THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 7, September, 1973
Printed in U.S.A.

Optimum Network Call-Carrying Capacity

By R. L. FRANKS and R. W. RISHEL

(Manuscript received September 8, 1972)

A telephone network with switching and trunk congestion is considered. An optimization problem expressed in terms of mean numbers of calls and mean rates of flow of calls in various categories of service throughout the network is formulated. The maximum mean number of talking calls given by this optimization problem is an upper bound on the mean number of talking calls which could be carried by the network using theoretically optimum network management. Examples are given suggesting that the upper bound is close to values which actually can be attained.

The optimum of the problem is achieved by controls which (i) restrict the number of calls coming into the network from the end offices and (ii) route appropriate fractions of the remaining calls over the various possible routes.

I. INTRODUCTION

Telephone communication facilities are designed to adequately handle peak traffic loads of an average day. In many instances the system is subject to higher loads. Classic examples of situations in which overloads occur are during holidays such as Christmas or Mother's Day, after disasters such as earthquakes or hurricanes, and during facility failures. Because of the time lag necessary to install new equipment, high overloads can also occur in normal operation in cases in which predicted traffic growth is greatly exceeded by actual traffic growth. An interesting observed phenomenon is that under certain high-load situations fewer calls may be completed than during normal load periods. Recognition of this gave rise to the subject of network management. One objective of network management is to control the handling of calls so that the maximum number of calls is put through the network.

An interesting discussion of the network management problem and early work to understand the phenomena involved in it is contained in Ref. 1. In response to the problems mentioned in Ref. 1, a simulation

study of the network management problem was carried out in Ref. 2. Simulations[2] of a modest-sized telephone network were developed and were used to evaluate the performance of various control techniques.

A large number of control techniques have been suggested for managing the network under overload conditions. Some papers which are representative of these controls are Refs. 3 through 5.

For given point-to-point calling rates, and given network management controls, let us measure the steady-state performance of the network by the expected total number of talking calls carried by the network. Define the capacity of the network (with the given calling rates) to be the maximum performance that could be obtained by any network management controls. Our objective is to set up techniques for computing the capacity of the network. The capacity of a network can be used as a benchmark in evaluating network management controls. That is, the performance of a given network management control system could be computed and compared against the capacity to tell how effective the given control system is.

The calls carried on a telephone network can be considered as a stochastic process described by a very large number of variables. There are techniques which apply to optimization of stochastic systems;[6–9] however, these would not give practical methods for computation of the optimum capacity.

For simplicity, we will consider only the steady-state situation. Our method will be to establish inequalities and equations which must be satisfied by the mean steady-state values of numbers of calls in various categories of being set up and mean rates of flow to calls into and out of these categories. From these equations and inequalities an optimization problem will be formulated. It will not be claimed that the set of inequalities and equations obtained is an exhaustive set. Hence the optimum value of the criterion of the optimization problem will only be an upper bound on the optimum value for a corresponding criterion for the real system. Later, examples will be given to show that in these cases the optimum value of the optimization problem can be nearly obtained by an appropriate choice of controls incorporated in a simulation of the network.

While the situations are quite different, our treatment of this problem follows in spirit a corresponding technique in optimal control theory, called "relaxation" of the problem.[10,11] However, the interest there is in finding an optimal control, while we are not after an optimal control, but merely want a good upper bound for the message carrying capacity of the network.

II. BACKGROUND

The factors underlying the decrease in the number of calls carried by the telephone network as it became highly overloaded were already well understood in the early work of Ref. 1. As a call is being set up, it uses equipment in one switching machine until the next switching machine on its route accepts the call and receives the destination of the call from the previous machine. If a switching machine becomes overloaded, machines adjacent to it will have to wait longer to have their calls accepted and the destinations passed on. This causes an increase in the service time for putting a call through these machines. This in turn may cause the adjacent machines to become congested. A current device to relieve this congestive phenomenon is that calls that must wait longer than a fixed time-out time for a subsequent machine are given a no-circuit announcement. However, even with time-outs, switching machine congestion can back up throughout the network.

Calls being set up occupy trunks on the partial route over which they have progressed. If a large number of calls are attempting on routes which are blocked, a portion of the capacity of certain links could be used by these ineffective attempts trying to set up. This would use capacity that could be used by talking calls. These ineffective attempts also use switching capacity in machines preceding the blockage. Most blocked attempts try again. These retrials increase the congestion.

The model which will be set up will incorporate the features mentioned above. Based on these observations, the model must take into account both trunking congestion and switching machine congestion. Since the route a call may take can be controlled and calls which are given busy signals free the entire partial route they occupied, the model should keep track of the partial routes over which calls have progressed.

