1989-04

A system for routing and capacity assignment in computer communication networks

Gavish, Bezalel; Neuman, Irina

IEEE

Gavish, Bezalel, and Irina Neuman. "A system for routing and capacity assignment in computer communication networks." IEEE Transactions on Communications 37.4 (1989): 360-366.

http://hdl.handle.net/10945/62179

A System for Routing and Capacity Assignment in Computer Communication Networks

BEZALEL GAVISH, SENIOR MEMBER, IEEE, AND IRINA NEUMAN

Abstract—**The joint problem of selecting a primary route for each communicating pair and a capacity value for each link in computer communication networks is considered. The network topology and traffic characteristics are given; a set of candidate routes and of candidate capacities for each link are also available. The goal is to obtain the least costly feasible design where the costs include both capacity and queuing components.**

Lagrangean relaxation and subgradient optimization techniques were used in order to obtain verifiable good solutions to the problem. The method was tested on several topologies, and in all cases good feasible solutions, as well as tight lower bounds were obtained.

I. Introduction

As a result of the important advantages they offer, both the number and the range of applications supported by communication based computer systems have significantly increased. A variety of computer networks, such as SNA [17], BNA [18], and DECNET [7] architectures, TELENET [25], TYMNET [26], TRANSPAC [6], and DATAPAC [4] are currently available.

This paper deals with the following problem faced by the network designer whenever a new network is set up or when an existing network is to be expanded: how to simultaneously select the link capacities and the routes to be used by the communicating nodes in the network, such as to ensure an acceptable performance level at a minimum cost. The topology of the network and estimates of the external traffic requirements are given. Messages in the network follow static, nonbifurcated routes, a routing strategy adopted by many operational networks. The effectiveness of fixed routing methods is also supported by the simulation results presented in [15], suggesting that at steady state there is no significant difference between the delays induced in a network by good static and adaptive routing strategies. Statis routing policies are implemented by providing each pair of communicating nodes in the network with an ordered set of routes, out of which the first available route is chosen whenever a session is initiated. Such is, for instance, the general framework for routing in SNA-based networks (see [1]). Recently, the model presented in [13] has been implemented by IBM in a commercial product NETDA [23]. Consistent with this approach, we concentrate here on the choice of the primary route, i.e., the recommended one in the candidate set.

Though some attempts at a formal treatment of the backbone network design problem in a general setting exist (see [3], [5], [16], [19], and more recently, [9], [10], and [24]), much of the existing literature deals with the two imbedded subproblems independently. This is often inappropriate, since the close interplay between the capacity value of a link, and the delay incurred by a given flow on that link, makes it difficult to claim that a truly good solution has been found for either of these subproblems when considered separately.

The literature focusing on the capacity and flow assignment (CFA) problem is very limited. In [21], the authors incorporate the heuristic methods for capacity assignment developed in [20], into a more general procedure. Using several initial flow assignments as starting points, the procedure iterates between the cost minimizing capacity assignment algorithms, and a flow assignment phase in which a measure of the average delay is minimized, until a local optimum is reached. In addition, a priority assignment scheme is also considered. Using a similar iterative approach, Gerla and Kleinrock present in [16] four heuristic methods for solving the CFA problem based on their flow deviation algorithm [8]. A weakness common to all existing attempts to solve the CFA problem is that no means, either theoretical or empirical, are provided in order to evaluate the quality of the heuristic solution generated. This may seriously hamper their usefulness for real life applications. For local access networks, the situation is significantly better due to the development of heuristics with constant error guarantees [2], [12], unfortunately, no such heuristics have been discovered for backbone network design problems.

The remainder of the paper is organized as follows. In Section II, the CFA problem is defined; Sections III and IV describe the methods used for generating tight lower and upper bounds on the value of the objective function. The main results of computational tests are presented in Section V, while Section VI briefly discusses possible generalizations of the model. We conclude by discussing some related open problems and suggesting further research.

II. Problem Formulation

The model to be presented here is a generalization of the one introduced on [13]. Since the two models share some common assumptions, and there are similarities in the solution procedure as well, the reader is refered to the earlier work for further details.

The queuing phenomena are captured by modeling each link as a server whose service rate is determined by its capacity, and by viewing messages on the link as customers competing for its service. The resulting model is that of a network of queues, and standard assumptions are used for modeling the queuing phenomena (see [13]).

Since the model deals in a unified way with both the flow and the capacity assignment issues, the following two distinct types of costs are considered.

- *Capacity Costs*, comprised of a fixed setup cost (including a base monthly charge and a term proportional to the distance between the two nodes), and a variable cost, which is a function of the traffic on the line;
- *Queueing Costs*, associated with the delay incurred by messages in the network.

Paper approved by the Editor for Wide Area Networks of the IEEE Communications Society. Manuscript received July 23, 1984; revised August 8, 1986, February 6, 1987, August 13, 1987, and February 25, 1988. This work was supported in part by an IBM—Management of Information Systems grant. This work was also performed in part at the Department of Administrative Sciences, Naval Postgraduate School, Monterey, CA 93943.

B. Gavish is with the Owen Graduate School of Management, Vanderbilt University, Nashville, TN 37203.

I. Neuman is with the Graduate School of Business Administration, New York University, New York, NY 10006.

IEEE Log Number 8926521.

