The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design

L. Fratta
Instituto di Elettrotecnia ed Elettronica
Politecnico di Milano, Italy

M. Gerla
Network Analysis Corporation
Glen Cove, New York

L. Kleinrock
University of California
Los Angeles, California

ABSTRACT

Two problems relevant to the design of a store-and-forward communication network (the message routing problem and the channel capacity assignment problem) are formulated and are recognized to be essentially non-linear, unconstrained multicommodity (m.c.) flow problems. A "Flow Deviation" (FD) method for the solution of these non-linear, unconstrained m.c. flow problems is described which is quite similar to the gradient method for functions of continuous variables; here the concept of gradient is replaced by the concept of "shortest route" flow. As in the gradient method, the application of successive flow deviations leads to local minima. Finally, two interesting applications of the FD method to the design of the ARPA Computer Network are discussed.

1. INTRODUCTION

In this paper we consider a procedure (the "flow deviation" method) for assigning flow within store-and-forward communication networks so as to minimize cost and/or delay for a given topology and for given external flow requirements. We begin by defining the basic model below and follow that with some examples. We then discuss various approaches to the problem and then introduce and describe the "flow deviation" method. This method is evaluated under some further restrictions and is then applied to various problem formulations for the ARPA network [6], [7].

Suppose we have a collection of nodes N_i, (i=1,...,n), and are required to route a quantity r_{ij} of type (i,j) commodity from N_i to N_j through a given network (Figure 1).

Networks, 3: 97-133

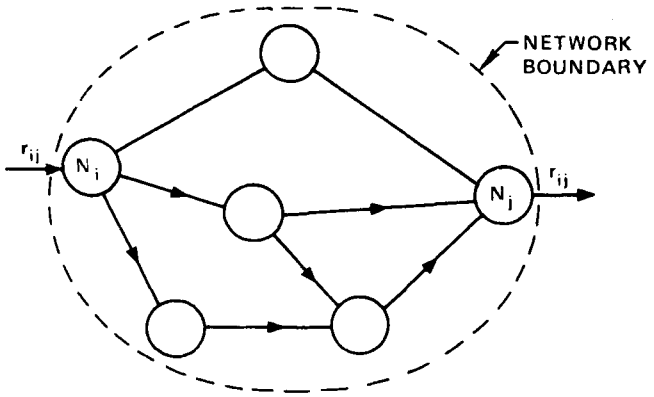


Fig. 1 Example of routing of the (i,j) commodity.

The multicommodity (m.c.) flow problem consists of finding the routes for all such commodities, which minimize (or maximize) a well-defined performance function (e.g., cost or delay), such that a set of constraints (e.g., channel capacity constraints) are satisfied.

The most general *multicommodity problem* can be expressed formally in the following way:

Given:	A network of n nodes and b arcs An $n \times n$ matrix $R = [r_{ij}]$, called the requirement matrix, whose entries are non-negative
Minimize: over Φ	(or maximize)* $P(\Phi)$ where Φ is the flow configuration and P is a well-defined performance function

Furthermore, Φ must satisfy the following constraints:

Constraints:

1. Φ must be a multicommodity flow satisfying requirement R. For this, the following conditions must be verified:
Conservation of the flow at nodes, commodity by commodity:

$$\sum_{k=1}^{n} f_{k\ell}^{(ij)} - \sum_{m=1}^{n} f_{\ell m}^{(ij)} = \begin{cases} -r_{ij} & \text{if } \ell = i \\ +r_{ij} & \text{if } \ell = j \\ 0 & \text{otherwise} \end{cases} \quad \forall i,j \qquad (1.1)$$

*Without loss of generality, only the minimum problem is considered in the following.

Non-negativity of flow in directed arcs:

$$f_{k\ell}^{(ij)} \geq 0 \quad \forall i,j,k,\ell \tag{1.2}$$

where $f_{k\ell}^{(ij)}$ is the portion of commodity (i,j) flowing on arc (k,ℓ).

2. Φ must satisfy some *additional* constraints,* different from problem to problem (e.g., capacity constraints on each channel and/or cost constraints).

Let us define the *(i,j) commodity flow* $\underset{\sim}{f}^{(ij)}$ as:

$$\underset{\sim}{f}^{(ij)} \triangleq \left(f_1^{(ij)}, f_2^{(ij)}, \ldots, f_b^{(ij)}\right)$$

where $f_m^{(ij)}$ is the portion of (i,j) commodity flowing in arc m, and define *the global flow* $\underset{\sim}{f}$ as:

$$\underset{\sim}{f} = \sum_{i=1}^{n} \sum_{j=1}^{n} \underset{\sim}{f}^{(ij)}$$

In the sequel, we restrict our analysis to m.c. problems in which the performance depends solely on the global flow:

$$P(\Phi) \equiv P(\underset{\sim}{f}) \tag{1.3}$$

However, most of the arguments and techniques presented in the paper can be extended to the general case of $P(\Phi)$ explicitly depending upon various types of commodities.

So far, we represented the flow configuration Φ in terms of $\underset{\sim}{f}^{(ij)}$, $\forall i,j$.

An equivalent representation is obtained by providing for each commodity (i,j) a set of routes π_{ij}^k, $k = 1, \ldots, K_{ij}$, from node i to node j, associated with some weights α_{ij}^k ($\alpha_{ij}^k > 0$, $\sum_{k=1}^{K_{ij}} \alpha_{ij}^k = 1$): by this we mean that commodity (i,j) is transferred from i to j along K_{ij} routes, and route π_{ij}^k carries an amount $\alpha_{ij}^k \cdot r_{ij}$ of commodity (i,j).

*If an m.c. flow problem has no additional constraints, we define it to be an unconstrained m.c. flow problem; such a definition will be motivated in one of the following sections.

As a third representation, we can consider the global flow $\underset{\sim}{f}$. It can very easily be seen that $\underset{\sim}{f}$ does not completely characterize Φ: for instance, two different sets of routes might yield the same $\underset{\sim}{f}$. However, from Equation (1.3), it turns out that such a representation is sufficient for many considerations, and is certainly more compact than the previous two. In the following we use whichever of these representations is most convenient.

It can be seen that the set of m.c. flows satisfying constraints (1.1) and (1.2) is convex. In particular, if we let $F \triangleq \{\underset{\sim}{f} | \underset{\sim}{f}$ is an m.c. flow satisfying constraints (1.1) and (1.2)$\}$, we have that F is a convex polyhedron. The global flows corresponding to the "corners" (extreme points) of F have an interesting property: they are shortest route* flows [9].

2. MULTICOMMODITY PROBLEMS IN THE DESIGN OF S/F NETWORKS

Let us now consider a store-and-forward (S/F) communication network [1]. In such a network, messages traveling from N_i to N_j are "stored" in queue at any intermediate node N_k, while awaiting transmission, and are sent "forward" to N_ℓ, the next node in the route from N_i to N_j, when channel (k,ℓ) permits. Thus, at each node there are different queues, one for each output channel. The message flow requirements between nodes arise at random times and the messages are of random lengths; therefore the flows in the channels and the queue lengths in the nodes are random variables. Under appropriate assumptions,† an analysis of the system can be carried out [1]; in particular, it is possible to relate the average delay T suffered by a message traveling from source to destination (the average is over time and over all pairs of nodes) to the average flows in the channels.

The result of the analysis is:

$$T = \sum_{i=1}^{b} \frac{\lambda_i}{\gamma} T_i \qquad (2.1)$$

A shortest route flow is an m.c. flow whose routes can be described by a shortest route matrix, computed for an arbitrary assignment of lengths to the arcs.

†*Assumptions: Poisson arrivals at nodes, exponential distribution of message length, independence of arrival processes at different nodes, independence assumption of service times at successive nodes* [1].

where

T = total average delay per message [sec/messg]
b = number of arcs in the network
λ_i = message rate on channel i [messg/sec]

$$\gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} = \text{total message arrival rate from external sources [messg/sec]}$$

T_i = average delay suffered by a message waiting for channel i [sec/messg]

T_i is the sum of two components:

$$T_i = T_i' + T_i''$$

where

$$T_i' = \frac{1}{\mu C_i - \lambda_i} = \text{transmission and queueing delay}$$

$$T_i'' = p_i = \text{propagation delay}$$

and

C_i = capacity of channel i [bits/sec]
$1/\mu$ = average message length [bits/messg]

We can rewrite Equation (2.1) as follows:

$$T = \frac{1}{\gamma} \sum_{1}^{b} \left\{ \frac{\lambda_i/\mu}{C_i - \lambda_i/\mu} + (\lambda_i/\mu)\mu p_i \right\} \tag{2.2}$$

Letting $\lambda_i/\mu = f_i$, Equation (2.2) becomes:

$$T = \frac{1}{\gamma} \sum_{1}^{b} \left\{ \frac{f_i}{C_i - f_i} + f_i p_i' \right\} \tag{2.3}$$

where

f_i = average bit rate on channel i [bits/sec]
$p_i' = \mu p_i$

The average delay T is the most common performance measure for S/F networks, and the multicommodity problem consists of finding that routing, or flow pattern F, which minimizes T.

We may now pose two problems:

Problem A: "Routing assignment"

Given: Topology, channel capacities and a requirement matrix R

Minimize: $T(\underset{\sim}{f}) = \frac{1}{\gamma} \sum_{i=1}^{b} \left(\frac{1}{C_i - f_i} + p_i'\right) f_i$

over $\underset{\sim}{f}$

Constraints: (i) $\underset{\sim}{f}$ is an m.c. flow

(ii) $\underset{\sim}{f}_i \leq C_i$, $i = 1, \ldots, b$

The problem is in the standard multicommodity form* and the *additional* constraints are capacity constraints. Let F_A be the set of feasible flows for Problem A: $F_A = F \cap \{\underset{\sim}{f} | \underset{\sim}{f} \leq \underset{\sim}{C}\}$. Clearly F_A is a convex set (intersection of convex sets).

