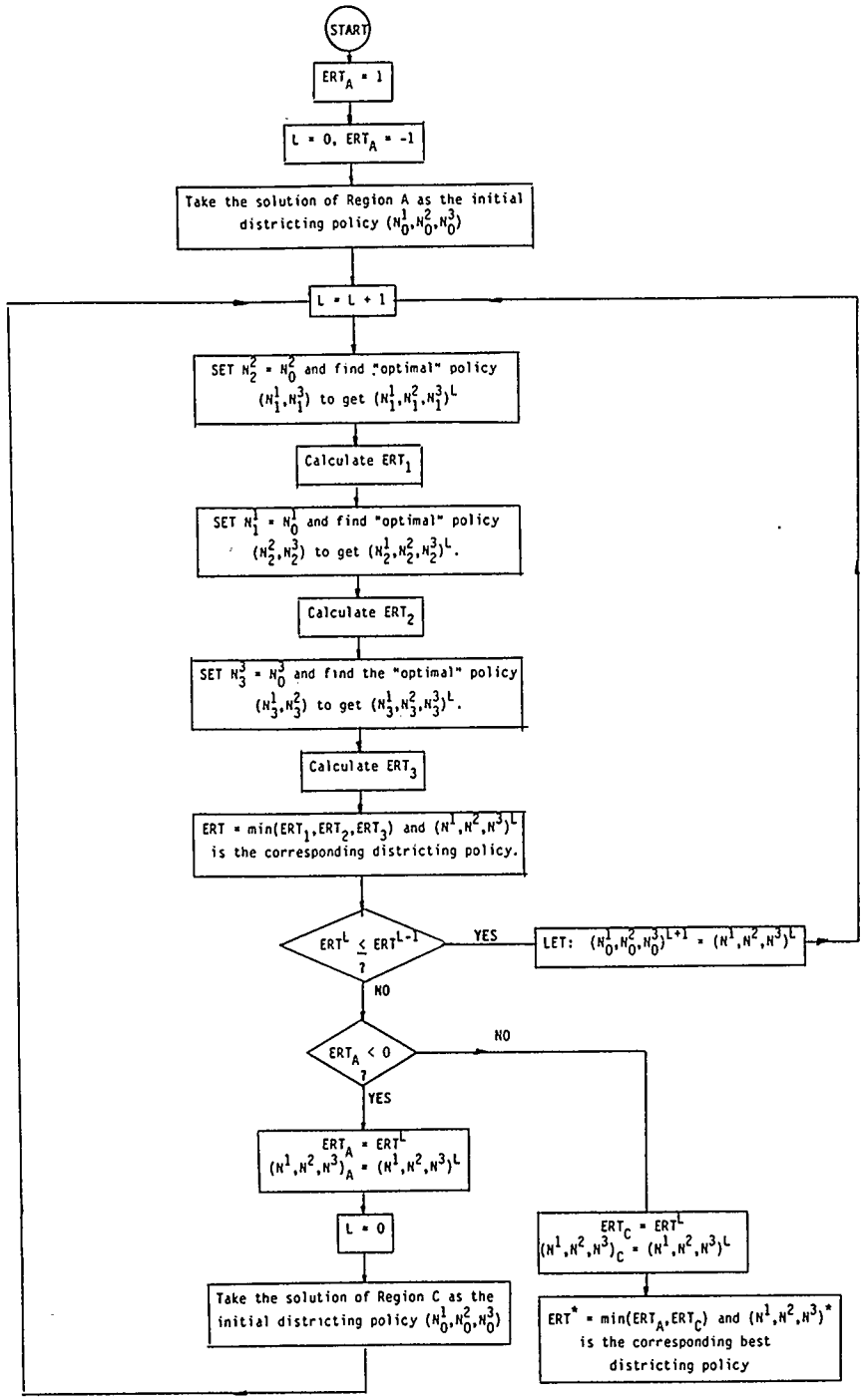


FIGURE 3.1 3-Server Districting Flowchart

3.3.2 Summary of Results

A code in APL was formulated and provided the results presented in this section. The three networks illustrated in Figures 2.2, 2.3 and 2.4 were used as examples for the heuristic and for each network optimal districting policies were found for given facility locations and varying demand rates. These results are presented in Tables 3.2, 3.3 and 3.4 for the 5, 10 and 25 node networks, respectively.

For the 5-node network, the range of λ values in Region A is relatively small. For this network and at the given facility locations there is one "optimal" policy for all feasible demand rates above λ_A ($0.006 < \lambda_A < 0.007$) which is the same as the unique optimal policy found for Region C. The best districting policy for low demand rates is based on the proximity of nodes to servers as expected.

As the network gets congested with increasing demand there is a change in node allocation. Over 50% of calls for service originate at Node 2 which is served exclusively by the server located at Node 2. The nodes allocated to the server at Node 3 are Nodes 1 and 3 which is similar to the allocation for low rates of demand. The server at Node 5 which is the most isolated node of the network continues to serve Node 5 itself but also has Node 4 allocated to it when congestion occurs. Due to the high proportion of calls occurring at Node 2, the overall network expected response time under congestion conditions is

TABLE 3.2 3-Server Districting for the 5-Node Network

Demand Rate λ	Facility Locations (x^1,x^2,x^3)	Best Districting Policy $(N^1,N^2,N^3)^*$	Best Expected Response Time ERT^*
0.002	(NODE 2, NODE 3, NODE 5)	(2,4, 1,3, 5)	0.309798
0.006	(NODE 2, NODE 3, NODE 5)	(2,4, 1,3, 5)	0.312920
0.007	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.313703
0.01	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.316055
0.04	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.340251
0.1	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.392653
0.3	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.619367
0.45	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	0.874009
0.65	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	1.461891
0.80	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	2.404566
1.00	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	8.620167
1.0835	(NODE 2, NODE 3, NODE 5)	(2, 1,3, 4,5)	120929628.96

TABLE 3.3 3-Server Districting for the 10-Node Network

Demand Demand λ	Facility Locations (X^1,X^2,X^3)	Best Districting Policy $(N^1,N^2,N^3)^*$	Best Expected Response Time ERT^*
0.01	(NODE 1, NODE 4, NODE 5)	(1,2; 3,4,8,9,10; 5,6,7)	1.877868
0.1	(NODE 1, NODE 4, NODE 5)	(1,2; 3,4,8,9,10; 5,6,7)	3.158937
0.11	(NODE 1, NODE 4, NODE 5)	(1,2; 3,8,9,10; 4,5,6,7)	3.358176
0.125	(NODE 1, NODE 4, NODE 5)	(1,2; 3,8,9,10; 4,5,6,7)	3.640244
0.15	(NODE 1, NODE 4, NODE 5)	(1,2; 3,8,9,10; 4,5,6,7)	4.216321
0.20	(NODE 1, NODE 4, NODE 5)	(1,2; 3,4,8,10; 5,6,7,9)	5.715255
0.25	(NODE 1, NODE 4, NODE 5)	(1,2; 3,4,7,8,9; 5,6,10)	7.830401
0.35	(NODE 1, NODE 4, NODE 5)	(1,2,3; 4,7,9,10; 5,6,8)	14.321850
0.45	(NODE 1, NODE 4, NODE 5)	(1,2,9; 3,4,7,8; 5,6,10)	32.543407
0.475	(NODE 1, NODE 4, NODE 5)	(1,2,9; 3,4,7,8; 5,6,10)	43.532804
0.50	(NODE 1, NODE 4, NODE 5)	(1,2,6,9; 3,4,7,8; 5,10)	62.033854
0.5494	(NODE 1, NODE 4, NODE 5)	(1,2,6,9; 3,4,7,8; 5,10)	14013.719

TABLE 3.4 3-Server Districting for the 25-Node Network

Demand Rate λ	Facility Locations (X^1,X^2,X^3)	Best Districting Policy $(N^1,N^2,N^3)^*$	Best Expected Response Time ERT^*
0.01	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	4.637452
0.03	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	5.135768
0.05	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	5.713128
0.075	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	6.578128
0.10	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	7.65845
0.15	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	10.901296
0.20	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	17.473365
0.225	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	24.291865
0.25	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9,13 / 11,12,,15,16,17,18,19,20 / 10,14,21,22,23,24,25)	39.095461
0.2909	(NODE 2, NODE 17, NODE 22)	(1,2,3,4,5,6,7,8,9,11 / 12,15,16,17,18,19,20,21 / 10,13,14,22,23,24,25)	399.367879

improved by relieving the server at Node 2 of Node 4 and assigning it to the next closest server which is located at Node 5.

For the 10 node network λ_A occurs at $0.10 < \lambda < 0.11$ and Region B includes several "optimal" policies depending on the values of λ. For the three servers located at Nodes 1, 4 and 5 the system explodes at $\lambda_{max} = 0.54945$ and λ_C occurs at $0.45 < \lambda < 0.5$. Figure 3.2 illustrates the changes in best districting policies with changes in the demand rate. It is interesting to note that in the case of $\lambda = 0.15$ the server at Node 4 serves a district to which Node 4 itself does not belong. In this case, Node 4 is served by the server at Node 5 rather than by the server located at Node 4 itself. Figure 3.2 illustrates that even for a relatively simple problem with ten nodes and three servers, intuition would not suffice to determine "optimal" districting policies under congestion at the various rates of demand. For λ values in Region B there are at least five "optimal" districting policies that are different from those for Regions A and C.