0090-6778/89/0400-0360$01.00

The following notation will be used throughout the paper:

L = the total number of links in the network,

I_l = the index set of line types available for link l, $l \in L$,

Q_{lk} = the capacity [bps] of line type k, $k \in I_l$,

S_{lk} = the fixed cost [\$/month] of line type k, $k \in I_l$,

C_{lk} = the variable cost [\$/month/bits/s] of line type k, $k \in I_l$ per unit of traffic on link l,

D = unit cost of delay [\$/month/message],

R = the set of candidate routes. It may be obtained through various route generation procedures and/or may be provided by the users,

Π = the set of communicating origin-destination pairs in the network,

S_p = the set of candidate routes for p, $p \in \Pi$. We assume that $S_p \cap S_q = \varnothing$ for $p \neq q$,

λ_r = the message arrival rate [messages/sec] of the unique origin-destination pair associated with route r, $r \in R$. We define $\lambda_p = \lambda_r$, $\forall r \in S_p$,

γ = the total external arrival rate [messages/s],

δ_{rl} = an indicator function, taking the value one if link l is used in route r, and zero otherwise,

$1/\mu$ = the average message length [bits/message],

x_r = a decision variable, which is one if route r is chosen to carry the flow of its associated origin-destination pair, and zero otherwise, and

y_{lk} = a decision variable, which is one if line type k is assigned to link l, and zero otherwise.

The CFA problem can then be defined as finding the x_r and y_{lk} values that satisfy the following.

Problem P:

$$Z_{P1} = \min \left\{ \sum_{l \in L} D \frac{F_l}{\sum_{k \in I_l} Q_{lk} y_{lk} - F_l} + \sum_{\substack{l \in L \\ k \in I_l}} S_{lk} y_{lk} + \sum_{\substack{l \in L \\ k \in I_l}} C_{lk} F_l y_{lk} \right\} \quad (1)$$

subject to

$$\sum_{r \in R} \lambda_r \delta_{rl} x_r / \mu \leq \sum_{k \in I_l} Q_{lk} y_{lk} \ \forall l \in L \quad (2)$$

$$\sum_{k \in I_l} y_{lk} = 1 \ \forall l \in L \quad (3)$$

$$\sum_{r \in S_p} x_r = 1 \ \forall p \in \Pi \quad (4)$$

$$x_r = 0, 1 \ \forall r \in R \quad (5)$$

$$y_{lk} = 0, 1 \ \forall k \in I_l, \ l \in L \quad (6)$$

where F_l, the average bit rate [bits/s] on link l is expressed in terms of the decision variables as $\Sigma_{r \in R} \lambda_r \delta_{rl} x_r / \mu$, and is used here for ease of notation.

The first term in the objective function captures the total cost of delay, while the second and third terms refer to the total fixed and variable costs associated with the links in the network, respectively. The constraints in (2) guarantee the feasibility of the flow on each link in terms of the capacity assigned to it. Constraints in (3) and (4) guarantee that only one line type is chosen for each link, and only one route for each origin-destination pair, respectively.

To better evidence the underlying structure of the problem, a set of derived decision variables is introduced. f_l is defined as the utilization of link l, i.e., $f_l = F_l / \Sigma_{k \in I_l} Q_{lk} y_{lk}$. The CFA problem now becomes the following.

Problem P:

$$Z_P = \min \left\{ \sum_{l \in L} \frac{Df_l}{1 - f_l} + \sum_{l \in L} \sum_{k \in I_l} S_{lk} y_{lk} + \sum_{l \in L} \sum_{k \in I_l} C_{lk} Q_{lk} f_l y_{lk} \right\}$$

$$\sum_{r \in R} \lambda_r \delta_{rl} x_r / \mu \leq f_l \sum_{k \in I_l} Q_{lk} y_{lk} \ \forall l \in L \quad (7)$$

$$0 \leq f_l \leq 1 \ \forall l \in L \quad (8)$$

and (3)–(6).

The fact that the objective function is increasing in f_l allows for the constraints in (7), which correspond to the very definition of the f_l variables, to be rewritten as inequalities. Notice also that the third term in Z_P follows from the requirement that y_{lk} may be one for only one k in each I_l set.

The following approach is used to get good solutions to the problem. First, the set of constraints in (7) are relaxed, and the corresponding Lagrangean problem is constructed; next, a subgradient optimization procedure is used in order to improve on the quality of the Lagrangean lower bound. The solutions to the Lagrangean problem obtained at each of the iterations of the subgradient procedure are used as a basis for generating feasible solutions to the CFA problem. Since the value of the optimal solution lies somewhere between the lower bound and the value of the best feasible solution available, the quality of the heuristic solution generated can thus be evaluated.

III. Lagrangean Relaxation

In spite of its increased complexity, Problem P has the same fundamental structure as the problem discussed in [13]. After the relaxation of the constraints in (7), the resulting Lagrangean problem can be decomposed into $|L| + |\Pi|$ subproblems, one for each link and for each origin-destination pair, respectively.

The origin-destination subproblems are identical to the ones obtained for the earlier model, namely, the following.