A second interesting problem in S/F networks is formulated below. Assume that we have a given network topology in which the channel capacities have to be assigned. A cost is associated with the values of the capacities, and the total cost of the network is given. In addition, the flow routes must be determined. The problem statement is:

Problem B': "Routing and capacities assignment, general cost-capacity function"

Given: Topology, requirement matrix R, number of dollars available D

Minimize: $T(\underset{\sim}{C},\underset{\sim}{f}) = \frac{1}{\gamma} \sum_{i=1}^{b} \left(\frac{1}{C_i - f_i} + p_i'\right) f_i$

over $\underset{\sim}{C},\underset{\sim}{f}$

Constraints: (i) $\underset{\sim}{f}$ is an m.c. flow

(ii) $f_i \leq C_i$, $i = 1, \ldots, b$

(iii) $\sum_{i=1}^{b} d_i(C_i) \leq D$

where

$\underset{\sim}{C} = (C_1, C_2, \ldots, C_b)$

$d_i(C_i)$ = arbitrary cost-capacity function for arc i

The minimization can be carried out first on $\underset{\sim}{C}$, keeping $\underset{\sim}{f}$ fixed, and then on $\underset{\sim}{f}$.

**The possibility of formulating the routing problem as a multicommodity flow problem was already recognized by Frank and Chou in [24]. An interesting linear programming approach is presented there.*

If the cost-capacity functions are linear (i.e., $d_i(C_i) = d_i C_i$), then the minimization over $\underset{\sim}{C}$ can easily be performed by the method of Lagrange multipliers and we get the following optimum capacities as functions of the flows [1]:

$$C_i = f_i + \frac{D_e}{d_i} \frac{\sqrt{f_i d_i}}{\sum_{j=1}^{b} \sqrt{f_j d_j}} \tag{2.4}$$

where

$$D_e = D - \sum_{i=1}^{b} f_i d_i$$

By introducing Equation (2.4) into the expression of $T(\underset{\sim}{C},\underset{\sim}{f})$ we have:

$$T(\underset{\sim}{C},\underset{\sim}{f}) = T(\underset{\sim}{f}) = \frac{\left(\sum_{i=1}^{b} \sqrt{f_i d_i}\right)^2}{\gamma D_e} + \frac{1}{\gamma} \sum_{i=1}^{b} f_i p_i' \tag{2.5}$$

Since

$$D \geq \sum_{i=1}^{b} d_i C_i \qquad \text{for (iii)}$$

and

$$\sum_{i=1}^{b} d_i C_i \geq \sum_{i=1}^{b} d_i f_i \qquad \text{for (ii)}$$

then

$$D \geq \sum_{i=1}^{b} d_i f_i$$

and

$$D_e = D - \sum_{i=1}^{b} d_i f_i \geq 0 \qquad \text{(iv)}$$

It is easy to see from Equation (2.4) that (iv) implies also (ii) and (iii); hence both (ii) and (iii) can be replaced by (iv).

By introducing Equation (2.5) into Problem B' and using result (iv), we obtain:

Problem B: "Routing and capacities assignment, linear cost-capacity function"

Given: Topology, requirement matrix R, number of dollars D

Minimize: over $\underset{\sim}{f}$

$$T(\underset{\sim}{f}) = \frac{\left(\sum_{i=1}^{b} \sqrt{f_i d_i}\right)^2}{\gamma D_e} + \frac{1}{\gamma} \sum f_i p_i'$$

Constraints: (i) $\underset{\sim}{f}$ m.c. flow
(ii) $D_e \geq 0$

Again the problem is reduced to an optimal flow problem of the standard multicommodity form. The *additional* constraint is now a cost constraint. Let F_B be the set of feasible flows for Problem B:

$$F_B = F \cap \{\underset{\sim}{f} \mid D - \sum_{i=1}^{b} d_i f_i \geq 0\}$$

Clearly F_B is convex.

The inspection of Problems A and B motivates the following important observation:

Observation:

In both Problems A and B, the performance $T(\underset{\sim}{f})$ goes to ∞ whenever f approaches the boundaries defined by the additional constraints (i.e., when any channel becomes saturated in A, or when the excess dollars D_e reduce to zero in B).

Using mathematical programming terminology, the performance $T(\underset{\sim}{f})$ incorporates the additional constraints as *penalty functions*. From a practical point of view, such a property is very important: it guarantees the feasibility of the solution (with respect to the additional constraints) during the application of usual non-linear minimization techniques, provided a feasible starting flow is found.

The property is quite general for S/F networks: when the additional constraints are satisfied with equality, usually some saturation occurs, the queues at nodes grow large and the delay T increases rapidly.

As a consequence of the above observation, if we assume that a feasible starting solution can be found,* we can disregard

*Techniques for finding feasible starting solutions are shown in the applications section.

the additional constraints and approach Problems A and B as *unconstrained* m.c. flow problems. Problems A and B will be investigated further in later sections.

3. THE FD METHOD AS AN APPROACH TO THE SOLUTION OF NON-LINEAR M.C. FLOW PROBLEMS

In order to place the Flow Deviation (FD) method in the proper perspective in relation to the existing methods, it is convenient to classify the various m.c. flow problems into categories; for each category, the solution techniques available in the literature are reviewed and the contribution of the FD method is discussed.

a) Unconstrained M.C. Flow Problems

a.1) Linear performance. The linear min cost flow problem with no constraints on capacity has the well known shortest route solution (where the arc length is equivalent to the linear cost of the arc) [9,12]. Very efficient techniques are available for the evaluation of all shortest routes on a graph and for the routing of the commodities along such routes [9,16]; therefore it appears convenient to reduce complicated flow problems (i.e., non-linear, or constrained) to the linear, unconstrained form, which can be solved efficiently.

a.2) Non-linear performance. The most natural thing to do is to linearize the problem. Problems which are separable* and convex can be linearized by approximating the convex functions with piecewise linear functions and by introducing one supplementary variable and one constraint equation for each linearized segment [11,15,24]. This method has two serious drawbacks: first, it can be applied only to separable and convex problems; secondly, the number of variables and constraints becomes prohibitively large for large networks.

Another method, which applies to differentiable problems, consists of approximating the performance function with the tangent hyperplane, which is expressed in terms of the partial derivatives $\{\partial P/\partial f_i\}$. The min cost solution of the linearized problem is the shortest route flow, where the length of arc i is defined as $\partial P/\partial f_i$. As it will be shown later, such shortest route flow represents the direction of the *steepest descent* flow deviation.

**A separable m.c. flow problem has the form:*

$$P(\underset{\sim}{f}) = \sum_{i=1}^{b} P_i(f_i)$$

The above idea is the essence of the *FD method*, which consists of repeated evaluations of steepest descent directions and of one variable minimizations along such directions; the method (described in Section 5) is conceptually very similar to the gradient method applied to non-linear minimization problems. If the problem is differentiable, the FD method is clearly superior to the supplementary variables method mentioned before: it does not add new variables and constraints, and can be applied to non-convex, non-separable cases.

In fact, the idea of using shortest routes (computed with partial derivatives) for the solution of non-linear problems is not new: using such techniques, Dafermos [17] solved various traffic problems, formulated as unconstrained, convex m.c. flow problems, and Yaged [18] solved a min cost capacity assignment for a communications network, which was formulated as an unconstrained, concave m.c. flow problem.

Dafermos stated the conditions for the optimality of the solution and proposed an algorithm for finding the optimal routing in the convex case; the algorithm, however, is impractical for large nets, as it requires the bookkeeping of all paths for all commodities [17]. Yaged's results, on the other hand, are very restricted: they apply only to a separable, concave problem [18].

In this paper, we attempt a more general, systematic investigation of the method; we introduce the main results in a more straightforward way and in a simpler formulation than in [17]. We indicate an algorithm which is applicable to non-separable problems and which has been efficiently applied to large nets.

b) Constrained M.C. Flow Problems

b.1) Linear performance, linear constraints. The classical, and most efficient, approach is the Dantzig-Wolfe decomposition [13,14], which reduces the solution of the main problem to the repeated solution of a Master Problem and a Subproblem. The Master is a linear program containing the additional constraints, and the Subproblem, which generates new columns to introduce into the Master, is an *unconstrained* linear min cost flow problem.

b.2) Non-linear performance, non-linear constraints. The general theory of non-linear problems with non-linear constraints is very hard. The special case of convex performance and concave non-negativity constraints, however, can be attacked efficiently with the Dantzig-Wolfe decomposition for convex programs [11]; the Master Problem is a linear program, and the column generating Subproblem is an *unconstrained* convex min cost flow problem. Here is another important area of application for the FD method.

We showed that the two design problems considered in the paper can be regarded as unconstrained m.c. flow problems; therefore, in the sequel, unless otherwise specified, we refer to unconstrained problems.