Throughout the entire paper we will be interested only in the expected values of the various variables in the steady state. We will assume throughout that the processes are ergodic.

III. SWITCHING MACHINE MODEL

A block diagram of the operations of the switching machine which will be modeled is given in Fig. 1. This is a simplified model of the Bell System No. 4A-ETS switching machine. In the model a call coming into a switching machine enters a queue to wait for a vacant sender. When a call gets a sender it inpulses its destination information to the sender and releases the sender it had in the previous machine.

The call then gets into a queue for a decoder-marker combination. This decoder-marker decides on the machine the call should be routed

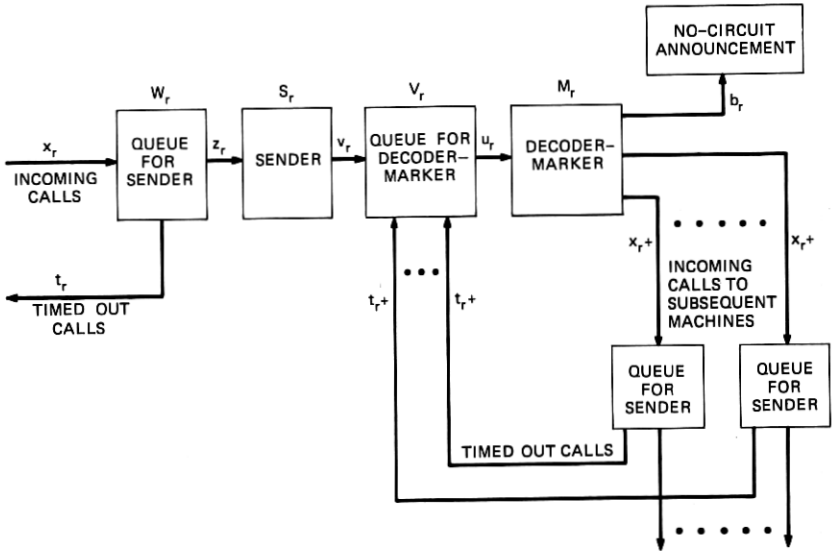

Fig. 1—Switching machine block diagram.

to next, tests for a vacant trunk, and sets up the connection, if possible. If there are no vacant trunks to appropriate subsequent machines, a no-circuit announcement is given. After a no-circuit announcement, the trunks on all the links that the call had progressed through become vacant again. If it is routed to a subsequent machine, it enters a queue for a sender in that machine.

The sender in the current machine is occupied by the call from the time it begins processing the call until the call has transmitted its destination information to the sender which processes the call in the subsequent machine. If a call waits longer than a fixed time to get a sender in the subsequent office, it is timed out. If it is timed out, the call is sent back to the marker-decoder which then connects it to a no-circuit announcement.

The process of a call being connected for service to a sender is accomplished by a sender link controller and a sender link connector. These devices test for an idle sender and an idle path to the sender. Then a path through the switch from the incoming trunk to the sender is selected and the connection established by closing appropriate switches.

The process of connecting a call for service to a marker is carried out in a similar fashion by a marker connector.

IV. COMPLETE AND PARTIAL ROUTES

A telephone network can be thought of as a collection of call switching machines connected by communication links. In the operation of the network, a talking call which hangs up frees a trunk on all the links of the route it occupied. A call which is in some state of being set up will occupy a partial route. If it is given a no-circuit announcement, it will free trunks on all the links of the partial route it occupied.

To model routes and partial routes, we will make the following definitions. Assume there are k switching machines of the network labeled by the integers $1, \cdots, k$. We will use the letter R to denote a complete route. A complete route

$$R = (i_1, \cdots, i_n)$$

is a succession of switching machines. A call occupying a complete route is always considered to be a talking call and it is understood that the call is occupying trunks on the links connecting switching machines in adjacent positions in the expression R.

A partial route describes the route occupied by a call in the process of being set up and the destination of the call. Let r designate a partial route. If

$$r = (i_1 \cdots i_{n-1}, i_n),$$

it is understood that the call has passed through machines $i_1 \cdots i_{n-2}$, is waiting to get into i_{n-1} or is in i_{n-1} being processed, and its destination is machine i_n. It occupies trunks on links connecting adjacent machines from i_1 to i_{n-1}. When a partial route r is of the above form, we shall say the partial route "terminates" at machine i_{n-1}.