Subproblem $P_p(\alpha)$:

$$L_p(\alpha) = \min \left\{ \sum_{r \in S_p} a_r x_r \right\}$$

subject to

$$\sum_{r \in S_p} x_r = 1$$

$$x_r = 0, 1, \ \forall r \in S_p$$

where $a_r = \Sigma_{l \in L} - \alpha_l \lambda_r \delta_{rl} / \mu$.

$P_p(\alpha)$ is solved by setting $x_r = 1$ for that index $b \in S_p$ that satisfies

$$a_b = \min_{r \in S_p} a_r.$$

As a result of the introduction of the y_{lk} link capacity variables, the link subproblems become the following.

Subproblem $P_l(\alpha)$:

$$L_l(\alpha) = \min \left\{ \frac{Df_l}{1 - f_l} + \sum_{k \in I_l} S_{lk} y_{lk} + \sum_{k \in I_l} Q_{lk} f_l y_{lk} (C_{lk} + \alpha_l) \right\}$$

subject to

$$0 \leq f_l \leq 1$$

$$\sum_{k \in I_l} y_{lk} = 1 \quad (9)$$

$$y_{lk} = 0, 1 \ \forall k \in I_l. \quad (10)$$

To solve the above subproblem, we take advantage of the fact that the set of candidate capacities for each link is generally of small cardinality, and exhaustively search the I_l set. Thus, for any given values of the y_{lk} variables that satisfy the constraints in (9) and (10), the subproblem becomes the following.

Subproblem $P_l(\alpha, k)$:

$$L_l(\alpha, k) = \min \left\{ D \frac{f_l}{1-f_l} + Q_{lk}(C_{lk}+\alpha_l)f_l \right\} + S_{lk}$$

subject to

$$0 \le f_l \le 1$$

where the k index corresponds to the y_{lk} variable chosen to be one.

The solution to the subproblem is

$$f_l(k) = \begin{cases} 1 - \sqrt{\dfrac{-D}{(C_{lk}+\alpha_l)Q_{lk}}} & \text{if } \dfrac{-D}{(C_{lk}+\alpha_l)Q_{lk}} < 1 \text{ and } \\ & \alpha_l < -C_{lk} \\ 0 & \text{otherwise.} \end{cases}$$

The value of the Lagrangean for any α will be equal to the sum of the optimal solutions to the subproblems as follows:

$$L(\alpha) = \sum_{l \in L} L_l(\alpha) + \sum_{p \in \Pi} L_p(\alpha).$$

Where $L_l(\alpha) = \min_{k \in I_l} L(\alpha, k)$.

The best lower bound is provided by the vector α^* that corresponds to

$$L(\alpha^*) = \max_{\alpha \le 0} L(\alpha).$$

The following theorem states the relationship that exists between $L(\alpha^*)$ and the continuous relaxation of Problem P.

Theorem 1:

$$L(\alpha^*) = \bar{Z}_p$$

where $\bar{Z}_p$ is the objective function value obtained for Problem P when the x_r and y_{lk} variables are allowed to take fractional values.

The theorem indicates that the Lagrangean relaxation cannot provide better bounds than those obtainable from solving the corresponding continous program. On the other hand, the complexity of the original problem, together with the fact that the Lagrangean subproblems have a simple structure, and that the subgradient procedure to be presented in the next section is very effective in narrowing the gap between the lower and the upper bound, justify the use of the Lagrangean technique in this case. Moreover, for the cases when certain conditions on the values of the input parameters hold (e.g., the I_l set is of low cardinality, and there is no significant variance among the values of the external arrival rates λ_r), it is possible to further improve on the quality of the Lagrangean bound by incorporating the following restriction as part of the solution procedure. In a feasible solution, it must always be possible to represent the flow an any link l as a sum of the arrival rates of some of the origin-destination pairs that might use the link as part of their primary route, i.e., in reality, the f_l variables are defined over a discrete set of values.

IV. Subgradient Optimization and Heuristic Procedures

This section presents the methods used for obtaining tight lower and upper bounds on the value of the optimal solution.

A subgradient optimization procedure is used in order to estimate α^*, the vector of multipliers corresponding to the Lagrangean that provides the tightest lower bound. A detailed description of the main steps comprising the subgradient algorithm can be found in [13].

A price often to be paid for the ease with which the relaxed problem can be solved is that, even after applying the subgradient procedure, the resulting lower bound is still of poor quality. This is explained in our case by the fact that the relaxed constraints express the very connection between the two sets of decision variables. The lower bound was tightened by generating additional constraints (i.e., constraints that would be redundant in the original problem, but that may prove to be binding in the relaxed one) and thus reducing the feasible region over which the Lagrangean problem is defined. The main idea behind the redundant constraint generation is to try to make some of the structure of the set of candidate routes "known" to the link related subproblems, i.e., an attempt to recapture some of the meaning lost through relaxation.