4. STATIONARITY CONDITIONS

Let us assume that $P(\underset{\sim}{f})$ is continuous with its first partial derivatives. We want to establish necessary and sufficient conditions for $\underset{\sim}{f}$ to be stationary.*

The most general perturbation (which we define as *flow deviation*) around $\underset{\sim}{f}$ can be obtained as a convex combination of $\underset{\sim}{f}$ with any m.c. flow $\underset{\sim}{v}$. The result of such flow deviation, $\underset{\sim}{f}'$, is expressed as:

$$\underset{\sim}{f}' = (1 - \lambda)\underset{\sim}{f} + \lambda\underset{\sim}{v} = \underset{\sim}{f} + \lambda(\underset{\sim}{v} - \underset{\sim}{f})$$

where

$$\underset{\sim}{v} \in F, \quad 0 \leq \lambda \leq 1$$

If $\lambda \to 0$, the flow deviation is *infinitesimal*. For $\lambda = \delta\lambda << 1$, we have:

$$\delta P(\underset{\sim}{f}) \triangleq P(\underset{\sim}{f}') - P(\underset{\sim}{f}) \cong \delta\lambda \sum_{k=1}^{b} \ell_k(v_k - f_k) \qquad (4.1)$$

where

$$\ell_k = \frac{\partial P}{\partial f_k}$$

From Equation (4.1) and from the definition of stationarity, $\underset{\sim}{f}$ is stationary if:

$$\sum_{k=1}^{b} \ell_k(v_k - f_k) \geq 0, \quad \underset{\sim}{v} \in F \qquad (4.2)$$

We can also produce infinitesimal perturbations that involve only one of the commodities; $\underset{\sim}{f}$ must be stationary with respect to any one of them separately. It follows that $\underset{\sim}{f}$ is stationary if, for all (i,j) commodities:

$$\sum_{k=1}^{b} \ell_k(v_k^{(ij)} - f_k^{(ij)}) \geq 0, \forall \underset{\sim}{v} \in F^{(ij)} \qquad (4.3)$$

*$\underset{\sim}{f}$ *is defined as stationary if, for any infinitesimal perturbation* $\delta\underset{\sim}{f}$ *(such that* $\underset{\sim}{f} + \delta\underset{\sim}{f}$ *is also m.c. flow) we have*

$$P(\underset{\sim}{f} + \delta\underset{\sim}{f}) \geq P(\underset{\sim}{f})$$

A local minimum is always stationary; the opposite, however, is not true.

where $F^{(ij)}$ is the set of the feasible (i,j) commodity flows. In fact, Equations (4.2) and (4.3) are equivalent, as will be seen from the subsequent derivations. Condition (4.2) can be rewritten as:

$$\min_{\underset{\sim}{v}\in F} \sum_{k=1}^{b} \ell_k v_k \geq \sum_{k=1}^{b} \ell_k f_k \tag{4.4}$$

But, as $\underset{\sim}{f} \in F$, Equation (4.4) becomes:

$$\min_{\underset{\sim}{v}\in F} \sum_{k=1}^{b} \ell_k v_k = \sum_{k=1}^{b} \ell_k f_k \tag{4.5}$$

Similarly, Equation (4.3) becomes:

$$\min_{\underset{\sim}{v}^{(ij)}\in F^{(ij)}} \sum_{k=1}^{b} \ell_k v_k^{(ij)} = \sum_{k=1}^{b} \ell_k f_k^{(ij)} \tag{4.6}$$

Condition (4.5)* is easy to check: the right hand side can be directly evaluated, and the left hand side requires the computation of the shortest route flow under the metric $\{\ell_k\}$.

If we represent the m.c. flow as a collection of weighted routes (see Section 1), Equation (4.6) becomes:

$$\min_{\pi'} \sum_{k\in\pi'} \ell_k r_{ij} = \sum_{m=1}^{NP} \sum_{k\in\pi_m} \ell_k (\alpha_m r_{ij}) \tag{4.7}$$

where

π' is any (i,j) route

π_m, m = 1, ..., NP, are the (i,j) routes used by commodity (i,j)

α_m, m = 1, ..., NP, are the associated weights

NP is the total number of routes used by commodity (i,j)

Let $\ell(\pi) \triangleq \sum_{k\in\pi} \ell_k$; Equation (4.7) becomes:

$$\min_{\pi'} \ell(\pi') = \sum_{m=1}^{NP} \alpha_m \ell(\pi_m) \tag{4.8}$$

*A different derivation of Equation (4.5) is given in [19].

Recalling that $\alpha_m > 0, \forall m$, and $\sum_{m=1}^{NP} \alpha_m = 1$, we obtain, for all commodities (i,j):

$$\ell(\pi_1) = \ell(\pi_2) = \ldots \ell(\pi_{NP}) \leq \ell(\pi') \qquad (4.9)$$

where π' is any (i,j) route.

Condition (4.9) is stated also in [17]; a similar equilibrium condition was mentioned by Wardrop [20]. In fact, the condition is very intuitive: it states that all non-zero weight routes must have the same marginal "gain," whereas the zero-weight routes must be less (or, at most, equally) convenient than the weighted ones. For an immediate interpretation of Equation (4.9), suppose there are two paths, π_p and π_q, both with non-zero weight, which do not satisfy Equation (4.9), i.e., $\ell(\pi_p) > \ell(\pi_q)$, say. An infinitesimal deviation of commodity (i,j) from π_p to π_q produces a variation $\delta P < 0$; therefore, the initial flow configuration was not stationary.

Notice that test (4.5) is computationally more convenient than test (4.9), as (4.5) only requires the knowledge of the global flow, while (4.9) requires the knowledge of all the paths [19].

The question remains, whether the stationary point is a local (or global) minimum. If $P(\underset{\sim}{f})$ is strictly convex, the stationary point, if it exists, is unique and is a global min. If $P(\underset{\sim}{f})$ is not convex, further considerations are required.

5. DESCRIPTION OF THE FD METHOD

The results of the previous section indicate that, if $\underset{\sim}{f}$ is not a stationary flow, then the shortest route flow (evaluated under the metric $\ell_k = \partial P/\partial f_k$) represents the flow deviation of *steepest decrease* for P. This fact suggests a method, which we call *Flow Deviation method*, for the determination of stationary solutions of unconstrained, non-linear, differentiable flow problems $P(\underset{\sim}{f})$.

The FD can be regarded as an operator (denoted by $FD(\underset{\sim}{y},\lambda)\,\odot$) which maps an m.c. flow $\underset{\sim}{f}$ into another m.c. flow $\underset{\sim}{f}'$ and is defined as follows:

$$FD(\underset{\sim}{y},\lambda) \odot \underset{\sim}{f} \triangleq (1 - \lambda)\underset{\sim}{f} + \lambda\underset{\sim}{y} = \underset{\sim}{f}' \qquad (5.1)$$

where

$\underset{\sim}{y}$ is a properly chosen m.c. flow ε F

λ is the step size ($0 \leq \lambda \leq 1$)

Clearly FD is a map of $\underset{\sim}{F}$ onto itself:

$$FD(\underset{\sim}{v},\lambda): \quad \underset{\sim}{F} \rightarrow \underset{\sim}{F}$$

Now, for each $\underset{\sim}{f} \in \underset{\sim}{F}$, we want to determine a pair $(\underset{\sim}{v},\lambda)$ in such a way that the repeated application of $FD(\underset{\sim}{v},\lambda)$ (starting from any flow $\underset{\sim}{f}^0$), produces a sequence $\{\underset{\sim}{f}^n\}$ which converges to a stationary flow. If we can define such a $FD(\underset{\sim}{v},\lambda)$, then we have an algorithm for the determination of stationary flows.

It can be shown [21] that, for a function $P(\underset{\sim}{f})$ which is continous, nondegenerate* and lower bounded, the following conditions† are sufficient for the convergence of an FD-mapping to a stationary flow:

$$\begin{aligned} &\text{(i)} \quad \Delta P(\underset{\sim}{f}) \geq 0 \quad \forall\, \underset{\sim}{f} \in \underset{\sim}{F} \\ &\text{(ii)} \quad \Delta P(\underset{\sim}{f}) = 0 \Rightarrow \underset{\sim}{f} \text{ stationary} \end{aligned}$$

where $$\Delta P(\underset{\sim}{f}) = P(\underset{\sim}{f}) - P(FD \circledcirc \underset{\sim}{f})$$

Conditions (i) and (ii) require that the FD method be a true steepest descent method.

Again in [21] it was shown that under reasonable assumptions§ on $P(\underset{\sim}{f})$, the following definition of $FD(\underset{\sim}{v},\lambda)$ satisfies conditions (i) and (ii):

$$\begin{aligned} \underset{\sim}{v} &\triangleq \text{shortest route flow under metric } \ell_k\,\P \\ \lambda &\triangleq \text{minimizer of } P[(1-\lambda)\underset{\sim}{f} + \lambda\underset{\sim}{v}],\ 0 \leq \lambda \leq 1 \end{aligned} \tag{5.2}$$

*$P(\underset{\sim}{f})$ *is defined to be nondegenerate if, for any two distinct stationary flows, say* $\underset{\sim}{f}^1$ *and* $\underset{\sim}{f}^2$, *we have:*

$$P(\underset{\sim}{f}^1) \neq P(\underset{\sim}{f}^2).$$

†*Similar, but more restrictive conditions were stated by Dafermos in* [17].

§*The assumptions are:* $P(\underset{\sim}{f})$ *continuous and lower bounded; first partial derivatives continuous and nonnegative; second partial derivatives* $< +\infty$; $P(\underset{\sim}{f})$ *nondegenerate. The nonnegativity of the first partial derivatives is a reasonable assumption, as, in general, the performance that we want to minimize is an increasing function of the flow in each arc.*

¶*Notice that, by assumption,* $\ell_k = \partial P/\partial f_k \geq 0$; *this fact excludes the presence of negative cycles, which would have caused the failure of the shortest route computation (and therefore of the FD algorithm).*

Another valid definition of FD is the following.

Let:

$\pi_{ij}^{p} \triangleq$ shortest (i,j) path (under metric ℓ_k)

$\pi_{ij}^{q} \triangleq$ longest (i,j) path, with $\alpha_{ij}^{q} > 0$

Define (i,j) - deviation as the deviation of commodity (i,j) from π_{ij}^{q} to π_{ij}^{p}, which minimizes $P(\underset{\sim}{f})$. Define the FD operator as the composition of all (i,j) deviations: such a definition satisfies (i) and (ii).*

A general algorithm, based on the first definition of the FD operator, is outlined as follows:

1. Find a feasible starting flow $\underset{\sim}{f}^{0}$
2. Let n = 0
3. $\underset{\sim}{f}^{n+1} = FD(\underset{\sim}{v}^{n},\lambda^{n}) \circledcirc \underset{\sim}{f}^{n}$
4. If $\{P(\underset{\sim}{f}^{n}) - P(\underset{\sim}{f}^{n+1})\} < \varepsilon$, $\left(\text{or if } \sum_{k=1}^{b} \ell_k(f_k^n - v_k^n) < \varepsilon'\right)^{\dagger}$, where ε and ε' are acceptable positive tolerances, stop.