In the case of the 25 node network, the districting policy based on the proximity of nodes to servers is best for close to 70% of the range of feasible λ values. Region C includes a small range of λ values and only one "optimal" policy which is different from the policies for Regions A and C was found for demand rates in Region B.

3.4 3-Server Location-Allocation

The alternate location and allocation solution improvement procedure described in Chapter 2 is applicable to the 3-server case.

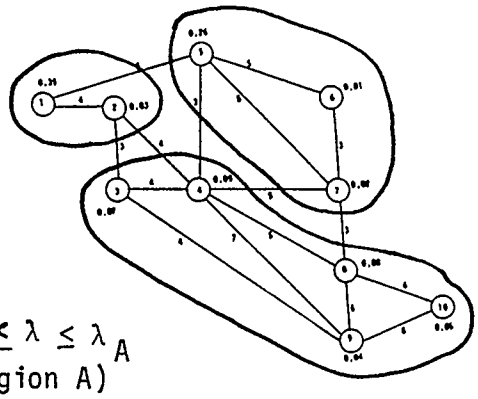
The 5, 10 and 25 node networks shown in Figures 2.2, 2.3 and 2.4, respectively, were used as illustrative examples.

The results for the 5 node network indicate that there are only two different "optimal" policies for all feasible λ values greater than zero (for $\lambda = 0$ the median-proximity policy is optimal). For low rates of demand the best location policy is not equal to the 3-median. Here, due to the fact that Nodes 1 and 3, each with the same fraction of demand, are both in one district, the best location is halfway between them. The best districting is based on proximity to the servers. The best location does not change for higher λ values but the best districting changes to the solution of Region C.

The results for the 10 node network which are shown in Table 3.6 indicate that the best location policy for all λ values is to locate the servers at the 3-median. Due to the unique location policy for all rates of demand, the variation of "optimal" districting policies with changes in demand rates resembles the typical results presented in the original paper by Berman and Larson[5]. The median-proximity policy is optimal for low rates of demand and the best solution for high rates of demand is locating servers at the 3-median and allocating nodes to servers according to the solution of Region C. For intermediate λ values there are at least two different "optimal" districting policies.

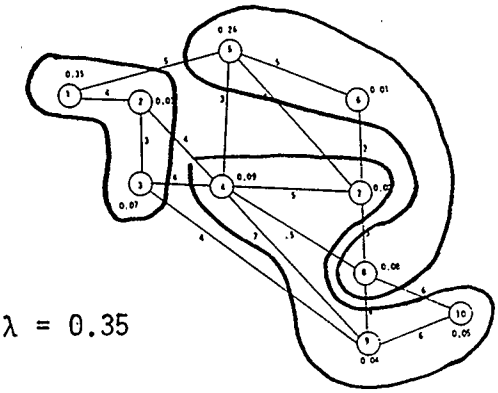
It is interesting to compare the solution to the LA problem for the 10-node network at $\lambda = 0.15$, with the best districting policy when the three servers are located at Nodes 1, 4 and 5, as shown in Table 3.3 and illustrated in Figure 3.2. Locating the servers at

1.



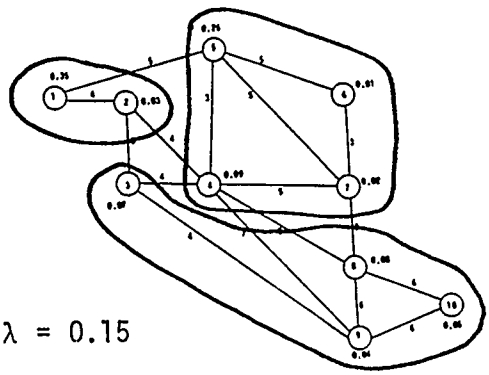
$0 \leq \lambda \leq \lambda_A$
(Region A)

5.



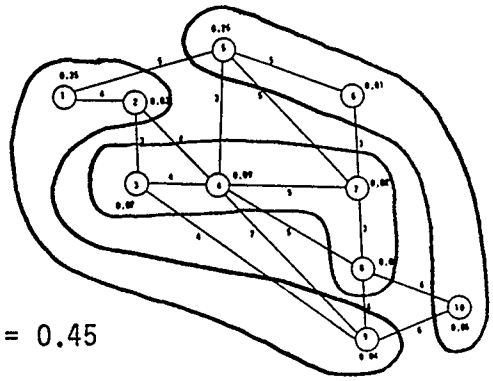
$\lambda = 0.35$

2.



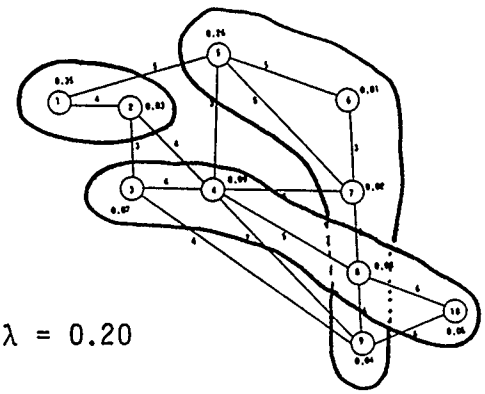
$\lambda = 0.15$

6.



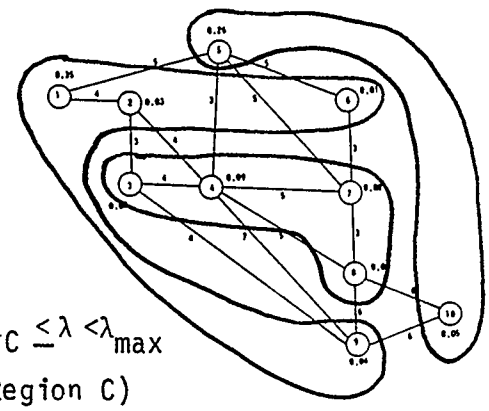
$\lambda = 0.45$

3.



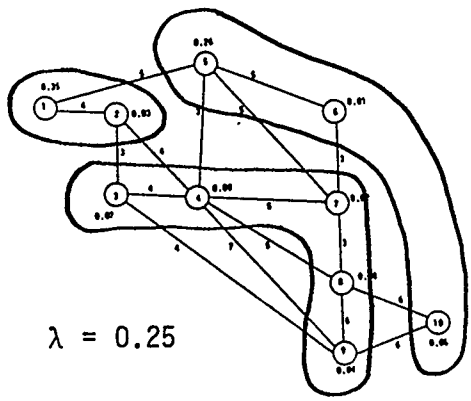
$\lambda = 0.20$

7.



$\lambda_C \leq \lambda < \lambda_{max}$
(Region C)

4.



$\lambda = 0.25$

FIGURE 3.2 "Optimal" 3-Server Districting Policies at Different Demand Rates for the 10-Node Network.

(The Servers are located at Nodes 1, 4 and 5)

Nodes 1, 4 and 5 is inferior to a location policy which places the servers at the 3-median, as in the second case the expected response time is approximately 67% of that in the first. Due to this "weak" location of servers the system becomes congested at a lower λ value (λ_{max} = 0.54945 for location policy $\begin{bmatrix}\text{NODE 1}\\ \text{NODE 4}\\ \text{NODE 5}\end{bmatrix}$ compared to λ_{max} = 0.689655 for location policy $\begin{bmatrix}\text{NODE 1}\\ \text{NODE 5}\\ \text{NODE 8}\end{bmatrix}$) and the best districting in the first case is very unusual and non-intuitive as described in Sub-Section 3.3.2.

For the 25 node network, although the best location changes from the 3-median at $0.01 < \lambda < 0.1$, the districting policy based on the proximity of nodes to servers remains the best policy for 50% of the range of feasible λ values and changes for high values of λ.

The results presented in Tables 3.5, 3.6 and 3.7 are all based on initial locations of the servers at the 3-median. For comparison purposes, results based on other initial locations which differ from those based on the 3-median are shown in Table 3.8. In the first example, using the 3-median as the initial location results in a better LA solution. In the second case the 3-median also provides the best solution but in the third case the final solution using the 3-median is sub-optimal. It is interesting to note that while in the second example a different initial location policy than the 3-median provided the best solution, the same initial location gave a worse solution than the 3-median in the third example. As previously stated, in general, computational experience has shown that using the p-median as the initial location in the LA heuristic results in the best solutions.