Define $A_l = \{p : \delta_{rl} = 1 \ \forall r \in S_p\}$, i.e., the set of origin-destination pairs whose primary route must use link l, and $B_l = \{p : \delta_{rl} = 1 \text{ for some } r \in S_p\}$ i.e., the set of origin-destination pairs that might use link l as part of their primary path. As a result, the following tighter formulation of subproblem $P_l(\alpha, k)$ can be obtained:

$$L_l(\alpha, k) = \min \left\{ D \frac{f_l}{1-f_l} + Q_{lk}(C_{lk}+\alpha_l)f_l \right\} + S_{lk}$$

subject to

$$0 \le L_{lk} \le f_l \le U_{lk} \le 1$$

where

$$L_{lk} = \sum_{p \in A_l} \sum_{\in S_p} \lambda_r / \mu Q_{lk}$$

$$U_{lk} = \sum_{p \in B_l} \sum_{\in S_p} \lambda_r / \mu Q_{lk}.$$

The solution to the subproblem is now

$$f_l(k) = \begin{cases} \bar{f}_l(k) & \text{if } -D/\{(C_{lk}+\alpha_l)Q_{lk}\} < 1, \text{ and} \\ & \alpha_l < -C_{lk}, \text{ and} \\ & L_{lk} \le \bar{f}_l(k) \le U_{lk} \\ L_{lk} & \text{if } -D/\{(C_{lk}+\alpha_l)Q_{lk}\} \ge 1, \text{ or} \\ & \alpha_l \ge -C_{lk} \text{ or} \\ & \bar{f}_l(k) < L_{lk} \\ U_{lk} & \text{if } \bar{f}_l(k) > U_{lk} \end{cases}$$

where

$$\bar{f}_l(k) = 1 - \sqrt{-D/(C_{lk}+\alpha_l)Q_{lk}}.$$

After applying the subgradient procedure, the reformulated Lagrangean problem produced significantly tighter lower bounds.

In order to bound the value of the optimal solution from above, the algorithm presented earlier was extended so that, using the solution to the Lagrangean problem obtained during the subgradient procedure as a starting point and with some additional computational effort, a sequence of feasible solutions is generated. At each iteration, the least costly capacity assignment is determined for the Lagrangean solution, as well as for several other randomly generated flow assignments that are "close" to it, in the sense of having similar reduced costs a_r. For chosen solutions, this is followed by a route improvement step, based on a modified version of the model. For a

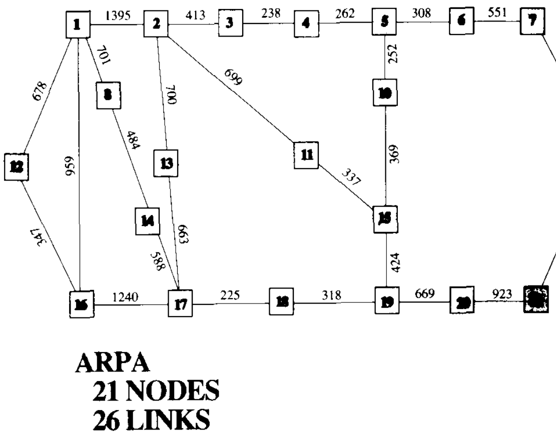

Fig. 1. Topology and distances for the ARPA network.

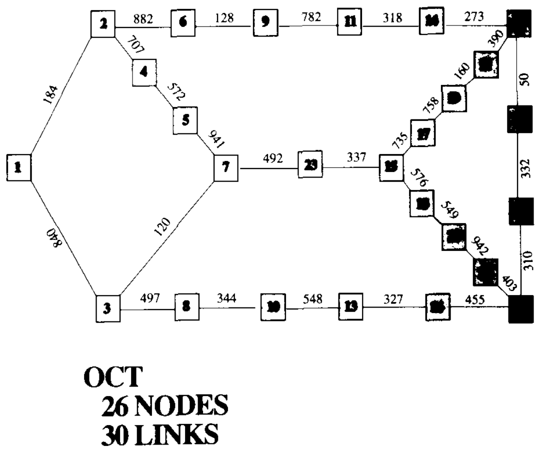

Fig. 2. Topology and distances for the OCT network.

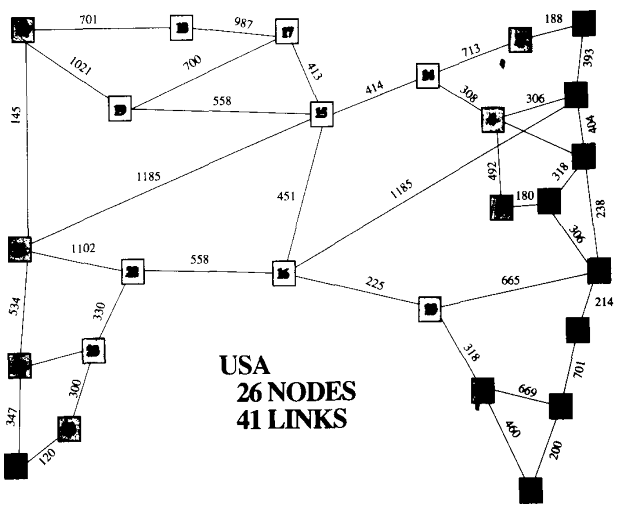

Fig. 3. Topology and distances for the USA network.

