Reprinted from –
AFIPS – Conference Proceedings, Volume 36

Montvale, New Jersey
07645

Topological considerations in the design of the ARPA computer network*

by H. FRANK, I. T. FRISCH, and W. CHOU

Network Analysis Corporation
Glen Cove, New York

INTRODUCTION

The ARPA Network will provide store-and-forward communication paths between a set of computer centers distributed across the continental United States. The message handling tasks at each node in the network are performed by a special purpose Interface Message Processor (IMP) located at each computer center. The centers will be interconnected through the IMPs by fully duplex telephone lines, of typically 50 kilobit/sec capacity.

When a message is ready for transmission, it will be broken up into a set of packets, each with appropriate header information. Each packet will independently make its way through the network to its destination. When a packet is transmitted between any pair of nodes, the transmitting IMP must receive a positive acknowledgement from the receiving IMP within a given interval of time. If this acknowledgement is not received, the packet will be retransmitted, either over the same or a different channel depending on the network routing doctrine being employed.

One of the design goals of the system is to achieve a response time of less than 0.2 seconds for short messages. A measure of the efficiency with which this criterion is met is the cost per bit of information transmitted through the network when the total network traffic is at the level which yields 0.2 second average time delay. The goal of the network design is to achieve the required response time with the least possible cost per bit. The final network design is subject to a number of additional constraints. It must be reliable, it must have reasonably flexible capacity in order to accommodate variations in traffic flow without significant degradation in performance, and it must be neatly expandable so that additional nodes and links can be added at later dates. The sequence and allowable variations with which the nodes are added to the network must also be taken into account. At any stage in the evolution of the network, there must be at least one communication path between any pair of nodes that have already been activated. In order to achieve a reasonable level of reliability, the network must be designed so that at least two nodes and/or links must fail before the network becomes disconnected.

To plan the orderly growth of the network, it is necessary to predict the behavior of proposed network designs. To do this, traffic flows must be projected and network routing procedures specified. The time delay analysis problem has been studied by Kleinrock[1,2] who considered several mathematical models of the ARPA Network. Kleinrock's comparison of his analysis with computer simulations indicates that network behavior can be qualitatively predicted with reasonable confidence. However, additional study in this area is needed before all the significant parameters which describe the system can be incorporated into the model. For the present, it appears that a combination of analysis and simulation can best be applied to determine a specific network's behavior.

Even if a proposed network can be accurately analyzed, the most economical networks which satisfy all of the constraints are not easily found. This is because of the enormous number of combinations of links that can be used to connect a relatively small number of nodes. It is not possible to examine even a small fraction of the possible network topologies that might lead to economical designs. In fact, the direct enumeration of all such configurations for a twenty node network is beyond the capabilities of the most powerful present day computer.

* This work was supported by the Advanced Research Projects Agency of the Department of Defense (Contract No. DAHC15-70-C-0120).

Topological considerations in the design of the ARPA computer network*

by H. FRANK, I. T. FRISCH, and W. CHOU

Network Analysis Corporation
Glen Cove, New York

INTRODUCTION

The ARPA Network will provide store-and-forward communication paths between a set of computer centers distributed across the continental United States. The message handling tasks at each node in the network are performed by a special purpose Interface Message Processor (IMP) located at each computer center. The centers will be interconnected through the IMPs by fully duplex telephone lines, of typically 50 kilobit/sec capacity.

When a message is ready for transmission, it will be broken up into a set of packets, each with appropriate header information. Each packet will independently make its way through the network to its destination. When a packet is transmitted between any pair of nodes, the transmitting IMP must receive a positive acknowledgement from the receiving IMP within a given interval of time. If this acknowledgement is not received, the packet will be retransmitted, either over the same or a different channel depending on the network routing doctrine being employed.

One of the design goals of the system is to achieve a response time of less than 0.2 seconds for short messages. A measure of the efficiency with which this criterion is met is the cost per bit of information transmitted through the network when the total network traffic is at the level which yields 0.2 second average time delay. The goal of the network design is to achieve the required response time with the least possible cost per bit. The final network design is subject to a number of additional constraints. It must be reliable, it must have reasonably flexible capacity in order to accommodate variations in traffic flow without significant degradation in performance, and it must be neatly expandable so that additional nodes and links can be added at later dates. The sequence and allowable variations with which the nodes are added to the network must also be taken into account. At any stage in the evolution of the network, there must be at least one communication path between any pair of nodes that have already been activated. In order to achieve a reasonable level of reliability, the network must be designed so that at least two nodes and/or links must fail before the network becomes disconnected.