TABLE 3.5 3 Server Location-Allocation for the 5-Node Network (3 Median = 1,2,5)

Demand Rate λ	Initial Location Policy (X^1,X^2,X^3)	"Optimal" Location-Allocation Policy $[(X^1,X^2,X^3),(N^1,N^2,N^3)]^*$	"Optimal" Expected Reponse Time ERT^*	Number of Iterations K
0.0001	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2,4, 5)]	0.308303	2
0.001	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2,4, 5)]	0.308831	2
0.005	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2,4, 5)]	0.311186	2
0.0075	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	0.312664	2
0.05	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	0.338678	2
0.10	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	0.371799	2
0.20	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	0.447849	2
0.50	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	0.809784	2
0.80	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	1.825810	2
1.00	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	6.163920	2
1.08	(NODE 1, NODE 2, NODE 5)	[(1,1,3, NODE 2, NODE 5), (1,3, 2, 4,5)]	143.094	2

TABLE 3.6 3 Server Location-Allocation
for the 10-Node Network
(3 Median = 1,5,8)

Demand Rate λ	Initial Location Policy (X^1,X^2,X^3)	"Optimal" Location-Allocation Policy $[(X^1,X^2,X^3),(N^1,N^2,N^3)]^*$	"Optimal" Expected Reponse Time ERT*	Number of Iterations K
0.0001	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2, 3,4,5,6, 7,8,9,10)]	1.530673	2
0.01	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2, 3,4,5,6, 7,8,9,10)]	1.598467	2
0.05	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	1.896215	2
0.10	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	2.327628	2
0.15	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	2.844104	2
0.20	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	3.474785	2
0.30	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	5.285523	2
0.40	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,3, 2,4,5, 6,7,8,9,10)]	8.406807	2
0.50	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,3, 2,4,5,6, 7,8,9,10)]	14.266762	2
0.60	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,3, 2,4,5,6, 7,8,9,10)]	31.252438	2
0.65	(NODE 1, NODE 5, NODE 8)	[(NODE 1, NODE 5, NODE 8) (1,3, 2,4,5,6, 7,8,9,10)]	64.290360	2

TABLE 3.7 3 Server Location-Allocation
for the 25-Node Network
(3 Median = 2,17,22)

Demand Rate λ	Initial Location Policy (X^1,X^2,X^3)	"Optimal" Location-Allocation Policy $[(X^1,X^2,X^3),(N^1,N^2,N^3)]^*$	"Optimal" Expected Reponse Time ERT*	Number of Iterations K
0.0001	(NODE 2, NODE 17, NODE 22)	[(NODE 2, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	4.415171	2
0.01	(NODE 2, NODE 17, NODE 22)	[(NODE 2, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	4.637452	2
0.04	(NODE 2, NODE 17, NODE 22)	[(NODE 2, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	5.413427	2
0.08	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	6.731615	2
0.10	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	7.586372	2
0.15	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9 / 11,12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	10.741405	2
0.20	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9,11 / 12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	17.178178	2
0.25	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9,11 / 12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	37.422912	2
0.29095	(NODE 2, NODE 17, NODE 22)	[(NODE 4, NODE 17, NODE 22) (1,2,3,4,5,6,7,8,9,11 / 12,13,15,16,17,18,19,20 / 10,14,21,22,23,24,25)]	384.16	2

EX.	NETWORK	DEMAND RATE (λ)	INITIAL LOCATION POLICY (3-MEDIAN)	NUMBER OF ITERATIONS K	BEST LOCATION-ALLOCATION POLICY	ERT	OTHER INITIAL LOCATION POLICIES	NUMBER OF ITERATIONS K	BEST LOCATION-ALLOCATION POLICY	ERT
1.	10 NODE	0.01	(NODE 1, NODE 5, NODE 8)	2	[(NODE 1, NODE 5, NODE 8) (1,2, 3,4,5,6, 7,8,9,10)]	1.598467	(NODE 1, NODE 4, NODE 5)	2	[(NODE 1, NODE 4, NODE 5), (1,2, 3,4,8,9,10, 5,6,7)]	1.877868
2.	10 NODE	0.05	(NODE 1, NODE 5, NODE 8)	2	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	1.896215	(NODE 2, NODE 4, NODE 9)	2	[(NODE 1, NODE 5, NODE 9), (1,2, 4,5,6,7, 3,8,9,10)]	1.899527
3.	10 NODE	0.10	(NODE 1, NODE 5, NODE 8)	2	[(NODE 1, NODE 5, NODE 8) (1,2,3, 4,5,6, 7,8,9,10)]	2.327628	(NODE 2, NODE 4, NODE 9)	2	[(NODE 1, NODE 5, NODE 9), (1,2, 3,4,5,6, 7,8,9,10)]	2.270688

TABLE 3.8 Best Location-Allocation Policies Based on Different Initial Locations

Finally, a comparison of results for three servers with those for two servers suggests that the improvement in performance of the system in terms of decrease in expected response time when 3 servers are employed is greater than the increase in resources, especially for high demand rates. For example, for a five node network with $\lambda = 0.5$, an increase of 50% in resources from 2 to 3 servers results in an expected response time which is approximately 14% of the expected response time for two servers. With three servers the demand rate at which the system explodes increases, i.e. the systems can operate at greater rates of demand than with two servers, as shown in Table 3.9. In addition, for demand rates that are feasible in both cases the advantage of three servers compared to two increases as the demand rate goes up, as shown in Figure 3.3 for the 5 node network.

NUMBER OF NODES	NUMBER OF SERVERS	λ_{max}
5	2	0.6501
5	3	1.0835
10	2	0.2747
10	3	0.6897
25	2	0.1284
25	3	0.2910

TABLE 3.9 λ_{max} Values for Two and Three Servers for the Three Networks

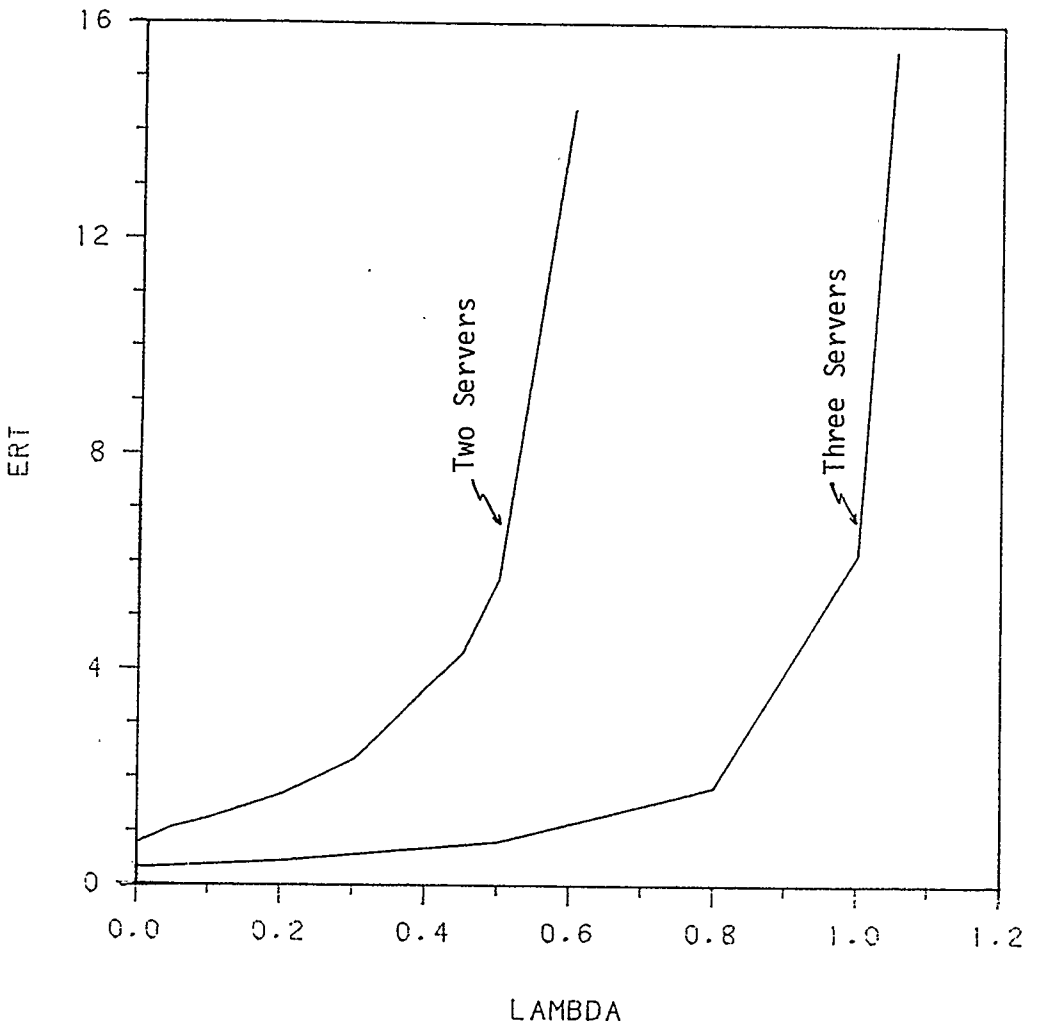


FIGURE 3.3 Comparison of Performance of Two and Three Servers at "Optimal" Location-Allocation Policies for the 5-Node Network

CHAPTER 4

m-SERVER LOCATION-ALLOCATION ON CONGESTED NETWORKS

4.1 Introduction and Outline

In Chapters 2 and 3 solution methods were derived for the LA problem with two and three servers, respectively, and with no restriction as to the size of the network (the number of nodes). The state-of-the-art theory in location and districting under conditions of congestion was utilized in developing the 2-server LA heuristic. An extension to the 2-server districting provided a heuristic method for districting with three servers, based on which the 3-server LA heuristic was developed. This chapter extends the treatment of the LA problem under congestion to the general case of m servers and n nodes.

In order to solve the m-server LA problem, the development of an m-server districting heuristic was required. Other than that, both the location algorithm used for finding each district's stochastic queue median and the alternate location and allocation solution improvement procedure described earlier remain practically unchanged.

The logic behind the m-server districting heuristic solution method is similar to that described for three servers in Chapter 3. Here, the iteration procedure improves the districting between two servers at a time, keeping the previous allocation of nodes to all the other servers constant. Each iteration uses a different pair of

servers and a cycle of iterations consists of all possible combinations of server pairs. The districting heuristic continues to search for a better solution as long as the best iteration of the last cycle is better than the best iteration of the previous one.