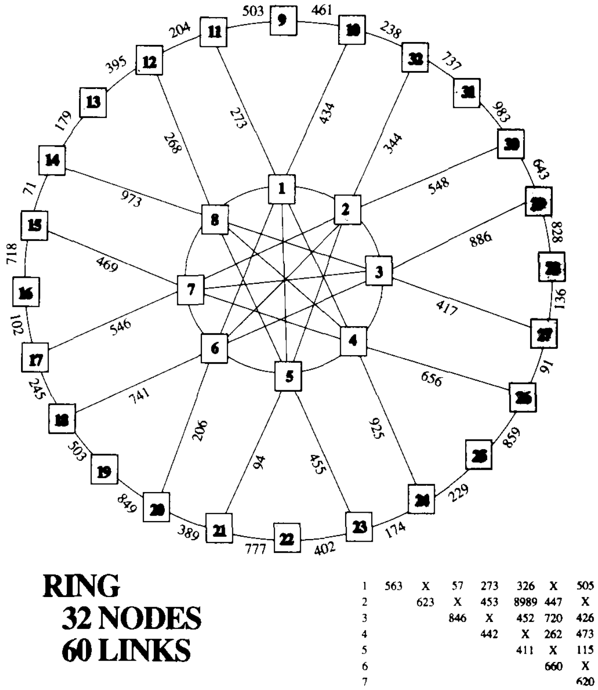

Fig. 4. Topology and distances for the RING network.

given capacity assignment, this procedure rearranges the flow on the links so as to minimize the sum of the queueing and variable costs. The algorithm iterates between the capacity and the route improvement steps until no further change in the overall cost can be achieved, and a local minimum is reached.

V. Computational Results

The model and the algorithm presented in this paper are currently implemented in a system that allows for an easy and flexible definition of the topologies to be used and of the model parameters. At the end of each major iteration (defined as a given number of subgradient iterations, to be specified by the user), control is returned to the user. At this point, the procedure can be stopped if a satisfactory solution was reached. At the beginning of each major iteration, the user has the possibility to change the values of some of the parameters that control the procedure. A comprehensive output corresponding to the best feasible solution generated so far is produced at the end of each major iteration.

Four different topologies (Figs. 1–4) were used in computational experiments with the model. In all cases, each node is allowed to communicate with each other node in the network. Two sessions are assumed to be active at each node, each session generating a traffic of 1 message/s on an average, resulting in an average traffic of 4 messages/s for both directions.

The set of candidate routes was obtained by the combined effect of two route generation algorithms, one which is based on a capacitated minimum cost flow algorithm, and another that uses a modified shortest path algorithm.

Extensive numerical experiments have been performed with the algorithm presented in this paper. Some of the results are presented here, while a more detailed presentation can be found in [14]. The experiments were conducted with two main purposes in mind: first, to test the performance of the algorithm, and second, to examine the impact of various parameters on the solutions generated, and thus to get a feeling

TABLE I
CAPACITY SET AND BASE COSTS USED IN COMPUTATIONAL EXPERIMENTS

CAPACITY [bps]	SETUP COST [dollars/month]	DISTANCE COST [dollars/month/mile]	VARIABLE COST [dollars/month/bps]
4800	650	0.4	.360
9600	750	0.5	.252
19200	850	2.1	.126
50000	850	4.2	.030
108000	2400	4.2	.024
230000	1300	21.0	.020
460000	1300	60.0	.017

Delay Cost = 2000 [dollars/month/message]
Average Message Length = 400 [bits]

TABLE II
SUMMARY OF COMPUTATIONAL RESULTS FOR DIFFERENT MESSAGE LENGTHS

Network ID	Message length	Lower bound	Upper bound	Queuing cost	Fixed cost	Variable cost	Upper/ Lower	Average message delay
ARPA	100	107838	132111	25586	90684	15840	1.225	7.6
ARPA	200	165543	186457	37505	121149	27802	1.126	11.2
ARPA	300	224499	245798	61014	145280	39504	1.095	18.2
ARPA	400	288567	311079	82854	177358	50867	1.078	24.7
ARPA	500	355536	377538	99620	216958	80960	1.061	29.6
ARPA	600	428101	446692	111768	263663	71261	1.043	33.3
OCT	300	399498	421346	93196	256673	71477	1.055	17.9
OCT	400	524964	560794	165875	303460	91459	1.068	31.9
OCT	500	652882	698284	145212	445740	107332	1.069	27.9
USA	300	342003	366102	87428	219264	59410	1.070	16.8
USA	400	435906	463281	95663	294975	72643	1.063	18.4
USA	500	534203	560667	114226	357093	88348	1.049	22.0
RING	300	453291	487288	106233	300166	80890	1.075	13.4
RING	400	571368	595285	138505	352472	104307	1.042	17.5
RING	500	686015	714269	161495	427962	124812	1.041	20.4

for the appropriateness of the model to be used as a flexible design tool.

The capacity and delay costs used as a base case are presented in Table I. For simplicity of exposition and without any loss of generality, the same set of candidate capacities was used for all links. The values for the capacity costs are the same as the ones used in [20] and [22]. The cost of delay is an estimate based on the value to the user of the time spent while waiting for an answer from the system. It is therefore determined by the user requirements, and by the type of applications using the network.

Table II shows the results obtained for different mean message lengths. Since the capacity cost components are always dominant in the overall cost, an increase in the total load almost always results in higher average message delays. Notice though that in the case of the OCT network, the average delay went down as a result of increasing the message length from 400 to 500 bits. As a result, the corresponding increase in the fixed capacity cost is even more significant now (32 percent, as opposed to roughly 18 percent in all other cases).