Otherwise, let n = n + 1 and go to 3.

The algorithm converges to stationary points; however, the only stationary points of stable equilibrium are the local minima, so we can assume that the algorithm converges to local minima.

In the case of $P(\underset{\sim}{f})$ strictly convex, the algorithm converges to the global min (see Appendix I for a proof of convergence and an upper bound on the error).

For $P(\underset{\sim}{f})$ non-convex, one should explore all local minima, in order to find the global minimum. However, a systematic search is impossible, for large-size networks, so heuristic approaches (like the repeated application of the FD algorithm to various initial flow configurations) have to be devised. In the case of $P(\underset{\sim}{f})$ concave (or quasi-concave [23]), the local minima correspond to extreme points of F, i.e., to shortest route flows [23]: this property, as shown later, greatly simplifies the FD algorithm and speeds up its convergence.

In the following sections, the FD method is applied to the solution of Problems A and B.

Such an FD operator is essentially the "equilibration operator" defined by Dafermos [17].

†*Such a test is obtained directly from the stationarity condition (3.5).*

6. THE ROUTING ASSIGNMENT

Let us consider Problem A in Section 2. The performance $T(\underset{\sim}{f})$ (see Equation (2.3)) is strictly convex (separable sum of strictly convex functions), and the feasible set F_A is a convex polyhedron. Therefore, if the problem is feasible, there is a unique stationary point, which is the global minimum. The additional constraints are included in $T(\underset{\sim}{f})$ as penalties; therefore, if we can find a feasible starting flow $\underset{\sim}{f}^0 \in F_A$, Problem A can be regarded as an unconstrained m.c. flow problem and solved with the FD method.

Let us check if $T(\underset{\sim}{f})$ satisfies the conditions for the convergence (see Section 5). The first and second partial derivatives are:

$$\frac{\partial T}{\partial f_i} = \frac{1}{\gamma}\left[\frac{C_i}{(C_i - f_i)^2} + p_i'\right] \tag{6.1}$$

$$\frac{\partial^2 T}{\partial f_i \partial f_j} = \begin{cases} 0 & \text{for } i \neq j \\ \dfrac{1}{\gamma}\dfrac{2C_i}{(C_i - f_i)^3} & \text{for } i = j \end{cases} \tag{6.2}$$

From Equation (2.3), the optimal solution $\underset{\sim}{f}^*$, if it exists (i.e., if the problem is feasible), satisfies the capacity constraints as strict inequalities ($f_i^* < C_i \ \forall i$). Therefore, we can find an $\varepsilon > 0$ s.t.:

$$\underset{\sim}{f}^* \in F_A' \triangleq F \cap \{\underset{\sim}{f} \mid f_i \leq C_i - \varepsilon\} \tag{6.3}$$

The application of the FD method can be restricted to $F_A' \subset F_A$; for $\underset{\sim}{f} \in F_A'$, the sufficient conditions on the first two derivatives of $P(\underset{\sim}{f})$ (as from Section 5) are satisfied; therefore the FD algorithm converges to the global minimum.

In order to find a flow $\underset{\sim}{f}^0 \in F_A$, several methods are available. One of them was described in [19]. Another method (applied below) consists of picking any $\underset{\sim}{f} \in F$, and then reducing the flows in all arcs by a scaling factor RE, until a feasible flow $\underset{\sim}{f}^0 = RE \cdot \underset{\sim}{f} \in F_A$ is obtained; $\underset{\sim}{f}^0$ satisfies a reduced requirement matrix $R_0 = RE \cdot R$. The FD method is applied using $\underset{\sim}{f}^0$ as

starting flow and R_0 as starting requirement; after each FD iteration, the value of RE is increased up to a level very close to saturation. The search for a feasible flow terminates when one of the two following cases occurs: either RE $\geq$ 1, and a feasible flow is found; or the network is saturated, $T(\underset{\sim}{f})$ is minimized and RE < 1. In the latter case the problem is infeasible and we are finished.

The FD algorithm for the solution of the routing problem consists of two phases, Phase 1 and Phase 2. In Phase 1 a feasible flow $\underset{\sim}{f}^0$ is found (if it exists), or the problem is declared infeasible. In Phase 2 the optimal routing is obtained. The algorithm is outlined as follows:

Phase 1:

0. With $RE_0 = 1$, let $\underset{\sim}{f}^0$ be the shortest route flow computed at $\underset{\sim}{f} = 0$, i.e. with metric $\ell_k \triangleq [\partial T/\partial f_k]_{f_k=0} = 1/\gamma(1/C_k + p_k')$.*
   Let n = 0.

1. Let $\sigma_n = \max_k \left(\frac{f_k^n}{C_k}\right)$.
   If $\sigma_n/RE_n < 1$, let $\underset{\sim}{f}^0 = \underset{\sim}{f}^n/RE_n$ and go to Phase 2. Otherwise, let $RE_{n+1} = RE_n(1 - \varepsilon\,(1 - \sigma_n))/\sigma_n$, where ε is a proper tolerance, $0 < \varepsilon < 1$.
   Let $\underset{\sim}{g}^{n+1} = \underset{\sim}{f}^n(RE_{n+1}/RE_n)$.† Go to 2.

2. Let $\underset{\sim}{f}^{n+1} = FD \odot \underset{\sim}{g}^{n+1}$
   where FD is defined as in Equation (5.2).

3. If n = 0, go to 5.

The shortest route π_{ij} is therefore the route for which $\sum_{k\in\pi_{ij}} (p_k' + 1/C_k)$ is minimum. Notice that $1/C_k$ is the transmission delay per bit on channel k and p_k' is the propagation delay. No queueing delay is considered as the traffic is zero ($f_k = 0$). So, as we expect, for $f_k \to 0$, the shortest route π_{ij} minimizes the sum of transmission + propagation delay.

†$\underset{\sim}{g}^{n+1}$ *is a feasible m.c. flow with requirement* RE_{n+1}.

4. If $\left| \sum_{k=1}^{b} \ell_k (v_k - g_k^{n+1}) \right| < \theta$ *and* $|RE_{n+1} - RE_n| < \delta$, where θ and δ are proper positive tolerances, and $\underset{\sim}{v}$ is the shortest route flow computed at $\underset{\sim}{g}^{n+1}$, stop: the problem is infeasible within tolerances θ and δ. Otherwise, go to 5.
5. Let $n = n + 1$ and go to 1.

Phase 2:

0. Let $n = 0$.
1. $\underset{\sim}{f}^{n+1} = FD \odot \underset{\sim}{f}^{n}$
2. If $\left| \sum \ell_k (v_k - f_k^n) \right| < \theta$, where θ is a proper positive tolerance, stop: $\underset{\sim}{f}^{n}$ is optimal within a tolerance θ. Otherwise, let $n = n + 1$ and go to 1.

The algorithm, in the form described above, provides only the optimum global flow $\underset{\sim}{f}$. If complete information about the routes taken by each commodity is required, a simple updating of routing tables at each FD iteration allows one to recover it at the end of the algorithm (see [19]).

7. NON-BIFURCATED ROUTING FOR LARGE AND BALANCED NETS

An m.c. flow is defined to be non-bifurcated if each commodity flows along one route only. Some applications require a non-bifurcated routing assignment; in some other applications the non-bifurcated solution is a very good approximation to the optimum bifurcated one, and is obtained with considerable saving in the amount of computation (see below). The above reasons motivate an investigation of the non-bifurcated routing assignment.

The introduction of the "non-bifurcation" constraint reduces the set of feasible m.c. flows to a discrete set: the number of elements in the set is equal to the number of all possible combinations of π_{ij} paths, $\forall$ i,j. Continuous techniques, like the FD method, cannot in general be used; discrete techniques, on the other hand, are very involved and computationally prohibitive already for networks of medium size (on the order of ten nodes). It is of interest to devise, therefore, efficient sub-optimum techniques. We will show that, in the important case of "large and balanced networks," a modification of the FD method can be successfully applied.

A network is said to be *large* if it has a large number of nodes; it is said to be *balanced* if the elements r_{ij} of the requirement matrix R are not highly diversified one from the other. For a more precise definition of "balanced," let r:

$$r \triangleq \frac{1}{(n-1)n} \sum_{ij} r_{ij}$$

be the average requirement per pair of nodes and let m:

$$m \triangleq \max_{(ij)} [r_{ij}/r]$$

be the ratio between the max and the average requirement.* Notice that $m \geq 1$ and that $m = 1$ corresponds to a uniform requirement matrix. A network is said to be balanced if m is close to 1.

We now combine these ideas into the notion of "large and balanced net." Let:

$$\eta \triangleq \frac{Km}{(n-1)\bar{p}'} \tag{7.1}$$

where: $K \triangleq b/n$, the average arc to node density of the graph.

$\bar{p}' \triangleq \left(\sum_{ij} r_{ij}p'_{ij}\right) \Big/ \sum_{ij} r_{ij}$, where p'_{ij} is the length of the shortest (i,j) path (length of a path $\triangleq$ number of arcs in the path); $\bar{p}'$ is therefore the average path length, when all commodities are routed along the shortest paths.