To plan the orderly growth of the network, it is necessary to predict the behavior of proposed network designs. To do this, traffic flows must be projected and network routing procedures specified. The time delay analysis problem has been studied by Kleinrock[1,2] who considered several mathematical models of the ARPA Network. Kleinrock's comparison of his analysis with computer simulations indicates that network behavior can be qualitatively predicted with reasonable confidence. However, additional study in this area is needed before all the significant parameters which describe the system can be incorporated into the model. For the present, it appears that a combination of analysis and simulation can best be applied to determine a specific network's behavior.

Even if a proposed network can be accurately analyzed, the most economical networks which satisfy all of the constraints are not easily found. This is because of the enormous number of combinations of links that can be used to connect a relatively small number of nodes. It is not possible to examine even a small fraction of the possible network topologies that might lead to economical designs. In fact, the direct enumeration of all such configurations for a twenty node network is beyond the capabilities of the most powerful present day computer.

* This work was supported by the Advanced Research Projects Agency of the Department of Defense (Contract No. DAHC15-70-C-0120).

TOPOLOGICAL OPTIMIZATION

As part of NAC's study of computer network design, a computer program was developed to find low cost topologies which satisfy the constraints on network time delay, reliability, congestion, and other performance parameters. This program is structured to allow the network designer to rapidly investigate the tradeoffs between average time delay per message, network cost, and other factors of interest.

The inputs to the program are:

1. Existing network configuration (i.e., lines and nodes already installed and ordered)
2. Estimated traffic between nodes
3. Maximum average delay desired for short messages

In addition, the user may specify to the program a maximum cost that no network design will be allowed to exceed.

The output of the program is a sequence of low cost networks. Each network is identified by the following information:

1. Network topology
2. Cost per month
3. Maximum throughput
4. Estimated average traffic
5. Message cost per megabit at maximum throughput
6. Average message delay for short messages

Each acceptable network design also conforms to the standard that at least two nodes and/or links must fail before all communication paths between any pair of nodes are disrupted.

APPROACH

The general design problem as stated above is similar to other network design problems for which computationally practical solutions have recently been obtained. These problems include the minimum cost design of survivable networks,[3] the minimum cost selection and interconnection of Telpaks in telephone networks,[4] the design of offshore natural gas pipeline networks,[5] and the classical Traveling Salesman problem.[6] These problems have long resisted exact solution; however, recent work on approximate methods has been extremely successful and has led to efficient methods of finding low cost solutions in practical computation times.

The design philosophy

By a "feasible" solution, we mean one which satisfies all of the network constraints. By an "optimal" network, we mean the feasible network with the least possible cost. Our goal is to develop a method that can handle realistically large problems in a reasonable computation time and which can find feasible solutions with costs close to optimal.

The method to be used has two main parts called the *starting routine* and the *optimizing routine*. The starting routine generates a feasible solution. The optimizing routine then examines networks derived from this starting network by means of local transformations applied to the network topology. When a feasible network with lower cost is found, it is adopted as a new starting network and the process is continued. In this way, a feasible network is eventually reached whose cost cannot be reduced by applying additional local transformations of the type being considered. Such a network is called a *locally optimum* network.

Once a locally optimum network is found, the entire procedure is repeated by again using the starting routine. The starting routine may incorporate suggestions made by a human designer. For example, the present tentative configurations for the ARPA Network have been used. Alternatively, if desired, the starting routine may generate feasible networks without such advice. At the present time, our starting routine is capable of generating about 100,000 low cost networks.

By finding local optima from different starting networks, a variety of solutions can be generated. Figure 1 shows a diagrammatic representation of the process.