Table III examines the solutions obtained for different costs of delay. As expected, when the unit cost of delay increases, the expected delay in the network goes down, but at the expense of an increase in the line and in the traffic flow costs. As the delay cost goes to zero, more significant increases in the average message delay are observed. Whenever the cost of a unit of delay is difficult to predict, the designer may easily generate several solutions corresponding to different values of this parameter. The resulting curve, that corresponds to the tradeoff between response time and link costs, can then be used by the decision maker as a basis for selecting the prefered alternative.

TABLE III
SUMMARY OF COMPUTATIONAL EXPERIMENTS FOR DIFFERENT DELAY COSTS

Network ID	Delay cost	Lower bound	Upper bound	Queuing cost	Fixed cost	Variable cost	Upper/ Lower	Average message delay
ARPA	1	170337	196582	236	142113	54234	1.154	140.2
ARPA	100	183236	207904	8606	145646	53651	1.134	51.2
ARPA	400	212030	231687	25876	152902	52909	1.092	38.5
ARPA	1000	246569	265822	47187	100047	52589	1.078	28.1
ARPA	2000	287428	311079	82854	177358	50867	1.082	24.7
ARPA	3000	313269	343919	89679	204730	49510	1.097	17.8
OCT	1	313863	361092	162	267606	93324	1.150	62.2
OCT	100	338797	386999	14162	277311	95520	1.142	54.5
OCT	1000	453570	474905	93978	288133	92794	1.047	36.1
OCT	2000	524964	560794	165875	303460	91459	1.068	31.9
OCT	3000	578460	620706	134092	396386	88227	1.073	17.2
USA	1	247415	299175	418	215361	833395	1.162	160.9
USA	100	283124	320249	16328	223691	80230	1.131	62.8
USA	1000	374321	404534	61726	266984	75824	1.080	23.7
USA	2000	435906	463281	95663	294975	71643	1.063	18.4
USA	3000	485875	510445	127416	319670	72358	1.050	16.3
RING	1	311688	385053	435	272311	112307	1.160	109.6
RING	100	356498	407808	17501	278205	112102	1.143	44.1
RING	1000	494690	518119	83122	327899	107098	1.047	20.9
RING	2000	571368	595285	138505	352472	104307	1.042	17.5
RING	3000	829823	664235	183058	379326	102851	1.054	15.3

It is important to keep in mind that in many cases the data used in the planning model are based on forecasts of the future behavior of the network users, which in many cases are only rough estimates of the actual external traffic requirements. It is then highly desirable to have a robust solution, i.e., a solution whose cost when used under real traffic conditions does not significantly differ from its estimated cost. The next set of experiments tested the sensitivity of the solution to this parameter.

Define the following:

Λ_e = the matrix of estimates of traffic requirement

Λ_a = the matrix of actual traffic requirements

A_e = the capacity and routing assignment obtained based on Λ_e

A_a = the capacity and routing assignment obtained based on Λ_a.

The following cost measures are then of interest. $C(\Lambda_e, A_e)$, the estimate of the solution cost, i.e., the cost value as it is determined by the algorithm during the design stage, $C(\Lambda_a, A_e)$, the actual cost of this solution when implemented, i.e., its cost under real traffic conditions, and $C(\Lambda_a, A_a)$, the cost of the solution that would have been generated, had the actual traffic conditions been known. An important ratio that can be used as a measure of the robustness of the solutions generated by the algorithm, is $C(\Lambda_a, A_e)/C(\Lambda_a, A_a)$. Notice that this ratio will not always be greater than one, since in both cases we deal only with heuristic solutions.

In testing, great uncertainty in estimating the external traffic requirements was allowed for, by randomly generating errors within intervals ranging from ± 10 to ± 50 percent. The results, shown in Table IV, are averages over 5 problems.

TABLE IV
IMPACT OF THE ERRORS IN ESTIMATING THE EXTERNAL ARRIVAL RATES

Network ID	Error range (%)	$C(\lambda_a, A_e)$	$C(\lambda_a, A_a)$	$C(\lambda_a, A_e)/C(\lambda_a, A_a)$
ARPA	(-10,+10)	248745	248438	1.001
ARPA	(-30,+30)	253863	249170	1.018
ARPA	(-50,+50)	249880	246193	1.015
OCT	(-10,+10)	421479	425655	0.990
OCT	(-30,+30)	427369	418767	1.020
OCT	(-50,+50)	414530	421591	0.983
USA	(-10,+10)	370602	368746	1.005
USA	(-30,+30)	368969	366882	1.005
USA	(-50,+50)	372657	375934	0.991
RING	(-10,+10)	486973	489526	0.994
RING	(-30,+30)	486055	484940	1.002
RING	(-50,+50)	489163	486900	1.004

Notice that in all cases, the ratio is very close to one, meaning that there will be no significant difference between the actual cost of the solution generated by the algorithm, and the cost of the solution that could have been obtained had the real values of the external arrival rates been known. In four cases, the ratio is even less than one, i.e., due to the heurisitic nature of the procedure, the solution based on the estimates is actually better! Therefore, for this set of problems characterized by symmetric errors, the solution generated are not very sensitive to variations in the external arrival traffic, definitely an encouraging fact.

VI. MODEL GENERALIZATIONS

It is possible to render the model more accurate by allowing the service time, which is a function of the message length and of the link capacity, to follow a general distribution. This causes the mathematical complexity of the model to increase considerably, but its general structure is nevertheless preserved.

