

ROUTING AND CAPACITY ASSIGNMENT IN BACKBONE
COMMUNICATION NETWORKS

DISSERTATION

Presented in Partial Fulfillment of The Requirements for the
Degree of Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Ali Amiri

The Ohio State University

1992

Dissertation Committee:

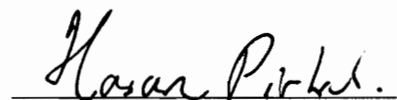
Dr. Hasan Pirkul

Dr. John Current

Dr. Varghese Jacob

Dr. Pai-Cheng Chu

Approved by

A handwritten signature in black ink, reading "Hasan Pirkul.", is written over a horizontal line.

Graduate Program in

Business Administration

To My Parents, Khemais and Mna

and to

Wife, Henda

ACKNOWLEDGEMENTS

I take this opportunity to express my sincere appreciation to Dr. H. Pirkul for his guidance and interest throughout the research. I also express my appreciation to the other members of my advisory committee Drs. J. Current, V. Jacob and P. C. Chu for their helpful suggestions and comments.

To my father, mother and brothers, I offer thanks for their support and trust. I also thank my wife Henda and our future child Msaddak.

VITA

January 26, 1962	Born, Grombalia, Tunisia
1981-1985	B.S., Institut des Hautes Etudes Commerciales in Tunis, Tunisia
1986-1988	MBA, The Ohio State University,
1988-1992	Graduate Teaching Assistant, Department of Management Science, The Ohio State University, Columbus, Ohio

FIELDS OF STUDY

Major Field: Business Administration

Minor Field: Management Information Systems

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
VITA	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix

CHAPTER

I.	INTRODUCTION	1
	1. Design of Communication Networks	1
	2. The Routing Problem in Backbone Communication Networks	4
	3. The Routing and Capacity Assignment Problem in Backbone Communication Networks	9
	4. The Lagrangean Relaxation Method	11
	5. Significance of the Research	16
II.	LITERATURE REVIEW	17
	1. The Routing problem in Backbone Communication Networks	17
	2. The Routing and Capacity Assignment Problem in Backbone Communication Networks.	19

III.	THE ROUTING PROBLEM IN BACKBONE COMMUNICATION NETWORKS: MINIMIZING THE MAXIMUM LINK QUEUEING DELAY	25
	1. Introduction	25
	2. Problem Formulation	26
	3. A Lagrangean Relaxation of the Problem	31
	4. Solution Techniques	32
	5. Computational Results	36
	6. Conclusion	37
IV.	THE ROUTING PROBLEM IN BACKBONE COMMUNICATION NETWORKS: MINIMIZING THE AVERAGE QUEUEING DELAY . . .	41
	1. Introduction	41
	2. Problem Formulation	42
	3. A Lagrangean Relaxation of the Problem	45
	4. Solution Techniques	48
	5. Computational Results	50
	6. Conclusion	52
V.	THE PRIMARY AND SECONDARY ROUTE SELECTION PROBLEM IN BACKBONE COMMUNICATION NETWORKS	58
	1. Introduction	58
	2. Problem Formulation	59
	3. A Lagrangean Relaxation of the Problem	65
	4. Solution Techniques	68
	5. Computational Results	72

6. Conclusion	73
VI. THE ROUTING AND CAPACITY ASSIGNMENT PROBLEM IN BACKBONE COMMUNICATION NETWORKS	76
1. Introduction	76
2. Problem Formulation	78
3. A Lagrangean Relaxation of the Problem	82
4. Solution Techniques	85
5. Computational Results	90
6. Effects of Cut Constraints on the Quality of the solutions	97
7. Conclusion	104
VII. CONCLUSIONS	109
LIST OF REFERENCES	112

LIST OF TABLES

Table	Page
3.1. Summary of Computational Results for ARPA Network	39
3.2. Summary of Computational Results for OCT Network	39
3.3. Summary of Computational Results for Networks With 10 Nodes	40
3.4. Summary of Computational Results for Networks With 15 Nodes	40
4.1 Summary of Computational Results	57
5.1. Summary of Computational Results	75
6.1. Link Capacity and Its Costs Components	91
6.2. Computational Results With Diffrent Message Lengths	93
6.3. Computational Results With Diffrent Delay Costs	94
6.4. Computational Results With Diffrent Fixed Costs.	95
6.5. Computational Results With Diffrent Variable Costs.	96
6.6. Computational Results With Diffrent Message Lengths Using Cut Constraints	105
6.7. Computational Results With Diffrent Delay Costs Using Cut Constraints . . .	106
6.8. Computational Results With Diffrent Fixed Costs Using Cut Constraints . . .	107
6.9 Computational Results With Diffrent Variable Costs Using Cut Constraints . .	108

LIST OF FIGURES

Figure	Page
1.1. Schematic Representation of a Centralized Communication Network	5
1.2. Schematic Representation of a Decentralized Communication Network	6
4.1. ARPA Network	53
4.2. OCT Network	54
4.3. USA Network	55
4.4. OCT Network	56

CHAPTER I

INTRODUCTION

1. Design of Communication Networks

Change is rapidly occurring throughout the information technology domain, but nowhere is this change more dramatic and more evident than in the area of telecommunications and networking. A communications revolution is taking place that is directly or indirectly affecting the performance of every business. To respond to the global and international orientation of many companies, a large number of firms are now competing to develop and market telecommunications equipment and services. Partially because of this increased competition, innovation in the telecommunications and networking area is developing quickly. Digital networks, fiber optics, and the ability to send both voice and data simultaneously over the same lines have been major components of the revolution.

In order to adjust more quickly to market opportunities and competitors' moves, many firms have decentralized and geographically deconcentrated their operations. This has resulted in a growing need for more reliable voice and data communication among the different parts of the companies and with the customers and suppliers.

There are different incentives for networking. One of them is that networking allows various users on the network to share important and often expensive resources. For example, it is very common for mainframes or supercomputers to share magnetic disk

devices and high-speed line or laser printers. Another incentive is that distributed data processing depends upon highly reliable telecommunication networks in order to provide an acceptable level of service and response time to users.

There are many applications that require telecommunications. First, the basic of these is simply the access of remote databases such as information and financial services available to personal computer users. More complex applications involve remote updating of databases, in addition to accessing the data. Airline reservation systems, automatic teller machines, inventory control systems provide a number of examples. Another application involves using a remote computer system for some computational task. This could happen if there is no local computer, or it is not operational or if the remote computer can carry out the task more efficiently.

Many factors affect the operations and performance of a telecommunication network. One major factor is the routing and capacity assignment policies for the network, main topic of our dissertation. In the design process tradeoffs have to be made between the response time to users and the cost of the network. If high capacities are assigned to the links in the network, connection costs will be high while response time will be low. On the other hand, if low capacity links are installed the reverse will be true. This argument shows that the tradeoffs between response time and connection costs are integral part of the network design. The main focus of this research is on issues concerning routing and capacity assignment in telecommunication networks. Currently, network designers use heuristic solution techniques during the design process. However, it is not possible using only such techniques to analyze the quality of the resulting design in terms of cost and response time. This dissertation develops mathematical programming techniques that

directly consider both costs as well as service quality to design and operate telecommunications networks.

Telecommunications networks can be classified as 1) local area networks, 2) metropolitan area networks or 3) wide area networks. A local area network is typically a network that connects different computing devices such as terminals or personal computers and other data processing devices such printers and file servers located in a small geographical area such as a university campus. A local area network may represent a subnetwork that connects a collection of terminals to a wide area network through various types of concentrators.

Metropolitan networks are a type of networks which are larger than local area networks. Typically, these networks are designed to connect different buildings of an organization located within a metropolitan area or to connect factories and offices located within a distance of several miles. In general this type of networks is used to fill the gap between local area networks and wide area networks.

Wide area networks are used to connect devices spread over a wide geographical area of hundreds or even thousands of square miles. These networks are designed to allow users to use the computing and storage capabilities of some remotely located mainframe host computers in the network. The users may employ simple terminals or devices with some limited data processing such as personal computers.

There are two classes of wide area networks 1) centralized computer networks and 2) distributed computer networks. Centralized computer networks usually include a single mainframe host computer. Figure 1.1 provides a schematic description of a centralized computer network. User nodes are either directly connected to the central computer or connected via concentrators and point-to-point lines or multi-point lines. Distributed

networks contain several mainframe host computers. A subnet or backbone of nodes or switches, usually connected with high speed communication links, is used as an integral part of the network to tie together the various computers and connect the users to those computers (figure 1.2). User nodes are usually connected to the switches through concentrators and point-to-point lines or multi-point lines. Messages originate at these user nodes, pass into the subnet, pass from a backbone switch to another on the communication links until they reach their final destination which is either a host computer or another user node. The switches of the backbone, usually computers in their own right, serve primarily to route the messages through the backbone network typically in store and forward fashion using packet switching techniques.

The cost of a wide area network is presently dominated by transmission costs as opposed to local or metropolitan area networks. Thus, it is critical to use the communication links efficiently, even at added computational costs. With the deployment of high-speed fiber optics, the efficient use of a transmission facility becomes less important. However, the issue of reliability of the network takes an increased importance during the design process. In this dissertation, we will discuss this issue in relation with costs and service quality in wide area networks. In the next sections, we will review in more details the design issues that are addressed in this dissertation.

2. The Routing Problem in Backbone Communication Networks

Routing policy in backbone communication networks is an important aspect of the design and operation of the networks. It has a significant effect on the response time experienced by the network users and on the utilization of the network resources (e.g. node

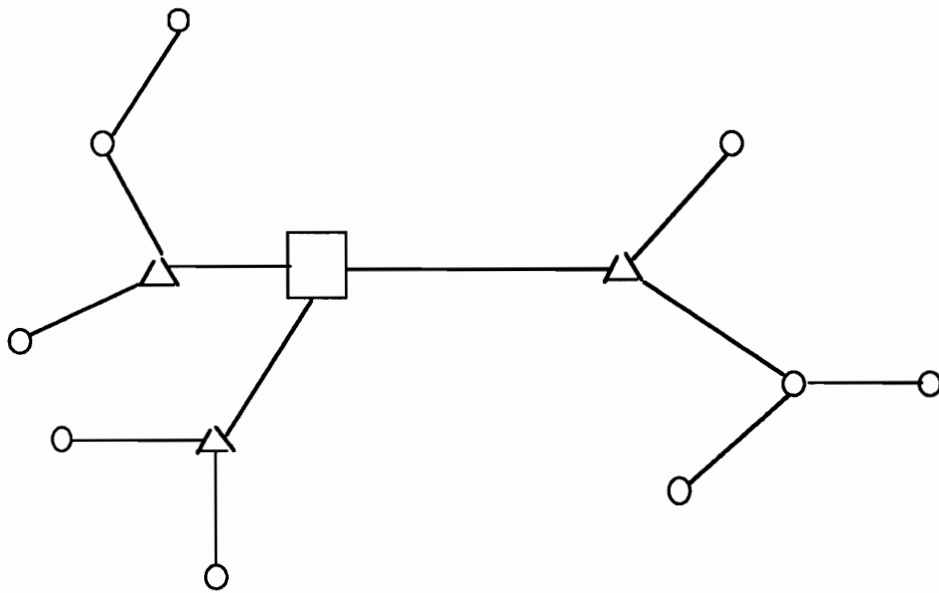


Figure 1.1: Schematic Representation of a Centralized Communication Network.

- Central Computer
- △ Concentrator
- Terminal

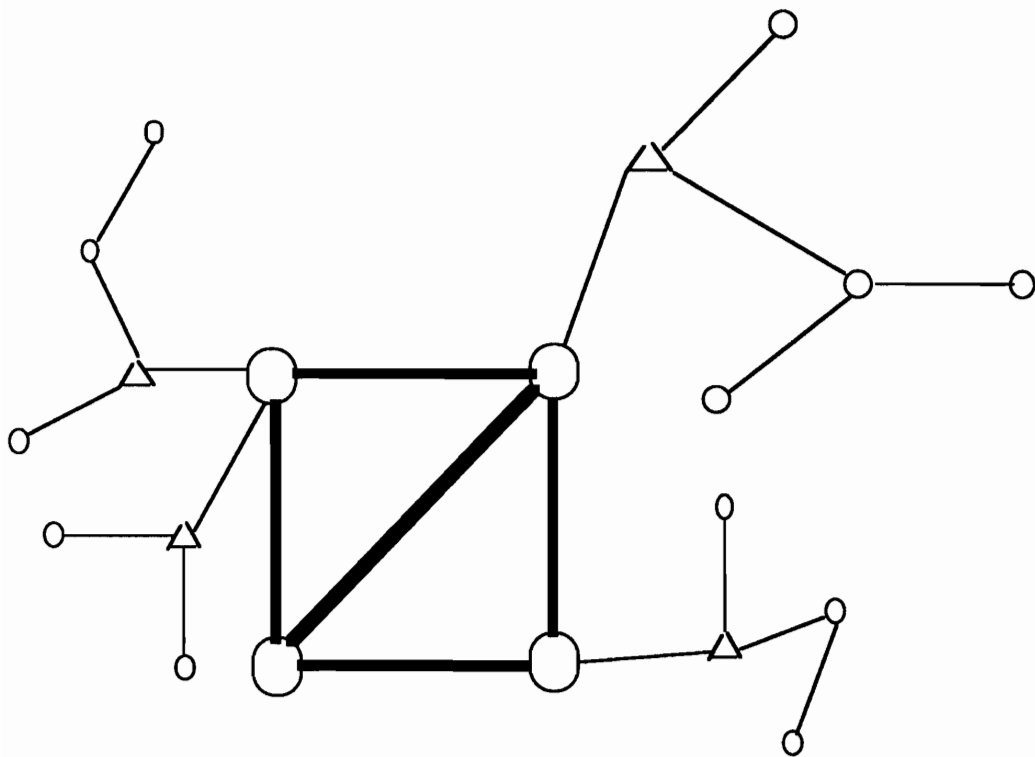
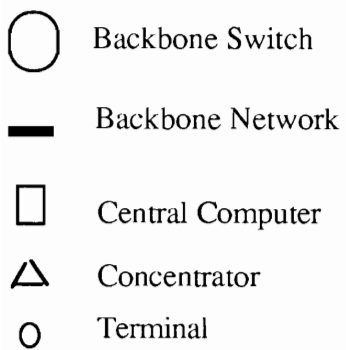


Figure 1.2: Schematic Representation of a Decentralized Communication Network.



buffers and link capacities). A good routing policy could also allow new users to use the network without significant deterioration of the quality of service to existing users and without incurring the costs of establishing new links or upgrading the capacities of existing links.

In backbone networks, messages or packets are transmitted from the source node to the destination node through intermediate links and nodes along paths which are either predetermined or determined dynamically. If a link is busy, messages sent through that link wait at the network node from which that link emanates. Given a particular network topology and link capacities, the queueing and transmission delays can be significantly reduced with improved routing schemes.

Routing policy is the policy of setting up a path between every pair of communicating nodes along which messages are transmitted. There are several alternative routing methods such as fixed routing, random routing and adaptive routing. In fixed routing, the path along which the messages are sent is fixed for every source/destination pair. The path does not reflect the dynamic conditions of the network, but it is adjusted and left unchanged for periods of time large compared to normal fluctuations in the network. In random routing, one of the potential routes from source to destination is randomly selected using probabilities of choosing a particular link emanating from a node based on traffic load, line capacity, and other network conditions. Finally adaptive routing techniques determine routes according to the instantaneous states of the queues at the various links in the network. Each node (switch) has a routing table that indicates the best route for any destination in the network depending on current network conditions such as traffic congestion and link or node failure.

Another method of classification of routing schemes is based on whether they are bifurcated or nonbifurcated. In bifurcated routing, messages between an origin node and destination node can be sent through several routes. In nonbifurcated routing, there is only one route over which messages between a communicating node pair are sent.

In this dissertation, the routing method is assumed to be static and nonbifurcated. Most of the commercially available networks such as DATAPAC [7,39], TELENET[57], TRANSPAC [12] and TYMNET [51,57] have adopted static or semidynamic routing schemes. Even in designing networks which use adaptive routing, fixed routing is usually assumed because network configurations optimized with fixed routing are (near) optimal for adaptive routing operations [29].

Given the network topology, link capacities, and traffic requirements between communicating node pairs, the routing problem to be studied here can be further described as that of selecting, for every source/destination pair, one route from among all possible routes over which all messages for that pair will be routed. In this dissertation, two objectives will be considered. The first one is to minimize the maximum link queueing delay encountered by messages at the network nodes. The second objective is to minimize the average queueing delay. The second objective is commonly used as a criterion for solving the routing problem. However, the first objective is also valid for two reasons. First, one fast way to minimize the average queueing delay in the network is to minimize the maximum link delay. Second, this criterion leads to a more balanced distribution of the traffic load over the network links. Thus, in the case of a failure of the most utilized link, less traffic would be disrupted than in the case of minimizing the average queueing delay.

We also examine the problem of selecting primary and secondary routes for each communicating node pair to minimize the average delay encountered by messages at

network switches. That is, we address the routing problem in backbone networks that are survivable under single-link failures. Survivability, in this context, is the ability to reroute messages through alternate routes in the network in case of a link failure. In our model, survivability is achieved by selecting a primary route and a link-disjoint alternate route. Because the two routes are link disjoint, one route will always be available in any single link failure scenario. Although capacity is reserved on the network links for both the primary and secondary routes, messages are normally routed via the primary routes. If the failed link is on the primary route of a communicating node pair, messages between that pair are rerouted over the secondary route. It is expected that all communicating node pairs are switched back to their primary routes as soon as the failed link is back in service.

The routing problem is very difficult to solve because of its combinatorial nature. In chapters III-V, mathematical programming formulations of the problem are presented and efficient solution procedures based on a Lagrangean relaxation of the problem are developed. The results of extensive computational experiments showing the effectiveness of the solution procedures are also reported.

3. The Routing and Capacity Assignment Problem in Backbone Communication Networks

Even though the routing problem which minimizes the maximum link delay or the average delay is very important, it is only one aspect of the design problem of a packet switched communication network. A more general problem is to determine simultaneously the link capacities and routes over which messages between communicating node pairs are transmitted. The goal is to design a network with minimum overall system costs. Systems

costs are composed of connection costs which depend on link capacities and delay costs incurred by users due to the limited capacities of the links and the resulting queueing at intermediate nodes.

The routing and capacity assignment problem captures the tradeoffs that should be made between connection and message delay costs during the design process. If high capacity links are used in the network, connection costs will be high while delay costs will be low. On the other hand, if low capacities are assigned to links, the opposite will be true. This shows that tradeoffs between delay and capacity costs are an important aspect of a proper design and operation of a communication network.

The routing and capacity assignment problem can be specifically described as follows. Given the network topology (location of the nodes and links), the traffic requirements between source/destination pairs, a set of link types with different capacities and costs, and a unit cost of delay, one must determine: (1) the capacity to assign to each link and (2) the best route between communicating node pairs. The objective is to minimize the overall system costs. The routing policy is assumed to be static and nonbifurcated. The route for each source/destination pair is to be chosen from among all possible routes for that pair, contrary to what has been done in most of the previous studies where the routes are chosen from among a prespecified subset of all possible routes.

The routing and capacity assignment problem is studied in chapter VI. A mathematical formulation of the problem is presented. Lagrangean relaxation embedded in a subgradient optimization procedure is used to obtain tight lower bounds. An effective solution method is developed and extensive computational results are reported.

4. Lagrangean Relaxation Technique

It is well known that there are two classes of combinatorial optimization problems. The first class, called P, includes "easy" problems that can be solved in time bounded by a polynomial function of the input length. The second class, called NP-Complete, includes "hard" problems for which the computing time required by any known exact solution algorithm grows exponentially with the problem size. It is believed that it is unlikely that exact solution procedures for this class of problems can be developed which can be used to provide optimal solutions in reasonable amount of computing time.

Many of the problems in class NP-Complete problems may be viewed as easy problems complicated by a number of side constraints. The Lagrangean relaxation method is based on this observation. It involves relaxing (dualizing) the side constraints by using dual multipliers. The new problem is called the Lagrangean problem whose optimal value is a lower bound (for minimization problems) on the optimal value of the original problem. Frequently the solution to the Lagrangean problem can be used to construct a feasible solution to the original problem. The quality of this feasible solution is estimated by comparing its value to the lower bound which is used as an estimate of the unknown optimal solution value of the original problem. This ability to evaluate the quality of the feasible solutions is an important feature of heuristics based on the Lagrangean relaxation technique.

The use of the generalized lagrangean multipliers was first suggested by Everett [15]. The successful application of the Lagrangean method to the travelling salesman problem by Held and Karp [33] has led to its use in a variety of other problems such as location problems [11,14,43,45], scheduling problems [16] computer network problems [24,44]. In this section we briefly present the Lagrangean relaxation method. For a survey of this method the reader is referred to [4,17,18].