Chapter 4 starts with a definition of the problem in Section 4.2. Section 4.3 deals with the m-server districting problem and includes a discussion of the heuristic solution method, a flowchart of the heuristic and an illustrative example. The m-server LA problem is treated in Section 4.4. A flowchart for the general case is provided and two illustrative examples are presented to emphasize the usefullness of this heuristic. A discussion related to the computational results is included at the end of this section. Computer codes in APL for the m-server location algorithm (one server in each district), the m-server districting heuristic and the LA heuristic are included in the appendix to this chapter.

4.2 Problem Definition

The formal problem definition which is included for the sake of completion follows the detailed description presented for the 2-server case in Sub-Section 1.3.2 and is similar to the problem definition for two and three servers. In the general case, m mobile service units are located when available at fixed locations on the network G(N,L). Let $X^i \varepsilon G(N,L)$ be the location of unit i, i=1,2,...,m. A districting policy is defined as any partition of the set N into m mutually exclusive sets $N^1, N^2 \ldots, N^m$ with N^i being the set of nodes in the

district of service unit i. A LA policy for m servers is a combination of a location decision $(X^1,X^2,\ldots,X^m)$ and a districting decision $(N^1,N^2,\ldots,N^m)$ to form a combined policy $[(X^1,X^2,\ldots,X^m),(N^1,N^2,\ldots,N^m)]$.

Given that server i is free and given a demand from node $j \in N^i$, the server is immediately dispatched to node j. No cooperation is allowed between the m districts and therefore, if server i is busy when a demand at node j occurs, the call enters a queue which is depleted according to a FCFS discipline.

For any location policy $(X^1,X^2,\ldots,X^m)$ and districting policy $(N^1,N^2,\ldots,N^m)$, the expected response time to a random demand is:

$$\bar{T}_R^{1,2,\ldots,m} = h^1\bar{T}_R^1 + h^2\bar{T}_R^2 + \ldots + h^m\bar{T}_R^m$$

and the problem is to find the optimal LA policy $[(X^1,X^2,\ldots,X^m),(N^1,N^2,\ldots,N^m)]^*$ that minimizes $\bar{T}_R^{1,2,\ldots,m}$ when a call for service may originate from any node in any one of the m districts.

4.3 m-Server Network Districting in the Presence of Queueing

4.3.1 The Method of Solution

For λ values in Regions A and C there are unique optimal districting solutions but as it is not known in advance to which region a given demand rate belongs, the following heuristic utilizes both of these two solutions as starting points for an iterative

sequence of solution improvements.

Starting with any given location of servers, the optimal districting policies for Regions A and C are found. Computational experience has shown that in some cases, although one of the two policies may be superior (with a lower expected response time) to the other, at a given rate of demand, the final solution based on this initially superior policy may be inferior to the other (this was discussed in Chapter 2). Therefore, although the amount of computation may largely increase compared to the use of only one initial policy, it is recommended to go through the heuristic twice, once starting with the solution of Region A and a second time with the solution of Region C.

For each initial solution the best districting policy for two servers is found while the other servers and the nodes allocated to them remain constant. This is done for all possible server pairs with the other districts remaining the same as under the initial districting policy. The number of combinations of server pairs is a function only of the number of servers and is equal to $\binom{m}{2} = \frac{m(m-1)}{2}$. if the network has five servers (X^1, X^2, X^3, X^4, X^5), there are ten possible server pairs as follows:

1. (X^1,X^2)	3. (X^1,X^4)	5. (X^2,X^3)	7. (X^2,X^5)	9. (X^3,X^5)
2. (X^1,X^3)	4. (X^1,X^5)	6. (X^2,X^4)	8. (X^3,X^4)	10. (X^4,X^5)

For twenty servers there are 190 combinations and for fifty servers there are 1225 possible server pairs.

Each iteration uses a different pair of servers for which the best

districting policy is found while the other districts remain the same as in the initial solution. A cycle of iterations consists of $\frac{m(m-1)}{2}$ possible iterations and at the end of a cycle the best iteration is taken as the initial policy for the next cycle. This sequence continues until a complete cycle does not provide a policy which is better than the best policy of the previous cycle.

In the case of a 5-server network, cycle L includes ten iterations as shown in Table 4.1 starting with policy $(N_0^1, N_0^2, N_0^3, N_0^4, N_0^5)^L$.

ITERATION NUMBER	FIXED DISTRICTS	2-SERVER DISTRICTING	5-SERVER DISTRICTING POLICY
1	N_0^3, N_0^4, N_0^5	$(N_1^1, N_1^2)^L$	$(N_1^1, N_1^2, N_0^3, N_0^4, N_0^5)^L$
2	N_0^2, N_0^4, N_0^5	$(N_2^1, N_2^3)^L$	$(N_2^1, N_0^2, N_2^3, N_0^4, N_0^5)^L$
3	N_0^2, N_0^3, N_0^5	$(N_3^1, N_3^4)^L$	$(N_3^1, N_0^2, N_0^3, N_3^4, N_0^5)^L$
4	N_0^2, N_0^3, N_0^4	$(N_4^1, N_4^5)^L$	$(N_4^1, N_0^2, N_0^3, N_0^4, N_4^5)^L$
5	N_0^1, N_0^4, N_0^5	$(N_5^2, N_5^3)^L$	$(N_0^1, N_5^2, N_5^3, N_0^4, N_0^5)^L$
6	N_0^1, N_0^3, N_0^5	$(N_6^2, N_6^4)^L$	$(N_0^1, N_6^2, N_0^3, N_6^4, N_0^5)^L$
7	N_0^1, N_0^3, N_0^4	$(N_7^2, N_7^5)^L$	$(N_0^1, N_7^2, N_0^3, N_0^4, N_7^5)^L$
8	N_0^1, N_0^2, N_0^5	$(N_8^3, N_8^4)^L$	$(N_0^1, N_0^2, N_8^3, N_8^4, N_0^5)^L$
9	N_0^1, N_0^2, N_0^4	$(N_9^3, N_9^5)^L$	$(N_0^1, N_0^2, N_9^3, N_0^4, N_9^5)^L$
10	N_0^1, N_0^2, N_0^3	$(N_{10}^4, N_{10}^5)^L$	$(N_0^1, N_0^2, N_0^3, N_{10}^4, N_{10}^5)^L$

TABLE 4.1 Iterations of One Cycle for a 5-Server Network

In iteration Number 8, for example, Districts 1, 2 and 5 remain the same as under the initial policy and the best districting policy is found between Servers 3 and 4.

In a similar manner to the procedure for three servers, prior to each iteration Z the fixed nodes are assigned zero demand and the new fraction of demand at node j, $j \in N^a + N^b$ (where X^a and X^b are the two servers to which the non-fixed nodes are allocated) becomes $h^L_{Zj} = h_j / \Sigma_{j \in N^a + N^b} h_j$. Also, the demand rate for these two districts reduces to $\lambda^L_Z = \lambda \cdot \Sigma_{j \in N^a + N^b} h_j$ due to the fixing of all the other nodes.

The flowchart in Figure 4.1 depicts the m-server districting heuristic for a congested network with queues. The purpose of assigning a negative value to ERT_A at the outset is to indicate initially that the best solution is based on the solution to Region A as a starting point, and once a non-negative value is assigned to ERT_A, the best solution is based on the solution of Region C as the starting point. After ERT_C is found, the best of the two solutions is taken as the "optimal" districting policy.

A computer code of this heuristic is included in the appendix to this chapter.

4.3.2 An Illustrative Example

The network used to illustrate the m-server districting heuristic is the 25-node network shown in Figure 2.4. In this example, there are five mobile servers and the five facilities to house these servers