For any given service distribution, T_l the average delay incurred on link l may be expressed as a function of F_l, the total bit flow on the link, and K_l, the link capacity, i.e., $T_l = g_1(F_l, K_l)$. The exact form this function will assume depends on the specific assumptions made about the characteristics of the service distribution. Since F_l and K_l can be expressed as functions of the x_r and y_{lk} decision variables, respectively, we have $T_l = g_2(x_r, r \in R; y_{lk}, k \in I_l)$.

The objective function of the general CFA problem becomes then

$$Z = \min \left\{ \sum_{l \in L} Dg(f_l, y_{lk}, k \in I_l) + \sum_{\substack{l \in L \\ k \in I_l}} S_{lk} y_{lk} + \sum_{\substack{l \in L \\ k \in I_l}} C_{lk} Q_{lk} y_{lk} f_l \right\}$$

while the constraint set remains unchanged.

The dependence of the average delay T_l on the x_r variables is only through the intermediary of F_l. The introduction of the new set of decision variables f_l, $l \in L$ allows therefore for T_l to be expressed only as a function of link related variables: $T_l = g_3(f_l, y_{lk}, k \in I_l)$.

After the Lagrangean relaxation of the connecting constraints, the resulting link related subproblems reduce to a minimization problem of the form

$$\min_{0 \le f_l \le 1} \{G(f_l) + Af_l\}$$

where G is a nondecreasing quasiconvex function, and A is a known parameter. For the general case, these problems cannot be solved analytically, but in principle, appropriate numerical methods can always be devised for their solution.

It is often the case that, instead of a single large capacity line, two or more smaller capacity ones are assigned to a link. The resulting solution may not only have a lower cost in terms of the cost components considered here, but may also be preferable from the point of view of some other desirable criteria, e.g., reliability. This is consistent with the concept of transmission groups, as defined for SNA networks. Though in current implementations transmission groups are generally made up of homogenous lines, in principle they may contain any number of lines of various capacities.

This situation can be modeled as a special case within the general framework presented here. For any candidate set of single capacity lines I_l, the set $\tilde{I}_l$ of all possible transmission groups that can be obtained based on these capacities can be constructed. Assigning a capacity to a link corresponds now to choosing one of the generalized capacities in $\tilde{I}_l$. If exponential message lengths are assumed, the underlying queuing model for this case corresponds to a $M/M/K$ system, with K being the number of lines in the transmission group, and where, in the general case, the servers may have different service rates. Notice that transmission groups can still be represented in the $M/M/1$ based model, simply by including different configurations as part of the candidate capacity set, but the mathematical intricacies resulting from having more than one server are ignored. Since certain approximations are inherent to the very nature of the high-level design problem addressed here, choosing to concentrate on the general method, rather than on the details of the queuing model, seems justified.

VII. CONCLUSIONS

A model and solution methods for the problem of capacity and primary route assignment in computer communication networks were presented. What we see as the main value of this approach is that the model as well as the optimization procedure deal simultaneously with both aspects of the problem, thus driving the solution towards a global optimum. Recently, the model introduced here was further refined in [11] where the sets of candidate capacities associated with each origin-destination pair are no longer part of the input, and are instead dynamically generated within the model. From the computational experience with the present procedure, it can be concluded that the procedure is both efficient and effective in identifying robust solutions that are satisfactorily close to the lower bound.

The present model can be generalized to deal with different classes of customers, characterized by different priorities, message lengths, and/or delay requirements. Work is currently in progress on modeling and developing the solution techniques for the case when the delay phenomena are represented as a network of nonpreemptive head of the line priority queues.

REFERENCES

[1] V. Ahuja, "Routing and flow control in systems network architecture," *IBM Syst. J.*, vol. 18, pp. 298-314, 1979.

[2] K. Altinkemer and B. Gavish, "Heuristics with constant error guarantees for the design of tree networks," *Manage. Sci*, vol. 34, pp. 331-341, 1988.

[3] R. Boorstyn and H. Frank, "Large scale network topological optimization," *IEEE Trans. Commun.*, vol. COM-25, pp. 29-47, 1977.

[4] P. M. Cashin, "Datapac network protocols," in *Proc. 3rd Int. Conf. Commun.*, Toronto, Ont., Canada, 1976, pp. 150–155.

[5] W. Chou, F. Ferrante, and M. Balagangaadhar, "Integrated optimization of distributed processing networks," in *Proc. Nat. Comput. Conf.*, 1978, pp. 795–811.

[6] R. Despres and G. Pichon, "The Transpac network status report and perspectives," in *Proc. Online Conf. Data Networks Develop. Use*, London, England, 1980, pp. 209–232.

[7] Digital Network Architecture, General Description. Digital Equip. Corp., Maynard, MA, aa-h202a-tk Ed., 1978.

[8] L. Fratta, M. Gerla, and L. Kleinrock, "The flow deviation algorithm: An approach to store-and-forward computer communication network design," *Networks*, vol. 3, pp. 97–133, 1973.

[9] B. Gavish, "A general model for the topological design of computer networks," in *Proc. GLOBCOM '86*, 1986, pp. 1584–1588.