Consider the following integer programming problem (P):

Problem (P):

$$Z_p = \text{Min } f(X) \quad (1)$$

subject to:

$$AX \leq b \quad (2)$$

$$HX \leq e \quad (3)$$

$$X \geq 0 \text{ and integer} \quad (4)$$

where X is $n \times 1$, b is $m \times 1$, e is $k \times 1$ and matrix A is of dimension $m \times n$, and matrix H is $k \times n$ and $f(X)$ is function of X .

We assume the constraints of problem (P) have been partitioned into sets (2) and (3) so that after the constraints in set (2) are dualized, the lagrangean problem (L) formulated below is relatively easy to solve.

Problem (L):

$$Z_L(\omega) = \text{Min } f(X) + \omega (AX - b) \quad (5)$$

subject to:

$$HX \leq e \quad (6)$$

$$X \geq 0 \text{ and integer} \quad (7)$$

where $\omega = (\omega_1, \omega_2, \dots, \omega_m)$ is a vector of nonnegative multipliers.

$Z_L(\omega)$ is a lower bound.

Even though the theory behind Lagrangean relaxation is straight forward and generally well understood [4], the use of this technique to solve optimization problems still remains a challenge. There are very important issues that must be dealt with when using the Lagrangean relaxation method. Among these issues are:

(1) How does one select an appropriate value for ω ? Can one find a value for ω for which $Z_L(\omega)$ is equal to or close to Z_P ?

(2) How does one choose among the various possible relaxations of problem (P) , i.e. different Lagrangean relaxations and LP relaxation?

(3) How might the solutions to problem (L) be used to obtain feasible solutions for (P) ? How good are these feasible solutions likely to be?

It is clear that the best choice for ω would be an optimal solution to the dual problem (D) :

$$Z_D = \max_{\omega} Z_L(\omega) \quad (D)$$

Most schemes for determining ω have as their objectives finding optimal or near optimal solutions to (D) .

In this section we review three approaches that have been used in Lagrangean relaxation applications to solve problem (D) : (1) the subgradient method, (2) various versions of the simplex method employing column generation techniques, and (3) multiplier adjustment method.

The subgradient method is a modified version of the gradient method in which subgradients replace gradients [34]. Given an initial multiplier vector ω^0 , usually set to zero, a sequence of multipliers is generated by updating the vector at the iteration k using the formula

$$\omega^{k+1} = \omega^k + t_k (AX^k - b),$$

where ω^{k+1} and ω^k are the multiplier vectors at iterations $k+1$ and k respectively, X^k is the optimal solution to the Lagrangean Problem L with multiplier vector ω^k , t_k is a positive scalar step size, and $AX^k \leq b$ is the set of constraints being relaxed. It is well known that $\limsup_{k \rightarrow \infty} Z_L(\omega^k)$ converges to $Z_L(\omega^*)$ if $t_k \rightarrow 0$ and $\sum_{k=0}^{\infty} t_k \rightarrow \infty$ [47]. Since in general these conditions are very difficult to

satisfy, this method is always used as a heuristic. In this study, we use the following step size that has been found to be satisfactory in practice

$$t_k = \lambda_k (Z_f - Z_L(\omega^k)) / \|AX^k - b\|^2,$$

where Z_f is the value of the best feasible solution found so far and λ_k is scalar satisfying $0 \leq \lambda_k \leq 2$. This scalar is set to 2 at the beginning of the algorithm and is halved whenever the bound does not improve in 20 consecutive iterations. The algorithm is terminated after a specified number of iterations unless an optimal solution is reached before that point. The algorithm is also terminated if the gap between the best lower bound and the best feasible solution found is less than 0.1% of the best lower bound, or the best lower bound does not improve in 100 consecutive iterations by at least 0.01%.

The second group of methods to solve problem (D) is based on applying a variant of the simplex method to (P) . The primal simplex method with column generation has been used frequently. However, this approach is known to converge very slowly and does not produce monotonically increasing lower bounds [20,21]. Generally, the simplex-based methods are harder to program and required more computational time than has the subgradient method.

The third set of procedures that have been used for determining Lagrangean multipliers consists of multiplier-adjustment methods. Multiplier-adjustment methods are heuristics for the dual problem of (D) that are developed for an application for which some special structure of the dual problem can be exploited. If a multiplier-adjustment method can be developed for some problem class, then one is usually able to improve on the subgradient method. However, because the subgradient method is easy to program and has performed robustly in a wide variety of applications it will be used in this dissertation.

Two properties are important in evaluating a particular relaxation: the tightness of the bounds produced and the amount of computation required to obtain these bounds. Usually selecting a relaxation involves making tradeoff between these two properties; tighter bounds usually require more time to compute. However, it is usually possible to at least compare the bounds and computational requirements for different relaxations and then choose the best.

It is important to state a sufficient condition for which Lagrangean relaxation and linear programming relaxation are equivalent [27], that is for which $Z_D = Z_{LP}$. Namely, $Z_D = Z_{LP}$ whenever $Z_D(\omega)$ is not decreased by removing the integrality restriction on X from the constraints of the Lagrangean problem. This is called the integrality property. It should be emphasized that the integrality property is not defined relative to a given problem class but relative to a given integer programming formulation of a problem class. This is an important distinction because problem (P) may have more than one formulation.

Solutions to problem (L) can be used to obtain feasible solutions for (P) . The method which is used for this purpose is called the Lagrangean based solution method. It is possible in the course of solving (L) that a solution to this problem will be discovered that is feasible in (P) . Because dualized constrained $Ax = b$ are equalities, this solution is also optimal for (P) . If the dualized constraints include some inequalities, a Lagrangean problem solution can be feasible but suboptimal for (P) . However it is rare that a solution of either type be identified. On the hand, it often happens that a solution to (L) obtained while solving (L) will be "nearly" feasible for (P) and can be made feasible with some judicious thinking.

5. Significance of the Research

The results of the models developed in this dissertation are significant in terms of their theoretical and practical contributions. Theoretically, we present new efficient approaches to solve several large scale telecommunication problems using mathematical programming techniques. These techniques are used to obtain good feasible solutions to the problems and to assess the qualities of these solutions. The results of this research are also beneficial to the analysis and design of networks in other areas such transportation and telephone networks. They also represent a significant contribution to the research on nonlinear combinatorial optimization. As we will see in the following chapters, the objective functions of the problems discussed in this dissertation are nonlinear. Previous research on nonlinear programming has mainly focused on solving small problem instances and studying the quadratic assignment problem.

Practically, the approaches developed here can be used by commercial network vendors to design new computer networks. They can also be used to reduce capacity costs of existing networks, specially when it has been noticed that many operational communication networks are characterized by excessive capacities. Network managers can also use these approaches to study the effect of adding new users on the current level of service to existing users.

CHAPTER II

LITERATURE REVIEW

1. The Routing Problem in Backbone Communication Networks

In this section the literature on routing in backbone network is reviewed. As mentioned in the previous chapter, one method of classifying routing schemes is based on whether they are bifurcated or nonbifurcated. In bifurcated routing, messages between an origin node and destination node are sent through several routes. In nonbifurcated routing, there is only one route over which messages between a communicating node pair are sent. Most of the previous research in computer communication networks has concentrated on the bifurcated routing. Early research efforts on bifurcated routing have been devoted to develop heuristic algorithms [21,28,38]. With the advances in mathematical programming techniques, some researchers such as Frank and Chou [20], Canter and Gerla [8], and Bertsekas [5] developed improved algorithms to determine the optimal routes in backbone networks with bifurcated routing. These algorithms are based on the formulation of a continuous mathematical programming problem with nonlinear convex objective function representing the average network delay. These algorithms have employed the gradient projection, flow deviation, and external flow methods among others.

Courtois and Semal [10] developed a heuristic based on a modified version of the flow deviation method to solve the nonbifurcated routing problem. They tested their heuristic on a variety of networks and were able to generate good solutions for lightly

loaded networks. Gavish and Hantler [23], Narasimhan et al. [42], and Tcha and Maruyama [56] have all used solution procedures based on mathematical programming techniques to solve the nonbifurcated routing problem. In [23] Lagrangean relaxation was used to obtain lower bounds as well as feasible solutions to minimize average message delays. They reported results of computational experiments with reasonable gaps between lower bounds and feasible solutions. Narasimhan et al. presented a new formulation for the same problem. This formulation led to a new relaxation which was shown to be capable of yielding tighter lower bounds than those reported in [23]. They report computational experiments that confirm their claims. Tcha and Maruyama [56] discussed a related problem of minimizing the maximum link utilization factor. They described a linear programming bound and outlined a technique to obtain feasible solutions and applied them to small problem instances with up to 34 links and 95 communicating node pairs.

Pirkul and Narasimhan [47] extended their model to include reliability considerations. For each communicating node pair a primary and secondary route must be selected from among a set of predetermined candidate routes. The model captures situations where a single link or node failure would divert traffic to the appropriate secondary routes. A mathematical programming formulation was presented and an effective solution procedure based on Lagrangean relaxation of the problem was developed.

The studies reported in [23,42,47,56] share one shortcoming. In all of them, a set of prespecified candidate routes is assumed to be given for every communicating node pair. Obviously, the quality of the solutions obtained by these methods depends heavily on the choice of the candidate route set generated before the procedure is applied. Gavish and Altinkemer extended the algorithm in [23] to overcome this shortcoming by considering all possible routes for every communicating node pair [26]. This routing scheme is used as an

integral part of a procedure to solve the routing and capacity assignment problem. One drawback of their algorithm is that when link utilization exceeds moderate levels the procedure frequently terminates without a feasible routing scheme. In this dissertation we present new formulations of the routing problem and discuss improved heuristics for generating feasible solutions for even heavily loaded networks. With these new methods the gaps between the lower bounds and the feasible solutions are generally very small.

2. The Routing and Capacity Assignment Problem in Backbone Communication Networks

Many heuristic solution procedures have been suggested to solve the topological backbone design problem. Frank, et. al. [19] were the first to describe a branch (link) exchange technique (BXC) based on branch deletion and addition. Starting with an arbitrary topology, the BXC method iteratively adds, deletes or exchanges links. If there is improvement in terms of cost and throughput, the modified topology is accepted. This process continues until no more improvement is possible. The criterion for selecting the links to add and/or delete is somewhat arbitrary. The use of information about network properties in deciding which links to add and/or is limited. A large number of transformations is required before reaching a local minimal solution, which makes this approach computationally inefficient.

The Concave Branch Elimination (CBE) method is another heuristic technique for the optimization of topological network design [6,30]. The CBE method assumes that the discrete costs can be reasonably approximated by concave curves. The method starts from a fully connected topology and eliminates the most uneconomical link. The flow deviation

method [21] is then used to reroute the traffic. This process of deletion and rerouting continues until a local minimum is reached or the removal of a link would disconnect the network.

An extension and improvement of the BXC and the CBE in terms of the quality of the solution and computational efforts is the Cut Saturation (CS) algorithm proposed by Gerla et. al. [31]. The CS algorithm uses a basic property of the network flow problem called "cut-saturation". A cut is a set of links whose removal will disconnect the network. A cut is saturated if the flows in all the links in the cut equal the capacities. The traffic flow in the network can increase until a cut is saturated, and in this situation, the only way to increase traffic flow in the network is to increase the capacity of the cut. The CS algorithm is based on this principle. Thus, the idea behind the CS algorithm is to relieve the most heavily congested portion of the network (which corresponds to the saturated cut set). Adding new links in the neighborhood of that cut set or increasing the capacities of its links should be more effective in improving the network throughput than adding links in other portions of the network. While maintaining the throughput within a specified range, the algorithm tries to reduce the communication costs subject to the demand, connectivity and response time constraints.

The algorithm has been shown to perform better than the BXC in terms of closeness to the optimal solution and computational efficiency. However, it has some serious drawbacks as noted by Sapir [53]. These include

1. It lacks generality because it was exclusively developed for the ARPANET.
2. CS requires a good starting topology to perform efficiently. A lot of human effort is needed to design a reasonable initial topology since no criteria for a good one were given.

3. Parallel links are not allowed, even though parallel links can be cost effective and increase network connectivity and therefore reliability.

A new Generalized Cut Saturation (GCS) algorithm has been proposed by Sapir [53] and Chou and Sapir [9] to overcome the above limitations.

A related problem to the routing and capacity assignment problem to be studied in chapter VI is the general concave cost network flow problem where the total cost is a concave function of the flow along the arcs. Some of the methods used to solve this problem include branch and bound [54], dynamic programming [13,61] and other heuristic approaches [45,50]. The authors in [2] studied the special case of piecewise linear costs in directed networks where each link is assigned a capacity and a path is identified for each source/destination pair. They formulated the problem as a mixed integer program and developed a composite algorithm to generate both lower bounds and feasible solutions. The model does not take into account the delay issue that arises when link capacity utilization reaches certain levels.

Gershet and Weihmayer [32] studied the problem of assigning capacities to network switches and potential links in order to accommodate traffic demand between nodes and satisfy a performance requirement that specifies an upper bound on the link utilization. The goal is to minimize switch and link capacity costs which are assumed to be continuous. A solution procedure which alternates between solving an uncapacitated design/routing subproblem and a capacity assignment subproblem is presented. The procedure was applied to a real network with 20 nodes and two levels of link capacities.

Monma and Shallcross [40] studied the problem of designing a minimum cost communication network subject to reliability constraints requiring node-disjoint paths between every pair of communicating nodes. Heuristic solution methods have been

presented and computational results have been reported. A drawback of this study is that no procedure to check the quality of the feasible solutions was developed.

Ng and Hoang [43] examined a special case of the routing and capacity assignment problem where 1) each link's capacity must equal a multiple of fixed capacity C corresponding to a specified line type and 2) full-duplex capacity is approximated by modeling separate one-way capacities and then choosing the larger of these. They formulated the problem using continuous link capacity variables, and used the flow deviation method for solution. A procedure based on dynamic programming was used to discretize the link capacities. Numerical examples of networks including six and eighteen nodes were given.

Rosenberg [52] studied the problem of computing link capacities in a multi-hour alternate routing communication network. For each source/destination, only two predetermined candidate routes are assumed to be given. The first one is formed by direct link from source to destination and the second includes only two links. Obviously this assumption is unrealistic in the case of backbone network. Moreover, the model developed in [52] has a nonlinear constraint for each source/destination pair (for each hour of the day, since it is a multihour model). This makes the problem very difficult to solve for large network with hundreds or thousands of source/destination pairs.

Gavish and Newman [25] formulated the routing and capacity assignment problem as a mixed integer programming model. The routing policy is assumed to be static and nonbifurcated. They developed a solution procedure based on Lagrangean relaxation technique. Narasimhan [41] studied a similar problem where the goal is to design a minimal cost network subject to some performance criterion specifying an upper limit on the average delay experienced by messages in the network. He presented a nonlinear

integer programming formulation of the problem and used Lagrangean relaxation to obtain feasible solutions and lower bounds. Computational results across a variety of networks show that the solution procedure yields solutions with gaps ranging between 0.0 and 24.5%.

The models in the previous three studies share one shortcoming. They assume that a set of prespecified candidate routes is given for every communicating node pair. Obviously, the quality of the solutions obtained by using these models depends heavily on the choice of the candidate route set generated before the procedures are applied.

LeBlanc and Simmons [37] formulated the routing and capacity assignment problem using continuous link capacity variables. They also suggested a new convex delay function and showed that, for their assumed message length distribution, this new function predicts delay more accurately than the conventional delay function when flow-capacity ratios are less than 0.80. Computational results for networks with up to 100 nodes are reported.

Gavish and Altinkemer [26] extended the work in [25] by considering all possible routes for every communicating node pair. They formulated the problem and used Lagrangean relaxation embedded in a subgradient optimization procedure to obtain lower bounds as well as feasible solutions to the problem. They included cut constraints which are redundant in the original problem to improve the lower bounds. These cut constraints are assumed to be defined before the solution procedure starts. Obviously the quality of the solutions depends heavily on the number and choice of the cuts.

In this dissertation we present a new formulation of the routing and capacity assignment problem that treats the routing and capacity assignment policies simultaneously. Our model overcomes the shortcomings in the previous methods, namely the assumptions

that a candidate route set for every source/destination pair is given and that a set of predetermined cut constraints is required. A procedure that generates improved solutions when compared to solutions produced by previous methods is developed. We study also the effects of adding the cut constraints in the formulation on the quality of the solutions to the routing and capacity assignment problem.

The remaining of this dissertation is organized as follows. In chapters III and IV, we consider the routing problem in backbone communication networks in which the maximum link delay and average queueing delay are minimized, respectively. Chapter V addresses the reliability issue in the case of link failures. For each communicating node pair, a primary route and a link disjoint secondary route have to be selected in order to minimize the weighted average queueing delay. The routing and capacity assignment problem is considered in chapter VI. Finally, some concluding remarks are presented in chapter VII.

CHAPTER III

THE ROUTING PROBLEM IN BACKBONE COMMUNICATION NETWORKS: MINIMIZING THE MAXIMUM LINK QUEUEING DELAY

1. Introduction

Route selection in backbone communication networks is an important factor in determining the response time experienced by network users and affects the efficiency of the utilization of network resources such as node buffers and link capacities. A good routing scheme can allow the addition of new network users without significant degradation of the service level to current users and without increasing link capacities.

In this chapter we study the problem of determining the optimal set of routes for all communicating node pairs in a backbone communication network with static, nonbifurcated routing. Given the topological configuration of the network, link capacities, and the traffic requirements between communicating node pairs, the problem can be more specifically described as that of determining, for each source/destination pair, one route over which all messages for that pair of nodes will be routed. This route is to be chosen from among all possible routes. The objective is to minimize the maximum link queueing delay encountered by messages at the network nodes. This criterion of minimizing the maximum link delay is appropriate for two reasons. First, one fast way to minimize the average queueing delay in the network is to minimize the maximum link delay. Second, this

proposed criterion leads to a more balanced distribution of the traffic load over the network links. Thus, in the case of a failure of the most utilized link, less traffic would be disrupted than it would be if the network had been designed to minimize the average queueing delay.

The remaining of this chapter is organized as follows. We present two mixed integer nonlinear programming formulations of the problem in section 2. A Lagrangean relaxation of the problem obtained by dualizing a subset of the constraints in the second formulation is presented in section 3. In section 4, we discuss a method for solving the relaxed problem and present a procedure to generate feasible solutions to the original problem using the information obtained from the solution of the relaxed problem. In section 5, we present results of computational experiments on four network topologies to show the effectiveness of the solution procedures. Finally, some concluding remarks are presented in section 6.

2. Problem Formulation

As noted earlier, response time (defined as an average source-to-destination packet delay) is an important factor in the performance of packet-switched networks. In these networks, messages of different sizes between pairs of communicating nodes arrive at random intervals. Packets of these messages travel over the network forming queues at intermediate nodes waiting for an outgoing channel to become available. Thus, it is possible to model packet-switched networks as networks of queues [35].

In order to formulate the problem of minimizing the maximum link delay in the backbone network, we assume that the network topology, the capacities of the links, and the traffic requirements between every pair of communicating nodes are given. We also

make the usual assumptions which are used in modeling the queueing phenomena in backbone networks. Specifically, we assume that nodes have infinite buffers to store messages waiting for transmission links, that the arrival process of messages to the network follows a Poisson distribution, and that message lengths follow an exponential distribution. We further assume that the propagation delay in the links is negligible, that there is no message processing delay at the nodes, and that there is only a single class of service for each communicating node pair.

The backbone network is modeled as a network of independent M/M/1 queues [35,36] in which links are treated as servers with service rates proportional to the link capacities. The customers are messages whose waiting areas are the network nodes. The queueing and transmission delay in link (i,j) is $1/(\mu Q_{ij} - A_{ij})$ where $1/\mu$ is the average message length, Q_{ij} is the capacity of link (i,j) , and A_{ij} is the arrival rate of messages to link (i,j) .