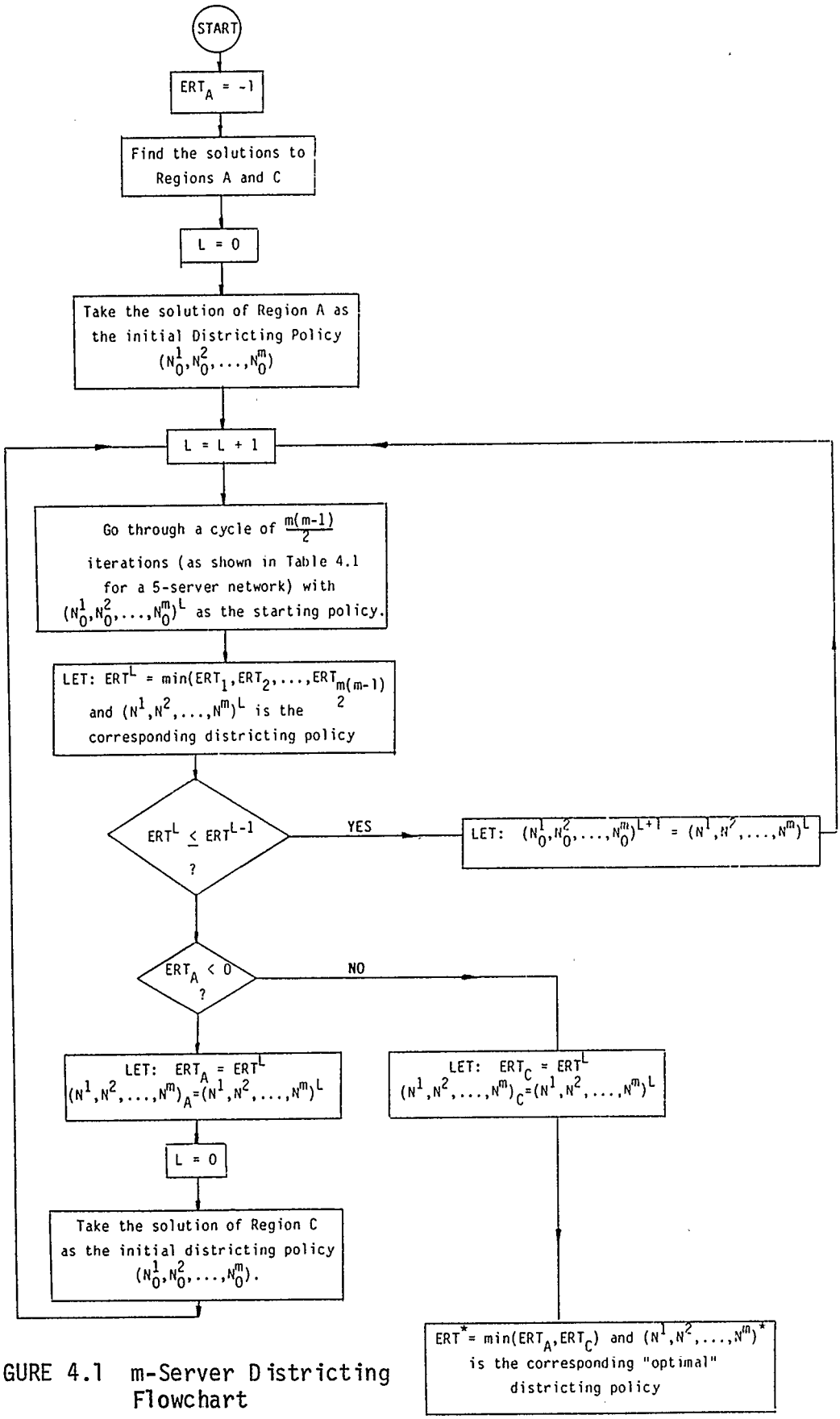


FIGURE 4.1 m-Server Districting Flowchart

are located according to the following location policy: $\begin{bmatrix} \text{NODE 2} \\ \text{NODE 8} \\ \text{NODE 14} \\ \text{NODE 17} \\ \text{NODE 22} \end{bmatrix}$.

The overall network demand rate λ is 0.01, and a value of one is assigned to $(R_j^i+W_j^i)$ and v.

The heuristic starts with the optimal districting policy of Region A as the initial solution. With five servers a cycle consists of ten iterations and at the end of the first cycle the best iteration of the cycle is not better than the initial solution. Therefore, the initial solution is taken as the best interim districting policy $(N^1,N^2,N^3,N^4,N^5)_A$.

Now, the optimal districting policy of Region C is taken as the initial solution. As a demand rate of 0.01 is relatively low, this initial solution is inferior to the solution of Region A. Nevertheless, after five cycles the "optimal" solution $(N^1,N^2,N^3,N^4,N^5)_C$ is found to be equal to that of $(N^1,N^2,N^3,N^4,N^5)_A$ and is taken as the overall "optimal" districting policy. This solution process is illustrated in Table 4.2.

The strength of the iteration procedure is shown in the above example where an inferior solution (in this case the optimal districting policy of Region C) converges after a number of cycles onto the "optimal" solution.

CYCLE L	SOLUTION TO REGION A	ERT	$(N^1,N^2,N^3,N^4,N^5)^L$	ERT^L	$(N^1,N^2,N^3,N^4,N^5)_A$	ERT_A
0	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605				
1			1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605

CYCLE L	SOLUTION TO REGION C	ERT	$(N^1,N^2,N^3,N^4,N^5)^L$	ERT^L	$(N^1,N^2,N^3,N^4,N^5)_C$	ERT_C
0	1,2,3,4,5,9,25 6,7,8,12,13,15 10,14,20,21,22,23 11,16,17,18,19 24	3.836450				
1			1,2,3,4,5,9,25 6,7,8,12,13,15 10,14,20,21 11,16,17,18,19 22,23,24	3.497636		
2			1,2,3,4,5,9,25 6,7,8,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24	3.326290		
3			1,2,3,4,5,9 6,7,8,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.277349		
4			1,2,3,4,5, 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605		
5			1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605

TABLE 4.2 Districting Solution for the 5-Server, 25-Node Example

4.4. m-Server Location-Allocation

4.4.1 Flowchart and Illustrative Examples

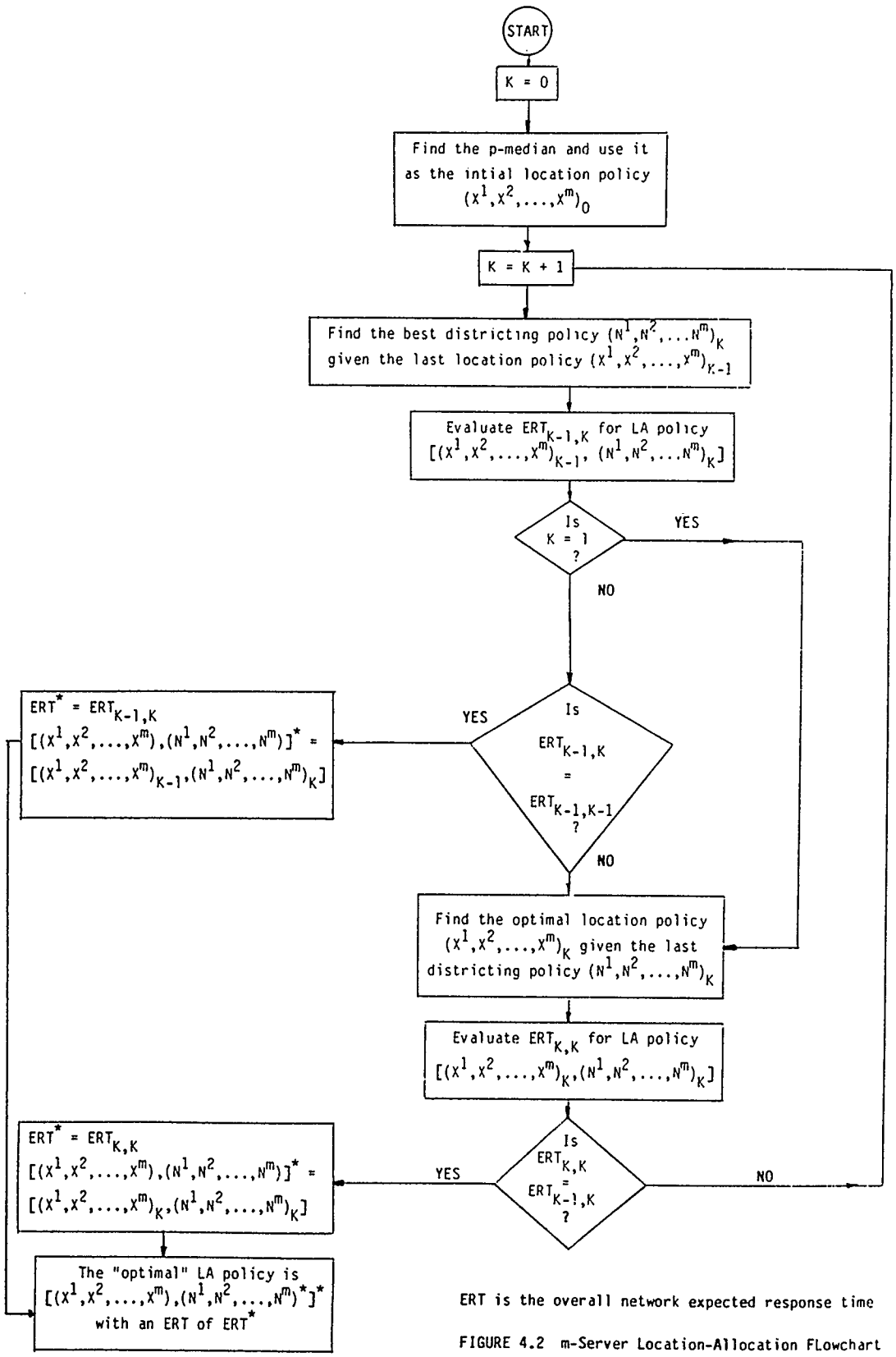
The flowchart shown in Figure 4.2 represents the LA heuristic for the general case of m servers and n nodes in a congested network with queues. Although it is similar to the flowchart in Figure 2.1, the notation in Figure 4.2 refers to the general problem. A computer code in APL for solving the general LA problem is included in the appendix to this chapter.

Two examples are presented here to illustrate the alternate location and allocation solution procedure in detail. In both examples the servers are initially located at the 5-median $\begin{bmatrix} \text{NODE 2} \\ \text{NODE 8} \\ \text{NODE 14} \\ \text{NODE 17} \\ \text{NODE 24} \end{bmatrix}$.

In one example the "optimal" LA policy is found at a low demand rate and in the second example a relatively high demand rate for this network is used and the solution procedure involves a greater number of cycles to reach the optimal solution.

Tables 4.3 and 4.4 include best location and districting policies as the heuristic converges towards the optimal solution for demand rates of 0.01 and 0.5, respectively.