[10] B. Gavish, "Models for the topological design of computer communication networks." Working Paper, Dep. Admin. Sci., Naval Postgrad. School, Monterey, CA, 1987.

[11] B. Gavish and K. Altinkemer, "Backbone network design tools with economic tradeoffs," Working Paper, Krannert Grad. School Manage., Purdue Univ., West Lafayette, IN, 47907, 1987.

[12] —, "A parallel savings heuristic for the topological design of local access tree networks," in *Proc. IEEE INFOCOM '86*, Miami, FL, 1986, pp. 130–147.

[13] B. Gavish and S. L. Hantler, "An algorithm for optimal route selection in SNA networks," *IEEE Trans. Commun.*, vol. COM-31, pp. 1154–1161, 1983.

[14] B. Gavish and I. Neuman, "Capacity and flow assignment in large computer networks," in *Proc. IEEE INFOCOM '86*, Miami, FL, Apr. 1986, pp. 275–284.

[15] M. Gerla, "Deterministic and adaptive routing policies in packet switched computer networks," in *ACM IEEE 3rd Data Commun. Symp.*, Tampa, FL, 1973.

[16] M. Gerla and L. Kleinrock, "On the topological design of distributed computer networks," *IEEE Trans. Commun.*, vol. COM-25, pp. 48–60, 1977.

[17] J. P. Gray and T. B. McNeill, "SNA multiple system networking," *IBM Syst. J.*, vol. 18, pp. 263-297, 1979.

[18] S. Johari, "Burroughs network architecture: An overview," in *Proc. 2nd Joint Conf. IEEE Comput. Commun. Soc.*, San Diego, CA, 1983, pp. 503-511.

[19] K. Maruyama, "Designing reliable packet switched computer networks," IBM Thomas J. Watson Res. Cen., Yorktown Heights, NY, Tech. Rep. RC6973, 1978.

[20] K. Maruyama, K. Fratta, and D. T. Tang, "Heuristic design algorithm for computer communication networks with different classes of customers," *IBM J. Res. Develop.*, pp. 360-369, 1977.

[21] K. Maruyama and D. T. Tang, "Discrete link capacity and priority assignment in communication networks," IBM Thomas J. Watson Res. Cent., Yorktown Heights, NY, Tech. Rep. RC 6121, 1976.

[22] K. Maruyama and D. T. Tang, "Discrete link capacity assignment in communication networks," in *Proc. 3rd ICC*, Toronto, Ont., Canada, 1976, pp. 92–97.

[23] Network design and analysis (NETDA) program 5664-202. IBM, Armonk, NY, 1985.

[24] H. Pirkul and S. Narasimhan, "A new algorithm for the design of backbone networks," Working Paper, College of Business, Ohio State Univ., 1987.

[25] TELENET Communications Corporation—Packet switching network, Auerbach, NY, 1978.

[26] L. R. W. Thymes, "Routing and flow control in Tymnet," *IEEE Trans. Commun.*, vol. COM-29, pp. 298-392, 1981.

★

Bezalel Gavish (M'86–SM'87) received the B.Sc. degree in industrial engineering and management in 1967, and the M.Sc. and Ph.D. degrees in operations research from the Technion—Israel Institute of Technology, Haifa, Israel, in 1970 and 1975, respectively.

He is the 1988 Grace Murrey Hopper Professor of Computer Science at the Naval Postgraduate School in Monterey, CA, and is a Professor of Business Administration at the Owen Graduate School of Management, Vanderbilt University. From 1968 to 1973, he was Department Head of Systems Analysis and Scientific programming in charge of developing large-scale on-line logistics and transportation management computerized systems. From 1973 to 1976, he was Project Head of developing and testing a system for computer-assisted medical diagnosis at the IBM Israel Scientific Center. He joined the University of Rochester, in 1976. From 1976 to 1982 and in 1984 he was the Coordinator of the Computer and Information Systems area, being responsible for the development of the MBA and Ph.D. programs. He was a Visiting Faculty Member at the IBM Thomas J. Watson Research Center, Bell Laboratories, GTE Laboratories, and the Technion—Israel Institute of Technology. He is a consultant to numerous corporations and government agencies, and is on the editorial board of *Computers and Operations Research, Information Resources Management Journal, Annals of Mathematics of Artificial Intelligence, ORSA Journal on Computing, INFOR, Management Science,* and *Pesquisa Operacional.* His research interests span over the design and analysis of computer communication networks, design and analysis of distributed computing systems, systems analysis and design, combinatorial optimization, and scheduling and routing in logistic systems.

Dr. Gavish is a member of the Association for Computing Machinery, American Institute of Industrial Engineers, the Institute of Management Sciences, Operations Research Society of America, IEEE Computer Society, and IEEE Communications Society. He is the recipient of the 1972 IPA award and the 1982 AIIE Transactions Development and Applications Award and is the President of ORSA SIG on Telecommunications.

★

Irina Neuman received the Master degree in economic cybernetics from the Academy of Economic Sciences, Bucharest, Romania, in 1975, and the M.S. and Ph.D. degrees in computer information systems from University of Rochester in 1982 and 1987, respectively.

Since January 1985 she has been an Assistant Professor in the Information Systems Department, Graduate School of Business Administration, New York University. Her current research interests are mainly in the areas of design and analysis of computer communication systems, distributed data bases, and combinatorial optimization.