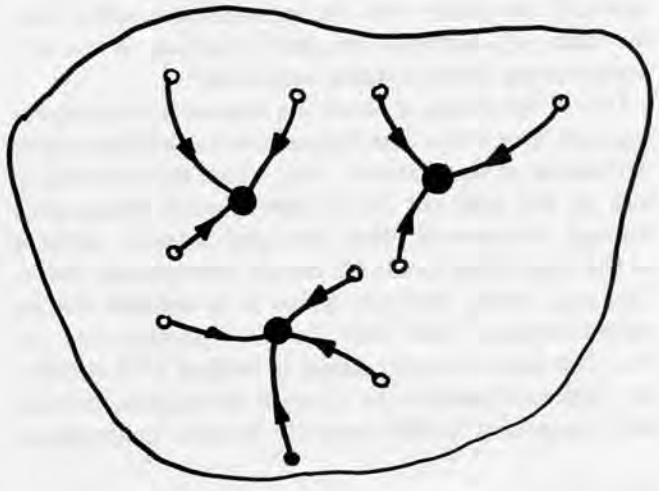

Figure 1—Diagrammatic representation of the optimization procedure

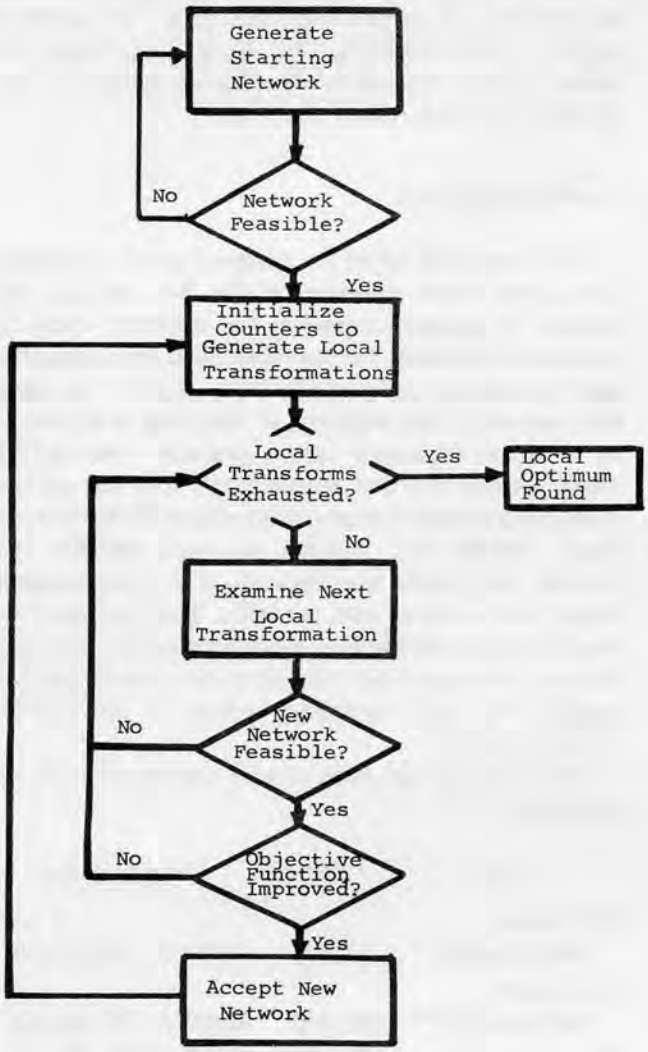

Figure 2—Block diagram of optimization procedure

The space of feasible solutions is represented by the area enclosed by the outer border of the figure; starting solutions are represented by light circles and local optima by dark circles. The practicality of the approach is based on the assumption that with a high probability some of the local optima found are close in cost to the global optimum. Naturally, this assumption is sensitive to the particular transformation used in the optimizing routine. A block diagram of the optimization procedure is shown in Figure 2.

Local transformations

A local transformation on a network is generated by identifying a set of links, removing these links, and adding a new set to the network. The method of selection of the number and location of the links to be removed and added determines the usefulness of the transformation and its applicability to the problem in hand. For example, in the problem of economically designing offshore natural gas pipeline networks, dramatic cost reductions were achieved by removing and adding one link at a time.[5] On the other hand, in a problem of the minimum cost design of survivable networks, the most useful link exchange consisted of removing and adding two links at a time.[3] In general, it is not necessary that the same number of links be added and removed during each application of the transformation.

DESIGN CONSTRAINTS

The preceding section has a given general approach for the design of low cost feasible networks. To implement this approach, a number of specific problems must be considered. These include:

1. The distribution of network traffic.
2. Network Route Selection.
3. Link capacity assignment.
4. Node and Link Time Delays.