At $\lambda = 0.01$ the optimal solution is reached after one complete iteration. Starting at the 5-median of this network the best allocation policy is found to be that of Region A. The optimal location of servers, given a districting policy that is equal to the



ERT is the overall network expected response time

FIGURE 4.2 m-Server Location-Allocation FLowchart

solution of Region A, is the 5-median and therefore, no further improvement is possible. The best LA policy at this low demand rate is the median-proximity LA policy $\left[\begin{pmatrix}\text{NODE 2}\\ \text{NODE 8}\\ \text{NODE 14}\\ \text{NODE 17}\\ \text{NODE 24}\end{pmatrix}, \begin{pmatrix}1,2,3,4,5\\ 6,7,8,9,11,13\\ 10,14,20,21,22\\ 12,15,16,17,18,19\\ 23,24,25\end{pmatrix}\right]$.

While in the first example the "optimal" LA policy is equal to the best solution obtained at the end of the first iteration, for the higher demand rate of 0.5 the heuristic converges onto the "optimal" solution after several iterations. The improvement in expected response time between the initial solution and the "optimal" LA policy is significant. This is due to the greater sensitivity of the system to changes in server location or network districting at high demand rates.

As shown in Table 4.4 the initial expected response time, following the determination of the first districting policy with the servers located at the 5-median, is equal to 17.172809. The best location policy changes in the next step, moving the second server from Node 8 to a position close to Node 11 and the fifth server is relocated from Node 24 to a location between Nodes 23 and 24. This change in location policy in itself reduces the expected response time by about 17% to 14.274396. Following a total of three alternate districting policy solutions (the last districting solution being equal to the previous districting policy) and three location policies (including the initial 5-median) the "optimal" LA policy is found to be

$\left[\begin{pmatrix}\text{NODE 2}\\ \text{NODE 11}\\ \text{NODE 14}\\ \text{NODE 17}\\ 23,2.175150,24\end{pmatrix}, \begin{pmatrix}1,2,3,4,5,6\\ 8,11,12,13,15\\ 7,9,10,14,20,21\\ 16,17,18,19\\ 22,23,24,25\end{pmatrix}\right]$ with an expected response time

$(X^1,X^2,X^3,X^4,X^5)_0$	$(N^1,N^2,N^3,N^4,N^5)_1$	$ERT_{0,1}$	$(X^1,X^2,X^3,X^4,X^5)_1$	$ERT_{1,1}$
NODE 2 NODE 8 NODE 14 NODE 17 NODE 24	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21,22 12,15,16,17,18,19 23,24,25	2.719937	NODE 2 NODE 8 NODE 14 NODE 17 NODE 24	2.719937

TABLE 4.3 Location-Allocation Solution Sequence for $\lambda = 0.01$

$(X^1,X^2,X^3,X^4,X^5)_0$	$(N^1,N^2,N^3,N^4,N^5)_1$	$ERT_{0,1}$	$(X^1,X^2,X^3,X^4,X^5)_1$	$ERT_{1,1}$	$(N^1,N^2,N^3,N^4,N^5)_2$	$ERT_{1,2}$
NODE 2 NODE 8 NODE 14 NODE 17 NODE 24	1,2,3,4,5, 6,7,8,9,11,12 10,14,15,20,21 13,16,17,18,19 22,23,24,25	17.172809	NODE 2 8,6.786144,11 NODE 14 NODE 17 23,2.175150,24	14.274396	1,2,3,4,5,6 8,11,12,13,15 7,9,10,14,20,21 16,17,18,19 22,23,24,25	13.827382

$(X^1,X^2,X^3,X^4,X^5)_2$	$ERT_{2,2}$	$(N^1,N^2,N^3,N^4,N^5)_3$	$ERT_{2,3}$
NODE 2 NODE 11 NODE 14 NODE 17 23,2.175150,24	13.813671	1,2,3,4,5,6 8,11,12,13,15 7,9,10,14,20,21 16,17,18,19 22,23,24,25	13.813671

TABLE 4.4 Location-Allocation Solution Sequence for $\lambda = 0.5$

of 13.813671. This "optimal" expected response time is approximately 20% smaller than the expected response time under the initial LA policy.

4.4.2 Some Further Observations

Observation of some interim results in the example of $\lambda = 0.5$ provides further insight into the problem.

The median-proximity policy which is the best LA policy for $\lambda = 0.01$ proves to be disastrous at a demand rate of 0.5; the system explodes and the expected response time becomes infinite.

Although this is the case for a location of servers according to the 5-median and districting by proximity of nodes to servers (the solution of Region A), given other location policies at $\lambda = 0.5$ the solution of Region A as the initial policy of the districting heuristic is superior to the solution of Region C. Also, as previously mentioned, an initial solution for the districting heuristic that is superior to another may provide an inferior final districting policy and therefore both the solutions to Regions A and C should be used as initial solutions at each districting policy evaluation.

Finally, to once again exhibit the superiority of the 5-median as the initial location policy, Tables 4.5 and 4.6 show the LA solution sequence for the 25-node network, with the initial location being different from the 5-median. The only change made to the initial location compared to the previous two examples is that the fifth

server is now located at Node 22 instead of Node 24. For $\lambda = 0.01$ the computational effort doubles although the "optimal" solution is ultimately reached and in the case of $\lambda = 0.5$, two more iterations are required in order to reach the "optimal" solution.

$(X^1,X^2,X^3,X^4,X^5)_1$	$(N^1,N^2,N^3,N^4,N^5)_1$	$ERT_{0,1}$	$(X^1,X^2,X^3,X^4,X^5)_1$	$ERT_{1,1}$	$(N^1,N^2,N^3,N^4,N^5)_2$	$ERT_{1,2}$
NODE 2 NODE 8 NODE 14 NODE 17 NODE 22	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21 12,15,16,17,18,19 22,23,24,25	3.261605	NODE 2 NODE 8 NODE 14 NODE 17 NODE 24	2.870753	1,2,3,4,5 6,7,8,9,11,13 10,14,20,21,22 12,15,16,17,18,19 23,24,25	2.719937

$(X^1,X^2,X^3,X^4,X^5)_2$	$ERT_{2,2}$
NODE 2 NODE 8 NODE 14 NODE 17 NODE 24	2.719937

TABLE 4.5 Location-Allocation Solution Sequence for λ=0.01
(Initial Location Not at 5-Median)

$(X^1,X^2,X^3,X^4,X^5)_0$	$(N^1.N^2.N^3.N^4,N^5)_1$	$ERT_{0,1}$	$(X^1,X^2,X^3,X^4,X^5)_1$	$ERT_{1,1}$	$(N^1,N^2,N^3,N^4,N^5)_2$	$ERT_{1,2}$
NODE 2 NODE 8 NODE 14 NODE 17 NODE 22	1,2,3,4,5 6,7,8,11,12,13 10,14,20,21,22,23 15,16,17,18,19 24,25	31.065704	NODE 2 8,3.982149,11 NODE 14 NODE 17 NODE 24	21.185316	1,2,3,4,5,6 8,11,12,15 7,9,10,14,20,21 13,16,17,18,19 22,23,24,25	15.029241

$(X^1,X^2,X^3,X^4,X^5)_2$	$ERT_{2,2}$	$(N^1,N^2,N^3,N^4,N^5)_3$	$ERT_{2,3}$	$(X^1,X^2,X^3,X^4,X^5)_3$	$ERT_{3,3}$
NODE 2 8,6.786144,11 NODE 14 NODE 17 23,2.175150,24	14.274396	1,2,3,4,5,6 8,11,12,13,15 7,9,10,14,20,21 16,17,18,19 22,23,24,25	13.827382	NODE 2 NODE 11 NODE 14 NODE 17 23,2.175150,24	13.813671

$(N^1,N^2,N^3,N^4,N^5)_4$	$ERT_{3,4}$
1,2,3,4,5,6 8,11,12,13,15 7,9,10,14,20,21 16,17,18,19 22,23,24,25	13.813671

TABLE 4.6 Location-Allocation Solution Sequence for λ=0.5
(Initial Location Not at 5-Median)

APPENDIX

```
    ∇ LOC_M
[1]    ⍝  A PROGRAM TO FIND OPTIMAL ONE-SERVER LOCATIONS IN M DISTRICTS
[2]    ⍝  INPUT: LA (DEMAND RATE),XX(DISTRICTING POLICY),DISTMAT (DIST. MATRIX)
[3]    ⍝         H (FRACTIONS OF DEMAND),M (NUMBER OF SERVERS)
[4]    ⍝         N (NUMBER OF NODES),V (VEHICLE SPEED),AL (R+W)
[5]    ⍝  OUTPUT: X (OPTIMAL LOCATIONS),ERTS (OVERALL NETWORK ERT)
[6]    '            LAMBDA IS EQUAL TO ';LA
[7]    '      THE DISTRICTS ARE'
[8]    XX
[9]    HH←(M,N)⍴0
[10]   EXP←LAM←M⍴0
[11]   X←(M,3)⍴0
[12]   ERTS←0
[13]   H1←H
[14]   LA1←LA
[15]   DISTANCE        ⍝ A SUBROUTINE FOR MINIMUM DISTANCE MATRIX
[16]   I←0
[17]  BIC:I←I+1
[18]   J←0
[19]  CIC:J←J+1
[20]   HH[I;J]←H1[J]×(XX[I;J]=1)
[21]   →(J<N)/CIC
[22]   LAM[I]←LA1×(+/HH[I;])
[23]   →(I<M)/BIC
[24]   Q←0
[25]  DIC:Q←Q+1
[26]   H←HH[Q;]÷(+/HH[Q;])
[27]   LA←LAM[Q]
[28]   LOC             ⍝ OPTIMAL LOCATION SUBROUTINE
[29]   X[Q;1]←NODA
[30]   X[Q;2]←XS
[31]   X[Q;3]←NODB
[32]   EXP[Q]←TRS
[33]   ERTS←ERTS+EXP[Q]×(+/HH[Q;])
[34]   →(Q<M)/DIC
[35]   ' THE OPTIMAL LOCATION IS '
[36]   X
[37]   'WITH AN ERT OF ';ERTS
    ∇
```