The symbol r^+ will be used to designate a route subsequent to r. If $r = (i_1 \cdots i_{n-1}, i_n)$, r^+ may be the complete route

$$R = (i_1 \cdots i_{n-1}, i_n)$$

in the case in which there is a link between machine i_{n-1} and the destination i_n; or it may be a route of the type

$$r^+ = (i_1, \cdots, i_{n-1}, i_m, i_n),$$

that is, a partial route in which the call has passed to one further machine i_m on its way to its destination i_n than the call on partial route r had.

V. MODEL DESCRIPTION

In a given switching machine, calls can be distinguished by the partial route they occupy and the stage of processing they are in.

Completed talking calls can be distinguished by the complete route which they occupy. To describe the operation of the switching machines, we define the variables:

W_r = mean number of calls on partial route r waiting for a sender
S_r = mean number of calls on partial route r which are inpulsing into a sender
V_r = mean number of calls on partial route r waiting for service by a decoder-marker
M_r = mean number of calls on partial route r being serviced by a decoder-marker
N_R = mean number of talking calls on complete route R
x_r = mean rate of flow of calls on partial route r into the sender queue
z_r = mean rate of flow of calls on partial route r into the sender
v_r = mean rate of flow of calls on partial route r from the sender into the decoder-marker queue
u_r = mean rate of flow of calls on partial route r into decoder-marker
t_r = mean rate at which calls in sender queue on partial route r are timed out
b_r = mean rate at which calls on partial route r are blocked
z_R = mean rate at which calls are being completed through the network on complete route R.

We shall show that the following statements must be satisfied by these mean rates and mean numbers.

1. The mean rate at which calls flow into senders in a given machine is less than or equal to a constant times the mean number of senders which are not currently processing calls.
2. The mean rate at which calls arrive at the marker queue from the senders is equal to a constant times the mean number of senders which are processing calls which have not yet entered a marker queue.
3. The mean rate at which calls flow into the marker-decoder is less than or equal to a constant times the mean number of markers not currently processing calls.
4. The mean rate at which calls leave the markers is equal to a constant times the mean number of markers which are processing calls.

Consider statement 1. Let us define a call to be in the process of connecting to a sender from the time an idle sender is found until the connection has been completed to that sender. Defining the process this way, no time-outs occur during it. While a call is undergoing this connection process, we will say it is in the connector. We use this definition:

number of calls connecting to some sender
$\leqq$ number of free senders.

The above inequality also must hold for the mean values of both quantities.

Now the number connecting to some sender is the number in a queuing system (the connector) whose service time is the time required to make a connection. Applying Little's Theorem[12,13] gives:

E {number of calls connecting to some sender}
= (mean rate at which calls are flowing into the connector)
$\times$ (mean time to make a connection).

Since we are considering a steady-state situation, the mean rate at which calls are flowing into the connector equals the mean rate at which calls are connected to some sender. Hence we obtain statement 1. The constant in statement 1 is the reciprocal of the mean work time required to connect a call to a sender.

Consider statement 2. Calls attach to a sender, impulse their destination information, and then enter a marker queue. Applying Little's law to the number of calls inpulsing into a sender gives,

mean number of calls inpulsing into a sender
= (mean rate at which calls arrive at senders)
$\times$ (mean inpulsing time).

Since the system is in steady state, the mean rate at which calls arrive at senders equals the mean rate at which calls arrive at the marker queue. The mean number of calls inpulsing into a sender is the mean number of calls in the sender which have not yet entered the marker queue. Hence statement 2 is established.

Statements 3 and 4 are statements concerning markers similar to statements 1 and 2 concerning senders. They can be established by using Little's law in a similar manner to statements 1 and 2.

Next, statements 1 through 4 will be expressed more formally as equations and inequalities in the variables defined previously. For a

given switching machine, let

I_i = all partial routes terminating in machine i
O_i = all partial routes immediately subsequent to a partial route terminating in machine i
$\mathcal{S}_i$ = total number of senders in machine i
$\mathfrak{M}_i$ = total number of decoder-markers in machine i.

The total rate at which calls are entering senders in machine i is the sum over all the partial routes which terminate at machine i of the rates at which calls on those partial routes are entering senders. In symbols this is

$$\sum_{r \in I_i} z_r.$$

The mean number of calls occupying senders in machine i is

$$\sum_{r \in I_i} (S_r + V_r + M_r) + \sum_{r^+ \in O_i} W_r{}^+.$$

Hence, the mean number of free senders is

$$\mathcal{S}_i - \left\{ \sum_{r \in I_i} (S_r + V_r + M_r) + \sum_{r^+ \in O_i} W_r{}^+ \right\}.$$