Distribution of traffic

At the present time, it is difficult to estimate the precise magnitude and distribution of the Host-to-Host traffic. However, one design goal is that the amount of flow that can be transmitted between nodes should not significantly vary with the locations of sender and receiver. Hence, two users several thousand miles apart should receive the same service as two users several hundred miles apart. A reasonable requirement is therefore that the network be designed so that it can accommodate equal traffic between all pairs of nodes. However, it is known that certain nodes have larger traffic requirements to and from the University of Illinois' Illiac IV than to other nodes. Consequently, information of this type is incorporated into the model.

The magnitude of the network traffic is treated as variable. A "base" traffic requirement of $500 \cdot n$ bits per second (n is a positive real number) between all nodes is assumed. An additional $500 \cdot n$ bits per second is then added to and from the University of Illinois (node No. 9) and nodes 4, 5, 12, 18, 19, and 20. The base traffic is used to determine the flows in each link and the link capacities as discussed in the following sections. n is then increased until the average time delay exceeds .2 seconds. The average number of bits per second per

node at average delay equal .2 seconds is taken as a measure of performance and the corresponding cost per bit is taken as a measure of efficiency of the network.

Route selection

In order to avoid the prohibitively long computation times required to analyze dynamic routing strategies, a fixed routing procedure is used. This procedure is similar to the one which will be used in the operating network but it has the advantage that it can be readily incorporated into analysis procedures which do not depend on simulation.

The routing procedure is determined by the assumption that for each message a *path which contains the fewest number of intermediate* nodes from origin to destination is most desirable.* Given a proposed network topology and traffic matrix, routes are determined as follows: For each i ($i = 1, 2, \cdots, N = 20$):

1. With node i as an initial node, use a labelling procedure[7] to generate all paths containing the *fewest* number of intermediate nodes, to all nodes which have non-zero traffic from node i. Such paths are called *feasible paths.*
2. If node i has non-zero traffic to node j ($j = 1, 2, \cdots, N, j \neq i$) and the feasible paths from i to j contain more than seven nodes, the topology is considered infeasible.
3. Nodes are grouped as follows:
   (a) All nodes connected to node i.
   (b) All nodes connected to node i by a feasible path with one intermediate node.
   (c) All nodes connected to node i by a feasible path with two intermediate nodes.
   (d) — — — — — — — — — —
   (e) — — — — — — — — — —
   (f) All nodes connected to node i by a feasible path with five intermediate nodes.

Traffic is first routed from node i to any node j which is directly connected to i over link (i, j). Consequently, after this stage, some flows have been assigned to the network. Each node in group (b) is then considered. For any node j in this group, all feasible paths from i to j are examined, and the maximum flow thus far assigned in any link in each such path is found. All paths with the smallest maximum flow are then considered. The path whose total length is minimum is then selected and all traffic originating at i and destined for j is routed over this path.* All nodes in group (b) are treated in this matter. The same procedure is then applied to all nodes in group (c), (d), (e) and (f) in that order.

Capacity assignment

Link capacities could be assigned prior to routing. Then after route selection, if the flow in any link exceeds its assigned capacity, the network would be considered infeasible. On the other hand, link capacities may be assigned *after* all traffic is routed; we adopt this approach. The capacity of each link is chosen to be the least expensive option available from AT&T which satisfies the flow requirement. The line options which are presently being considered are: 50,000 bits/sec (bps), 108,000 bps, 230,400 bps, and 460,000 bps. Monthly link costs are the sum of a fixed terminal charge and a linear cost per mile. Thus, to satisfy a requirement of 85,000 bps, depending on the length of the link it is sometimes cheaper to use two 50,000 bps parallel links and sometimes cheaper to use a single 108,000 bps link.

The following line options and costs have been investigated:

Type	*Speed*	*Cost Per Month*
Full Group (303 data set)	50 KB	\$850 + \$4.20/mile
Full Group (304 data set)**	108 KB	\$2400 + \$4.20/mile
Telpak C	230.4 KB	\$1300 + \$21.00/mile
Telpak D	460 KB	\$1300 + \$60.00/mile

Link and node delays

Response time T is defined as the average time a message takes to make its way through the network from its origin to its destination. Short messages are considered to correspond to a single packet which may be as long as 1008 bits or as short as few bits, plus the header. If T_i is the mean delay time for a packet passing through the ith link, then

$$T = r^{-1} \sum_{i=1}^{M} y_i T_i,$$

* A node $j \neq s, t$ is called an *intermediate node* with respect to a message with origin s and destination t if the path from s to t over which the message is transmitted contains node j.