FIGURE 4.A.1: A Computer Code for Optimal m-Server Location (one server in each district)

FIGURE 4.A.2: A Computer Code for m-Server Districting.

```
    ∇ ALLOC_M
[1]   ⍝  A PROGRAM TO FIND THE OPTIMAL M-SERVER DISTRICTING POLICY
[2]   ⍝  INPUT: LA (DEMAND RATE),X (LOCATION POLICY),DISTMAT (DIST. MATRIX),
[3]   ⍝          H (FRACTION OF DEMAND),M (NUMBER OF SERVERS),
[4]   ⍝          N (NUMBER OF NODES),V (VEHICLE SPEED),AL (R+W)
[5]   ⍝  OUTPUT: XX (OPTIMAL DISTRICTING POLICY),ERSS (OVERALL NETWORK ERT)
[6]   'LAMBDA IS EQUAL TO ';LA
[7]   ' THE SERVERS ARE LOCATED AT '
[8]   X
[9]   COMBIN     ⍝ SUBR. TO FIND ALL SERVER PAIR COMBINATIONS
[10]  XXAR←(PN,M,N)⍴0
[11]  H1←H
[12]  LA1←LA
[13]  ERSS←10E17
[14]  HZ←N⍴0
[15]  XA←X
[16]  DISTANCE  ⍝ SUBR. FOR MINIMUM DISTANCE MATRIX
[17]  CREATEDISTM   ⍝ SUBR. FOR MATRIX OF DISTANCES OF NODES FROM SERVERS
[18]  KAP←-1
[19]  AAM        ⍝ SUBR. TO FIND THE OPTIMAL SOLUTION OF REGION A
[20]  XXZ←XX←AAX
[21]  '      THE INITIAL POLICY AA IS'
[22]  AAX
[23]  CYCLE←0
[24]  →AOA
[25] ABA:KAP←KK←0
[26]  ERSS←10E17
[27]  CCM        ⍝ SUBR. TO FIND THE OPTIMAL SOLUTION OF REGION C
[28]  XXZ←XX←CCX
[29]  '     THE INITIAL POLICY CC IS'
[30]  CCX
[31]  CYCLE←0
[32] AOA:KK←0
[33]  ERTM       ⍝ SUBR. TO CALCULATE THE ERT
[34]  ERTZZ←ERTT
[35]  →(ERSS=10E17)/BDA
[36]  CYCLE←CYCLE+1
[37]  '                    CYCLE ';CYCLE
[38] AAA:KK←KK+1
[39]  '              ITERATION ';KK
[40]  XX←XXZ
[41]  XXT←XX[(PERM[KK;1]);]+XX[(PERM[KK;2]);]
[42]  J←0
[43] ABB:J←J+1
[44]  HZ[J]←H[J]×XXT[J]=1
[45]  →(J<N)/ABB
[46]  H←HZ÷(+/HZ)
[47]  LA←LA×(+/HZ)
[48]  X←2 3⍴(X[(PERM[KK;1]);]),(X[(PERM[KK;2]);])
[49]  '  THE LOCATION OF SERVERS IS'
[50]  X
[51]  CREATEDIST
[52]  KOP←-1
[53]  LINCC
[54]  XX←CC
[55]  →BONA
[56] RUNA:AA2
[57]  XX←AA
[58]  KOP←0
[59] BONA:ALLOC      ⍝ OPTIMAL 2-SERVER DISTRICTING SUBR.
[60]  →(KOP=-1)/FIH
[61]  XXAA←XX
[62]  ERT
[63]  ERA←ERTT
[64]  →FIT
```

cont'd....

```
[65] FIH:XXCC←XX
[66]  ERT
[67]  ERS←ERTT
[68]  →RUNA
[69] FII:(ERA≤ERS)/FIJ
[70]  XX←XXCC
[71]  ERTT←ERS
[72]  →FIK
[73] FIJ:XX←XXAA
[74]  ERTT←ERA
[75] FIK:'       THE BEST ALLOCATION IS'
[76]  XX
[77]  ' WITH AN ERT OF ';ERTT
[78]  J←0
[79] ACA:J←J+1
[80]  XXZ[(PERM[KK;1]);J]←XX[1;J]×XXT[J]=1
[81]  XXZ[(PERM[KK;2]);J]←XX[2;J]×XXT[J]=1
[82]  →(J<N)/ACA
[83]  XX←XXAR[KK;;]←XXZ
[84]  X←XA
[85]  H←H1
[86]  LA←LA1
[87]  CREATEDISTM
[88]  ERTM
[89]  PERM[KK;3]←ERTT
[90]  ' ALLOCATION '
[91]  XX
[92]  ' ERT ';ERTT
[93]  →(KK<PN)/AAA
[94]  ORD←⍋PERM[;3]
[95]  XXZ←XX←XXAR[(ORD[1]);;]
[96]  ERTZZ←PERM[(ORD[1]);3]
[97]  ' THE BEST ALLOCATION FOR THIS CYCLE IS '
[98]  XX
[99] BDA:' WITH AN ERT OF ';ERTZZ
[100] →(ERTZZ=ERSS)/ADA
[101] ERSS←ERTZZ
[102] →AOA
[103] ADA:→(KAP=0)/ZAX
[104] EAAT←ERSS
[105] XAAX←XX
[106] ' THE OPTIMAL DISTRICTING WITH AA AS THE INITIAL POLICY IS'
[107] XAAX
[108] '  WITH AN ERT OF ';EAAT
[109] →ABA
[110] ZAX:ECCT←ERSS
[111] XCCX←XX
[112] ' THE OPTIMAL DISTRICTING WITH CC AS THE INITIAL POLICY IS'
[113] XCCX
[114] '  WITH AN ERT OF';ECCT
[115] →(EAAT≤ECCT)/ZBX
[116] XX←XCCX
[117] ERSS←ECCT
[118] →ZIZ
[119] ZBX:XX←XAAX
[120] ERSS←ECCT
[121] ZIZ:'       THE OPTIMAL ALLOC_M IS'
[122] XX
[123] '     WITH AN ERT OF ';ERSS
    ∇
```

```
    ∇ LOCALLOC_M
[1]    ⍝  A PROGRAM TO FIND THE OPTIMAL LOCATION-ALLOCATION POLICY
[2]    ⍝  INPUT: X (INITIAL LOCATION POLICY), LA (DEMAND RATE)
[3]    ⍝         DISTMAT (DIST. MATRIX),H (FRACTION OF DEMAND)
[4]    ⍝         M (NUMBER OF SERVERS),N (NUMBER OF NODES)
[5]    ⍝         V (VEHICLE SPEED), AL (R+W)
[6]    K←1
[7]    RES←10E17
[8]    H1←H
[9]    LA1←LA
[10] RUN:'        LOCATION-ALLOCATION ITERATION ';K
[11]  ALLOC_M
[12]  RES←ERSS
[13]  →(K=1)/BIA
[14]  TIP←(XX=XXP)
[15]  (((+/,TIP)÷(M×N))=1)/SOF
[16] BIA:XXP←XX
[17]  LOC_M
[18]  →(K=1)/PIL
[19]  TIZ←(X=XP)
[20]  →(((+/,TIZ)÷(3×M))=1)/SOF
[21] PIL:XP←X
[22]  H←H1
[23]  LA←LA1
[24]  K←K+1
[25]  →RUN
[26] SOF:' THE OPTIMAL LOCATION-ALLOCATION POLICY IS'
[27]  '                        LOCATION'
[28]  X
[29]  '                        ALLOCATION'
[30]  XX
[31]  'WITH AN ERT OF ';ERTS
    ∇
```

FIGURE 4.A.3: A Computer Code for m-Server Location-Allocation.

CHAPTER 5

SUMMARY AND CONCLUSIONS

This chapter provides a brief summary of the previous chapters and reviews the main conclusions. Chapter 1 presented the problem and emphasized the importance of LA decisions under conditions of congestion and queueing of service demands. Chapter 1 also included a brief literature review, a more detailed presentation of two papers that led to this study and a summary and outline of the following chapters.

In Chapter 2, the previously formulated one server location algorithm and the 2-facility districting heuristic were combined into a 2-facility LA heuristic. The heuristic is based on an alternate location and allocation solution improvement procedure which is a monotone-decreasing convergent process. Although the final result of the heuristic may or may not be equal to the global optimum, computational results indicated that the best results are in most cases obtained when the 2-median is used as the initial location policy.

Using a code of the heuristic in APL "optimal" 2-server LA policies were found for three networks: with 5, 10 and 25 nodes, and at various demand rates. For low values of λ, the median-proximity policy was found to be "optimal". For high demand rates, the best location of servers is not necessarily at the 2-median but the best districting policy is equal to the unique solution of Region C

given the best location policy. For intermediate λ values it was found that there may be many "optimal" solutions as shown in the case of the 25-node network.