We use the following notation:

N	the set of nodes in the network
E	the set of undirected links in the network
M	the set of communicating node pairs
A^m	the message arrival rate for communicating node pair $m \in M$
$O(m)$	the source node for communicating node pair $m \in M$
$D(m)$	the destination node for communicating node pair $m \in M$

The decision variables are

$$Y_{ij}^m = \begin{cases} 1 & \text{if the route for communicating node pair } m \text{ traverses link} \\ & (i,j) \text{ in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

X_{ij}^m = flow of communicating node pair m on link (i,j)

D = maximum link delay

If T is defined as $T = \frac{1}{\mu} \sum_{m \in M} A^m$, then the problem can be formulated as :

$$\text{Min } D \quad (1)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \text{ and } m \in M \quad (2)$$

$$\frac{1}{\mu} A^m (Y_{ij}^m + Y_{ji}^m) \leq X_{ij}^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (3)$$

$$\frac{1}{\mu} \sum_{m \in M} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{m \in M} X_{ij}^m \quad \forall (i,j) \in E \quad (4)$$

$$X_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (5)$$

$$\sum_{m \in M} X_{ij}^m \leq Q_{ij} \quad \forall (i,j) \in E \quad (6)$$

$$\frac{1}{T} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} \leq D \quad \forall (i,j) \in E \quad (7)$$

$$X_{ij}^m \geq 0 \quad \forall (i,j) \in E \text{ and } m \in M \quad (8)$$

$$D \geq 0 \quad (9)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (10)$$

In this formulation, the objective function minimizes the maximum queueing delay for messages. Constraint set (2) contains the flow conservation equations which define a route (path) for each communicating node pair. Constraints in set (3) link together the X_{ij}^m and Y_{ij}^m variables. They ensure that the flow for communicating node pair m on link (i,j) is at least equal to the traffic requirement for that pair if its assigned route uses link (i,j) . Constraints in set (3) hold as equalities at the optimum. Constraints in set (4) represent an aggregate form of the constraints in set (3). Even though these constraints are redundant in, they are helpful in obtaining better lower bounds in the Lagrangean relaxation suggested in the next section. Constraint set (5) guarantees that the flow for communicating node pair m on arc (i,j) does not exceed its traffic requirement. Constraint set (6) enforces the capacity limitations on the links. Constraints in set (7) define D , the maximum link delay in the network. Constraint sets (8) and (9) restrict the X_{ij}^m and D variables to be nonnegative, respectively and constraint set (10) enforces integrality conditions on the Y_{ij}^m variables.

This problem can be reformulated as the following integer linear programming model.

Problem P:

$$Z_P = \text{Max } V \quad (11)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \text{ and } m \in M \quad (12)$$

$$\frac{1}{\mu} A^m (Y_{ij}^m + Y_{ji}^m) \leq X_{ij}^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (13)$$

$$\frac{1}{\mu} \sum_{m \in M} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{m \in M} X_{ij}^m \quad \forall (i,j) \in E \quad (14)$$

$$X_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (15)$$

$$\sum_{m \in M} X_{ij}^m + Q_{ij} S_{ij} = Q_{ij} \quad \forall (i,j) \in E \quad (16)$$

$$S_{ij} \geq V \quad \forall (i,j) \in E \quad (17)$$

$$X_{ij}^m \geq 0 \quad \forall (i,j) \in E \text{ and } m \in M \quad (18)$$

$$S_{ij} \geq 0 \quad \forall (i,j) \in E \quad (19)$$

$$V \geq 0 \quad (20)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (21)$$

If S_{ij} is interpreted as the residual utilization factor of link (i,j) , then the goal is to maximize the minimum residual utilization factor. This means that minimizing the maximum link delay is equivalent to maximizing the minimum residual utilization factor.

By allowing the best route for each communicating node pair to be chosen from the set of all possible routes, our solution method based on the above formulation eliminates a shortcoming that the methods presented in [23,41,56] suffer which is the theoretical possibility of generating lower bound values that are higher than the optimal solution values to the routing problem when all possible routes are considered.

3. A Lagrangean Relaxation of the Problem

Problem studied in [34] is a special case of problem P and is known to be NP-hard. Consequently, it is highly unlikely that real world instances of P can be solved optimally in reasonable computation time. Therefore, we propose a composite upper and lower bounding heuristic solution procedure based on a Lagrangean relaxation of the problem. Consider the Lagrangean relaxation of problem P obtained by dualizing constraint set (13) and (14) using nonnegative multipliers α_{ij}^m and β_{ij} for all $(i,j) \in E$ and $m \in M$, respectively.

Problem L:

$$Z_L = \text{Max } V + \sum_{(i,j) \in E} \sum_{m \in M} (\alpha_{ij}^m + \beta_{ij}) X_{ij}^m - \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (22)$$

subject to constraint sets (12), (15)-(21).

Problem L can be decomposed into two subproblems:

Problem L1:

$$Z_{L1} = \text{Max } V + \sum_{(i,j) \in E} \sum_{m \in M} (\alpha_{ij}^m + \beta_{ij}) X_{ij}^m \quad (23)$$

subject to constraint sets (15)-(20).

and

Problem L2:

$$Z_{L2} = \text{Max } - \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (24)$$

subject to constraint set (12) and (21).

problem $L2$ can be decomposed into $|M|$ subproblems (one for each sourced/destination pair) as follows:

$$\text{Max} - \frac{1}{\mu} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (25)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \quad (26)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \quad (27)$$

In the next section, we discuss the methods for solving the Lagrangean problem L and the original problem P .

4. Solution Techniques

4.1 Solution of Problem $L1$

For any given vector of multipliers α_{ij}^m and β_{ij} , problem $L1$ is a continuous linear programming problem that can be solved using an LP package. However, if the capacities of the links as well as the traffic requirements are equal, then problem $L1$ can be solved more efficiently using the following greedy type procedure.

Procedure-LAG1

Step 1: Reorder the X_{ij}^m variables by sorting them in nonincreasing order of α_{ij}^m for all links; assume that the variables are reindexed in this order and that the indices are the same for all links. Let $m=0$ and set $S=0$.

Step 2 : Let $m=m+1$ and set for every link (i,j)

$$X_{ij}^m = \begin{cases} X_0 & \text{if } \sum_{(i,j) \in E} (\alpha_{ij}^m + \beta_{ij}) > 1/Q \\ 0 & \text{otherwise} \end{cases}$$

where $X_0 = \min\{ \frac{1}{\mu} A^m, (Q - S) \}$, Q is the value of capacities of the links.

Set $S = S + X_0$

Step 3: If $m=|M|$ stop; If $X_0 < \frac{1}{\mu} A^m$ then stop and set $X_{ij}^k = 0$ for

$k=m+1, \dots, |M|$, and every link (i,j) . Otherwise go to step 2.

The value assigned to X_{ij}^m in step 2 is determined in a such a way that it increases the objective function of the problem the most.

4.2 Solution of Problem $L2$

Each subproblem of problem $L2$ is equivalent to a shortest path problem from $O(m)$ to $D(m)$ with $(\alpha_{ij}^m + \beta_{ij})$ as the nonnegative costs on the links. In our study Dijkstra's algorithm [55] is used for solving these subproblems.

4.3. Heuristic Solution Methods

In this section we introduce a heuristic solution procedure for solving problem P . This is a two phase procedure which first generates an initial routing schedule (possibly infeasible) and then improves this schedule by reallocating traffic from overloaded links to links that are lightly utilized. This procedure is used to generate a feasible solution which is used as the starting solution for the the subgradient optimization procedure. Additionally at every iteration of the subgradient optimization procedure feasibility of the Lagrangean solution is checked. In fact, the improvement phase of the heuristic procedure can be used

at every iteration of the subgradient procedure to either convert an infeasible solution to a feasible one or to improve a feasible solution. We have chosen not to do this as it was found to be very time consuming. Instead, the improvement phase is applied at the termination of the subgradient search to the best feasible solution generated during that search. In our computational experiments this step typically leads to significant improvement in solution quality as it will be seen in the computational results reported in the next section.

The heuristic used to generate an initial feasible solution can be described as follows.

Procedure-Init

- Step 1: For each communicating node pair determine a route with the minimum number of links.
- Step 2: If for every link the total flow on the link does not exceed the link capacity, then a feasible solution is at hand (stop); otherwise go to step 3.
- Step 3: Pick the most violated link (i,j) . Among all communicating node pairs using that link, reroute the traffic requirement of the communicating node pair k which, using the "alternative path" would decrease the following cost function the most:

$$Z_{\text{init}} = \text{Max}_{(i,j) \in E} \{cost(i,j)\}, \text{ where}$$

$$cost(i,j) = \begin{cases} \frac{1}{T} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} & \text{if } \sum_{m \in M} X_{ij}^m < Q_{ij} \\ \mathfrak{M} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij}} & \text{otherwise} \end{cases}$$

(\mathfrak{M} is a large positive number.)

This artificial cost function decreases when the capacity utilization of link (i,j) (and possibly of other links on the original route for commodity k) decreases.

The "alternative path" is the path from $O(k)$ to $D(k)$ such that the most utilized link on the path has a lower utilization than the most utilized link on any other path from $O(k)$ to $D(k)$. The problem of finding such path is known as the bottleneck shortest path problem and can be solved using a modified version of Dijkstra's algorithm [55] and therefore can be solved in $O(|N|^2)$ operations. If the traffic requirement of no communicating node pair can be rerouted, then stop since no initial feasible solution can be obtained; otherwise, repeat step 3 until no link capacity is violated.

Although Procedure-Init was created to simply find a feasible solution, it can be modified to attempt to improve a given feasible solution. Consequently at the termination of the subgradient optimization procedure the best feasible solution may be further improved by utilizing Procedure-Init which would include only step3 with some modifications. The only modifications to be made to step 3 are : (1) to consider "the most

utilized link" rather than "the most violated link" and (2) to change the stopping criterion to "repeat step 3 until no further improvement can be effected".

5. Computational Results

The solution procedures presented in the previous section were coded in Pascal. A number of computational experiments were performed using IBM-3081D running under MVS/XA 2.1.7. Two networks (ARPA and OCT) used in previous studies [23,25,26,42] were utilized in the tests. In ARPA there are 21 nodes and 26 links and in OCT there are 21 nodes and 29 links (figures 4.1 and 4.2 in chapter IV). In addition, a set of randomly generated networks with 10 and 15 nodes were used.

The following test problem generator was used to generate the random networks. First, the generator locates the specified number of nodes on a 100X100 grid. Each node has a degree equal to 2, 3 or 4 with probability of 0.6, 0.3 and 0.1, respectively. We repeat the following procedure for each node $i \in N$. Determine node i 's closest neighbor (in terms of Euclidean distance) with unsatisfied degree requirement, call this node j . Add arc (i,j) and repeat this until 1) node i 's degree requirement is satisfied or 2) all the nodes with unsatisfied degree requirements have been considered. In the latter case, connect node i to its closest neighbors to which it is not already connected until the degree requirement of node i is satisfied. At the end check to see if the network is connected; if not, add links necessary to make it connected. For each network size tested (10 and 15 nodes) we generated 5 different instances by changing the seed value used in the random number generator. The average number of links in the networks with 10 nodes is 15 and it is 22 for networks with 15 nodes.

In all of the networks used in this study each node communicates with every other node and sends one message. In both ARPA and OCT networks the capacity of the links is set equal to 50000 bps. For networks with 10 nodes, the capacity of each link is set equal to 25000 bps. For networks with 15 nodes, the capacity is 30000 bps. Tables 3.1 and 3.2 summarize the computational results for ARPA and OCT, respectively. Tables 3.3 and 3.4 give the average (over the five problem instances) performance measures for the networks with 10 and 15 nodes, respectively. The results of the experiments are described by providing the values of average message length (in bits), the percentage gap between the value of the best feasible solution found and the lower bound, and the average and maximum link utilizations. The percentage gap is computed as $(\text{Feasible Solution Value} - \text{lower Bound}) / \text{Feasible Solution Value}$.

The mean message length capture a variety of traffic loads for all networks tested. They range from light loads to cases where the load is beyond the normal operating levels. In general, the gap between the feasible solution values and lower bounds is small and ranges from 0.95 to 7.83%. This gap usually increases as the traffic load increases. The computing times were of the order of 0.9 and 1.5 seconds per iteration of the subgradient optimization technique.

6. Conclusion

In this chapter we studied the nonbifurcated static routing problem in backbone networks. In this problem, a route for every pair of communicating nodes is to be identified in order to minimize the maximum link delay faced by messages. A mathematical programming formulation of the problem is presented. An efficient solution procedure

based on Lagrangean relaxation of the problem is developed. Computational results across a variety of networks are reported. These results indicate the procedure to be very effective.

Table 3.1. Summary of Computational Results for ARPA Network.

Average Message Length	Gap	Average link Utilization	Maximum Link Utilization
50	0.95	5.7	8.4
100	2.09	11.4	16.3
150	1.89	17.7	24.0
200	2.61	23.0	32.0
250	3.83	30.0	40.0
300	4.21	35.1	47.2
350	7.83	41.8	56.0

Table 3.2. Summary of Computational Results for OCT Network.

Average Message Length	Gap	Average link Utilization	Maximum Link Utilization
100	0.97	20.1	23.2
150	3.66	29.9	36.0
200	5.87	39.7	48.0
250	4.19	50.0	58.0
300	3.45	60.5	68.4

Table 3.3. Summary of Computational Results for Networks With 10 Nodes.

Average Message Length	Gap	Average link Utilization	Maximum Link Utilization
300	1.72	15.6	27.6
350	2.16	18.7	31.5
400	2.68	20.7	36.8
450	3.29	23.3	38.3
500	4.04	26.1	46.1
550	4.89	30.4	49.3
600	6.01	35.6	55.5

Table 3.4. Summary of Computational Results for Networks With 15 Nodes.

Average Message Length	Gap	Average link Utilization	Maximum Link Utilization
300	1.40	24.6	45.3
350	1.92	28.7	52.9
400	2.53	32.3	60.47
450	3.77	36.4	68.0
500	4.30	40.5	75.5

CHAPTER IV

THE ROUTING PROBLEM IN BACKBONE COMMUNICATION NETWORKS: MINIMIZING THE AVERAGE QUEUEING DELAY

1. Introduction

This chapter extends the problem studied in the previous chapter to the case where the objective is to minimize the average queueing delay in backbone communication networks. Given the network topology, link capacities, and traffic requirements between communicating node pairs, the problem studied in this chapter can be described as that of selecting, for every source/destination pair, one route over which all messages for that pair of nodes will be routed. The objective is to minimize the average queueing and transmission delay encountered by messages at the network nodes. This delay is the result of the finite transmission capacities of links and the resultant queueing at intermediate nodes.

The previous studies [23,42] which have treated this problem share one major shortcoming. Specifically, they assume that a set of prespecified candidate routes is given for every communicating node pair. Obviously, the quality of the solutions obtained by these methods depends heavily on the choice of the candidate route set generated before the procedure is applied. Gavish and Altinkemer [26] avoid this shortcoming by considering all possible routes for every communicating node pair. However, when link utilization

exceeds moderate levels, the procedure developed in [26] frequently terminates without a feasible routing scheme. In this chapter we present a new formulation of the problem and a new heuristic procedure for generating feasible solutions for even heavily loaded networks. With this new procedure the gaps between the lower bounds and the feasible solutions measuring the solution quality are generally very small.

The remaining of this chapter is organized as follows. We present a mixed integer nonlinear programming formulation of the problem in section 2. A Lagrangean relaxation of the problem obtained by dualizing a subset of the constraints is presented in section 3. In section 4, we discuss a method for solving the relaxed problem and present a heuristic procedure to generate feasible solutions to the original problem using the information obtained from the solution of the relaxed problem. In section 5, we present results of computational experiments on four network topologies to show the effectiveness of our solution procedures. Finally, some concluding remarks are presented in section 6.

2. Problem Formulation

In order to formulate the problem of minimizing the average end-to-end delay in the backbone network, we assume that the network topology, the capacities of the links, and the traffic requirements between every pair of communicating nodes are given. We also make the usual assumptions which are used in modeling the queueing phenomena in backbone networks. Specifically, we assume that nodes have infinite buffers to store messages waiting for available transmission links, that the arrival process of messages to the network follows a Poisson distribution, and that message lengths follow an exponential distribution. We further assume that the propagation delay in the links is negligible, that

there is no message processing delay at the nodes, and that there is only a single class of service for each communicating node pair.

We use the following notation:

- N the set of nodes in the network
- E the set of undirected links in the network
- M the set of communicating node pairs
- A^m the message arrival rate for communicating node pair $m \in M$
- $O(m)$ the source node for communicating node pair $m \in M$
- $D(m)$ the destination node for communicating node pair $m \in M$

The decision variables are

$$Y_{ij}^m = \begin{cases} 1 & \text{if the route for communicating node pair } m \text{ traverses link} \\ & (i,j) \text{ in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

X_{ij}^m = flow of communicating node pair m on link (i,j)

If T is defined as $T = \frac{1}{\mu} \sum_{m \in M} A^m$, then the problem can be formulated as :

Problem P:

$$Z_P = \text{Min} \quad \frac{1}{T} \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} \quad (1)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \text{ and } m \in M \quad (2)$$

$$\frac{1}{\mu} A^m (Y_{ij}^m + Y_{ji}^m) \leq X_{ij}^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (3)$$

$$\frac{1}{\mu} \sum_{m \in M} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{m \in M} X_{ij}^m \quad \forall (i,j) \in E \quad (4)$$

$$X_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (5)$$

$$\sum_{m \in M} X_{ij}^m \leq Q_{ij} \quad \forall (i,j) \in E \quad (6)$$

$$X_{ij}^m \geq 0 \quad \forall (i,j) \in E \text{ and } m \in M \quad (7)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (8)$$

In this formulation, the objective function minimizes the average queueing delay for all messages. Constraint set (2) contains the flow conservation equations which define a route (path) for each communicating node pair. Constraints in set (3) link together the X_{ij}^m and Y_{ij}^m variables. They ensure that the flow for communicating node pair m on link (i,j) is at least equal to the traffic requirement for that pair if its assigned route uses link (i,j) ; constraints in set (3) hold as equalities at the optimum. Constraints in set (4) can be seen as the aggregate form of the constraints in set (3). Even though these constraints are redundant in problem P , they are helpful in obtaining better lower bounds in the Lagrangean relaxation suggested in the next section. Constraint set (5) guarantees that the flow for communicating node pair m on arc (i,j) does not exceed its traffic requirement. Constraint set (6) represents the capacity constraints on the links. Constraint set (7)

restricts the X_{ij}^m variables to be nonnegative and constraint set (8) enforces integrality conditions on Y_{ij}^m .

By allowing the best route for each communicating node pair to be chosen from the set of all possible routes, our solution method based on the above formulation eliminates a shortcoming that the methods presented in [23,42] suffer which is the theoretical possibility of generating lower bound values that are higher than the optimal solution to the original routing problem when all possible routes are considered.

This formulation (1)-(8) may be viewed as a disaggregate formulation of the routing problem solved in [26]. Specifically, if one were to drop constraint sets (3) and (5) and substitute the terms $\sum_{m \in M} X_{ij}^m$ by variables which represent the total flows on link (i,j) , one would have the formulation in [26]. The variable set X_{ij}^m and constraint sets (3) and (5) are introduced to represent flows between each pair of communicating nodes separately. This disaggregate formulation leads to better lower bounds and feasible solutions than does the formulation in [26] but at the expense of added computational effort as shown in section 5. Other researchers have also observed a similar effect regarding the disaggregation of flows within the context of other classes of problems [3,59,60].

3. A Lagrangean Relaxation of the Problem

Problem studied in [25,42] is a special case of problem P and is known to be NP-hard. Consequently, it is highly unlikely that real world instances of P can be solved optimally in reasonable computation time. Therefore, we propose a procedure based on a Lagrangean relaxation of the problem to obtain lower bounds and feasible solutions to the

routing problem. Consider the Lagrangean relaxation of problem P obtained by dualizing constraint set (3) and (4) using nonnegative multipliers α_{ij}^m and β_{ij} for all $(i,j) \in E$ and $m \in M$, respectively.

Problem L:

$$Z_L = \text{Min} \quad \frac{1}{T} \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} - \sum_{(i,j) \in E} \sum_{m \in M} (\alpha_{ij}^m + \beta_{ij}) X_{ij}^m + \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (9)$$

subject to constraint sets (2), (5)-(8).

Problem L can be decomposed into two subproblems:

Problem L1:

$$Z_{L1} = \text{Min} \quad \frac{1}{T} \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} - \sum_{(i,j) \in E} \sum_{m \in M} (\alpha_{ij}^m + \beta_{ij}) X_{ij}^m \quad (10)$$

subject to constraint sets (5)-(7).

and

Problem L2:

$$Z_{L2} = \text{Min} \quad \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (11)$$

subject to constraint set (2) and (8).

Problem $L1$ can be decomposed into $|E|$ subproblems (one for each link) as follows:

$$\text{Min } \frac{1}{T} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} - \sum_{m \in M} (\alpha_{ij}^m + \beta_{ij}) X_{ij}^m \quad (12)$$

subject to

$$X_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall m \in M \quad (13)$$

$$\sum_{m \in M} X_{ij}^m \leq Q_{ij} \quad (14)$$

$$X_{ij}^m \geq 0 \quad \forall m \in M \quad (15)$$

Similarly, problem $L2$ can be decomposed into $|M|$ subproblems (one for each source/destination pair) as follows:

$$\text{Min } \frac{1}{\mu} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (16)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \quad (17)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \quad (18)$$

In the next section, we discuss the methods for solving the Lagrangean problem L and the original problem P .