*It is also possible to divide the traffic from i to j and send it over more than one feasible path, but for uniform traffic this is not an important factor.

**Not a standard AT&T offering.

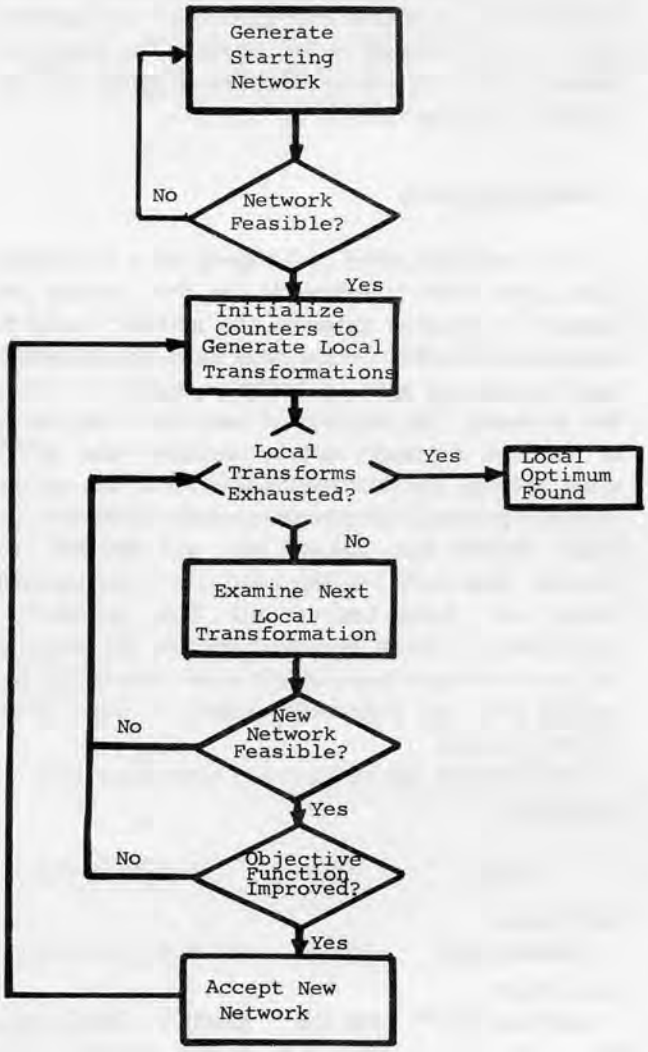

Figure 2—Block diagram of optimization procedure

The space of feasible solutions is represented by the area enclosed by the outer border of the figure; starting solutions are represented by light circles and local optima by dark circles. The practicality of the approach is based on the assumption that with a high probability some of the local optima found are close in cost to the global optimum. Naturally, this assumption is sensitive to the particular transformation used in the optimizing routine. A block diagram of the optimization procedure is shown in Figure 2.

Local transformations

A local transformation on a network is generated by identifying a set of links, removing these links, and adding a new set to the network. The method of selection of the number and location of the links to be removed and added determines the usefulness of the transformation and its applicability to the problem in hand. For example, in the problem of economically designing offshore natural gas pipeline networks, dramatic cost reductions were achieved by removing and adding one link at a time.[5] On the other hand, in a problem of the minimum cost design of survivable networks, the most useful link exchange consisted of removing and adding two links at a time.[3] In general, it is not necessary that the same number of links be added and removed during each application of the transformation.

DESIGN CONSTRAINTS

The preceding section has a given general approach for the design of low cost feasible networks. To implement this approach, a number of specific problems must be considered. These include:

1. The distribution of network traffic.
2. Network Route Selection.
3. Link capacity assignment.
4. Node and Link Time Delays.