In order to illustrate the changes in "optimal" LA policies with increasing λ values, 2-server LA policies were found for a 3-node network. Graphs of the expected response time as a function of the demand rate for the 5, 10 and 25 node networks were presented at the end of the chapter which suggest convexity of the expected response time as a function of λ.

Chapter 3 dealt with the 3-server LA problem on a congested network which required the development of a 3-server districting heuristic. An iteration procedure was introduced which improves the districting between two servers at each iteration, while keeping the allocation of nodes to the third server constant. This heuristic was coded in APL and computational results of the districting heuristic were presented.

The alternate location and allocation solution improvement procedure presented in Chapter 2 was used for obtaining the "optimal" LA policy with three servers. The superiority of the 3-median as an initial location of servers was shown by comparing final solutions of the heuristic using the 3-median as the initial location to final solutions that were based on a different initial location policy. In most cases the 3-median proved to be either as good or better than any other initial location policy.

Finally, a comparison of results for three servers with those for two servers showed that especially at high demand rates the percentage improvement in expected response time is greater than the percentage increase in resources (the number of servers). Also shown was the greater range of demand rates under which three servers can operate in various networks compared to two servers.

The use of the alternate location and allocation solution improvement procedure for two and three servers as well as the extension of the 2-server districting heuristic to the case of three servers, served as preliminary work to the treatment of the general case in Chapter 4. In order to solve the m-server LA problem, the development of an m-server districting heuristic was required. This heuristic extended the 3-server districting solution procedure to m servers by improving the districting between two servers at a time, keeping the previous allocation of nodes to all other servers constant.

The choice of an initial districting policy according to the original 2-server districting study was modified in this thesis. Rather than taking the better of the solutions of Regions A and C as the initial solution of the districting heuristic and running through it only once (as suggested by Berman and Larson[5]), it was found that both should be used as initial solutions in two separate runs. The better result of the two is then taken as the "optimal" solution. Even for a small number of servers (two or three), computational experience had shown that in some cases this procedure which largely

increases the computational effort, provides a better districting policy.

To illustrate the usefulness of the m-server LA heuristic, two examples were presented. Both were of a 5-server, 25-node network, one with a low demand rate and the second example with a relatively high demand rate. For the low demand rate, the best LA policy was found to be equal to the median-proximity policy. In the second case, the best LA policy was not at all obvious and required several iterations until the final solution was reached. Solving the same two problems but with initial location policies not equal to the 5-median provided the "optimal" solutions but required more iterations.

A general observation of the alternate location and allocation solution sequence refers to the relative improvement of the solution throughout this sequence. In general, the greatest relative improvement occurs in the first iteration (i.e. when an optimal location is found for the first districting policy), with the expected response time subsequently decreasing by smaller increments. With a "weak" initial location policy the initial improvement is usually larger than when the p-median is used.

The significance of the solution methods presented in this thesis increases with the degree of congestion in the networks analyzed. While for low demand rates the expected response time is not very sensitive to changes in LA policies, at high rates of demand even slight changes in server locations or in node allocations could be "disastrous". It was shown that even for simple networks optimal LA

policies at high demand rates may not be intuitive and that using the median-proximity LA policy at high demand rates can cause the system to explode.

All computations were done on a Honeywell Information Systems computer through the Multics operating systems running on DPS-8-M processors. Although most of the coding was in APL, the integer programming computations related to the examples in Chapter 4 were performed by the MPS (Mathematical Programming System) in Multics. The following table provides CPU computation times for problems of varying complexity at a demand rate of 0.1. In these examples, only the best of the solutions of Regions A and C was taken as the initial solution of the districting heuristic at each iteration, and the p-median was used as the initial location policy.

NUMBER OF NODES	NUMBER OF SERVERS	CPU TIME (minutes:seconds)
5	2	1:15
10	2	5:30
25	2	31:00
5	3	2:00
10	3	8:15
25	3	32:00

TABLE 5.1 Examples of CPU Computation Times for the Location-Allocation Problem

It is believed that further improvement is possible in the solution methods presented in this thesis and especially in the districting heuristics. A decrease in the number of iterations in each cycle would largely reduce the computational effort. Further study should also deal with the determination of a lower bound solution and a comparison of results obtained through heuristics to this lower bound. Work is currently being done to improve the efficiency of the computer codes in APL. Although computational results point in the direction of convexity of the expected response time as a function of the demand rate, this characteristic does not have an apparent use in the determination of optimal LA policies. Future study can treat the LA problem in congested networks with probabilistic link lengths. An extension to this study could also allow more than one mobile service unit at each facility. Another area of future research may treat cooperation between districts, but as the M/G/k queueing problem is not analytically solvable, some approximations would have to be introduced in a solution to this problem. Finally, it may be more appropriate for certain systems to solve the centre problem taking congestion into account. This change would affect both the location and districting solutions of the LA heuristic by changing the criterion of optimality from the minimization of average response time to the minimization of maximum response time.

REFERENCES

1. Ahituv, N. and O. Berman, "Quantitative Models for Policy Making in Urban Emergency Networks", International Journal of Policy and Information, 6, No. 2, December 1982, pp. 35-42.

2. Beaumont, J. R., "Location-Allocation Problems in a Plane. A Review of Some Models". Socio-Economic Planning Sciences, 15, No. 5, 1981, pp. 217-229.

3. Berman, O., R. C. Larson and S. S. Chiu. "Optimal Server Location on a Network Operating as an M/G/1 Queue", to appear in Operations Research.

4. Berman, O. and R. C. Larson, "The Median Problem With Congestion", Computers and Operations Research, 9, No. 2, 1982, pp. 119-226.

5. Berman, O. and R. C. Larson, "Optimal 2-Facility Network Districting in the Presence of Queuing", Submitted to Transportation Science.

6. Berman, O. and Y. Wand, "Service Decisions with Two Criteria", IEEE Transactions, SMC-12, No. 4, July-August 1982, pp. 549-551.

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

8. Cooper, L., "Location-Allocation Problems", Operations Research, 11, No. 3, May-June 1963, pp. 331-343.

9. Cooper, L., "Heuristic Methods for Location-Allocation Problems", S.I.A.M. Review, 6, No. 1, January 1964, pp. 37-53.

10. Cooper, L., "The Transportation-Location Problem", Operations Research, 20, No. 1, January-February 1972, pp. 94-108.

11. Handler, G. Y. and P. B. Mirchandani, Location On Networks, The M.I.T. Press, Cambridge, Massachusetts, 1979.

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

13. Jarvis, J. P., "Optimization in Stochastic Service Systems with Distinguishable Servers", TR-19-75, Innovative Resource Planning Project in Urban Public Safety Systems, Operations Research Centre, M.I.T., June 1975.

14. Keeney, R. L., "A Method of Districting Among Facilities", Operations Research, 20, No. 3, May-June 1972, pp. 613-618.

15. Klinrock, L., Queueing Systems I, Wiley, New York, Chapter 5, 1975.

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

17. Larson, R. C., "A Hypercube Queueing Model for Facility Location and Redistricitng in Urban Emergency Services", Computers and Operations Research, 1, No. 1, March 1974, pp. 67-95.

18. Love, R. F., "One Dimensional Facility Location-Allocation Using Dynamic Programming", Management Science, 22, No. 5, January 1976, pp. 614-617.

19. Madsen, O. B. G., "Methods for Solving Combined Two Level Location-Routing Problems of Realistic Dimensions", European Journal of Operational Research, 12, No. 3, March 1983, pp. 295-301.

20. Murtagh, B. A. and S. R. Niwattisyawong, "An Efficient Method for the Multi-Depot Location-Allocation Problem", Journal of the Operational Research Society, 33, No. 7, July 1982, pp. 629-634.

21. Nambiar, J. M., L. F. Gelders and L. N. Van Wassenhove, "A Large Scale Location-Allocation Problem in the Natural Rubber Industry", European Journal of Operational Research, 6, No. 2, February 1981, pp. 183-189.

22. Narula, S. C. and V. I. Ogbu, "A Hierarchial Location-Allocation Problem", OMEGA, The International Journal of Management Science, 7, No. 2, 1979, pp. 137-143.

23. Neebe, A. W., "A Branch and Bound Algorithm for the p-Median Transportation Problem", Journal of the Operational Research Society, 29, No. 10, October 1978, pp. 989-995.

24. Or, I. and W. P. Pierskalla, "A Transportation Location-Allocation Model for Regional Blood Banking", AIIE Transactions, 11, No. 2, June 1979, pp. 86-95.

25. Sheppard, E. S., "A Conceptual Framework for Dynamic Location-Allocation Analysis", Environment and Planning, 6, No.5, 1974, pp. 547-564.

26. Stidham, S., "Stochastic Design Models for the Location and Allocation of Service Facilities in a Network", Presented at the Operations Research Society of America, Dallas, Texas, U.S.A., 1971.

27. Sule, D. R., "Simple Methods for Uncapacitated Facility Location-Allocation Problems", Journal of Operations Management, 1, No. 4, May 1981, pp. 215-223.

28. Tansel, B. C., R. L. Francis and T. J. Lowe, "Location On Networks: A Survey", Management Science, 29, No. 4, April 1983, pp. 382-511.

29. Wagner, H. M., Principles of Operations Research, Prentice Hall Inc., Chapter 15, 1975.

30. Wesolowsky, G. O. and W. G. Truscott, "The Multi-Period Location-Allocation Problem with Relocation of Facilities", Management Science, 22, No. 1, September 1975, pp. 57-65.