A network is defined *large and balanced* if $\eta << 1$. In order to motivate such a definition, let us consider, for an arbitrary m.c. flow $\underset{\sim}{f}$, the ratio of the total flow f_k in arc k, versus the contribution $f_k^{(ij)}$ given by any commodity (i,j). Let us evaluate the average of this ratio, taken over all arcs:

$$\text{average } \left(\frac{f_k}{f_k^{(ij)}}\right) \triangleq \frac{1}{b} \sum_{k=1}^{b} \left(\frac{f_k}{f_k^{(ij)}}\right) \geq \frac{1}{bmr} \sum_{k=1}^{b} f_k \tag{7.2}$$

Many other appropriate definitions of m are possible, for example $m' = \left[\sum \left(1 - \frac{r_{ij}}{r}\right)^2\right]^{1/2}$, in which case $m' = 0$ corresponds to the uniform traffic requirement.

It was shown by Kleinrock [1] that:

$$\sum_{k=1}^{b} f_k = r(n-1)n \cdot \bar{p}$$

where: $\bar{p} \triangleq \left(\sum_{ij} r_{ij}p_{ij}\right) \Big/ \sum_{ij} r_{ij}$, and p_{ij} is the number of arcs in (i,j) route, relative to the routing assignment under consideration; $\bar{p}$ is therefore the average path length.*

Equation (7.2) becomes:

$$\text{average}\left(\frac{f_k}{f_k^{(ij)}}\right) \geq \frac{(n-1)n\cdot\bar{p}}{bm} \geq \frac{(n-1)\bar{p}'}{Km} = 1/\eta \qquad (7.3)$$

From (7.3) the following property holds:

Property (7.1): In a large and balanced net, on the average, the contribution of one single commodity in any arc can be considered infinitesimal, as compared to the total flow in that arc.

In order to show how the FD method applies to the non-bifurcated solution of large and balanced nets, let us consider a new version of flow deviation, defined as the composition of deviations involving only one commodity at a time. Suppose that the flow $\underset{\sim}{f}$ is non-bifurcated; that commodity (i,j) flows on π_{ij}; and that π'_{ij} is the shortest (i,j) route, under the usual metric $\{\ell_k\}$. The FD method deviates a proper amount $\lambda \cdot r_{ij}$, $(0 \leq \lambda \leq 1)$, of (i,j) commodity from π_{ij} to π'_{ij}, such that the performance $T(\lambda)$:

$$T(\lambda) \triangleq T(\underset{\sim}{f}(1-\lambda) + \underset{\sim}{v}\lambda) \qquad (7.4)$$

where: $\underset{\sim}{f}$ contains π_{ij}

$\underset{\sim}{v}$ contains π'_{ij}

is minimized. We can rewrite Equation (7.4) as follows:

$$T(\lambda) = T(0) + \lambda \sum_{k=1}^{b} \ell_k(v_k - f_k) + O[\lambda(\underset{\sim}{v} - \underset{\sim}{f})] \qquad (7.5)$$

*Notice that $\bar{p}$ depends on the particular routing assignment, while $\bar{p}'$ depends on requirement matrix and topology only; also notice that $\bar{p} \geq \bar{p}'$.

where 0( ) contains the terms of order higher than 1. Due to Property (7.1), the terms $(v_k - f_k)$ can be considered as infinitesimal, and the term 0( ) is infinitesimal of order higher than 1. Therefore, as long as θ, defined as:

$$\theta \triangleq \sum_{k=1}^{b} \ell_k (v_k - f_k)$$

is sufficiently negative, the term 0( ) can be disregarded and the minimizer of $T(\lambda)$ in Equation (7.5) is at the boundary ($\lambda_{min} = 1$); hence the FD method preserves the non-bifurcated characteristic of the flow. On the other hand, if θ vanishes, the higher order terms become important and it might happen that $\lambda_{min} < 1$; however, $\theta \simeq 0$ implies that $\underset{\sim}{f}$ is very close to optimum (see Appendix for bounds on the error). Therefore, the FD method provides non-bifurcated solutions which are very good approximations to the optimum bifurcated solution, and, as a consequence, very good approximations also to the optimum non-bifurcated solution.

The non-bifurcated FD algorithm is next introduced:

Non-Bifurcated FD Algorithm

Let $\underset{\sim}{f}^0$ be a starting feasible non-bifurcated flow.*
Let n = 0.

1. Compute $SR(\underset{\sim}{f}^n)$, defined as the set of shortest routes under metric $\{\ell_k\}$.

2. Let $\underset{\sim}{g} = \underset{\sim}{f}^n$.

   For each commodity (i,j):

   2.a Let $\underset{\sim}{v}$ be the flow configuration obtained from $\underset{\sim}{g}$ by deviating commodity (i,j) to the shortest route π'_{ij} given by $SR(\underset{\sim}{f}^n)$.
   2.b If [$\underset{\sim}{v}$ feasible *and* $T(\underset{\sim}{v}) < T(\underset{\sim}{g})$], go to 2.c. Otherwise, go to 2.d.
   2.c $\underset{\sim}{g} = \underset{\sim}{v}$
   2.d If all commodities (i,j) have been processed, go to 3. Otherwise, go to 2.a.

3. If $\underset{\sim}{g} = \underset{\sim}{f}^n$, stop. The FD method cannot improve the non-bifurcated solution any further. Otherwise, let $\underset{\sim}{f}^{n+1} = \underset{\sim}{g}$, n = n + 1 and go to 1.

Such a starting flow can be found with a Phase 1 procedure, similar to that described in Section 6.

The algorithm converges in a finite number of steps, as there are only a finite number of non-bifurcated flows, and repetitions of the same flow are excluded by the stopping condition.

An application of the algorithm to a large and balanced net is presented in the application section.

8. THE ROUTING AND CAPACITIES ASSIGNMENT

It was shown in Section 2, that F_B, the feasible set for Problem B, is a convex polyhedron; it was also shown that the additional constraint is included in the performance $T(\underset{\sim}{f})$ as penalty function, so that Problem B can be regarded as an unconstrained m.c. flow problem.

Let us now investigate the properties of $T(\underset{\sim}{f})$. Recall (see Equation 2.5):

$$T(\underset{\sim}{f}) = \frac{\left(\sum_{i=1}^{b} \sqrt{f_i d_i}\right)^2}{\gamma\left(D - \sum_{i=1}^{b} f_i d_i\right)} + \sum f_i p_i' \qquad (8.1)$$

Kleinrock, in [1], considered this case and also dealt extensively with a simplified version of Equation (8.1)* He showed that, whenever two routes, say π_{ij}^1 and π_{ij}^2, with the same number of intermediate arcs, are available for commodity (i,j), then $T(\underset{\sim}{f})$ is minimized when the entire commodity is routed on one of the two routes only. Such a result, obtained under restrictive assumptions, suggests the conjecture that the optimal flow be, in general, non-bifurcated. In fact, further research has been done [21], [22], and it can be shown that $T(\underset{\sim}{f})$ in in Equation (8.1) is *quasi-concave* on F_B, i.e., given any two feasible flows $\underset{\sim}{f}^1$ and $\underset{\sim}{f}^2$ [23]:

$$T(\underset{\sim}{f}^1) \leq T(\underset{\sim}{f}^2) \Rightarrow T(\underset{\sim}{f}^1) \leq T[(1-\lambda)\underset{\sim}{f}^1 + \lambda \underset{\sim}{f}^2]$$

where: $0 \leq \lambda \leq 1$.

More generally, $T(\underset{\sim}{f})$ can be shown to be quasi-concave for all "routing and capacities assignment" problems with concave cost-capacity functions [21]; the linear case is therefore a special case.

*Essentially, $d_i = 1$ and $p_i' = 0$, $\forall i$.

As a consequence of such a property, the local minima are at extreme points of F_B, i.e., they correspond to shortest route flows (see Section 3), which are a subclass of the class of non-bifurcated flows.

The FD method, when applied to Problem B, can be greatly simplified: the step size λ is always equal to 1 (if we find a downhill direction, we go all the way down, due to the quasi-concavity of $T(\lambda)$), and the flow patterns generated are completely defined by just one $(n \times n)$ matrix, the shortest route matrix.

A schematic description of the FD algorithm, as applied to Problem B, is as follows:

0. Suppose* $\underset{\sim}{f}^0 \in F_B$; let $n = 0$.
1. Let $\underset{\sim}{f}^{n+1} = FD \circledcirc \underset{\sim}{f}^n$.
2. If $(T(\underset{\sim}{f}^{n+1}) \geq T(\underset{\sim}{f}^n))$, stop; $\underset{\sim}{f}^n$ local minimum. Otherwise let $n = n + 1$ and go to 1.

The convergence of the algorithm is guaranteed by the fact that there are only a finite number of shortest route flows, and repetitions of the same flow are not possible, as $T(\underset{\sim}{f}^n)$ is strictly decreasing.