Distribution of traffic

At the present time, it is difficult to estimate the precise magnitude and distribution of the Host-to-Host traffic. However, one design goal is that the amount of flow that can be transmitted between nodes should not significantly vary with the locations of sender and receiver. Hence, two users several thousand miles apart should receive the same service as two users several hundred miles apart. A reasonable requirement is therefore that the network be designed so that it can accommodate equal traffic between all pairs of nodes. However, it is known that certain nodes have larger traffic requirements to and from the University of Illinois' Illiac IV than to other nodes. Consequently, information of this type is incorporated into the model.

The magnitude of the network traffic is treated as variable. A "base" traffic requirement of $500 \cdot n$ bits per second (n is a positive real number) between all nodes is assumed. An additional $500 \cdot n$ bits per second is then added to and from the University of Illinois (node No. 9) and nodes 4, 5, 12, 18, 19, and 20. The base traffic is used to determine the flows in each link and the link capacities as discussed in the following sections. n is then increased until the average time delay exceeds .2 seconds. The average number of bits per second per

where r is the total IMP-to-IMP traffic rate, y_i is the average traffic rate in the ith link, and M is the total number of links. T_i can be approximated with the Pollaczak-Khinchin formula as:

$$T_i = \frac{1}{\mu C_i}\left[1 + \frac{y_i(1 + a^2)}{2(\mu C_i - y_i)}\right]$$

where $1/\mu$ is the average packet length (in bits), C_i is the capacity of the ith link (in bits/second), a is the coefficient of variance for the packet length.

These parameters are evaluated as follows:

1. r is the sum of all elements in the traffic matrix after each element has been adjusted to include headers, parity check and requests for next message (RFNM).
2. y_i is determined by the routing strategy.
3. In calculating $1/\mu$, we consider three kinds of packets: (a) packets generated by short messages and all other packets (except RFNM's) with length less than 1008 bits; (b) full length packets of 1008 bits belonging to long messages; (c) RFNM's.

It is assumed that the packets of part (a) are uniformly distributed with mean length equal to 560 bits. The packet length for part (b) is a constant equal to 1008 bits. The average packet length is then calculated by first estimating the average number of packets with 1008 bits. It is assumed that each long message consists of an average of 4 packets. In many of our computations, we assume that 80% of the messages are short. The number of RFNM packets can then be estimated. Finally, since the average length of each type of packet is known and the number of each type of packet has been estimated, the average packet length can be estimated.

4. y_i is adjusted to include the increased traffic due to acknowledgments. C_i is then selected as already described.

5. The larger the value of a, the larger the delay time. For the exponential distribution $a = 1$; for a constant, $a = 0$; and for many distributions $0 < a < 1$. Since it is reasonable to assume that the packet length distribution being considered is very close to the combination of a uniform distribution and a constant, the value of a should be less than one. To avoid underestimating T, a is set equal to one in all calculations.

The above analysis is based on the assumption that the number of available buffers is unlimited. When the traffic is low, this assumption is very accurate. For high traffic, adjustments to account for the limitation of buffer space are necessary.

There are two roles for buffers in an IMP; one for reassembling messages destined for that IMP's Host and the other for store-and-forward traffic. At the present time, about one-half of the IMP's core is used for the operating program. The remainder contains about 84 buffers each of which can store a single packet. Up to $\frac{2}{3}$ of the buffers may be used for reassembly. Buffers not used for reassembly are available for store-and-forward traffic. When no buffer is available for reassembly, any arriving packet which requires reassembly but does not belong to any message in the process of reassembly will be discarded and no acknowledgment returned to the transmitting IMP. This packet must then be retransmitted, and the effective traffic in the link is therefore increased. In addition, each time a packet is retransmitted, its delay time is not only increased by the extra waiting and transmitting time, but also by the 100 ms time-out period. To account for these factors, an upper bound on the probability that no buffer is available is calculated for each IMP. The traffic between IMPs is then increased and extra delay time for the retransmitted packets is calculated. The increase in delay time is then averaged over all the packets.

When no buffer is available for store-and-forward traffic, all incoming links become inactive. Effectively, the average usable capacities of these links is lower than their actual capacities. The probability that no buffer is available for store-and-forward traffic is set equal to the average of an upper bound and a lower bound; the upper bound is calculated by assuming that the ratio of flow to capacity of each link into the IMP is equal to the maximum ratio for all links at that node while the lower bound is found by assuming that the ratio of flow to capacity for each link is equal to the minimum such ratio. Link capacities are then reduced to include this effect and the response time is then recalculated. An example of the effect of the above assumptions is shown in Figure 4. Figure 4 relates average time delay and throughput per node for the network shown in Figure 3. Two curves are shown. One is obtained by assuming that there are an infinite number of buffers at each node. The second curve is obtained by using the actual buffer limitations of the ARPA network.