Hence, statement 1 may be written as the inequality

$$\sum_{r \in I_i} z_r \leqq C_1 \left\{ \mathcal{S}_i - \sum_{r \in I_i} (S_r + V_r + M_r) - \sum_{r^+ \in O_i} W_r{}^+ \right\}. \tag{1}$$

A similar interpretation of statement 3 yields

$$\sum_{r \in I_i} u_r \leqq C_3 \left\{ \mathfrak{M}_i - \sum_{r \in I_i} M_r \right\}. \tag{3}$$

Statements 2 and 4 are expressed by

$$v_r = C_2 S_r, \tag{2}$$

$$u_r = C_4 M_r. \tag{4}$$

The number of calls on a given link either talking or being processed in some switching machine must be less than or equal to the number of trunks in that link. Hence, the expected number of calls of these types must be less than or equal to the number of trunks on the link. The inequalities expressing this are given by

$$\sum_{R \supset i,j} N_R + \sum_{r \supset i,j} [W_r + S_r + V_r + M_r] \leqq C_{ij}. \tag{5}$$

In this notation there is one inequality for each link connecting a machine i and a machine j, C_{ij} is the number of trunks on this link, and the notations $R \supset i,j$ and $r \supset i,j$ indicate that the sums are respectively over all complete routes or all partial routes which pass through link i,j.

In the steady state the expected rate at which calls flow into any category of processing in a switching machine must equal the rate at which they flow out. Thus the following flow-in equal flow-out equations must hold.

$$x_r = z_r + t_r \tag{6}$$

$$z_r = v_r \tag{7}$$

$$v_r + \sum_{r^+ \in 0_i} t_r^+ = u_r \tag{8}$$

$$u_r = \sum_{r^+ \in 0_i} x_r^+ + b_r. \tag{9}$$

Let A_R denote the probability that a call that completes through the network on route R will be answered by the customer. The rate at which calls are completing to talking calls on route R is $A_R z_R$. If the mean length of a talking call is $1/\nu$, calls will be hanging up at rate ν. To be in steady state, the equation

$$A_R z_R = \nu N_R \tag{10}$$

must hold.

Let λ_{ij} denote the mean rate at which calls wish to enter the network originating at machine i with destination machine j. Suppose that, if a call is placed and receives a no-circuit announcement, the customer decides to retry with probability P or to give up placing the call with probability $1 - P$. Suppose this is true independently for every call irrespective of how many times the customer may have tried previously to place the call and failed.*

Let a_{ij} denote the rate at which calls including retrials are being placed from i to j. Let r_{ij} denote the rate at which no-circuit announcements are being given and s_{ij} the rate at which calls are being completed from i to j. Then

$$a_{ij} = s_{ij} + r_{ij} = \lambda_{ij} + P r_{ij}. \tag{11}$$

* Customer retrial behavior is discussed by Wilkinson in Ref. 14. The model considered here can be considered as an idealized approximation to the more complicated behavior reported in Ref. 14.

Solving for r_{ij},

$$r_{ij} = \frac{\lambda_{ij} - s_{ij}}{1 - P}. \tag{12}$$

Now the rate at which calls are being completed between i and j is the sum over all the complete routes joining i and j of the rate at which calls are flowing onto these complete routes. Hence,

$$s_{ij} = \sum_{R=(i,\cdots,j)} A_R z_R. \tag{13}$$

Using (11), (12), and (13) gives

$$a_{ij} = \frac{\lambda_{ij} - P \sum_{R=(i,\ldots,j)} A_R z_R}{1 - P}. \tag{14}$$

Calls entering the network at machine i will be assumed to originate through an end office which leads into machine i. Since this is so, there

---

$$\sum_{r\in I_i} z_r \leqq C_1\{S_i - \sum_{r\in I_i} (S_r + V_r + M_r) - \sum_{r^+\in O_i} W_{r^+}\} \tag{2.1}$$

$$\sum_{r\in I_i} u_r \leqq C_3\{\mathfrak{M}_i - \sum_{r\in I_i} M_r\} \tag{2.2}$$

$$x_{(i,j)} = \frac{\lambda_{ij} - P \sum_{R=(i,\cdots,j)} A_R z_R}{1 - P} - b_{(i,j)} \tag{2.3}$$

$$\sum_{R\supset i,j} N_R + \sum_{r\supset i,j} (W_r + S_r + V_r + M_r) \leqq C_{ij} \tag{2.4}$$

$$v_r = C_2 S_r \tag{2.5}$$

$$u_r = C_4 M_r \tag{2.6}$$

$$x_r = z_r + t_r \tag{2.7}$$

$$z_r = v_r \tag{2.8}$$

$$A_R z_R = \nu N_R \tag{2.9}$$

$$u_r = v_r + \sum_{r^+\in O_i} t_{r^+} = \sum_{r^+\in O_i} x_{r^+} + b_r \tag{2.10}$$

Fig. 2—Equations and inequalities described in Section V.

will be a possibility that a call which wishes to enter the network at machine i with destination machine j may be blocked in the end office prior to its getting into the network. Let $b_{(i,j)}$ denote this rate at which calls from i to j are blocked in the originating end office.