The partial derivatives, used for the shortest route computation, have the following expression:

$$\frac{\partial T}{\partial f_i} = \frac{1}{\gamma}\left(\frac{\sum\sqrt{f_j d_j}}{D_e}\right)\sqrt{\frac{d_i}{f_i}} + \frac{1}{\gamma}\left(\frac{\sum\sqrt{f_j d_j}}{D_e}\right)^2 d_i + \frac{p_i'}{\gamma}$$

Notice that $\frac{\partial T}{\partial f_i} \geq 0$; negative loops cannot exist. Also notice that:

$$\lim_{f_i \to 0} \frac{\partial T}{\partial f_i} = \infty$$

which means that, whenever the flow (and therefore the capacity, from Equation (2.4)) of an arc is reduced to zero at the end of

The problem of finding a feasible starting flow is discussed later in the section.

an FD iteration, then in such an arc, the flow and capacity are zero for all subsequent iterations, as the incremental cost of restoring the flow ($\equiv \partial T/\partial f_i$) is infinity.*

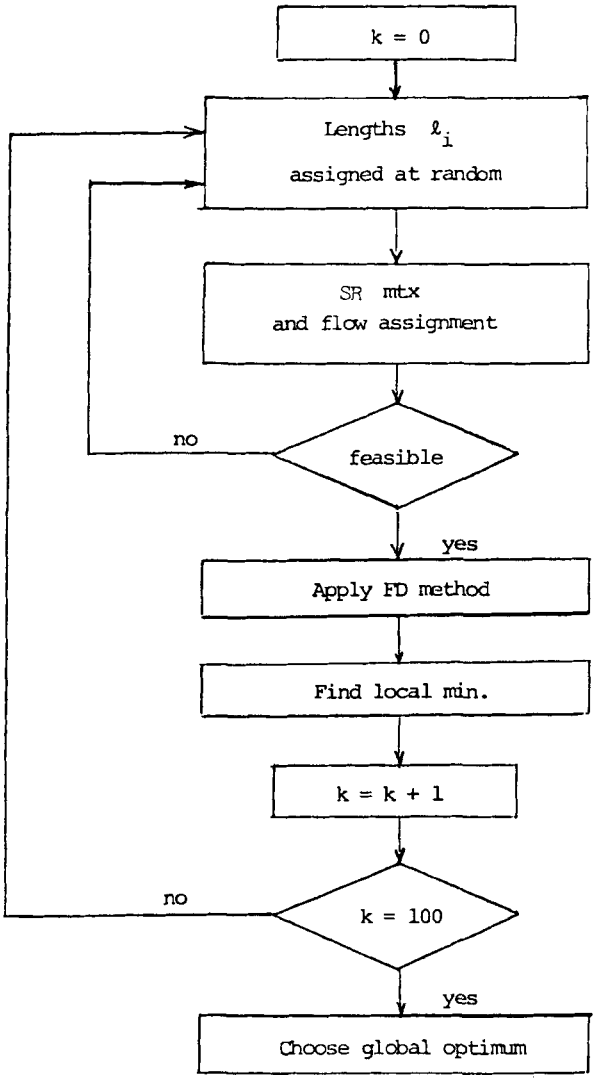


Fig. 2 Block diagram of the FD algorithm for Problem B.

*This property suggests a method for the design of the topology: we can start from a topology which is highly connected, and eliminate arcs with the FD method, until a suboptimal configuration is obtained [21]. A similar approach is used by Yaged in [18].

The FD method leads to a local minimum, which depends on the choice of the feasible starting flow. In order to find several local minima, a mechanism that produces a large variety of feasible flows is required. We propose the following randomized procedure for the generation of feasible flows:*

1. Assign initial equivalent lengths $\{\ell_i^0\}$ to the arcs *at random*.
2. Compute the shortest route flow $\underset{\sim}{f}^0$ according to the metric $\{\ell_i^0\}$.
3. If $D - \sum_{i=1}^{b} f_i d_i > 0$, $\underset{\sim}{f}^0$ is feasible and can be used to start the FD algorithm. Otherwise $\underset{\sim}{f}^0$ is rejected.

The initial random choice of the lengths guarantees a certain randomness in the starting feasible flow, thus providing a method for finding several local minima. After a convenient number of iterations, the global minimum is chosen as the minimum of the local minima. This provides a "suboptimal" solution.

A block diagram of the method is given in Figure 2.

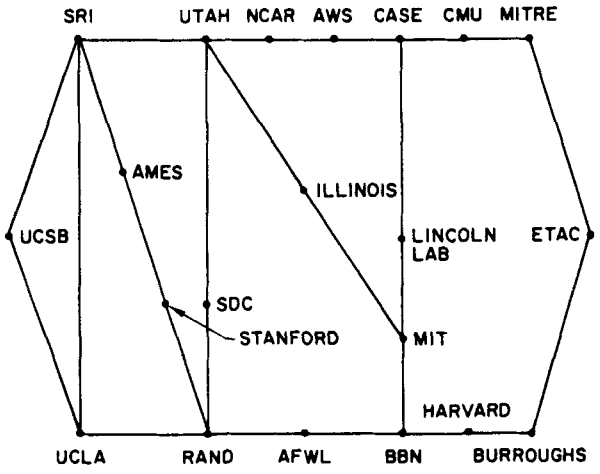


Fig. 3 A 21-node ARPA topology.

9. APPLICATIONS

As an application of the FD method, Problems A and B are solved for the ARPA Computer Network. The ARPA Computer Network is a S/F communication network connecting several computer

*Another procedure was proposed by Yaged [18].

facilities in the United States. A detailed description of the network is given in [3] - [8], [25] - [29]. Due to the fact that new computer centers are continually joining the network, its topology has been changing quite rapidly; in these applications we refer to one of the earlier proposed topologies, with 21 nodes connected by 26 full duplex channels (see Figure 3). We also assume that the traffic requirement is uniform between all pairs of nodes.

9.1 ARPA Network: The Routing Assignment

The traffic requirement $R = \{r_{ij}\}$ is assumed uniform:

$$r_{ij} = \begin{cases} r = 1.187 \text{ [kbits./sec.]}^* & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}$$

First, we show that, for the 21 node ARPA net with uniform requirement, the "large and balanced net" condition holds. From Equation (7.1), η is given by:

$$\eta = \frac{mb}{n(n-1)\bar{p}'}$$

In the present case:

$n = 21$

$\bar{p}' > 1$

$b = 52$ (each full duplex channel represents a pair of directed arcs: hence $26 \times 2 = 52$).

Hence: $\eta < 0.12 << 1$

The condition is satisfied. We can therefore apply both optimal and non-bifurcated FD algorithms and compare the results.

The result of the optimal FD algorithm is: $T_{min} = 0.2406$ sec., obtained after 80 shortest route computations, with an accuracy of 10^{-4} on T. The result of the non-bifurcated FD algorithm is: $T_{min} = 0.2438$ sec., obtained after 12 shortest path computations. The algorithms were programmed in Fortran and run on an IBM 360/91; the execution time was 30 sec. for

**The traffic requirement at saturation is $r_{sat} = 1.250$ [kbits./sec.] (see Figure 4). We chose $r = 0.95\, r_{sat} = 1.187$ in order to have a feasible, but difficult, requirement.*

the optimal algorithm and 4 sec. for the non-bifurcated one.* The error of the suboptimal non-bifurcated solution, with respect to the optimum, is less than 2 percent; the fact shows how powerful the non-bifurcated algorithm is for large and balanced nets, and suggests that a convenient modification of it could be useful for the solution of very large nets [21].

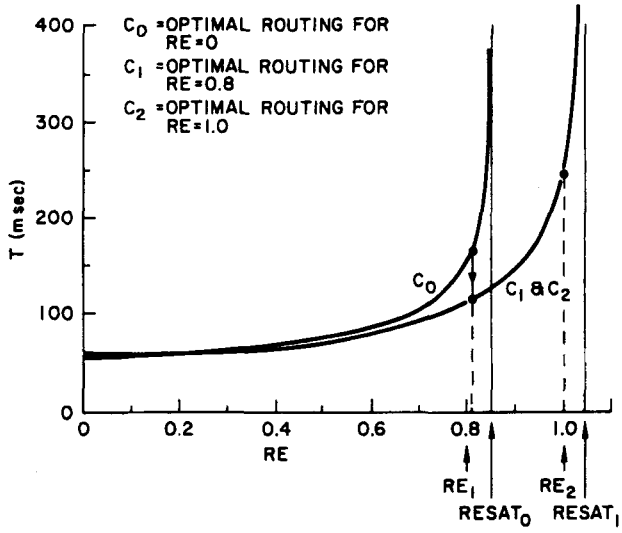


Fig. 4 Average delay T versus normalized traffic RE, using various routing schemes.

Figure 4 illustrates the application of the non-bifurcated algorithm. Recall that RE is the traffic level normalized to r = 1.187 kbits./sec. The traffic is first routed along the shortest routes computed for $RE_0 = 0$; curve C_0 plots the delay T versus RE, using such a routing scheme (which we refer to as RS_0). With RS_0, the saturation level for the traffic is $RESAT_0 = .85 < 1$; RE = 1 is infeasible, and therefore we are still in Phase 1. Let $\underset{\sim}{f}^1$ be the flow obtained by routing traffic level $RE_1 = .95\ RESAT_0 \simeq .8$, according to RS_0, and apply to $\underset{\sim}{f}^1$ the FD algorithm; a new routing scheme RS_1 is obtained, which improves $T(RE_1)$. Curve C_1, corresponding to RS_1, saturates at

*We expect to be able to reduce considerably the computation time by optimizing the code and by improving some hard working subroutines, like the shortest route and flow assignment routines [16].

$RESAT_1 = 1.05 > 1$; RE = 1 is feasible and Phase 2 is initiated, with $RE_2 = 1$. At the end of Phase 2, the sub-optimal, non-bifurcated routing scheme RS_2 is found; curve C_2 corresponding to RS_2 practically coincides with curve C_1, in Figure 4, as the scale of T is not detailed enough to show differences in values. Notice that, as expected, the routing RS_0 gives the best results at low traffic levels; in fact, RS_0 is almost optimal up to RE = 0.5.

9.2 ARPA Network: Routing and Capacities Assignment

The set of channel capacities available for the ARPA Network is discrete: Table 1 contains the list of capacity options and corresponding costs considered in the present application [6]. In order to be able to apply the FD method, the discrete cost-capacity curves have been approximated with continuous, piece-wise linear curves (see Figure 5). We do not discuss the details of the approximation, but merely mention that they must be concave.* The concavity of the cost-capacity curves implies that the local minima are shortest route flows (see Section 8). The FD method can, therefore, be applied in a form similar to the one presented in Section 8; a few modifications are required due to the non-linearity of the cost-capacity curves.