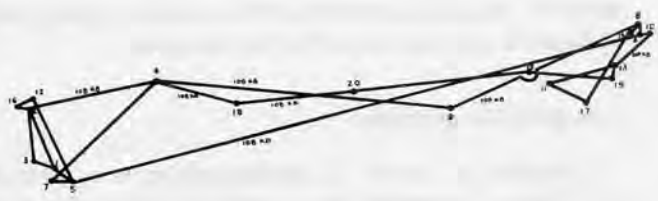

Figure 3

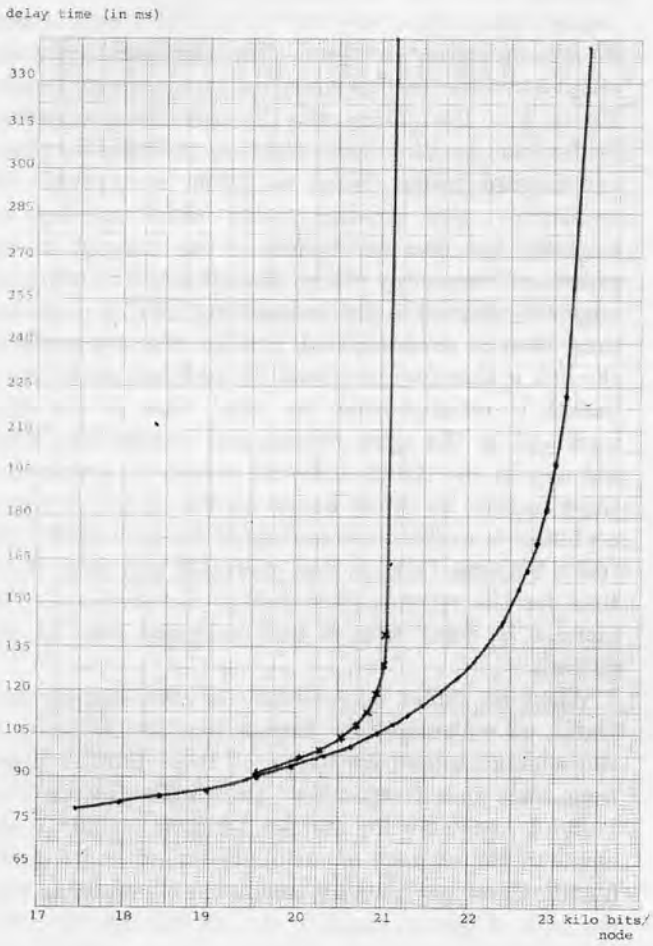

Figure 4

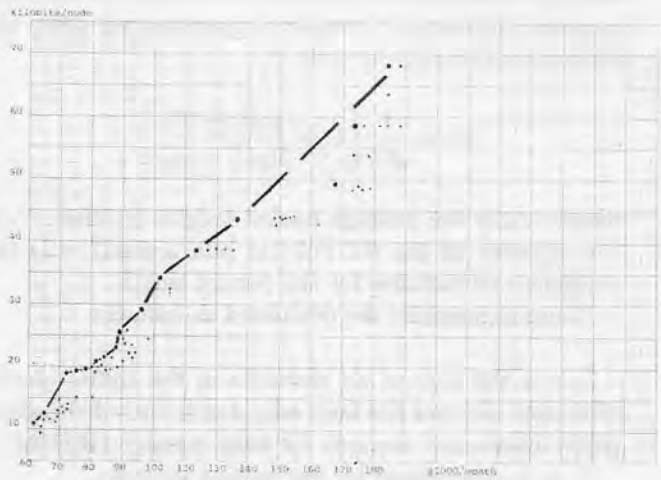

Figure 5

PRELIMINARY COMPUTATIONAL RESULTS

The optimization procedures were employed to design many thousand twenty node networks. The parameters of the best of these networks were then plotted as scatter diagrams as indicated in Figure 5. The coordinate of the horizontal axis on the graph is cost in dollars. The coordinate of the vertical axis is the average throughput per node* in bits per second for a specified distribution of traffic. The graph shown is for an average message delay of .2 seconds for short messages. Each point in the graph corresponds to a network generated, evaluated, and optimized by the computer.