If $r = (i, j)$, that is, r is the partial route of a call just starting at i whose destination is j, then the rate of flow onto r is given by

$$x_r = a_{ij} - b_{(i,j)}.$$

Since all the variables of the problem are mean numbers of calls or mean rates at which calls are flowing, all variables must be non-negative.

The equations and inequalities which have been described so far are gathered together in Fig. 2.

Rewriting the equations of Fig. 2, it can be seen that x_r, v_r, u_r, N_R, S_r, and M_r can be expressed in terms of z_r, t_r, and b_r. Eliminating these variables from the problem, we arrive at the equations expressed in Fig. 3.

$$z_r = b_r + \sum_{r^+ \in O_i} z_{r^+} \tag{3.1}$$

$$\sum_{r \in I_i} \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_4} \right) z_r + \sum_{r \in I_i} V_r + \sum_{r \in O_i} W_{r^+} + \frac{1}{C_4} \sum_{r \in I_i} \sum_{r^+ \in O_i} t_{r^+} \leqq S_i \tag{3.2}$$

$$\left(\frac{1}{C_3} + \frac{1}{C_4} \right) \left[\sum_{r \in I_i} z_r + \sum_{r \in I_i} \sum_{r^+ \in O_i} t_{r^+} \right] \leqq \mathfrak{M}_i \tag{3.3}$$

$$\sum_{R \supset i,j} \frac{A_R}{\nu} z_R + \sum_{r \supset i,j} \left(W_r + \frac{1}{C_2} z_r + V_r + \frac{1}{C_4} Z_r + \frac{1}{C_4} \sum_{r^+ \in O_i} t_{r^+} \right) \leqq C_{ij} \tag{3.4}$$

$$z_{(ij)} + t_{(ij)} + b_{(ij)} = \left[\lambda_{ij} - P \sum_{R = (i, \cdots, j)} A_R z_R \right] [1 - P]^{-1} \tag{3.5}$$

Fig. 3—Formulas of the relaxed optimization problem.

VI. AN OPTIMIZATION PROBLEM

Up to this point we have been considering the very complex stochastic process of calls progressing through a telephone network. For any such process, the constraints in Fig. 3 must be satisfied by the appropriate mean values. Next we will consider the relaxed optimization problem mentioned in Section I.

The total expected number of talking calls at a given time is

$$\sum_R N_R. \tag{15}$$

Suppose it is desired to maximize this quantity. From (10) this is equivalent to maximizing

$$\sum_R A_R z_R. \tag{16}$$

Let us first assume that it is possible to control the variables

$$z_R,\ z_r,\ V_r,\ W_r,\ b_r,\ t_r.$$

We formulate the optimization problem of maximizing (16) subject to the formulas of Fig. 3.

This is the relaxed optimization problem mentioned in Section I. Notice that while we have argued that the formulas of Fig. 3 must hold, it seems apparent that other relationships than those given in Fig. 2 will have to be satisfied for corresponding variables in the real network. However, we shall treat this relaxed optimization problem as a "model" for the network and investigate the optimum mean number of calls carried by this model. Later, for an example, the calls carried by the model and those carried by a call-by-call simulation will be compared to show that these call-carrying capacities are close. In this sense, this indicates that the model mirrors the major features of the call-carrying capacity of the network.

Notice that the variables which are to be controlled contain mean rates of flow of calls throughout the network. Implicit in this is the assumption that the routing of calls is to be chosen in the relaxed optimization problem. This will result in a routing different than conventional alternate routing. The relaxed optimization problem as formulated is a linear programming problem. If rules for alternate routing were imposed in addition to the formulas of Fig. 2, this linearity would be destroyed. For this reason the more flexible type of routing will be allowed rather than insisting on conventional alternate routing.

Notice that it might be felt that the variables to be controlled by the program are not really subject to control in a real network. It will

be shown that the optimum solution only adjusts a subset of those variables and it does appear that the variables which must be adjusted in the optimum solution are subject to adjustment in a real network.

VII. GRADE OF SERVICE CONSTRAINT

In the relaxed optimization problem that has just been formulated, no provision has been made to assure maintenance of an appropriate level of service. For instance, it is conceivable that the solution of the optimization problem would deny service completely to some point-to-point pair if the facilities on the route it used could be better utilized by other traffics. A provision should be made to insure at least a certain level of calling between each point-to-point pair.