4. Solution Techniques

4.1 Solution of Problem $L1$

For any given vector of multipliers α_{ij}^m and β_{ij} , each of the $|E|$ subproblems of problem $L1$ is a continuous knapsack problem with nonlinear objective function that can be solved optimally using the following greedy type procedure.

Procedure-LAG1

Step 1: Reorder the X_{ij}^m variables by sorting them in nonincreasing order α_{ij}^m ; assume that the variables are reindexed in this order. Let $m=0$.

Step 2: Let $m=m+1$ and set

$$X_{ij}^m = \begin{cases} X_0 & \text{if } \alpha_{ij}^m + \beta_{ij} > 0 \text{ and } X_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } X_0 = \min \left\{ \frac{1}{\mu} A^m, (Q_{ij} - S) - \left[\frac{Q_{ij}/T}{\alpha_{ij}^m + \beta_{ij}} \right]^{1/2} \right\} \text{ and}$$

$$S = \sum_{k < m} X_{ij}^k$$

Step 3: If $m=|M|$ stop; If $X_{ij}^m < \frac{1}{\mu} A^m$ then stop and set $X_{ij}^k = 0$ for

$k=m+1, \dots, |M|$. Otherwise go to step 2.

The value assigned to X_{ij}^m in step 2 is determined in a such a way that it decreases the objective function of the subproblem the most.

The subproblems of problem $L1$ are different from those corresponding to the routing problem treated in [26]. In the former, the total flow passing through each link depends on the individual communicating node pairs because the objective function of each

subproblem contains terms related to the communicating node pairs and constraint set (5) is included in the model. In the latter, the total flow is determined independently of the communicating node pairs because no disaggregate constraints similar to those in sets (3) and (5) are present. It is this disaggregation of flows included in our formulation that leads to better feasible solutions and tighter lower bounds but requires more computational time.

4.2 Solution of Problem $L2$

Each subproblem of problem $L2$ can be solved as a shortest path problem from $O(m)$ to $D(m)$ with $(\alpha_{ij}^m + \beta_{ij})$ as the nonnegative costs on the links. In our study Dijkstra's algorithm [55] is used for solving them. These subproblems are similar to those in [13] with one major difference. In [26] the "lengths" of the links are the same for all communicating node pairs, hence all the subproblems can be solved with one run of Floyd's algorithm [55] which finds shortest paths between all pairs of nodes simultaneously. In our subproblems, the "length" of a link depends on the commodity flowing through it. This requires that shortest paths be determined separately for each communicating node pair. This is a direct result of the disaggregated treatment of the flows. Even though it is more time consuming it results in better solutions and helps solve problems that could not be solved using the aggregated flow formulation of the problem.

4.3 Complexity of Solving Problem L

The complexity of procedure-LAG1 is $O\{|M|\log|M|\}$, and since this procedure must be applied $|E|$ times (once for each link), the complexity of solving problem $L1$ is $O\{|E||M|\log|M|\}$. Each of the $|M|$ subproblems of problem $L2$ can be solved in $O\{|N|^2\}$

using Dijkstra's algorithm. Thus, the complexity of solving problem $L2$ is $O\{|M||N|^2\}$. Therefore, the complexity of solving problem L is $O\{|M|\max(|E|\log|M|, |N|^2)\}$.

4.4. Heuristic Solution Method

The heuristic solution procedure for solving the routing problem which minimizes the average queueing delay in the network is similar to the one used to solve the routing problem which minimizes the maximum link delay. The only modification is that the cost function used in step 3 should be

$$Z_{\text{init}} = \sum_{(i,j) \in E} \text{cost}(i,j) \text{ instead of } Z_{\text{init}} = \max_{(i,j) \in E} \{\text{cost}(i,j)\}$$

5. Computational Results

The solution procedures presented in the previous section were coded in Pascal. A number of computational experiments were performed using IBM-3081D running under MVS/XA 2.1.7. A variety of previously used problems were utilized in the tests. We tested the procedures on the four networks shown in figures 4.1-4.4, i.e. ARPA, OCT, USA and RING. These networks and traffic parameters are similar to those tested by Gavish and Hantler [23], Narasimhan et al. [42], and Gavish and Altinkemer [26]. In all four networks each node communicates with every other node. In ARPA network (Fig. 1) there are 420 communicating node pairs with 4 messages per second being sent between each pair. The corresponding values are 650 and 1 for OCT, and 650 and 4 for USA and 992 and 1 for RING.

Table 4.1 summarizes the results of the computational tests. The results of our solution procedure and of those of the procedure described in [26] are reported. The Gavish and Altinkemer results were obtained using a code provided by the those authors. The results of the experiments are described by providing the values of average message length, best feasible solution (upper bound), best Lagrangean bound (lower bound), the "gap" between the upper and lower bounds, and the maximum and average percentage link utilizations. The mean message length is measured in bits and the lower and upper bounds in milliseconds.

In all the cases our solution procedure produced better feasible solutions and smaller "gaps" between the upper and lower bounds than those produced by the procedure reported in [26]. The improvement in the feasible solutions obtained by our procedure is between 0.1% and 13.5%. But more importantly, the procedure reported in [26] does not generate feasible solutions beyond moderate levels of traffic. For example the procedure reported in [26] did not identify feasible solutions beyond average link utilization of 51% for ARPA, 71.4% for OCT, 34.5% for USA and 24.2% for RING networks. However, our procedure finds good feasible solutions even for heavily loaded networks. This improved effectiveness is obtained at the expense of increased computational time. For the test problems one iteration of the procedure reported in [26] takes between 0.05 and 0.2 seconds whereas our procedure takes between 1 and 5 seconds.

The routing problem is a subproblem in the overall solution procedure for the routing and capacity assignment problem. Consequently, the results obtained here clearly have significant implications for not only solving the routing problem but also solving the routing and capacity assignment problem. The ability to find feasible routing schedules for

networks with higher link utilization levels implies that better capacity assignment can be made.

It is worthwhile to note that when we minimize the maximum link delay the maximum link utilization is reduced by approximately 3% and the average link utilization is increased by approximately 2% compared to the utilizations obtained when the average queueing delay in the network is minimized.

6. Conclusion

In this chapter we studied the nonbifurcated static routing problem in backbone networks. In this problem, a route for every pair of communicating nodes is to be identified in order to minimize the mean delay faced by messages. The route is to be chosen among all possible routes. A mathematical programming formulation of the problem is presented. An efficient solution procedure based on Lagrangean relaxation of the problem is developed. Computational results across a variety of networks are reported. These results indicate the procedure to be very effective in generating feasible solutions for even heavily loaded networks.

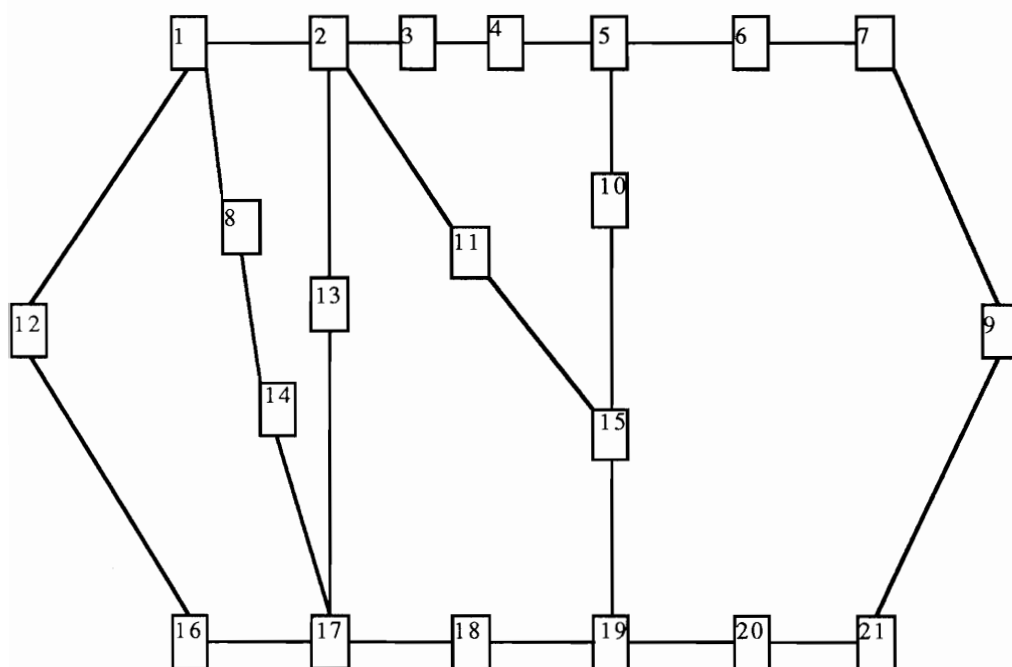


Figure 4.1: The ARPA network.

— 50 Kbits/sec

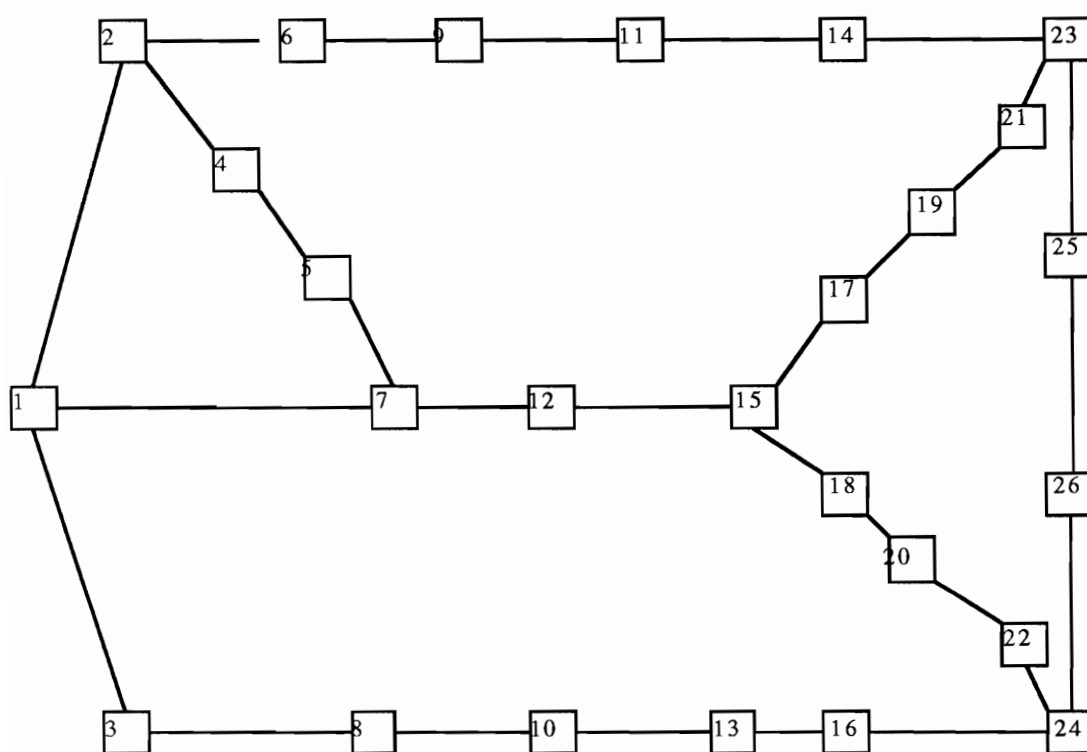


Figure 4.2: The OCT network.

———— 50 Kbits/sec

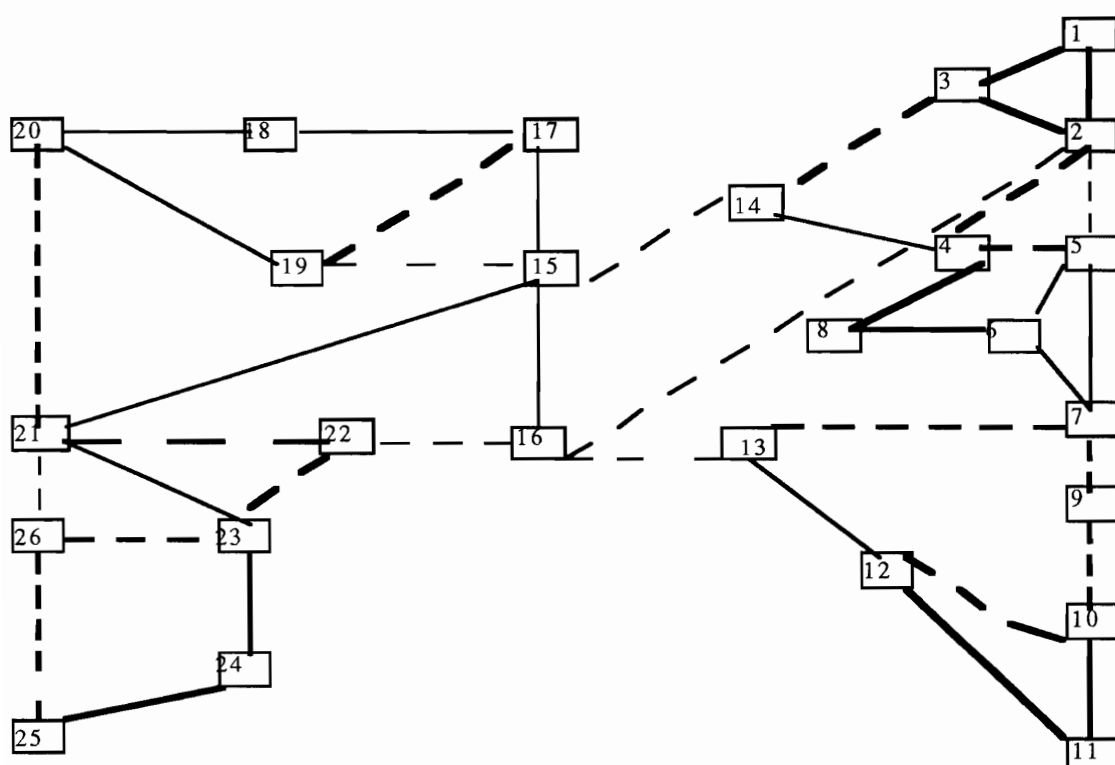


Figure 4.3: USA Network

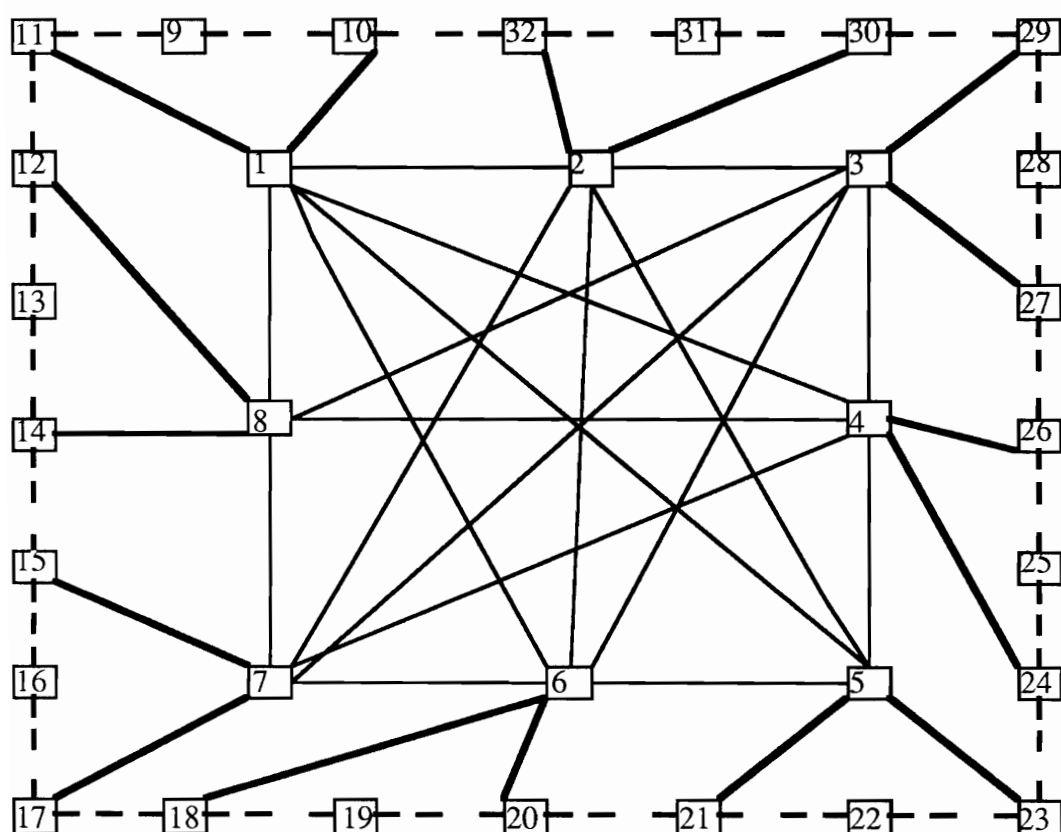


Figure 4.4: The RING network.

- 35 Kbits/sec
- 33 Kbits/sec
- 30 Kbits/sec

Table 4.1. Summary of Computational Results

Net-work	Mess-age Length	Upper bound	Lower bound	Gap*	Gavish and Altinkemer's Method						
					Max. % link utili.	Ave. % link utili.	Upper bound	Lower bound	Gap*	Max. % link utili.	Ave. % link utili.
ARPA	75	8.53	8.53	0.02	52.80	33.40	8.54	8.53	0.18	54.00	33.41
ARPA	100	14.87	14.86	0.11	65.60	45.20	15.06	14.86	1.37	68.80	44.86
ARPA	112	19.65	19.61	0.20	71.70	50.80	20.02	19.61	2.10	75.30	50.57
ARPA	113	20.11	20.09	0.10	74.10	51.20	20.53	20.01	2.61	75.90	51.03
ARPA	125	27.45	27.41	0.13	80.00	56.80	no feasible solution				
ARPA	130	31.80	31.72	0.25	83.20	59.10	no feasible solution				
ARPA	140	45.43	45.01	0.94	89.60	63.80	no feasible solution				
OCT	225	36.25	36.24	0.02	57.60	44.00	36.28	36.23	0.16	56.70	44.40
OCT	250	44.64	44.62	0.04	62.00	49.00	44.66	44.61	0.10	63.00	49.00
OCT	275	55.11	55.09	0.03	68.20	54.10	55.13	55.08	0.10	68.20	54.12
OCT	300	68.61	68.57	0.06	73.20	59.00	68.68	68.57	0.16	72.00	59.03
OCT	325	86.98	86.82	0.18	78.00	63.90	87.03	86.81	0.25	79.30	63.97
OCT	350	113.43	113.23	0.18	84.00	68.90	113.71	113.21	0.44	84.00	68.94
OCT	360	127.78	127.58	0.16	85.00	70.90	128.26	127.51	0.59	86.40	70.92
OCT	362	131.23	130.76	0.36	85.40	71.30	131.56	130.76	0.61	86.90	71.40
OCT	375	156.25	155.58	0.44	88.50	74.00	no feasible solution				
OCT	400	242.12	240.75	0.57	92.80	79.30	no feasible solution				
USA	25	5.25	5.24	0.11	48.00	23.30	5.33	5.24	1.71	48.30	23.70
USA	30	6.94	6.93	0.17	55.70	28.20	7.15	6.86	4.12	57.20	29.30
USA	31	7.33	7.31	0.16	56.50	29.20	7.63	7.08	7.81	58.10	30.30
USA	32	7.75	7.71	0.47	59.40	30.30	8.10	6.90	17.49	60.40	31.20
USA	33	8.16	8.13	0.38	60.20	31.30	9.42	2.17	334	76.95	34.50
USA	35	9.08	9.03	0.60	63.80	33.30	no feasible solution				
USA	40	11.90	11.71	1.56	71.70	38.50	no feasible solution				
USA	45	15.74	15.47	1.71	80.60	43.50	no feasible solution				
USA	50	22.09	21.56	2.46	89.60	48.80	no feasible solution				
RING	150	17.73	17.61	0.66	33.60	22.30	17.86	17.61	1.39	33.91	22.30
RING	160	19.34	19.19	0.79	36.80	23.80	19.50	19.18	1.64	37.10	23.90
RING	161	19.52	19.51	0.05	38.30	24.10	19.68	19.34	1.76	38.40	24.15
RING	162	19.67	19.51	0.84	38.30	24.10	19.84	19.51	1.71	38.42	24.20
RING	200	26.45	26.21	0.92	43.60	29.80	no feasible solution				
RING	250	37.51	37.11	1.06	54.50	37.30	no feasible solution				
RING	300	52.38	51.55	1.60	63.60	45.00	no feasible solution				
RING	325	61.90	60.71	1.96	67.00	48.90	no feasible solution				
RING	350	73.24	71.64	2.23	72.10	52.90	no feasible solution				
RING	375	87.28	84.92	2.78	68.60	57.40	no feasible solution				

CHAPTER V

THE PRIMARY AND SECONDARY ROUTE SELECTION PROBLEM IN BACKBONE COMMUNICATION NETWORKS

1. Introduction

Given the topology of a network (location of switches and links), link capacities and external traffic (messages) characteristics, we study the problem of selecting primary and secondary routes for each communicating node pair in order to minimize the average delay encountered by messages at network switches. We consider static, nonbifurcated routing since this routing scheme is used in most operational networks [23].