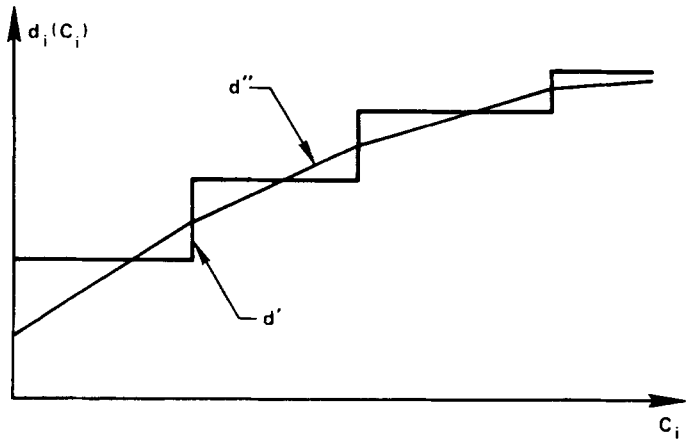
CHANNEL CAPACITIES AND CORRESPONDING COSTS USED IN THE OPTIMIZATION

Capacity [kbits/sec]	Termination Cost [$/month]	Line Cost [$/month/mile]
7.2	810	.35
19.2	850	2.10
50	850	4.20
108	2400	4.20
230.4	1300	21.00

Table 1

Note: The total cost per month of a channel is given by: total cost = termination cost + (line cost) × (length in miles).

**Other concave approximations can be considered: see [6], [18].*



d': staircase corresponding to discrete capacity levels.
d": piece-wise linear approximation

Fig. 5 Cost-capacity curves for arc i.

A schematic description of the algorithm follows here:

0. Let D_0 be the total dollar investment.
   Let $\underset{\sim}{f}^0 \in F_B$
   Let $\underset{\sim}{C}^0$ be the optimal capacities assignment for fixed $\underset{\sim}{f}^0$.*
   Let $T_0(\underset{\sim}{f})$ be as from Equation (8.1), using linear approximations of the cost-capacity curves around $\underset{\sim}{C}^0$.
   Let n = 0.

1. Let:
   $\underset{\sim}{f}^{n+1}$ = shortest route flow computed at $\underset{\sim}{f}^n$
   (using metric $\ell_k = [\partial T_n(\underset{\sim}{f})/\partial f_k]_{\underset{\sim}{f}=\underset{\sim}{f}^n}$).

2. Let $\underset{\sim}{C}^{n+1}$ be the optimal capacities assignment for fixed $\underset{\sim}{f}^{n+1}$, and let $T_{n+1}(\underset{\sim}{f})$ be as from Equation (8.1), using linear approximations of the cost-capacity curves around $\underset{\sim}{C}^{n+1}$.

3. If $\left(T_{n+1}(\underset{\sim}{f}^{n+1}) \geq T_n(\underset{\sim}{f}^n)\right)$, stop; $\underset{\sim}{f}^n$ is a local minimum.
   Otherwise, let n = n + 1 and go to 1.

*The optimal assignment of capacities, given the flows and the total dollar investment, for concave cost-capacity functions, has been discussed by Kleinrock [6].

The result of the above described algorithm is a local minimum for the continuous cost-capacity problem. In order to get a solution for the discrete problem, the capacities and flows given by the algorithm are "adjusted" in the following manner: in all arcs, the capacity is increased to the upper value of discrete capacity available (thus increasing the total investment to $D > D_0$); then, the routing is optimized once again with the FD routing algorithm.

The above described technique is clearly suboptimal. We cannot guarantee that the solutions so found are local minima; in fact, it is not even possible to *define* a local minimum in a discrete problem. Other suboptimal techniques have been proposed [7,10,21]; however, the optimization of a network with discrete capacities still remains a formidable (and basically unsolved) problem.*

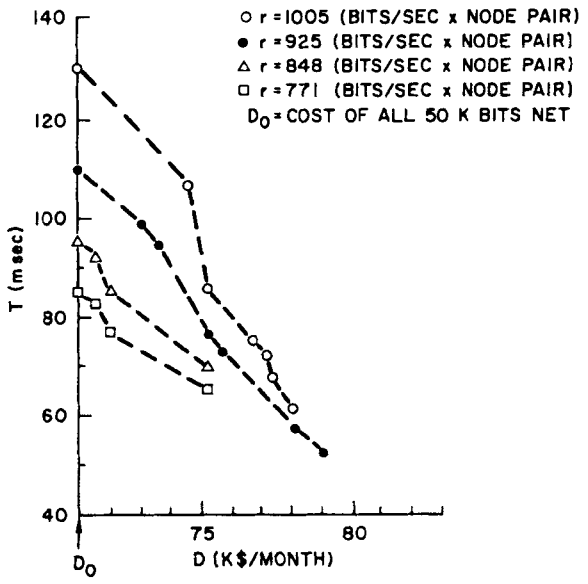


Fig. 6 Delay T versus cost D of various undominated capacity assignments for different traffic levels.

*The optimum solution can be obtained, with dynamic programming techniques, in the special case of a centralized network [30]. In fact, for such a case, the problem reduces to the optimal assignment of capacities only, as the flows are already determined by the tree-structure topology.

The technique has been applied to the design of the ARPA Network. Four cases have been run, each with a different value of uniform requirement r (see Figure 5). The initial cost D_0 was made equal to the cost of the proposed network with all 50 kbit channels ($D_0 \simeq 71{,}000$ \$/month). In order to be able to compare the 50 kbit capacities assignment to the assignments found with the FD method, the minimum delay T, with all 50 kbit capacities (i.e., with total cost $D = D_0$), was reported on the graph for each value of r (T was obtained from the curves in Figure 4). The delay T and the total cost D of the undominated* solutions are plotted in the graph of Figure 6.

10. CONCLUSION

The FD method can be applied to any *unconstrained* m.c. flow problem when some reasonable assumptions on $P(\underset{\sim}{f})$ are satisfied. It also can be applied to *constrained* flow problems: in particular to problems that include the constraints as penalties in $P(\underset{\sim}{f})$, or that have been decomposed with the Dantzig-Wolfe method. Local minima are in general attained; for convex problems, the global minimum is found.

The FD method seems to be an efficient tool for the design of S/F networks: for example, if we consider the optimal routing problem, it can be shown [19] that the amount of computation per iteration required by the FD method is comparable to that of the heuristic techniques so far proposed [16,24].† A general statement, however, about the effectiveness of the FD method as compared to other methods would not be appropriate: many factors, which depend on the specific application (like trade-off between precision and computational speed) should be considered in order to select the proper approach.

APPENDIX: CASE OF $P(\underset{\sim}{f})$ STRICTLY CONVEX

If $P(\underset{\sim}{f})$ is strictly convex, a direct proof of convergence of the FD algorithm, defined in Section 5, is available and a lower bound can be established.

A solution (T_i, D_i) is said to be dominated by (T_j, D_j) if:

$$(D_j < D_i) \text{ and } (T_j < T_i)$$

A solution is undominated if it is not dominated by any other solution.

†*The two bottlenecks, common to both approaches, are the shortest route computation and the flow assignment* [16].

Convergence

We want to show that:

$$\lim_{n\to\infty} \underset{\sim}{f}^n = \underset{\sim}{f}^* \tag{A.1}$$

where $\underset{\sim}{f}^*$ is the global minimum of $P(\underset{\sim}{f})$ on F, and $\{\underset{\sim}{f}^n\}$ is the sequence generated by recursive application of the FD operator on a given starting flow $\underset{\sim}{f}^0$. The associated sequence $\{P(\underset{\sim}{f}^n)\}$ is monotonically non-increasing and lower bounded by $P^* \triangleq P(\underset{\sim}{f}^*)$, therefore it must converge:

$$\lim_{n\to\infty} P(\underset{\sim}{f}^n) = P' \geq P^* \tag{A.2}$$

Also, recalling that:

$$P(\underset{\sim}{f}^n) - P' = \sum_{\ell=n}^{\infty} \Delta P(\underset{\sim}{f}^\ell)$$

where

$$\Delta P(\underset{\sim}{f}^\ell) \triangleq P(\underset{\sim}{f}^\ell) - P(FD \odot \underset{\sim}{f}^\ell) = P(\underset{\sim}{f}^\ell) - P(\underset{\sim}{f}^{\ell+1})$$

and recalling that:

$$\Delta P(\underset{\sim}{f}^\ell) \geq 0 \;\forall \ell$$

we have, from Equation (A.2):

$$\lim_{n\to\infty} \Delta P(\underset{\sim}{f}^n) = 0 \tag{A.3}$$

Suppose (A.1) is false; this implies, since $P(\underset{\sim}{f})$ is strictly convex, that $P' > P^*$. However, in such a case, we are able to establish a relation which contradicts Equation (A.3) as follows.