Since the situation to be considered is one in which the network is already overloaded, it will not be possible to keep the blocking for each point-to-point pair below desired levels. However, it is possible to require that each point-to-point pair has the capability of completing through the network a fixed minimal number of calls per unit time. The total mean rate at which calls are completed through the network between point-to-point pair (i, j) is

$$\sum_{R=(i,\cdots,j)} z_R.$$

The inequalities

$$\sum_{R=(i,\cdots,j)} z_R \geqq K_{ij} \tag{17}$$

assure that calls between each pair (i, j) will be completed through the network at a mean rate greater than or equal to K_{ij}. In the future inequalities, (17) will always be added to the formulas of the relaxed optimization problem given in Fig. 3.

VIII. REDUCTION OF THE OPTIMIZATION PROBLEM

Notice that in this problem if z_R, z_r, b_r, t_r, V_r, W_r is an optimal solution, then there is another optimal solution with the same z_R and z_r, the same b_r for r not of the form $r = (ij)$, new $b(ij)$ equal to the previous $b(ij) + t(ij)$, and with $V_r = W_r = t_r = 0$. This is so since the equations and inequalities are still satisfied and the quantity (16) to be maximized is unchanged.

Notice also that eq. (3.1) in Fig. 3 implies that

$$z_r \geqq \sum_{r \supset R} z_R. \tag{18}$$

In this, the notation $r \supset R$ means that the sum is to be taken over all

complete routes which have the partial route r as their common beginning and destination.

Let

$$z_R,\ z_r,\ b_r,\ 0,\ 0,\ 0$$

be any optimal solution of the type described above. If

$$b_r' = 0, \qquad r \neq (i, j),$$

$$b_{(ij)} = \left[\lambda_{ij} - P \sum_{R=(i,\cdots,j)} A_R z_R\right][1 - P]^{-1} - \sum_{R=(i,\cdots,j)} z_R, \tag{19}$$

$$z_r' = \sum_{r \supset R} z_R. \tag{20}$$

Then

$$z_R,\ z_r',\ b_r',\ 0,\ 0,\ 0$$

is an optimal solution since (20) implies equation (3.1) in Fig. 3 is satisfied, (18) and (19) imply that the equations or inequalities (3.2) through (3.5) are satisfied, and (16) and (17) are unchanged. This implies that the optimization problem may be rewritten in terms of only the variables $b_{(i,j)}$ and z_R through using eqs. (19) and (20). The optimization problem is restated in this form in Fig. 4.

IX. CONSEQUENCES OF THE OPTIMIZATION PROBLEM

Notice that the variables of the final optimization problem are

$b_{(i,j)}$ = the fraction of calls given a no-circuit announcement in their originating office

z_R = rate at which calls are being completed through the network on complete route R

and that it has been shown that if $b_{(ij)}$ and z_R are optimal values for the final relaxed optimization problem, then

$$z_R,\ z_r = \sum_{R \supset r} z_R, \quad V_r = 0, \quad W_r = 0, \quad t_r = 0$$

$$b_r = \begin{cases} b_{(i,j)} & \text{if } r = (i, j) \\ 0 & \text{if } r \neq (i, j) \end{cases}$$

are optimal values for the original optimization problem.

This implies that the solution of the relaxed optimization problem may be taken to be of the following form: An appropriate fraction of point-to-point attempts are blocked in their originating end office. The remainder is appropriately divided among the various complete routes

Maximize

$$\sum_{R} A_R Z_R \tag{4.1}$$

subject to

$$\sum_{r \in I_i} \sum_{r \supset R} z_R$$

$$\leqq \text{Min} \left[S_i \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_4} \right)^{-1}, \mathfrak{M}_i \left(\frac{1}{C_3} + \frac{1}{C_4} \right)^{-1} \right] \tag{4.2}$$

$$\sum_{R \supset i,j} \frac{A_R}{\nu} z_R + \left(\frac{1}{C_2} + \frac{1}{C_4} \right) \sum_{r \supset i,j} \sum_{r \supset R} Z_R \leqq C_{ij} \tag{4.3}$$

$$\sum_{R=(i,\cdots,j)} (1 - P + PA_R) z_R = \lambda_{ij} - (1 - P) b_{(ij)} \tag{4.4}$$

$$\sum_{R=(i,\cdots,j)} z_R \geqq K_{ij} \tag{4.5}$$

Fig. 4—Reduction of the optimization problem.

joining the origin and the destination. No further blocking takes place. The fractions of traffic assigned to each complete route are chosen so that no queues build up in senders or markers and no blocking occurs on trunks. They also maximize the total expected number of talking calls. The fraction of each point-to-point traffic which is blocked at its source and the fractions which are routed over each complete route joining a point-to-point pair can be considered to be the optimal controls chosen by the relaxed optimization problem.