The model developed in this chapter addresses the routing problem in backbone networks that are survivable under single-link failures. Survivability, in this context, is the ability to reroute messages through alternate routes in the network in case of a single link failure. In our model, survivability is achieved by selecting a primary route and a link-disjoint alternate route. Because the two routes are link disjoint, one will always be available in any single link failure scenario. Although sufficient capacity to handle all messages between communicating node pairs is reserved on the links for both the primary and secondary routes, messages are normally routed via the primary routes. If the failed link is on the primary route of a communicating node pair, messages between that pair are

rerouted over the secondary route. It is expected that all communicating node pairs are switched back to their primary routes as soon as the failed link is back in service.

Pirkul and Narasimhan [47] addressed the problem of selecting primary and secondary routes for every pair of communicating nodes. They developed a mathematical programming model which captures situations where a single link or node failure would divert interrupted traffic to the designated secondary routes. They assumed that a predefined set of route pairs is given for each origin/destination pair. Each route pair consists of a primary route and a secondary route. In this chapter we present a new mathematical programming model which addresses the problem of a single link failure. We also extend the work in [47] by considering all possible route pairs for each origin/destination pair.

The remaining of this chapter is organized as follows. We present a mixed integer nonlinear programming formulation of the problem in section 2. A Lagrangean relaxation of the problem obtained by dualizing a subset of the constraints is presented in section 3. In section 4, we discuss a method for solving the relaxed problem and present a procedure to generate feasible solutions to the original problem using the information obtained from the solution of the relaxed problem. In section 5, we present results of computational experiments on different network topologies to show the effectiveness of our solution procedures. Finally, some concluding remarks are presented in section 6.

2. Problem Formulation

The backbone network is modeled as a network of independent M/M/1 queues [35,36] in which links are treated as servers with service rates proportional to the link

capacities. The customers are messages whose waiting areas are the network nodes. We assume a link failure follows an i.i.d. Poisson process. When such failure occurs, the traffic on the affected primary routes is diverted to the secondary routes. Even though the failure processes are not Poissonian because of the single link failure assumption, we use the Poisson process approximation to develop the model.

We use the following notation:

N	the set of nodes in the network
E	the set of undirected links in the network
M	the set of communicating node pairs
A^m	the message arrival rate for communicating node pair $m \in M$
$O(m)$	the source node for communicating node pair $m \in M$
$D(m)$	the destination node for communicating node pair $m \in M$
$1/\mu$	the mean of the exponential distribution from which the message lengths are drawn

The decision variables are

$$w_{ij}^m = \begin{cases} 1 & \text{if the primary route for communicating node pair } m \\ & \text{traverses link}(i,j) \text{ in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

$$v_{ij}^m = \begin{cases} 1 & \text{if the secondary route for communicating node pair } m \\ & \text{traverses link}(i,j) \text{ in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

X_{ij}^m = flow of communicating node pair m on link (i,j) and (i,j) is used in the primary route for pair m

Y_{ij}^m = flow of communicating node pair m on link (i,j) and (i,j) is used in the secondary route for pair m

$Z_{(ij,kl)}^m$ = flow of communicating node pair m on link (i,j) and (i,j) is used in the secondary route for pair m and link (k,l) in the primary route for pair m fails

$U_{(ij,kl)}^m$ = flow of communicating node pair m on link (i,j) and (i,j) is used in the primary route for pair m and link (k,l) not in the primary route for pair m fails

In terms of the above notation, the average queueing delay in the network, when there is no

link failure, equals
$$\frac{1}{T} \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m}$$

where $T = \frac{1}{\mu} \sum_{m \in M} A^m$ is the total arrival rate of messages to the network. The expected

network delay when a link (k,l) fails is

$$\frac{1}{T} \sum_{(i,j) \neq (k,l)} \frac{\sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}{Q_{ij} - \sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}$$

In order to obtain the expression for average delay in the network, we multiply each of the above two expressions by their respective weights, i.e. $(1 - \sum_{(k,l) \in E} t_{kl})$ and t_{kl} , where t_{kl}

represents the fraction of time link (k,l) is not operational.

The primary and secondary route selection problem can now be formulated as follows

Problem P:

$$Z_P = \text{Min} \quad \frac{1}{T} (1 - \sum_{(k,l) \in E} t_{kl}) \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} +$$

$$\frac{1}{T} \sum_{(k,l) \in E} t_{kl} \sum_{(i,j) \neq (k,l)} \frac{\sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}{Q_{ij} - \sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)} \quad (1)$$

subject to

$$\sum_{j \in N} W_{ij}^m - \sum_{j \in N} W_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \text{ and } m \in M \quad (2)$$

$$\sum_{j \in N} V_{ij}^m - \sum_{j \in N} V_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \text{ and } m \in M \quad (3)$$

$$W_{ij}^m + W_{ji}^m + V_{ij}^m + V_{ji}^m \leq 1 \quad \forall (i,j) \in E \text{ and } m \in M \quad (4)$$

$$\frac{1}{\mu} A^m (W_{ij}^m + W_{ji}^m) \leq X_{ij}^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (5)$$

$$\frac{1}{\mu} A^m (V_{ij}^m + V_{ji}^m) \leq Y_{ij}^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (6)$$

$$X_{kl}^m + Y_{ij}^m - \frac{1}{\mu} A^m \leq Z_{(ij,kl)}^m \quad \forall (i,j) \in E \text{ } (i,j) \neq (k,l) \text{ and } m \in M \quad (7)$$

$$X_{ij}^m - X_{kl}^m \leq U_{(ij,kl)}^m \quad \forall (i,j) \in E \text{ } (i,j) \neq (k,l) \text{ and } m \in M \quad (8)$$

$$X_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (9)$$

$$Y_{ij}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E \text{ and } m \in M \quad (10)$$

$$Z_{(ij,kl)}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E, (i,j) \neq (k,l) \text{ and } m \in M \quad (11)$$

$$U_{(ij,kl)}^m \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E, ((i,j) \neq (k,l) \text{ and } m \in M \quad (12)$$

$$\sum_{m \in M} X_{ij}^m \leq Q_{ij} \quad \forall (i,j) \in E \quad (13)$$

$$\sum_{m \in M} Y_{ij}^m \leq Q_{ij} \quad \forall (i,j) \in E \quad (14)$$

$$\sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m) \leq Q_{ij} \quad \forall (i,j) \in E, (k,l) \neq (i,j) \text{ and } m \in M \quad (15)$$

$$W_{ij}^m, W_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (16)$$

$$V_{ij}^m, V_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (17)$$

$$X_{ij}^m \geq 0 \quad \forall (i,j) \in E \text{ and } m \in M \quad (18)$$

$$Y_{ij}^m \geq 0 \quad \forall (i,j) \in E \text{ and } m \in M \quad (19)$$

$$Z_{(ij,kl)}^m \geq 0 \quad \forall (i,j) \in E, ((i,j) \neq (k,l) \text{ and } m \in M \quad (20)$$

$$U_{(ij,kl)}^m \geq 0 \quad \forall (i,j) \in E, ((i,j) \neq (k,l) \text{ and } m \in M \quad (21)$$

The first term in the objective function measures the weighted queueing delay in the network when all links are operational and the second term captures the weighted message delay when link failure occurs. Constraint sets (2) and (3) contain the flow conservation equations which define, for each communicating node pair, a primary and secondary route, respectively. Constraints in set (4) specify that a link can be used either in the primary or in

the secondary route but not both for every communicating node pair. This guarantees that the primary and secondary routes are link disjoint. Constraints in set (5) ensure that the flow for communicating node pair m on any link (i,j) is at least equal to the traffic requirement for that pair if its assigned primary route uses link (i,j) . Similarly, constraints in set (6) ensure that the flow for communicating node pair m on any link (i,j) is at least equal to the traffic requirement for that pair if its assigned secondary route uses link (i,j) . However, constraints in sets (5) and (6) hold as equalities at the optimum because of the nature of the objective function. Constraint set (7) reserves space for secondary routes on each link in case any other single link fails. A constraint in set (7), for communicating node pair m , ensures that the traffic flow from pair m on link (i,j) is equal to the traffic requirement for pair m whose primary route has become unavailable due to failure of link (k,l) and consequently must use its secondary route which happens to use link (i,j) . Constraints in set (8) serve two purposes. First, they specify, for each communicating node pair m , the normal flow on link (i,j) due to the primary route using it when the failed link (k,l) is not on the primary route for that pair. Second, they guarantee that when link (k,l) fails there will be no flow on link (i,j) from pair m which uses both (i,j) and (k,l) on its primary route. Constraints in sets (9)-(12) state that the flow on any link for the different communicating node pairs can not exceed their traffic requirements. Constraint sets (13)-(15) ensure that the total flow on each link does not exceed its capacity when no link failure occurs, when the link is used only on the secondary routes, and when there is one single link failure. Constraints in sets (16)-(21) are the nonnegativity and integrality constraints on the decision variables.

By allowing the best route for each communicating node pair to be chosen from the set of all possible routes, our solution method based on the above formulation eliminates a

shortcoming that the method presented in [47] suffer which is the theoretical possibility of generating lower bound values that are higher than the optimal solution to the original routing problem when all possible routes are considered.

3. A Lagrangean Relaxation of the Problem

Problem studied in [47] is a special case of problem P and is known to be NP-hard. Consequently, it is highly unlikely that real world instances of P can be solved optimally in reasonable computation time. Therefore, we propose a composite upper and lower bounding heuristic solution procedure based on a Lagrangean relaxation of the problem. Consider the Lagrangean relaxation of problem P obtained by dualizing constraints in sets (4)-(8) using nonnegative multipliers γ_{ij}^m , P_{ij}^m , S_{ij}^m , $\alpha_{(ij,kl)}^m$ and $\beta_{(ij,kl)}^m$ for all $(i,j) \in E$, $(k,l) \neq (i,j)$ and $m \in M$, respectively.

Problem L:

$$Z_L = \text{Min} \quad \frac{1}{T} (1 - \sum_{(k,l) \in E} t_{kl}) \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} \\ + \frac{1}{T} \sum_{(k,l) \in E} t_{kl} \sum_{(i,j) \neq (k,l)} \frac{\sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}{Q_{ij} - \sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}$$

$$\begin{aligned}
& + \sum_{(i,j) \in E} \sum_{m \in M} \left(\sum_{(k,l) \neq (i,j)} (\alpha_{(ij,kl)}^m + \beta_{(ij,kl)}^m - \beta_{(kl,ij)}^m) - P_{ij}^m \right) X_{ij}^m \\
& + \sum_{(i,j) \in E} \sum_{m \in M} (-S_{ij}^m + \sum_{(k,l) \neq (i,j)} \alpha_{(ij,kl)}^m) Y_{ij}^m \\
& + \sum_{(i,j) \in E} \sum_{(k,l) \neq (i,j)} \sum_{m \in M} (-\alpha_{(ij,kl)}^m Z_{(ij,kl)}^m - \beta_{(ij,kl)}^m U_{(ij,kl)}^m) \\
& + \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (P_{ij}^m + \gamma_{ij}^m) (W_{ij}^m + W_{ji}^m) \\
& + \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (S_{ij}^m + \gamma_{ij}^m) (V_{ij}^m + V_{ji}^m) \\
& + \sum_{(i,j) \in E} \sum_{m \in M} (-\gamma_{ij}^m) + \frac{1}{\mu} \sum_{(i,j) \in E} \sum_{(k,l) \neq (i,j)} \sum_{m \in M} (\alpha_{(ij,kl)}^m A^m) \quad (22)
\end{aligned}$$

subject to

$$(2), (3), (9)-(21)$$

Problem L can be decomposed into five subproblems:

Problem $L1$:

$$\begin{aligned}
Z_{L1} = \text{Min} \quad & \frac{1}{T} (1 - \sum_{(k,l) \in E} t_{kl}) \sum_{(i,j) \in E} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij} - \sum_{m \in M} X_{ij}^m} \\
& + \sum_{(i,j) \in E} \sum_{m \in M} \left(\sum_{(k,l) \neq (i,j)} (\alpha_{(ij,kl)}^m + \beta_{(ij,kl)}^m - \beta_{(kl,ij)}^m) - P_{ij}^m \right) X_{ij}^m \quad (23)
\end{aligned}$$

subject to

$$(9), (13) \text{ and } (18)$$

Problem L2:

$$\begin{aligned}
 Z_{L2} = \text{Min } & \frac{1}{T} \sum_{(k,l) \in E} t_{kl} \sum_{(i,j) \neq (k,l)} \frac{\sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)}{Q_{ij} - \sum_{m \in M} (Z_{(ij,kl)}^m + U_{(ij,kl)}^m)} \\
 & + \sum_{(i,j) \in E} \sum_{(k,l) \neq (i,j)} \sum_{m \in M} (-\alpha_{(ij,kl)}^m Z_{(ij,kl)}^m - \beta_{(ij,kl)}^m U_{(ij,kl)}^m) \quad (24)
 \end{aligned}$$

subject to

$$(11), (12), (15), (20) \text{ and } (21)$$

Problem L3:

$$Z_{L3} = \text{Min } \sum_{(i,j) \in E} \sum_{m \in M} (-S_{ij}^m + \sum_{(k,l) \neq (i,j)} \alpha_{(ij,kl)}^m) Y_{ij}^m \quad (25)$$

subject to

$$(10), (14) \text{ and } (19)$$

Problem L4:

$$Z_{L4} = \text{Min } \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (P_{ij}^m + \gamma_{ij}^m) (W_{ij}^m + W_{ji}^m) \quad (26)$$

subject to

$$(2) \text{ and } (16)$$

and

Problem L5:

$$Z_{L5} = \text{Min} \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (S_{ij}^m + \gamma_{ij}^m) (V_{ij}^m + V_{ji}^m) \quad (27)$$

subject to

(3) and (17)

The initial values of all the the multipliers are set equal to zero except the P_{ij}^m

multipliers whose values are determined as follows. First, the routing problem addressed in chapter IV is solved. If we let g_{ij}^m and h_{ij} denote the values of the multipliers that yield the best lower bound to the routing problem, then we set $P_{ij}^m = g_{ij}^m + h_{ij}$ for $(i,j) \in E$ and $m \in M$.

For ease of the presentation in the remaining of the chapter, let

$$B_{ij}^m = \sum_{(k,l) \neq (i,j)} (\alpha_{(ij,kl)}^m + \beta_{(ij,kl)}^m - \beta_{(kl,ij)}^m) - P_{ij}^m,$$

$$C_{ij}^m = -S_{ij}^m + \sum_{(k,l) \neq (i,j)} \alpha_{(ij,kl)}^m \text{ and } D = \frac{1}{T} (1 - \sum_{(k,l) \in E} t_{kl})$$

In the next section, we discuss the methods for solving the Lagrangean problem L and the original problem P .

4. Solution Techniques

4.1 Solution of Problem $L1$

Problem $L1$ can be decomposed into $|E|$ subproblems (one for each link). For any given vector of multipliers B_{ij}^m each of the $|E|$ subproblems of problem $L1$ is a continuous knapsack problem with nonlinear objective function that can be solved using the following greedy type procedure.

Procedure-LAG1

Step 1: Reorder the X_{ij}^m variables by sorting them in nondecreasing order of B_{ij}^m ; assume that the variables are reindexed in this order. Let $m=0$.

Step 2: Let $m=m+1$ and set

$$X_{ij}^m = \begin{cases} X_0 & \text{if } B_{ij}^m < 0 \text{ and } X_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $X_0 = \min\left\{\frac{1}{\mu} A^m, (Q_{ij} - S) - \left[-\frac{D Q_{ij}}{m}\right]^{1/2}\right\}$ and

$$S = \sum_{k < m} X_{ij}^k$$

Step 3: If $m=|M|$ stop; If $X_{ij}^m < \frac{1}{\mu} A^m$ then stop and set $X_{ij}^k = 0$ for

$k=m+1, \dots, |M|$. Otherwise go to step 2.

The value assigned to X_{ij}^m in step 2 is determined in a such a way that it decreases the objective function of the subproblem the most. The complexity of solving problem $L1$ is $O(|E||M|\text{Log}|M|)$.

4.2 Solution of Problem $L2$

Problem $L2$ can be decomposed into $|E| |E|-1$ subproblems (one for every link and every other link). For any given vector of multipliers $\alpha_{(ij,kl)}^m$ and $\beta_{(ij,kl)}^m$ each of those subproblems of problem $L2$ is a continuous knapsack problem with nonlinear objective function that can be solved similarly to subproblems of problem $L2$. The complexity of solving problem $L2$ is $O(|E|^2 |M|\text{Log}|M|)$.

4.3 Solution of Problem $L3$

Problem $L3$ can be decomposed into $|E|$ subproblems (one for each link). For any given vector of multipliers C_{ij}^m each of the $|E|$ subproblems of problem $L1$ is a continuous knapsack problem that can be solved using the following greedy type procedure.

Procedure-LAG2:

Step 1: Sort the commodities in nonincreasing order of C_{ij}^m ; assume that the commodities are reindexed in this order. Let $m=0$.

Step 2: Let $m=m+1$ and set

$$Y_{ij}^m = \begin{cases} A_I & \text{if } C_{ij}^m < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $S = \sum_{k < m} Y_{ij}^k$, $A_I = \text{Min} \{ \frac{1}{\mu} A^m, Q_{ij} - S \}$

Step 3: Repeat step 2 until $m=|M|$.

The complexity of solving problem $L3$ is $O(|E||M|\text{Log}|M|)$.

4.4 Solution of Problem $L4$ and $L5$

Both problems $L4$ and $L5$ can be further decomposed into $|M|$ subproblems (one for each communicating node pair). Each subproblem can be solved as a shortest path problem from $O(m)$ to $D(m)$. In our study Dijkstra's algorithm [55] is used for solving the subproblems. The complexity of solving problem $L4$ and $L5$ is $O(|M||N|^2)$.

4.5 Heuristic Solution Procedure

In the beginning of the procedure an attempt is made to generate an initial feasible solution using the following procedure.

Procedure-Init:

- Step 1: Solve the routing problem addressed in the previous chapter. Let g_{ij}^m and h_{ij} , for $(i,j) \in E$ and $m \in M$, be the multipliers that yield the feasible solution with the lowest objective function value.
- Step 2: Identify, for each communicating node pair m , a primary route as the shortest path from $O(m)$ to $D(m)$ using $g_{ij}^m + h_{ij}$ as the lengths of the links in the network.
- Step 3: Identify, for each communicating node pair m , a secondary route using only links that are not included in the primary route for pair m . The secondary route is defined as the shortest path from $O(m)$ to $D(m)$ using $g_{ij}^m + h_{ij}$ as the lengths of the links. If, for a particular pair m , a secondary route can not be identified, then stop; no initial feasible solution is generated. Otherwise, go to step 3.
- Step 3: If the capacities of the links can support the routes selected in steps 2 and 3 in cases of a link failure and no link failure, then an initial feasible solution is obtained.

In every iteration of the subgradient optimization algorithm an attempt is made to generate a feasible solution to problem (P) using procedure-Grad described below. The best feasible solution is retained when the subgradient algorithm is terminated.

- Step 1: Use the solution of problem $(L4)$ as the primary routes for the communicating node pairs.

Step 2: Identify a secondary route, for every pair m , as the shortest path using only the links which are not included in the primary route with $(S_{ij}^m + \gamma_{ij}^m)$ as the lengths of the links.

Step 3: If a secondary route is identified for each communicating node pair and the links have enough capacities to support the traffic flows in cases of a link failure and no link failure, then a feasible solution is available and it is retained if its value is the best so far. Otherwise, the next subgradient iteration is executed without generating a feasible solution at the current iteration.

5. Computational Results

The solution procedures presented in the previous section were coded in Pascal. A number of computational experiments were performed using IBM-3090D running under MVS/XA 2.1.7. To evaluate the effectiveness of those procedures, we used the same set of randomly generated networks with 10 and 15 nodes that were used in chapter III. We used also networks with 20 nodes generated in the same way the previous networks were generated. The average number of links in the networks with 20 nodes is 28. The capacity of all links is set equal to 25000, 30000 and 35000 bps for networks with 10, 15 and 20 nodes, respectively. In all networks used in this study each node communicates with every other node and sends one message. The parameter t_{kl} is set equal to 0.1% for all links in the networks.