Let us first establish a lower bound on $\Delta P(\underset{\sim}{f})$. Let:

$$P(\lambda) \triangleq P[(1-\lambda)\underset{\sim}{f} + \lambda \underset{\sim}{y}], \quad 0 \leq \lambda \leq 1$$

where: $\underset{\sim}{y}$ is the shortest route flow computed at $\underset{\sim}{f}$. Using Taylor's expansion:

$$P(\lambda) = P(0) + \lambda\left[\frac{dP}{d\lambda}\right]_{\lambda=0} + \frac{1}{2}\lambda^2\left[\frac{d^2P}{d\lambda^2}\right]_{\lambda=\xi} \tag{A.4}$$

where ξ is a proper value in the interval $(0,\lambda)$ as usual. By assumption, the second partial derivatives of $P(\underset{\sim}{f})$ are upper bounded; therefore, the second directional derivative is also upper bounded, and Equation (A.4) becomes:

$$P(\lambda) - P(0) \leq \lambda\,\theta + \frac{1}{2}\lambda^2 M \tag{A.5}$$

where

$$\theta \triangleq \sum_{k=1}^{b} \ell_k(v_k - f_k) \leq 0 \tag{A.5}'$$

$$M: \text{ upperbound on } d^2P/d\lambda^2.*$$

After minimizing both sides of Equation (A.5) over λ, and recalling that $\min [P(\lambda) - P(0)] \triangleq -\Delta P(\underset{\sim}{f})$, we get:

$$\Delta P(\underset{\sim}{f}) \geq \begin{cases} \theta^2/2M & \text{if } -\theta/M < 1 \\ M/2 & \text{if } -\theta/M \geq 1 \end{cases} \tag{A.6}$$

Equation (A.6) can be rewritten as follows:

$$\Delta P(\underset{\sim}{f}) \geq \frac{M}{2} \min \left\{ \frac{\theta^2}{M^2}, 1 \right\} \tag{A.6}'$$

Inequality (A.6)' represents a useful lower bound on $\Delta P(\underset{\sim}{f})$.

Consider now:

$$P(\lambda) \triangleq P[(1-\lambda)\underset{\sim}{f}^n + \lambda \cdot \underset{\sim}{f}^*]$$

where: $0 \leq \lambda \leq 1$

$P(\lambda)$ is strictly convex, therefore it lies above its tangent line at $\lambda = 0$:

$$P(\lambda) \geq P(\underset{\sim}{f}^n) + \lambda \left[\sum_{k=1}^{b} \ell_k(f_k^* - f_k^n) \right] \tag{A.7}$$

where: $\ell_k = \left[\dfrac{\partial P}{\partial f_k} \right]_{\underset{\sim}{f}^n}$

Letting $\lambda = 1$ in (A.7) and recalling from (A.2) that $P(\underset{\sim}{f}^n) \geq P'$:

$$P(\underset{\sim}{f}^*) = P^* \geq P' + \sum_{k=1}^{b} \ell_k(f_k^* - f_k^n) \tag{A.8}$$

**Notice that $M > 0$ as $P(\lambda)$ is strictly convex.*

Let $\underset{\sim}{v}$ be the shortest route flow computed at $\underset{\sim}{f}^n$; we have, from Equation (A.8):

$$P^* \geq P' + \sum_{k=1}^{b} \ell_k (v_k - f_k^n) \tag{A.9}$$

From (A.9), using definition (A.5)', we have:

$$P' - P^* \leq |\theta| \tag{A.10}$$

Introducing (A.10) into (A.6)' we get:

$$\Delta P(\underset{\sim}{f}^n) \geq \frac{M}{2} \min \left\{ \frac{(P' - P^*)^2}{M^2}, 1 \right\} > 0 \tag{A.11}$$

The r.h.s. of Equation (A.11) is independent of n and strictly positive, therefore:

$$\lim_{n \to \infty} \Delta P(\underset{\sim}{f}^n) > 0 \tag{A.12}$$

Equation (A.12) contradicts Equation (A.3). Therefore (A.1) is true.

Lower Bound

By replacing $\underset{\sim}{f}^n$ with a generic $\underset{\sim}{f} \in F$ in (A.7), and letting $\lambda = 1$, we get, after a few steps:

$$P(\underset{\sim}{f}^*) \geq P(\underset{\sim}{f}) + \sum_{k=1}^{b} \ell_k (v_k - f_k) \tag{A.13}$$

where: $\underset{\sim}{f}^*$ is the global minimum

$\underset{\sim}{v}$ is the shortest route flow computed at $\underset{\sim}{f}$

From (A.13), lower and upper bounds on $P(\underset{\sim}{f}^*)$ are readily available:

$$P(\underset{\sim}{f}) \geq P(\underset{\sim}{f}^*) \geq P(\underset{\sim}{f}) + \sum_{k=1}^{b} \ell_k (v_k - f_k)$$

Notice that the test for optimality based on $\left| \sum_{k=1}^{b} \ell_k (v_k - f_k) \right|$ (see Section 5) is very powerful in the case of $P(\underset{\sim}{f})$ strictly convex, as it provides an upper bound on the optimal value error.

ACKNOWLEDGMENT

The authors are pleased to acknowledge Dr. D. G. Cantor for his valuable advice and his assistance in the development of the optimal routing algorithm.

REFERENCES

1. Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill Book Co., New York, 1964.

2. Steiglitz, K., P. Weiner and D. J. Kleitman, "The Design of Minimum Cost Survivable Networks," *Trans. on Circuit Theory*, November 1969, pp. 455-460.

3. Roberts, L., "Multiple Computer Networks and Intercomputer Communications," *ACM Symposium on Operating Systems Principles*, Gatlinburg, Tennessee, October 1967.

4. Roberts, L. and B. Wessler, "Computer Network Development to Achieve Resource Sharing," *AFIPS Conference Proc.*, SJCC, May 1970, pp. 543-549.

5. Heart, R., R. E. Kahn, S. M. Ornstein, W. R. Crowther and D. C. Walden, "The Interface Message Processor for the ARPA Computer Network," *AFIPS Conference Proc.*, SJCC, May 1970, pp. 551-567.

6. Kleinrock, L., "Analytic and Simulation Methods in Computer Network Design," *AFIPS Conference Proc.*, SJCC, May 1970, pp. 569-579.

7. Frank, H., I. T. Frisch and W. Chou, "Topological Considerations in the Design of the ARPA Computer Network," *AFIPS Conference Proc.*, SJCC, May 1970, pp. 581-587.

8. Carr, S., S. Crocker and V. Cerf, "HOST-to-HOST Communication Protocol in the ARPA Network," *AFIPS Conference Proc.*, SJCC, May 1970, pp. 589-597.

9. Hu, T. C., *Integer Programming and Network Flows*, Addison-Wesley, 1969.

10. Meister, B., H. R. Mueller, and H. R. Rudin, Jr., "Optimization of a New Model for Message-Switching Networks," *ICC'71*, pp. 39-16 to 39-21.

11. Dantzig, G. B., *Linear Programming and Extensions*, Princeton University Press, Princeton, New Jersey, 1962.

12. Ford, L. K. and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, New Jersey, 1962.

13. Ford, L. K. and D. R. Fulkerson, "A Suggested Computation for Maximal Multicommodity Network Flows," *Management Science*, Vol. 5, 1958, pp. 97-101.

14. Tomlin, J. A., "Minimum Cost Multi-Commodity Network Flows," *Operations Research*, Vol. 14, No. 1, January 1966, pp. 45-47.

15. Charnes, A. and W. W. Cooper, "Multicopy Traffic Network Models," *Proc. of the Symposium on the Theory of Traffic Flow*, (R. Herman, ed.), 1961.

16. Network Analysis Corporation, "Research in Store and Forward Computer Networks," *Fourth Semiannual Technical Report*, December 1971, (Contract No. DAHC 15-70-C-0120).

17. Dafermos, S. C. and F. T. Sparrow, "The Traffic Assignment Problem for a General Network," *Journal of Research of the National Bureau of Standards - B*, Vol. 73B, No. 2, April 1969.

18. Yaged, B. Jr., "Minimum Cost Routing for Static Network Models," *Networks*, Vol. 1, No. 2, 1971, pp. 139-172.

19. Cantor, D. and M. Gerla, "The Optimal Routing of Messages in a Computer Network via Mathematical Programming," *IEEE Comp. Conference*, San Francisco, September 1972.

20. Wardrop, J. G. "Some Theoretical Aspects of Road Traffic Research," *Proc. Inst. Civ. Eng. Part II*, I, 1952, pp. 325-378.

21. Gerla, M., "The Design of Store-and-Forward Networks for Computer Communications," Ph.D. Dissertation, Computer Science Department, School of Engineering and Applied Science, University of California, Los Angeles, California, January 1973.

22. Fratta, L. and M. Gerla, "The Synthesis of Computer Networks: Properties of the Optimum Solution," *ACM-International Computing Symposium*, Venice, Italy, April 1972.

23. Mangasarian, O., *Nonlinear Programming*, McGraw-Hill Book Co., New York, 1969.

24. Frank, H. and W. Chou, "Routing in Computer Networks," *Networks*, Vol. 1, No. 2, 1971, pp. 99-112.

25. Ornstein, S. M., F. E. Heart, W. R. Crowther, H. K. Rising, S. B. Russell and A. Michel, "The Terminal IMP for the ARPA Computer Network," *AFIPS Conference Proc.*, SJCC, 1972, pp. 243-254.

26. Frank H., R. E. Kahn and L. Kleinrock, "Computer Communication Network Design -- Experience with Theory and Practice," *AFIPS Conference Proc.*, SJCC, 1972, pp. 255-270.

27. Crocker, S. D., J. Heafner, J. Metcalfe and J. Postel, "Function-Oriented Protocols for the ARPA Computer Network," *AFIPS Conference Proc.*, SJCC, 1972, pp. 271-279.

28. Thomas, R. H. and D. A. Henderson, Jr., "McROSS -- A Multi-Computer Programming System," *AFIPS Conference Proc.*, SJCC, 1972, pp. 281-293.

29. Roberts, L. G., "Extension of Packet Communication Technology to a Hand-Held Personal Terminal," *AFIPS Conference Proc.*, SJCC, 1972, pp. 295-298.

30. Frank, H., I. T. Frisch, W. Chou and R. Van Slyke, "Optimal Design of Centralized Computer Networks," *Networks*, Vol. 1, No. 1, pp. 43-57.

This work was supported in part by the Advanced Research Projects Agency of the Department of Defense, Contract No. DAHC-15-69-C-0285, and in part by a 1971-72 John Simon Guggenheim Fellowship awarded to L. Kleinrock.

Paper received December 2, 1971.