Of course, the above paragraph refers to the solution of the relaxed optimization problem and not to an optimal control for a real network. Recall that the relaxed optimization problem contained some, but not all, of the constraints of a real network. As a result, its optimal solution may violate some of those additional constraints. However, the message-carrying capacity found by the relaxed optimization problem must be at least as large as the capacity of the corresponding real network.

The optimal solution of the relaxed problem is suggestive of a good, but suboptimal, control for a real network: Code block calls in the end offices by the same amount as used in the relaxed optimization problem

and route the calls as indicated. Do not intentionally block calls at any other point. Since the real system is stochastic, this will result in some queues forming and, depending on the method of implementation, in some blocking internally in the network. Because of this, the control might be improved by choosing slightly different values of code blocking.

X. JUSTIFICATION OF THE OPTIMIZATION PROBLEM

It is natural and crucial to ask if the capacity of the model will approximate the capacity of the stochastic network. It might be thought that the optimal solution of the model is an "Alice in Wonderland" solution due to the relaxed nature of this model, especially since it chooses just the right amounts of traffic to route over each complete route so that there is never any internal blocking or any queues in the switching mathines. We will try to demonstrate that this is not so by the following argument.

It has been shown in establishing the equations and inequalities of Fig. 4 that corresponding variables for any real network must satisfy these conditions. Hence, these variables for a real network are a feasible set for the optimization problem. Thus the capacity of the model is an upper bound on the capacity of a real network. We will show in typical examples that controls similar to those of the optimization problem when incorporated in a stochastic simulation of these networks give a carried load close to this upper bound.

The first example is shown in Fig. 5. The small network shown there was subjected to severe overloads. An overload factor of one corre-

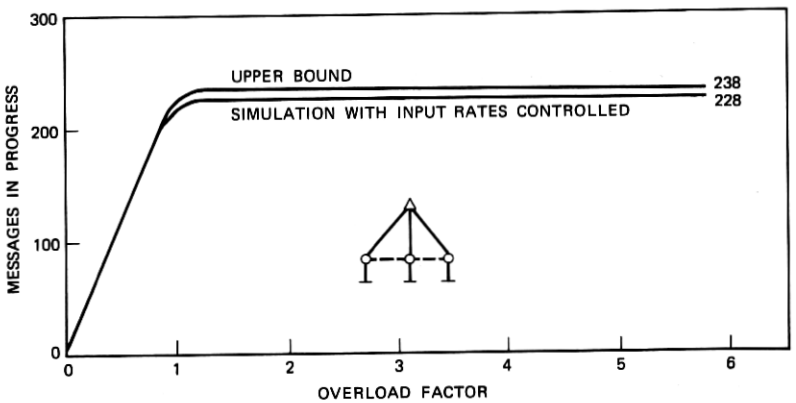

Fig. 5—Comparison of upper bound and achieved carrier loads with severe overloads.

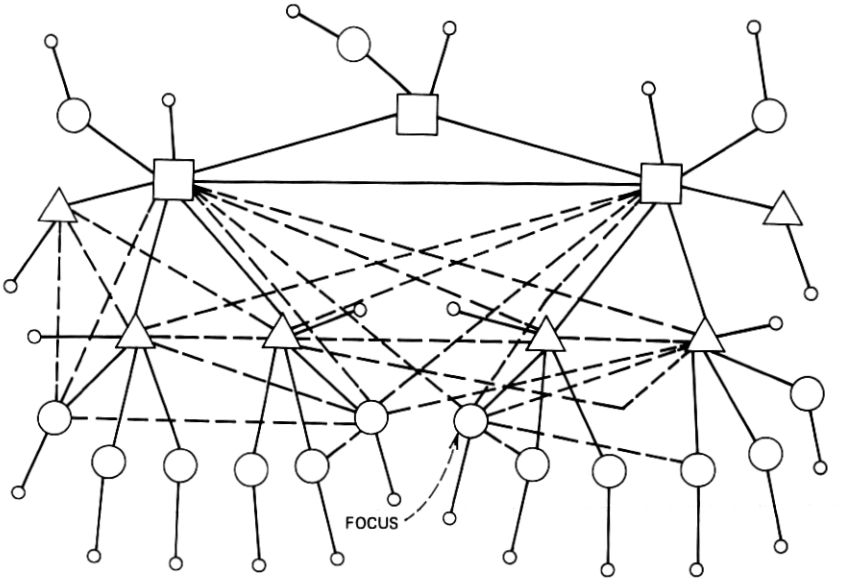

Fig. 6—Network used in focused overload example.

sponds to the traffic load for which the network was designed, an overload factor of two corresponds to twice the design load, and so forth. For all the overloads shown the upper bound on the network capacity was 238 messages. A Monte Carlo simulation of the network was run restricting the offered traffic at the end offices. The steady-state average number of messages carried was 228.