Table 5.1 gives the average (over the five problem instances) performance measures for the different networks. The results of the experiments are described by

providing the number of nodes in the network, the values of average message length (in bits), the percentage gap between the feasible solution value and the lower bound and the average and maximum link utilizations. These utilizations represent link utilizations when there is no link failure. The maximum link utilizations when there is a single link failure are also reported in the last column of table 5.1. The percentage gap is computed as $(\text{Feasible Solution Value} - \text{Lower Bound}) / \text{Feasible Solution Value}$.

The mean message length captures a variety of traffic loads for all networks tested. They range from light loads to cases where the load is beyond the normal operating levels. In general, the gap between the feasible solution values and lower bounds is small and ranges from 0.17% to 0.73%. The actual CPU computation time taken by the solution procedure is on average 115, 150 and 250 seconds for networks with 10, 15 and 20 nodes, respectively.

6. Conclusion

In this chapter we studied the primary and secondary route selection problem in backbone networks. In this problem, a primary route and a secondary link disjoint route for every pair of communicating nodes are to be identified in order to minimize the weighted average delay faced by messages. The selection of two link disjoint routes for each communicating node pair is a popular way to consider reliability issue in communication networks. The primary route is normally used for sending messages. However, in case a primary route is unavailable due to link failure, its corresponding secondary route is used.

A mathematical programming model of the problem was presented. An efficient solution procedure based on Lagrangean relaxation of the problem was developed. Computational results using two types of randomly generated networks are reported. These results indicate the procedure to be effective. Future work can be directed at developing fast procedures to solve the routing and capacity assignment problem which considers the reliability issue in backbone communication networks.

Table 5.1. Summary of Computational Results.

Number of Nodes	Average Message Length	Percent Gap	Average link Utilization	Maximum Link Utilization	Maximum Link Util. in Case of a Link Failure
10	300	0.17	15.5	27.6	51.0
10	350	0.20	18.1	32.2	59.5
10	400	0.29	20.6	36.8	70.5
10	450	0.19	23.2	41.4	76.5
10	500	0.31	25.8	46.1	80.4
15	50	0.53	4.1	8.3	15.3
15	100	0.54	8.4	15.4	28.1
15	150	0.47	12.4	24.0	42.6
15	200	0.45	17.1	34.2	60.3
15	250	0.33	21.3	44.7	77.8
20	50	0.35	6.8	13.2	23.0
20	100	0.37	11.5	23.0	39.5
20	150	0.45	17.2	33.9	59.3
20	200	0.57	23.0	44.1	77.5
20	250	0.73	27.1	53.3	87.0

CHAPTER VI

THE ROUTING AND CAPACITY ASSIGNMENT PROBLEM IN BACKBONE COMMUNICATION NETWORKS

1. Introduction

In this chapter, we consider the problem of determining simultaneously the link capacities and routes over which messages between communicating node pairs are transmitted in a backbone communication network. The network topology (location of the nodes and links), the traffic requirements between source/destination pairs, a set of link types with different capacities, and costs and a unit cost of delay are given. The goal is to design a communication network with minimum overall system costs composed of connection costs which depend on link capacities and delay costs incurred by users. Delays are caused by the finite capacities of the links which result in queueing at the intermediate nodes of a communication route.

This study overcomes three serious shortcomings of previous methods suggested in past research. The first one is the treatment of route selection and link capacity assignment separately. Given link capacities, the routing problem seeks to select the best routes for communicating node pairs in order to minimize the average queueing delay in the network. In the capacity assignment problem, the routing strategy is assumed to be given and the goal is to assign capacities to links in the network at minimum costs. Here the routing and link capacity assignment decisions are made simultaneously. The second shortcoming is

that these methods assume that a set of prespecified candidate routes chosen from among all possible routes is given for every communicating node pair. Obviously, the quality of the solutions obtained by these methods depends heavily on the choice of the candidate route sets generated before the methods are applied. The use of only a subset of all possible routes by these methods results in a theoretical limitation which is the possibility of generating lower bounds higher than the values of the optimal solutions to the routing and capacity assignment problem. Our solution method eliminates this theoretical limitation by considering all possible routes for every communicating node pair. The third drawback is that one of the previous methods assumes that a prespecified set of cut constraints is given. However, it is very difficult to generate good cut constraints. Our method does not use such cut constraints. The elimination of the three limitations by our method makes it easier to use by communication network designers.

The remaining of this chapter is organized as follows. We present a mixed integer nonlinear programming formulation of the problem in section 2. A Lagrangean relaxation of the problem obtained by dualizing a subset of the constraints is presented in section 3. In section 4, we discuss a method for solving the relaxed problem and present a procedure to generate feasible solutions to the original problem. In section 5, we present results of computational experiments on four network topologies to show the effectiveness of our solution procedure. In section 6 we study the effect of adding the cut constraints on the quality of the solutions. Finally, some concluding remarks are presented in section 7.

2. Problem Formulation

In order to formulate the routing and capacity assignment problem in backbone networks, we assume that the network topology, the capacities of the links, and the traffic requirements between every pair of communicating nodes are given. We also make the usual assumptions which are used in modeling the queueing phenomena in backbone networks. Specifically, we assume that nodes have infinite buffers to store messages waiting for available transmission links, that the arrival process of messages to the network follows a Poisson distribution, and that message lengths follow an exponential distribution. We further assume that the propagation delay in the links is negligible, that there is no message processing delay at the nodes, and that there is only a single class of service for each communicating node pair. Under these assumptions, the backbone network is modeled as a network of independent M/M/1 queues [35,36] in which links are treated as servers with service rates proportional to the link capacities and messages whose waiting areas are the network nodes as customers.

The organization using the network incurs a cost associated with the queueing delay encountered by messages in the network. This cost can be estimated by the cost of the waiting time of its employees waiting in front of the terminals for network services and the cost associated with decisions not made on time. Alternatively, artificial delay costs can be used to achieve a desired level of tradeoff between network costs and average response time. Line costs consist of 1) fixed costs which include a fixed setup cost per unit of time and a cost proportional to the length of the link, 2) variable costs which are assumed to be proportional to the traffic carried over the line and correspond to the estimated processing

cost of messages at the network computers. The objective of the problem addressed in this chapter is to minimize the total connection and message delay costs.

We use the following notation:

N	the set of nodes in the network
E	the set of undirected links in the network
R	the set of available capacity levels
M	the set of communicating node pairs
A^m	the message arrival rate for communicating node pair $m \in M$
$O(m)$	the source node for communicating node pair $m \in M$
$D(m)$	the destination node for communicating node pair $m \in M$
C_{ij}^r	variable cost for capacity level r for link (i,j) (\$/month/bps)
F_{ij}^r	fixed cost for capacity level r for link (i,j) (\$/month)
Q_{ij}^r	capacity of type r for link (i,j) (bps)
D	unit cost of delay (\$/month/message)

The decision variables are

$$Y_{ij}^m = \begin{cases} 1 & \text{if the route for communicating node pair } m \text{ traverses link} \\ & (i,j) \text{ in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^{mr} = \text{flow of communicating node pair } m \text{ on link } (i,j) \text{ with capacity level } r$$

$$W_{ij}^r = \begin{cases} 1 & \text{if link } (i,j) \text{ is assigned capacity level } r \\ 0 & \text{otherwise} \end{cases}$$

The problem can now be formulated as follows:

Problem P:

$$\begin{aligned}
 Z_P = \text{Min } D \sum_{(i,j) \in E} & \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} C_{ij}^r \sum_{m \in M} X_{ij}^{mr} + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r
 \end{aligned} \quad (1)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \quad (2)$$

$$\frac{1}{\mu} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{r \in R} X_{ij}^{mr} \quad \forall (i,j) \in E \text{ and } m \in M \quad (3)$$

$$\frac{1}{\mu} \sum_{m \in M} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{m \in M} \sum_{r \in R} X_{ij}^{mr} \quad \forall (i,j) \in E \quad (4)$$

$$X_{ij}^{mr} \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (5)$$

$$\sum_{m \in M} X_{ij}^{mr} \leq Q_{ij}^r W_{ij}^r \quad \forall (i,j) \in E \text{ and } r \in R \quad (6)$$

$$\sum_{r \in R} W_{ij}^r = 1 \quad \forall (i,j) \in E \quad (7)$$

$$X_{ij}^{mr} \geq 0 \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (8)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (9)$$

$$W_{ij}^r \in (0,1) \quad \forall (i,j) \in E \text{ and } r \in R \quad (10)$$

In this formulation, the objective function minimizes the sum of the delay cost and the fixed and variable costs. Constraint set (2) contains the flow conservation equations which define a route (path) for each communicating node pair. Constraint set (3) links together the X_{ij}^{mr} and Y_{ij}^m variables. They ensure that the flow for communicating node pair m on link (i,j) with capacity level r is at least equal to the traffic requirement for that pair if its assigned route uses link (i,j) . Constraints in set (3) hold as equalities at the optimum. Constraints in set (4) represent an aggregate form of the constraints in set (3). Even though these constraints are redundant, they are helpful in obtaining better lower bounds in the Lagrangean relaxation suggested in the next section. Constraint set (5) guarantees that the flow for every communicating node pair m on arc (i,j) with capacity level r does not exceed its traffic requirement. Constraint set (6) enforces the capacity limitations on the links. Constraint set (7) specifies that exactly one capacity level must be selected for each link. Constraint set (8) restricts the X_{ij}^m variables to be nonnegative and constraint sets (9) and (10) enforce integrality conditions on the Y_{ij}^m and W_{ij}^r variables, respectively.

By allowing the best route for each communicating node pair to be chosen from the set of all possible routes, our solution method based on the above formulation eliminates a shortcoming that the method presented in [25] suffers which is the theoretical possibility of generating lower bound values that are higher than the optimal solution to the original routing problem when all possible routes are considered.

This formulation is related to the one used in [26]. It may be viewed as a disaggregate formulation of the routing and capacity assignment problem solved in [26]. Specifically, if one were to drop constraint sets (3) and (5) and substitute the terms

$\sum_{m \in M} X_{ij}^{mr}$ by variables which represent the total flow on link (i,j) , one would obtain the formulation in [26] without the predetermined cut constraints. The variable set X_{ij}^{mr} together with constraint sets (3) and (5) is used to represent flows between each pair of communicating nodes separately. This disaggregate formulation leads to better lower bounds and feasible solutions than does the formulation in [26] but at the expense of added computational effort as shown in section 5.

Another advantage of the method presented here is that it does not require identification of cut constraints. Since cut constraints are specific to a problem instance, and since determining good cut constraints is not a trivial task, network designers may find our procedure easier to utilize than the one outlined in [26].

3. A Lagrangean Relaxation of the Problem

Problem P is a combinatorial optimization problem with a nonlinear objective function. Consequently, it is highly unlikely that real world instances of P can be solved optimally in reasonable computation time. Therefore, we propose a composite upper and lower bounding procedure based on a Lagrangean relaxation of the problem. Consider the following Lagrangean relaxation of problem P obtained by dualizing constraint set (3) and (4) using nonnegative multipliers α_{ij}^m and β_{ij} for all $(i,j) \in E$ and $m \in M$, respectively.

Problem L:

$$\begin{aligned}
 Z_L = \text{Min } D \sum_{(i,j) \in E} & \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r \\
 & + \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m)
 \end{aligned} \tag{11}$$

subject to constraint sets (2), (5)-(10).

Problem *L* can be decomposed into the following two subproblems:

Problem L1:

$$\begin{aligned}
 Z_{L1} = \text{Min } D \sum_{(i,j) \in E} & \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r
 \end{aligned} \tag{12}$$

subject to constraint sets (5)-(8), (10).

and

Problem L2:

$$Z_{L2} = \text{Min } \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \tag{13}$$

subject to constraint set (2) and (9).

Problem $L1$ can be decomposed into $|E|$ subproblems (one for each link) as follows:

$$\begin{aligned} \text{Min } D \quad & \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\ & + \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{r \in R} F_{ij}^r W_{ij}^r \end{aligned} \quad (14)$$

subject to

$$X_{ij}^{mr} \leq \frac{1}{\mu} A^m \quad \forall m \in M \text{ and } r \in R \quad (15)$$

$$\sum_{m \in M} X_{ij}^{mr} \leq Q_{ij}^r F_{ij}^r \quad \forall r \in R \quad (16)$$

$$\sum_{r \in R} W_{ij}^r = 1 \quad (17)$$

$$X_{ij}^{mr} \geq 0 \quad \forall m \in M \text{ and } r \in R \quad (18)$$

$$W_{ij}^r \in (0,1) \quad \forall r \in R \quad (19)$$

Similarly, problem $L2$ can be decomposed into $|M|$ subproblems (one for each sourced/destination pair) as follows:

$$\text{Min } \frac{1}{\mu} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (20)$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \quad (21)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \quad (22)$$

In the next section, we present methods for solving the Lagrangean problem L and the original problem P .

4. Solution Techniques

4.1 Solution of Problem $L1$

To solve a subproblem of problem $L1$ for link (i,j) , we observe that constraint set (17) requires that exactly one capacity level must be selected for link (i,j) . Let r be a trial value for the best level. In this case, $W_{ij}^r = 1$, $X_{ij}^{mp} = 0$, and $W_{ij}^p = 0$ for $p \neq r$. The optimal values of X_{ij}^{mr} can then be determined as follows. For any given vector of multipliers α_{ij}^m and β_{ij} , and a capacity level r , the subproblem for link (i,j) is a continuous knapsack problem with nonlinear objective function that can be solved optimally using the following greedy type procedure.

Procedure-PROC1

Step 1: Reorder the X_{ij}^{mr} variables by sorting them in nonincreasing order of α_{ij}^m ; assume that the variables are reindexed in this order. Let $m=0$.

Step 2 : Let $m=m+1$ and set

$$X_{ij}^{mr} = \begin{cases} X_0 & \text{if } \alpha_{ij}^m + \beta_{ij} - C_{ij}^r > 0 \text{ and } X_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } X_0 = \min\left\{ \frac{1}{\mu} A^m, (Q_{ij}^r - S) - \left[\frac{DQ_{ij}^r}{\alpha_{ij}^m + \beta_{ij} - C_{ij}^r} \right]^{1/2} \right\} \text{ and}$$

$$S = \sum_{k < m} X_{ij}^{kr}$$

Step 3: If $m=|M|$ stop; if $X_{ij}^{mr} < \frac{1}{\mu} A^m$ then stop and set $X_{ij}^{kr}=0$ for $k=m+1, \dots, |M|$. Otherwise go to step 2.

The value assigned to X_{ij}^{mr} in step 2 is determined in a such a way that it decreases the objective function of the subproblem the most. Suppose we solve the subproblem for each capacity level $r = 1, 2, \dots, |R|$. If r^* is capacity level with the minimum objective function value, then r^* is the capacity level to be selected for link (i, j) .

The subproblems of problem *L1* are different from those solved in [26]. In the former, the total flow passing through each link depends on the individual communicating node pairs because the objective function of each subproblem contains terms related to the communicating node pairs and constraint set (5) is included in the model. In [26], the total flow is determined independently of the communicating node pairs because no disaggregation constraints similar to those in sets (3) and (5) are present in the formulation of the problem. It is this disaggregation of flows included in our formulation that leads to better feasible solutions and tighter lower bounds.

4.2 Solution of Problem *L2*

Each subproblem of problem *L2* can be solved as a shortest path problem from $O(m)$ to $D(m)$ with $(\alpha_{ij}^m + \beta_{ij})$ as the nonnegative costs on the links. In our study Dijkstra's algorithm [55] is used for solving them. These subproblems are related to those in [26] with one major difference. In [26] the "lengths" of the links are the same for all communicating node pairs, hence all the subproblems can be solved with one run of Floyd's algorithm [55] which finds shortest paths between all pairs of nodes simultaneously. In our subproblems, the "length" of a link depends on the communicating node pair using it. This requires that shortest paths be determined separately for each

communicating node pair. This is a direct result of the disaggregated treatment of the flows.

4.3 Complexity of Solving Problem L

Since the ordering of the communicating node pairs is done only once for each link in the network, the complexity of solving problem $L1$ is $O\{|E||M|\max(|R|, \log|M|)\}$. Each of the $|M|$ subproblems of problem $L2$ can be solved in $O\{|N|^2\}$ using Dijkstra's algorithm. Thus, the complexity of solving problem $L2$ is $O\{|M||N|^2\}$. Therefore, the complexity of solving problem L is $O\{\max(|E||M|\max(|R|, \log|M|), |M||N|^2)\}$.

4.4. A Heuristic Solution Method

In this section, a three phase heuristic solution procedure to solve problem P is discussed. In phase 0, an initial feasible solution is generated using the following procedure similar to Procedure-Init presented in chapter III.

Procedure-Init1:

Step 1: For each communicating node pair determine a route with the minimum number of links.

Step 2: Each link is assigned the smallest capacity r that supports the flow along the link. A link is assigned the largest available capacity $Q_{ij}^{|R|}$ if the flow along the link exceeds this capacity. If, for every link, the total flow on the link does not exceed the maximum available capacity, then a feasible solution is at hand (stop); otherwise go to step 3.

Step 3: Pick the link (i,j) with the largest amount of flow which exceeds the maximum available capacity. Among all communicating node pairs using that link, reroute the traffic requirement of the communicating node pair k which, using the "alternative path" would decrease the following total cost function the most:

$$Z_{\text{init}} = \sum_{(i,j) \in E} \text{cost}(i,j), \text{ where}$$

$$\text{cost}(i,j) = \begin{cases} \frac{1}{T} \frac{\sum_{m \in M} X_{ij}^m}{Q_{ij}^r - \sum_{m \in M} X_{ij}^m} & \text{if } \sum_{m \in M} X_{ij}^m < Q_{ij}^r \\ \mathfrak{M} \frac{\sum_{m \in M} X_{ij}^m}{Q^{|R|}} & \text{otherwise} \end{cases}$$

(\mathfrak{M} is a large positive number.)

This artificial cost function decreases when the capacity utilization of link (i,j) (and possibly of other links on the original route) for commodity k decreases. The "alternative path" is the path from $O(k)$ to $D(k)$ such that the most utilized link on the path has a lower utilization than the most utilized link on any other path from $O(k)$ to $D(k)$. The problem of finding such path is known as the bottleneck shortest path problem and can be solved using a modified version of Dijkstra's algorithm [55] in $O(|N|^2)$ operations. If the traffic requirement of no

communicating node pair can be rerouted, then stop since no initial feasible solution can be obtained; otherwise, repeat step 3 until no link capacity is violated.

Phase 1 uses the information provided by the Lagrangean relaxation. In every iteration of the subgradient optimization algorithm an attempt is made to generate a feasible solution to problem P . The best feasible solution is retained when the subgradient algorithm terminates. Phase 1 can be outlined as follows.

Step1: Check the feasibility of the Lagrangean solution for problem P .

Step2: For each link (i,j) , find the smallest capacity which accommodates the traffic determined in the Lagrangean solution of problem $L2$. If the maximum available capacity is exceeded for one or more links stop, no feasible solution is generated. Otherwise, a feasible solution is determined.

Step3: For each link (i,j) , check whether the next higher capacity level would reduce the overall cost of the network, if so increase the capacity. Stop when all links are examined.

Phase 2 of the heuristic solution procedure is only applied at the end of the subgradient optimization algorithm. It solves a routing problem which consists of finding the best possible route for each communicating node pair using the capacities selected at the termination of phase 1. This routing problem can be solved very efficiently using the method described in chapter IV. It should be noted that phase 2 of the heuristic solution procedure can be used at every iteration of the subgradient optimization algorithm to improve a given feasible solution. We have chosen not to do this as it was found to lead to excessive computational efforts.

5. Computational Results

The solution procedures presented in the previous section were coded in Pascal. A number of computational experiments were performed using IBM-3081D running under MVS/XA 2.1.7. A variety of previously used problems were utilized in the tests. We studied the four topologies shown in figures 4.1-4.4, viz. ARPA, OCT, USA and RING in chapter IV. These networks along with traffic parameters and cost structure are similar to those tested in [25,26]. Note that the appropriate line capacities would most likely correspond to T1, T2, T3, ect. lines. The capacity/cost structure used here is chosen to facilitate comparison with previously developed procedures. In all four networks each node communicates with every other node. In ARPA network there were 420 communicating node pairs with 4 messages per second being sent along the chosen route. The corresponding values were 650 and 1 for OCT, and 650 and 4 for USA and 992 and 1 for RING. The average message length was set at 400 bits and the unit cost of delay was assumed to be \$2000 per month per message for the base case. The different capacities used in the base case and their corresponding cost components are presented in table 6.1. The fixed cost (F_{ij}^r) for link (i,j) with capacity level r is computed as the sum of the initial setup cost and the distance cost which is proportional to the length of the link.