Interpretation of results

Consider any point P_1 corresponding to a network N_1. Draw a horizontal line starting at P_1 to the right of P_1 and a vertical line down from P_1. Any point say P_2 which falls within the quadrant defined by the two lines is said to be *dominated by* P_1, since in a sense, network N_1 is "better than" network N_2. Similarly N_1 is said to be a dominant network. That is, for the same delay N_1 provides at least as much throughput as N_2 at no higher cost. Horizontal and vertical lines can be drawn through certain points $P_1, \cdots, P_n$ so that all other points are dominated by at least one of these. $P_1, \cdots, P_n$ thus represent, in one sense, the best networks.

One must be cautious, however, in that a network which is dominant for one time delay may not be dominant for another. Many networks with this property have been found in our studies.

Furthermore, in some cases a network may be dominated but might still be preferable to the network which dominates it because of other factors such as the order of leasing lines and plans for future growth. As an example, P_1 is a dominant point and yet there are many points which it dominates which are very close to it and might well be preferable.

Some other conclusions can be drawn from the graphs. Examining the set of dominant points it appears that there are significant savings due to economies of scale in the range of costs of \$64,000 to \$80,000. That is, small increases in cost yield large gains in throughput. Similar savings are observed in the \$90,000–\$100,000 cost range for average throughputs in the 30,000 bits/second range. These savings are due to the utilization of 108 kilobit lines which have the same line cost as 50 kilobit lines but a higher data set cost. This means that for a modest additional cost, the capacities of cross country lines can be more than doubled. To see

* throughput is the average number of bits/second *out* of each node.

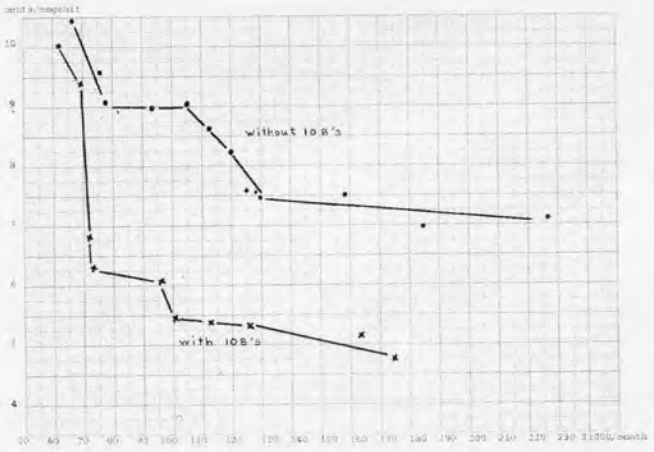

Figure 6

the effect of eliminating the 108 kilobit line option (which is not a standard AT&T offering), the cost per megabit of transmitted data is plotted against the total monthly line cost in Figure 6 for low cost networks designed with and without this option. Each point in this figure represents a feasible network. The points are connected by straight lines for visual convenience.

Additional investigations are presently under way to better understand the relationship between cost, delay and throughput, and the effect of the number of nodes on these parameters. Furthermore, alternative routing schemes will be considered as well as the cost-throughput tradeoffs that can be obtained by increasing the number of buffers at appropriate nodes.

REFERENCES

1 L KLEINROCK
Models for computer networks
Proceedings of the International Conference on Communications pp 21.9–21.16 June 1969

2 L KLEINROCK
Analytic and simulation methods in computer network design.
See paper this conference

3 K STEIGLITZ P WEINER D KLEITMAN
Design of minimum cost survivable networks
IEEE Transactions on Circuit Theory 1970

4 B ROTHFARB M GOLDSTEIN
Unpublished work

5 H FRANK B ROTHFARB D KLEITMAN K STEIGLITZ
Design of economical offshore natural gas pipeline networks
Office of Emergency Preparedness Report No R-1 Washington D C January 1969

6 S LIN
Computer solutions of the traveling salesman problem
Bell System Tech Journal Vol 44 No 10 pp 2245–2269 December 1965

7 H FRANK I T FRISCH
Communication, transmission, and transportation networks
Addison–Wesley 1971