Table 6.1. Link Capacity Set and Its Cost Components

Capacity [bps]	Setup Cost [\$ /month]	Distance Cost [\$ /month/mile]	Variable Cost [\$ /month/bps]
4800	650	0.4	0.360
9600	750	0.5	0.252
19200	850	2.1	0.126
50000	850	4.2	0.030
108000	2400	4.2	0.024
230000	1300	21.0	0.020
460000	1300	60.0	0.017

The results of the experiments are reported by providing the values of the best Lagrangean bound (Lower bound), the best feasible solution (upper bound), the different cost components of the feasible solution and the gap between the upper and lower bounds expressed as a percentage of the lower bound. In order to compare the effectiveness of our procedure to that described in [26], we also report the lower bound, upper bound, and the gap obtained in [26], the percentage improvements produced by our procedure in the values of lower bounds (ILB) and the values of the upper bounds (IUB) as well as the percentage gaps closed (GC) by our solution procedure.

Table 6.2 shows the results for various message lengths. Results for different delay costs are presented in table 6.3. Tables 6.4 and 6.5 demonstrates the impact of the changes in the fixed setup cost and variable cost, respectively on the solution quality. The fixed costs tested have been set to 0%, 10%, 50%, 150% and 200% of the fixed costs used in the base case and the variable costs have been set to 0%, 50% and 150% of the variable

costs used in the base case. Results reported in tables 5, 6, 7 and 8 in [26] are for the same experiments that are summarized in tables 6.2, 6.3, 6.4 and 6.5 in this chapter, respectively.

As shown in table 6.2, when the network load increases as a result of higher message length, the delay cost increases indicating a deterioration in the response time to users. This can be explained by the dominance of the fixed cost in the overall network cost which prevents the use of higher link capacities. On average, delay cost, fixed cost and variable cost represent 22%, 61% and 17% of the total cost, respectively.

The effects of changes in the delay cost are shown in table 6.3. As the delay cost increases a tradeoff between the fixed cost and the response time is made. The fixed cost increases when the unit delay cost increases as the networks switch to higher link capacities to provide better response time to users. The improvement in the response time is reflected in the decrease in the ratio of total delay cost by the unit delay cost. On average, this ratio decreases from 543 when the unit delay cost is 100 to 173 when the unit delay is 3000. The variable cost does not vary significantly with the change in the delay cost.

The impacts of variations in the fixed cost are examined in tables 6.4. As the fixed cost increases, its share of the total cost increases. When the multiplier of the fixed cost is 0.1, the delay, fixed and variable costs represent 23%, 35% and 42% of the total cost, respectively. When the multiplier of the fixed cost is 2, these figures are 19%, 70% and 11%. In addition, as the fixed cost increases the delay cost increases indicating a deterioration in the response time to users. This is the result of the higher fixed costs which lead to the use of lower link capacities.

The effects of the changes in variable costs are reported in table 6.5. As the variable cost increases its share in the total cost increases slowly and the fixed cost

Table 6.2. Computational Results with Different Message Lengths

Network Type	Message Length	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Gavish and Altinkemer Method						IUB	GC
							Percent Gap	Lower Bound	Upper Bound	Percent Gap	ILB			
ARPA	200	38999	118634	27932	176513	185565	5.13	159307	185009	16.13	10.80	-0.30	68.21	
ARPA	300	54134	152107	39499	235370	245740	4.41	219270	243927	11.25	7.34	-0.74	60.82	
ARPA	400	72326	185137	51174	298361	308637	3.44	285440	309137	8.30	4.53	0.16	58.51	
ARPA	500	82298	232511	62624	362662	377433	4.07	351546	379516	7.96	3.16	0.55	48.81	
ARPA	600	110657	263658	71587	429600	445902	3.79	421598	446423	5.89	1.90	0.12	35.56	
OCT	300	90573	244334	71332	393967	406239	3.11	392320	418807	6.75	0.42	3.00	53.86	
OCT	400	133759	320930	89955	520221	544644	4.69	519018	558492	7.61	0.23	2.48	38.27	
OCT	500	154419	409684	108384	647417	672487	3.87	648600	690474	6.46	-0.18	2.61	40.02	
OCT	600	144641	548571	126672	778427	819884	5.33	780013	836509	7.24	-0.20	1.99	26.47	
USA	300	91666	210655	60331	345934	362652	4.83	326592	367194	12.43	5.92	1.24	61.13	
USA	400	107071	270166	74720	435920	451957	3.68	421763	458124	8.62	3.36	1.35	57.33	
USA	500	130260	329318	92200	530605	551778	3.99	520369	559836	7.58	1.97	1.44	47.39	
USA	600	141256	403471	105144	628804	649871	3.35	625054	673271	7.71	0.60	3.48	56.57	
RING	300	98906	285735	80176	436100	464817	6.58	400723	463343	15.63	8.83	-0.32	57.86	
RING	400	133858	335183	102144	550219	571185	3.81	516780	569141	10.13	6.47	-0.36	62.39	
RING	500	155114	418662	124192	663173	697968	5.25	634065	699065	10.25	4.59	0.16	48.82	
RING	600	168312	499155	144163	778604	811630	4.24	750420	815765	8.71	3.76	0.51	51.29	

Table 6.3. Computational Results with Different Delay Costs

Gavish and Altinkemer Method															
Network Type	Unit Delay	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Percent Gap		Lower Bound	Upper Bound	Percent Gap		ILB	IUB	GC
							Percent Gap	Gap			Percent Gap	Gap			
ARPA	0	0	140238	53990	166149	194228	16.90		161667	194838	20.52		2.77	0.31	17.63
ARPA	100	7636	146439	53632	190742	207707	8.89		184785	207801	12.46		3.22	0.05	28.59
ARPA	1000	45654	163873	52992	254867	262519	3.00		244592	265155	8.41		4.20	0.99	64.29
ARPA	2000	72326	185137	51174	298361	308637	3.44		285440	309137	8.30		4.53	0.16	58.51
ARPA	3000	87242	207101	49376	331002	343719	3.84		315010	344589	9.39		5.08	0.25	59.08
OCT	0	0	262210	93107	293080	355317	21.24		294687	359595	22.03		-0.55	1.19	3.59
OCT	100	13545	252911	94300	338316	360756	6.63		342723	388894	13.47		-1.29	7.24	50.76
OCT	1000	87701	280475	91796	445670	459972	3.21		449656	472426	5.06		-0.89	2.64	36.63
OCT	2000	133759	320930	89955	520221	544644	4.69		519018	558492	7.61		0.23	2.48	38.27
OCT	3000	129396	386050	87574	576230	603020	4.65		571609	616682	7.89		0.81	2.22	41.04
USA	0	0	205958	82236	247603	288194	16.39		241318	298279	23.60		2.60	3.38	30.55
USA	100	16477	217096	79491	286624	313064	9.22		283479	316012	11.48		1.11	0.93	19.62
USA	1000	60597	262153	75699	378206	398449	5.35		364858	402471	10.31		3.66	1.00	48.08
USA	2000	107071	270166	74720	435920	451957	3.68		412763	458124	10.99		5.61	1.35	66.52
USA	3000	139676	287233	74829	484854	501738	3.48		463837	510572	10.08		4.53	1.73	65.44
RING	0	0	258633	112294	312122	370927	18.84		294866	380447	29.02		5.85	2.50	35.09
RING	100	16528	261916	111225	358087	389669	8.82		337262	393960	16.81		6.17	1.09	47.54
RING	1000	81704	307883	104205	473529	493792	4.28		446966	495660	10.89		5.94	0.38	60.72
RING	2000	133858	335183	102144	550219	571185	3.81		516780	569141	10.13		6.47	-0.36	62.39
RING	3000	162469	370338	100800	611231	633607	3.66		570650	635437	11.35		7.11	0.29	67.76

Table 6.4. Computational Results with Different Fixed Costs

Network Type	Fixed Cost Multiplier	Delay Cost	Fixed Cost	Variable Cost	Gavish and Altinkemer Method						IUB	GC
					Lower Bound	Upper Bound	Percent Gap	Lower Bound	Upper Bound	Percent Gap		
ARPA	0.0	13607	0	39369	51912	52976	2.05	51676	51913	0.46	0.46	-346.90
ARPA	0.1	23857	40423	43664	106709	107944	1.16	94316	108419	14.95	13.14	92.26
ARPA	0.5	57329	104326	49177	204875	210832	2.91	192530	211582	9.90	6.41	70.62
ARPA	1.0	72326	185137	51174	298361	308637	3.44	285440	309137	8.30	4.53	58.51
ARPA	1.5	91309	248554	52992	377717	392855	4.01	366438	395886	8.04	3.08	50.13
ARPA	2.0	91309	328138	52992	456026	472439	3.60	439429	488376	11.14	3.78	67.69
ARPA	3.0	106156	480792	53062	602865	640010	6.16	576998	641771	11.23	4.48	45.11
OCT	0.0	30410	0	76921	107329	107331	0.00	107325	107332	0.01	0.00	71.43
OCT	0.1	39302	64940	80070	182596	184312	0.94	174070	187784	7.88	4.90	88.07
OCT	0.5	81490	198641	86512	357587	366643	2.53	348737	374396	7.36	2.54	65.58
OCT	1.0	133759	320930	89955	520222	544644	4.69	519018	558492	7.61	0.23	38.27
OCT	1.5	163072	436683	91190	664556	690945	3.97	668990	708312	5.88	-0.66	32.44
OCT	2.0	175403	560950	91795	801703	828148	3.30	809541	854113	5.51	-0.97	40.09
USA	0.0	21275	0	58803	79727	80078	0.44	79723	79835	0.14	0.01	-213.38
USA	0.1	36795	55031	64944	153656	156770	2.03	138933	158294	13.94	10.60	85.46
USA	0.5	76963	158236	72441	299344	307640	2.77	282152	311355	10.35	6.09	73.22
USA	1.0	107071	270166	74720	435920	451957	3.68	421763	458124	8.62	3.36	57.33
USA	1.5	108348	403873	74438	561721	586659	4.44	543087	596458	9.83	3.43	54.82
USA	2.0	123670	521116	76009	676349	720795	6.57	658150	729617	10.86	2.77	39.48
RING	0.0	25394	0	80076	104998	105470	0.45	104992	105569	0.55	0.01	18.20
RING	0.1	47379	70430	88819	200772	206628	2.92	174330	206686	18.56	15.17	84.28
RING	0.5	91158	207521	98886	383258	397565	3.73	350449	398942	13.84	9.36	73.02
RING	1.0	133858	335183	102144	550219	571185	3.81	516780	569141	10.13	6.47	62.39
RING	1.5	149902	474349	104406	699205	728657	4.21	662093	728977	10.10	5.61	58.30
RING	2.0	161876	616170	105292	835860	883338	5.68	797474	889918	11.59	4.81	51.00

Table 6.5. Computational Results with Different Variable Costs

Network Type	Var. Cost Multiplier	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Gavish and Altinkemer Method					IUB	GC
							Percent Gap	Lower Bound	Upper Bound	Percent Gap	ILB		
ARPA	0.0	74550	182792	0	245102	257342	4.99	238300	257363	8.00	2.85	0.01	37.57
ARPA	0.5	80247	174997	26374	271939	281618	3.56	261983	283415	8.18	3.80	0.64	56.49
ARPA	1.0	72326	185137	51174	298361	308637	3.44	285440	309137	8.30	4.53	0.16	58.51
ARPA	1.5	72715	185137	76454	323216	334306	3.43	308535	336065	8.92	4.76	0.53	61.55
ARPA	3.0	61200	202539	148608	400023	412347	3.08	375699	413546	10.07	6.47	0.29	69.42
OCT	0.0	120889	335639	0	427105	456528	6.89	431902	467438	8.23	-1.11	2.39	16.27
OCT	0.5	133341	320930	45195	474879	499466	5.18	475762	512735	7.77	-0.19	2.66	33.38
OCT	1.0	133759	320930	89955	520221	544644	4.69	519018	558492	7.61	0.23	2.54	38.27
OCT	1.5	134771	320930	134750	565897	590451	4.34	561904	602517	7.23	0.71	2.04	39.97
USA	0.0	105321	271327	0	361119	376648	4.30	349726	386451	10.50	3.26	2.60	59.05
USA	0.5	103579	272877	37539	399561	413995	3.61	384865	422992	9.91	3.82	2.17	63.53
USA	1.0	107071	270166	74720	435920	451957	3.68	421763	458124	8.62	3.36	1.36	57.33
USA	1.5	106684	270961	111446	474659	489091	3.04	456270	495347	8.56	4.03	1.28	64.50
RING	0.0	134544	332232	0	445748	466776	4.72	423376	462039	9.13	5.28	-1.01	48.34
RING	0.5	137407	330091	51381	498423	518879	4.10	471605	514147	9.02	5.69	-0.91	54.50
RING	1.0	133858	335183	102144	550219	571185	3.81	516780	569141	10.13	6.47	-0.36	62.39
RING	1.5	135762	327654	153816	600590	617232	2.77	562720	622355	10.60	6.73	0.83	73.85

continues to be the dominant component. The change in the variable cost did not produce significant changes in the delay and fixed cost. This may be the result of using the particular cost structure in this chapter.

Computational results indicate that the solution procedure introduced here is effective in solving the routing and capacity assignment problem. Compared to the procedure described in [26], our solution procedure improved lower bounds and feasible solution values in 88% and 89% of the test problems solved, respectively. The gap between the lower bound and the feasible solution was reduced in 98% of the problems. The average improvement in the lower bound was 3.76%. The average improvement in the feasible solution value was 1.09%. On the average 45.31% of the gap between the feasible solution value and the lower bound from [26] is closed using the procedure outlined here. The effectiveness of our solution method may be improved if constraints similar to the cut constraints used in [26] are included in the formulation of the problem. The effect of imposing such constraints is analyzed in the next section. For the test problems used in this study one iteration of the subgradient optimization procedure takes anywhere from two to ten seconds to run.

6. Effects of Imposing Cut Constraints

If a network cut, say i , is generated which separates the network into two disconnected sets of nodes A_i and B_i , then the sum of the capacities of the links which belong to that cut has to be able to support the flow over that cut. The minimal required flow over a cut can be easily determined from the traffic requirement matrix. If cuts which do not satisfy the flow requirements are identified and the Lagrangean solution is required to

have sufficient capacities to support the flow on these cuts, then the Lagrangean solution value (lower bound) can be improved.

In this section we study the effect of cut constraints on the quality of the solutions to the routing and capacity assignment problem. The following additional notation will be used to formulate the routing and capacity assignment problem with cut constraints.

- W_i set of minimum number of links which disconnects the network into two sets of connected nodes A_i and B_i .
- C_i set of communicating pairs with an origin node in A_i and a destination node in B_i or an origin node in B_i and a destination node in A_i .
- G index set of cuts which do not have any link common.
- E' set of links which are not included in any link.

The routing and capacity assignment problem with cut constraints can now be formulated as follows.

Problem P':

$$\begin{aligned}
 Z_{P'} = \text{Min} \quad & D \sum_{(i,j) \in E} \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r + \sum_{(i,j) \in E} \sum_{r \in R} C_{ij}^r \sum_{m \in M} X_{ij}^{mr}
 \end{aligned} \tag{23}$$

subject to

$$\sum_{j \in N} Y_{ij}^m - \sum_{j \in N} Y_{ji}^m = \begin{cases} 1 & \text{if } i = O(m) \\ -1 & \text{if } i = D(m) \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \quad (24)$$

$$\frac{1}{\mu} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{r \in R} X_{ij}^{mr} \quad \forall (i,j) \in E \text{ and } m \in M \quad (25)$$

$$\frac{1}{\mu} \sum_{m \in M} A^m (Y_{ij}^m + Y_{ji}^m) \leq \sum_{m \in M} \sum_{r \in R} X_{ij}^{mr} \quad \forall (i,j) \in E \quad (26)$$

$$X_{ij}^{mr} \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (27)$$

$$\sum_{m \in M} X_{ij}^{mr} \leq Q_{ij}^r W_{ij}^r \quad \forall (i,j) \in E \text{ and } r \in R \quad (28)$$

$$\sum_{r \in R} W_{ij}^r = 1 \quad \forall (i,j) \in E \quad (29)$$

$$\sum_{(i,j) \in E} \sum_{r \in R} Q_{ij}^r W_{ij}^r \geq \frac{1}{\mu} \sum_{m \in C_i} A^m \quad \forall i \in G \quad (30)$$

$$X_{ij}^{mr} \geq 0 \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (31)$$

$$Y_{ij}^m, Y_{ji}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (32)$$

$$W_{ij}^r \in (0,1) \quad \forall (i,j) \in E \text{ and } r \in R \quad (33)$$

The objective function of problem P' is the same as that of problem P . Constraint set (30) defines the new cut constraints and ensures that the sum of the capacities for the links included in each cut i is sufficient to accommodate the total flow between the nodes in A_i and B_i .

The Lagrangean relaxation of problem P' is obtained by dualizing constraint set (25) and (26) using nonnegative multipliers α_{ij}^m and β_{ij} for all $(i,j) \in E$ and $m \in M$, respectively.

Problem L' :

$$\begin{aligned}
 Z_L = \text{Min } D \sum_{(i,j) \in E} & \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r \\
 & + \frac{1}{\mu} \sum_{m \in M} \sum_{(i,j) \in E} A^m (\alpha_{ij}^m + \beta_{ij}) (Y_{ij}^m + Y_{ji}^m) \quad (34)
 \end{aligned}$$

subject to constraint sets (24), (27)-(33).

Problem L' can be decomposed into two subproblems. One subproblem L'_2 is identical to subproblem L_2 of the routing and capacity assignment problem without cut constraints. The other subproblem L'_1 is as follows:

Problem L'_I :

$$\begin{aligned}
 \min \quad & D \sum_{(i,j) \in E} \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in E} \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{(i,j) \in E} \sum_{r \in R} F_{ij}^r W_{ij}^r \quad (35)
 \end{aligned}$$

subject to constraint sets (27)-(32).

Problem L'_I can be separated into two sets of subproblems: one set includes a subproblem for each link not included in any cut and the other set includes a subproblem for each cut i . The subproblems in the first set are identical to subproblems of Problem L_I in section 3 and can be solved in the same way. The second set includes $|G|$ subproblems, one for each each cut. The subproblem associated with the i^{th} cut is

$$\begin{aligned}
 \min \quad & D \sum_{(i,j) \in W_i} \frac{\sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}}{\sum_{r \in R} Q_{ij}^r W_{ij}^r - \sum_{r \in R} \sum_{m \in M} X_{ij}^{mr}} \\
 & + \sum_{(i,j) \in W_i} \sum_{r \in R} \sum_{m \in M} (-\alpha_{ij}^m - \beta_{ij} + C_{ij}^r) X_{ij}^{mr} + \sum_{(i,j) \in W_i} \sum_{r \in R} F_{ij}^r W_{ij}^r \quad (36)
 \end{aligned}$$

subject to

$$X_{ij}^{mr} \leq \frac{1}{\mu} A^m \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (37)$$

$$\sum_{m \in M} X_{ij}^{mr} \leq Q_{ij}^r W_{ij}^r \quad \forall (i,j) \in E \text{ and } r \in R \quad (38)$$

$$\sum_{r \in R} W_{ij}^r = 1 \quad \forall (i,j) \in E \quad (39)$$

$$\sum_{(i,j) \in E} \sum_{r \in R} Q_{ij}^r W_{ij}^r \geq \frac{1}{\mu} \sum_{m \in C_i} A^m \quad (40)$$

$$X_{ij}^{mr} \geq 0 \quad \forall (i,j) \in E, r \in R \text{ and } m \in M \quad (41)$$

$$W_{ij}^m \in (0,1) \quad \forall (i,j) \in E \text{ and } m \in M \quad (42)$$

It should be noted that if we ignore constraint set (40), then the above subproblem is separated into $|W_i|$ subproblems, one for each link in the cut. Based on this observation, the subproblem for the i^{th} cut can be solved using the following procedure.

Step 1: Solve the subproblem for each link in the cut without the cut constraint in the same way a subproblem in the first set is solved.

Step 2: If the capacities assigned to the links in the cut can support the total flow to be transmitted between the nodes in A_i and B_i , then those capacities are chosen as the solutions for the subproblem corresponding to the i^{th} cut. If not, go to step 3.

Step 3: Examine all possible combinations of capacities that satisfy the cut constraint. The combination which has the minimum objective function value of the subproblem of that cut is chosen.

This approach to solve the subproblems corresponding to the cuts is efficient if only few links are included in each cut and a link is not included in more than one cut. The exhaustive search for all capacity combinations for each cut is not time consuming because the cardinality of the set R is small.

The following lemma establishes the relationship between $Z_L \leq Z_L'$, the Lagrangean value for the link subproblem with and without cut constraints respectively, using the same Lagrangean multipliers.

Lemma 1. $Z_L \leq Z_L'$.

Proof. Link subproblem with cut constraints is a restriction of the link subproblem without cut constraints.

It is important to note that the above lemma holds only when the same Lagrangean multipliers are used for both subproblems. As will see next, the solution procedure based on the formulation without the cut constraints on average slightly outperforms the one based on the formulation with the cut constraints because the former generates better multipliers which produce tighter lower bounds than does the latter procedure.

The procedure with cut constraints was tested on the same problems as was the procedure without them. The cut constraints presented in [26] were used. The computational results are reported in tables 6.6-6.9. They indicate that the addition of the cut constraints has little impact on solution quality. The average gap between lower bounds and feasible solution values using the formulation without the cut constraints is 4.82%. The same measure using the formulation with the cut constraints is 4.91%. The best result obtained by adding the cut constraints correspond to the ARPA network with unit delay cost equal to zero where the gap between the lower bound and the solution value was reduced from 16.9% to 13.82%. The worst result is obtained for USA network with fixed cost multiplier equal zero where the gap increased from 0.445 to 5.20%.

Based on the computational results, it is not worthwhile to include cut constraints in the formulation of the routing and capacity assignment problem. In addition to the small

effect of the cut constraints on obtaining better solutions, these constraints should be defined before the procedure is applied and good cuts are difficult to generate.

7. Conclusion

In this chapter we studied the routing and capacity assignment problem in backbone networks. In this problem, a route for every pair of communicating nodes is to be identified and a capacity is to be assigned to each link in the network in order to minimize the total line capacity and delay costs. A mathematical programming formulation of the problem is presented. An efficient solution procedure based on Lagrangean relaxation of the problem is developed. Computational results across a variety of networks are reported. These results indicate the procedure to be effective. We also studied the effect of including cut constraints in the formulation of the routing and capacity assignment problem and we concluded that these constraints have no significant effect on the quality of the feasible solutions.

**Table 6.6. Computational Results with Different
Message Lengths Using Cut Constraints**

Network Type	Message Length	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Percent Gap
ARPA	200	39074	118830	27817	174867	185721	6.21
ARPA	300	54134	152107	39499	235207	245740	4.48
ARPA	400	74419	182792	51507	296964	308718	3.96
ARPA	500	80434	236159	62224	362131	378817	4.61
ARPA	600	102519	272058	71707	430337	446284	3.71
OCT	300	90572	244334	71332	397519	406239	2.19
OCT	400	133677	320930	90042	520685	544649	4.60
OCT	500	138142	424894	108344	646947	671380	3.78
OCT	600	145187	548571	126888	777946	820646	5.49
USA	300	78183	226719	58711	343881	363613	5.74
USA	400	105240	271716	74579	436865	451535	3.36
USA	500	119569	345939	90204	529512	555712	4.95
USA	600	146369	399120	105010	630180	650499	3.22
RING	300	100126	280580	81374	439279	462080	5.19
RING	400	138101	331328	102298	549672	571727	4.01
RING	500	154289	410155	125084	663251	689528	3.96
RING	600	181761	482589	144792	779150	809142	3.85

Table 6.7. Computational Results with Different Delay Costs Using Cut Constraints

Network Type	Unit Delay Cost	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Percent Gap
ARPA	0	0	139806	55206	171338	195012	13.82
ARPA	100	8070	146439	53465	192879	207974	7.83
ARPA	1000	45654	163873	52992	253324	262519	3.63
ARPA	2000	74419	182792	51507	296965	308718	3.96
ARPA	3000	87242	207101	49376	330948	343719	3.86
OCT	0	0	253633	94118	292885	347751	18.73
OCT	100	13490	256424	94348	337634	364262	7.89
OCT	1000	87702	280475	91795	447146	459972	2.87
OCT	2000	133677	320930	90041	520686	544649	4.60
OCT	3000	129574	386050	87331	574379	602955	4.98
USA	0	0	213598	80585	248519	294183	18.37
USA	100	18229	211577	81206	286003	311012	8.74
USA	1000	55418	268166	74736	376681	398320	5.74
USA	2000	105240	271716	74579	436865	451535	3.36
USA	3000	138934	289141	74009	484193	502084	3.70
RING	0	0	265221	110476	311847	375697	20.47
RING	100	14859	266351	108960	356538	390170	9.43
RING	1000	86745	305813	104473	471772	497031	5.35
RING	2000	138101	331328	102298	549672	571727	4.01
RING	3000	170782	358940	101472	613202	631104	2.92

Table 6.8. Computational Results with Different
Fixed Costs Using Cut Constraints

Network Type	Fixed Cost Multiplier	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Percent Gap
ARPA	0.0	13610	0	39369	51911	52979	2.06
ARPA	0.1	28154	35644	44480	106311	108278	1.85
ARPA	0.5	57330	104326	49177	205892	210833	2.40
ARPA	1.0	74419	182792	51507	296965	308718	3.96
ARPA	1.5	91309	245810	52992	378278	390111	3.13
ARPA	2.0	89627	332828	52685	455116	475140	4.40
ARPA	3.0	98728	486969	53408	602472	639105	6.08
OCT	0.0	30410	0	76921	107330	107331	0.00
OCT	0.1	39302	64940	80070	182639	184312	0.92
OCT	0.5	83785	195521	87034	356403	366340	2.79
OCT	1.0	87702	280475	91795	447146	459972	2.87
OCT	1.5	163072	436683	91190	665134	690945	3.88
OCT	2.0	175403	560950	91795	801316	828148	3.35
USA	0.0	24089	0	59782	79727	83871	5.20
USA	0.1	34319	57251	64701	153649	156271	1.71
USA	0.5	77021	159470	72202	299177	308693	3.18
USA	1.0	105240	271716	74579	436865	451535	3.36
USA	1.5	106222	406198	74374	559554	586794	4.87
USA	2.0	115311	530416	75258	677505	720985	6.42
RING	0.0	26619	0	80326	104997	106945	1.86
RING	0.1	46261	71691	88176	200593	206128	2.76
RING	0.5	92408	206215	98569	382777	397192	3.77
RING	1.0	138101	331328	102298	549672	571727	4.01
RING	1.5	149681	477013	104304	697517	730998	4.80
RING	2.0	170784	601954	107261	840544	879999	4.69

**Table 6.9. Computational Results with Different
Variable Costs Using Cut Constraints**

Network Type	Var. Cost Multiplier	Delay Cost	Fixed Cost	Variable Cost	Lower Bound	Upper Bound	Percent Gap
ARPA	0.0	72234	185137	0	245032	257371	5.04
ARPA	0.5	72320	185137	25792	271617	283249	4.28
ARPA	1.0	74419	182792	51507	296965	308718	3.96
ARPA	1.5	72715	185137	76454	323041	334306	3.49
ARPA	3.0	61200	202539	148608	399416	412347	3.24
OCT	0.0	133328	320930	0	428378	454258	6.04
OCT	0.5	133548	320930	45120	475196	499598	5.14
OCT	1.0	133677	320930	90042	520686	544649	4.60
OCT	1.5	134574	320930	134726	565267	590230	4.42
USA	0.0	97923	281128	0	360149	379051	5.25
USA	0.5	104192	273266	37150	398933	414608	3.93
USA	1.0	105240	271716	74579	436865	451535	3.36
USA	1.5	105304	271716	111792	473974	488812	3.13
RING	0.0	135044	333306	0	447402	468050	4.62
RING	0.5	133708	336283	51581	498062	521572	4.72
RING	1.0	138101	331328	102298	549672	571727	4.01
RING	1.5	133485	333633	153413	601486	620531	3.17

CHAPTER VII

CONCLUSIONS

This dissertation has addressed several problems in the area of routing and capacity assignment in communication networks. The first problem dealt with primary route selection in backbone communication networks. Given the topology of the network (locations of the nodes and links), the traffic requirements between the communicating node pairs and the capacities of the links, a route has to be identified for each pair in order to minimize the maximum link queueing delay encountered by messages in the network. Two mathematical formulations of the problem were developed and an efficient solution procedure based on Lagrangean relaxation of the second formulation was presented.

The second problem is a variation of the first one where the objective is minimize the average queueing delay in the network. In this problem as well as the first one, the route to be selected for a communicating node pair is chosen from among all possible routes. This represents an improvement over most of the previous studies which dealt with the routing problem where the "best" route is to be selected from a predetermined subset of all possible routes. The second problem was formulated as a nonlinear mixed integer programming model. Employing this model, a tight bound was obtained and an effective solution procedure was developed. The performance of this procedure was tested and compared to a previous method presented in a study published recently. We concluded that while the previous method frequently terminates without a feasible routing scheme when link utilization exceeds moderate levels, our solution method generates very good feasible

solutions for even heavily loaded networks. We have also observed that when we minimize the maximum link delay the maximum link utilization is reduced by approximately 3% and the average link utilization is increased by approximately 2% compared to the utilizations obtained when the average queueing delay in the network is minimized.

The third problem dealt with the reliability issue in backbone communication networks. The model presented captured the effect of link failures. Reliability is achieved by selecting, for each communicating node pair, a primary route and a link-disjoint secondary route. Traffic is switched to secondary routes when primary routes are not available due to a link failure. The objective is to minimize the weighted average delay faced by messages in the network. Because the primary and secondary routes are to be chosen from among all possible routes, the problem is very large. A fairly effective solution procedure based on Lagrangean relaxation was developed.

The fourth problem addressed the routing and capacity assignment problem in backbone communication networks. Two tasks should be accomplished in this problem. The first one is to identify, for each communicating node pair, a route over which messages are to be transmitted. The routes are to be chosen from among all possible routes. The second task is to assign a capacity to each link in the network. The capacity of a link is to be selected from among a set of discrete levels of capacity. The objective is to minimize total system costs composed of delay costs and link connection costs. A solution procedure based on Lagrangean relaxation of the problem was developed. This solution procedure was tested and compared to a previous method described in a study published recently and it was shown that our solution approach performed better for different traffic loads and costs structures.

Some potential areas of research include :

- The incorporation of the effect of node failures in addition to link failures in the reliability issue in the routing problem in backbone communication network
- Including reliability considerations in the routing and capacity assignment problem
- Study the problem of designing local area networks and backbone networks simultaneously.

REFERENCES

- [1] V. Ahuja, Routing and Flow Control in System Architecture, *IBM System Journal* (18) (1979) 298-314.
- [2] A. Balakrishnan and S. C. Graves, A Composite Algorithm for a Concave-Cost network Flow Problem, *Networks* 19 (1989) 175-202.
- [3] A. Balakrishnan, T. L. Magnanti, and R. T. Wong, A Dual-Ascent Procedure for Large Scale Uncapacitated Network Design, *Operations Research* (37) (1989) 716-740.
- [4] M.S. Bazaraa and J.J. Goode, A Survey of Various Tactics for Generating Lagrangean Multipliers in the Context of Lagrangean Duality, *European Journal of Operational Research* (3) (1979) 322-328.
- [5] D.P. Bertsekas, A Class of Optimal Routing Algorithms for Communications Networks, *Proc. 1980 Int. Conf. Circuits Comput.*, Atlanta, GA, Nov, 1980.
- [6] R. R. Boorstyn and H. Frank, Large-Scale Network Topological Optimization, *IEEE Transactions on Communications* 25 (1) (1977) 29-46.
- [7] P.M. Cahin, Datapac Network Protocols, *Proc. 3rd International Conference on Computer Communication*, Toronto, Canada (1976) 150-155.
- [8] D.G. Cantor and M. Gerla, The Optimal Routing of Messages in a Computer Network Via Mathematical Programming, *IEEE Computer Conference Proceedings*, San Francisco, Sep. 1972.
- [9] W. Chou and D. L. Sapir, A Generalized Cut-Saturation Algorithm for Distributed Computer Communications Network Optimization, *Proceedings ICC* June 1982.
- [10] P.J. Courtois and P. Semal, An Algorithm for the Optimization of Nonbifurcated Flows in Computer Communication Networks, *Performance Evaluation* (1) (1981) 139-152.
- [11] G. B. Dantzig, All Shortest Routes in a Graph, *Theory of Graphs, International Symposium*, Rome 1966, Gordon and Breach, New York 1967.

- [12] R. Despres and G. Pichon, The TRANSPAC Network Status Report and Perspectives, *Proc. Online Conference on Data Networks - Development and Use*, London (1980) 209-232.
- [13] R. E. Erickson, C. L. Monma, and A. F. Veinot, Send-and-Split Method for Minimum Concave-Cost Network Flows, *Math. Oper. Res.* (12) (1987) 634-664.
- [14] D. Erlenkotter, A Dual based Procedure for Uncapacitated Facility Location, *Operations Research* (26) (1978) 992-1009.
- [15] H. Everett, Generalized Lagrange Multipliers Method for Solving Problems of Optimum Allocation of Resources, *Operations Research* (11) (1963) 399-417.
- [16] M. L. Fisher, Optimal Solution of Scheduling Problems Using Lagrangean Multipliers: Part 1, *Operations Research* (21) (1973) 1114-1127.
- [17] M. L. Fisher, The Lagrangean Relaxation Method for Solving Integer Programming Problems, *Management Science* (27) (1981) 1-18.
- [18] M. L. Fisher, An Application Oriented Guide to Lagrangean Relaxation, *Interfaces* (15) (1985) 10-21.
- [19] H. Frank, I. T. Frisch and W. Chou, Topological Considerations in the Design of the ARPA Computer Network, *Proceeding Spring Joint Computer Conference* (1970) 581-587.
- [20] H. Frank and W. Chou, Routing in Computer Networks, *Networks*, (1) (1971) 99-122.
- [21] L. Fratta, M. Gerla and L. Kleinrock, The Flow Deviation Method - An Approach to Store-and-Forward Communication Network Design, *Networks* (3) (1973) 97-133.
- [22] B. Gavish, On Obtaining the 'Best' Multipliers for a Lagrangean Relaxation for Integer Programming, *Computers and Operations Research* (5) (1978) 55-71.
- [23] B. Gavish and S. Hantler, An Algorithm for Optimal Route Selection in SNA Networks, *IEEE Transactions on Communications* (31) (1983) 1154-1161.
- [24] B. Gavish and H. Pirkul, Computer and Database Location in Distributed Computer Systems, *IEEE Transactions on Computers* (35) (1986) 583-590.
- [25] B. Gavish and I. Neuman, A system for Routing and Capacity Assignment in Computer Networks, *IEEE Transactions on Communications* (37) (1989) 360-366.
- [26] B. Gavish and K. Altinkemer, Backbone Network Design Tools with Economic Tradeoffs, *ORSA J. on Computing* (2) (1990) 226-252.

- [27] A. M. Geoffrion, Lagrangean Relaxation for Integer Programming, *Math. Programming Study* (2) (1974) 279-285.
- [28] M. Gerla, The Design of Store-and-Forward (S/F) Networks for Computer Communications, *Ph.D. Thesis*, University of California, Los Angeles 1973.
- [29] M. Gerla, Deterministic and Adaptive Routing Policies in Packet Switched Networks, presented at the *ACM-IEEE 3rd Data Communi. Symp.*, FL, Nov. 13-15, 1973.
- [30] M. Gerla and L. Kleinrock, On the Topological Design of Distributed Computer Networks, *IEEE Transactions on Computers* (25) (1977) 48-60.
- [31] M. Gerla, H. Frank, W. Chou and R. J. Eck, A Cut Saturation Algorithm for Topological Design of Packet-Switched Communication Networks, *Proceedings National Telecommunication Conference* (1974) 1074-1085.
- [32] H. Gershet and R. Weihmayer, Joint Optimization of Data Network Design and Facility Location, *IEEE Journal on Selected Areas in Communications* (8) (1990) 149-152.
- [33] M. Held and R. M. Karp, The Traveling Salesman Problem and Minimum Spanning Tree, *Operations Research* (18) (1970) 1138-1162.
- [34] M. Held, P. Wolfe and H.P. Crowder, Validation of Subgradient Optimization, *Mathematical Programming* (5) (1974) 62-68.
- [35] L. Kleinrock, *Communications Nets: Stochastic Message Flow and Delay*, New York, Dover, 1964.
- [36] L. Kleinrock, *Queuing Systems*, Volumes 1, 2, New York, Wiley-Interscience, 1975, 1976.
- [37] L. LeBlanc and R.V. Simmons, Continuous Models for Capacity Design of Large Packet-Switched telecommunication Networks, *ORSA Journal on Computing* (1) (1989) 271-286.
- [38] K. Maruyama, L. Fratta and D.T. Tang, Flow Assignment Algorithm for computer Communication Network Design with Different Classes of Packets, *Proc. of the Symposium on Computer Networks: Trends and Applications*, Gaithersburg, Maryland, Nov. 17, 1976.
- [39] C.H. McGibbon, H. Gibbs and S.C.K. Young, DATAPAC- Initial Experience with a Commercial Packet Network, *Proc. 4th International Conference on Computer Communication*, Kyoto, Japan (1978) 103-108.

- [40] C. Monma, and D. Shallcross, Methods for Designing Communication Networks With Certain Two-Connected Survivability Constraints, *Operations Research* (37) (1989) 531-541.
- [41] S. Narasimhan, Topological Design of Networks for Data Communications Systems, *Ph.D. Thesis, College of Business, Ohio State University*, 1987.
- [42] S. Narasimhan, H. Pirkul and P. De, Route Selection in Backbone Data Communication Networks, *Computer Networks and ISDN systems* (15) (1988) 121-133.
- [43] T. Ng and D. Hoang, Joint Optimization of Capacity and Flow Assignment in a Packet-Switched Communications Network, *IEEE Transactions on Communications* (35) (1987) 202-209.
- [44] R. M. Nauss, R. E. Markland, Theory and Application of an Optimization Procedure for Lock Box Location Analysis, *Management Science* (270) (1981) 855-865.
- [45] H. Pirkul, Configuring Distributed Computer Systems with Online Database Backups, *Decision Support Systems* (3) (1987) 37-46.
- [46] H. Pirkul, S. Narasimhan and P. De, Firm Expansion through Franchising: A Model and Solution Procedure, *Decision Sciences* (4) (1987) 631-645.
- [47] H. Pirkul and S. Narasimhan, Primary and Secondary Route Selection in Backbone Data Communication Networks, *ORSA Journal on Computing*, forthcoming, 1992.
- [48] B.T. Poljack, A General Method of Solving Extremum Problems, *Soviet Math. Doklady* (8) (1967) 593-597.
- [49] W. B. Powell and Y. Sheffi, The load Planning Problem of LTL motor Carriers: Problem Description and a Proposed Solution Approach, *Trans. Res.* (17A) (1983) 471-480.
- [50] W. B. Powell, A Local Improvement Heuristic for the Design of Less-Than Truckload Motor Carrier Networks. *Working Paper EES-85-3*, Princeton University, Princeton, NJ (1985).
- [51] A. Rajaram, Routing in TYMNET, *Proc. European Computation Congress* (1978).
- [52] E. Rosenberg, A Nonlinear Programming Heuristic for Computing Optimal Link Capacities in a Multi-Hour Alternate Routing Communications Network, *Operations Research* (35) (1987) 354-364.
- [53] D. L. Sapor, A Generalized Cut-Saturation Algorithm for Distributed Computer Network Optimization, *MS Thesis, Computer Studies Program, North Carolina State University*, Raleigh, NC 1978.

- [54] R. M. Soland, Optimal Facility Location With Concave Cost, *Operations Research*. (22) (1974) 373-382.
- [55] M. M. Syslo, N. Deo, and J.S. Kowalik, *Discrete Optimization Algorithms with Pascal Programs*, N.J., Prentice-Hall, 1983.
- [56] D. Tcha and K. Maruyama, On the Selection of Primary Paths for a Communication Network, *Computer Networks and ISDN Systems* (9) (1985) 257-265.
- [57] *TELENET Communication Corporation - Packet Switching Network* (Auerbach, New York, 1978).
- [58] L.R.W. Tymes, Routing and Flow Control in TYMNET, *IEEE Transactions on Communications* (8) (1981) 392-398.
- [59] R. T. Wong, Integer Programming Formulations of the Traveling Salesman Problem, *Proceedings of the 1980 IEEE International Conference on Circuits and Computers* 149-152, 1980.
- [60] _____, A Dual Based Procedure for Solving Steiner Tree Problems on a Directed Graph, *Math. Programm* (28) (1984) 271-287.
- [61] W. I. Zangwill, Minimum Concave Cost Flows in Certain Networks, *Management Sci.* (14) (1968) 429-450.