The second example was run on the considerably larger network shown in Fig. 6. This network received its design offered load except that all offered traffics destined for the node marked "Focus" were eight times their design levels and all traffics originating at that node were twice their design levels. Focused overloads of this type occur in the toll network. A similar pattern might occur following a natural disaster in the vicinity of the node marked "Focus."

Four cases were run on this network with this offered load. In all four cases only in-chain routing was used, i.e., only message paths which could exist under the current hierarchial routing scheme were allowed. The results are shown in Fig. 7. First, a Monte Carlo simulation was run using short sender timing as the only network management control. The average number of messages carried in steady state was 250. Second, the simulation was run using the network manage-

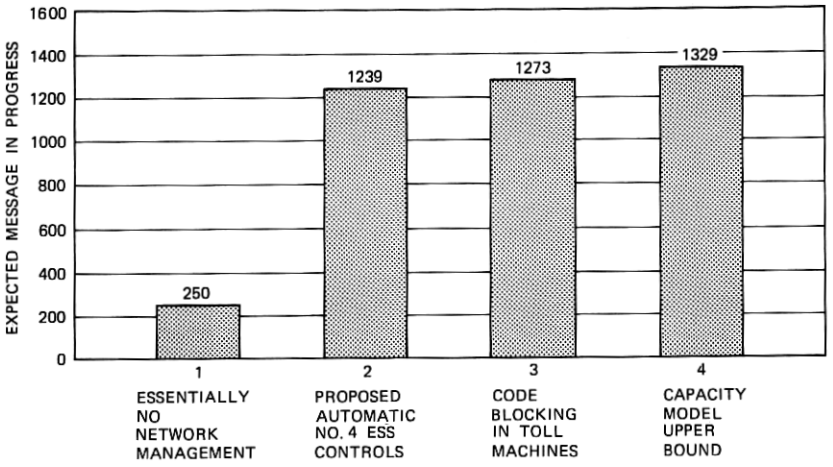

Fig. 7—Comparison of upper bound and achieved carrier loads with focused overload.

ment controls proposed for the No. 4 ESS switching machine, the next generation of Bell System toll switching machines. The steady-state average number of messages carried by the network was 1239. Third, an analytic model[15] was run with code blocking, not in the end offices but in their associated toll machines. The steady-state average number of messages carried was 1273. Finally, the upper bound on the network capacity was 1329.

In this example the proposed, economically feasible No. 4 ESS controls achieve most of the improvement between the essentially uncontrolled case and the upper bound on network capacity. The upper bound is more nearly attained by choosing code blocking in the toll centers, based on knowing the mean offered loads. Presumably the upper bound could be even more closely approached if code blocking was done in the end offices and some routing controls were included.

In summary, not only is it likely that controls based on complete knowledge of the underlying distributions can nearly achieve the upper bound on capacity, but also economically feasible control schemes can approach it.

XI. SOLUTION INFORMATION

Computer programs have been set up to solve the final linear programming problem. The program consists of two parts. The first part takes given data on the network, such as number of toll centers, which toll centers are connected by links, number of trunks on a link, and machine operating constants. Then it writes the equations given in

Fig. 4 in a form suitable for use in a linear programming algorithm. The second part uses a linear programming algorithm to solve the linear programming problem. These programs have been used to compute optimal controls for moderate-sized networks.

For a network with 17 nodes, 57 links, and 272 point-to-point offered traffics, it cost about $45 to run both programs on an IBM 360. For 21 nodes, 204 links, and 441 point-to-point offered traffics, it cost about $100 to run both programs on the same machine. By modifying the standard linear programming algorithm, taking into account the special structure of the model, it appears that it will be feasible to compute capacities for very large networks.

The output of the solution algorithm for the linear programming problem contains much additional information which can be important in evaluating the network's operation. The amount of slackness in the inequality constraints and the optimal variables for the dual linear program are printed out.

In Fig. 4, the left side of inequality (4.2) is the rate at which calls are processed in switching machine i. The left side of inequality (4.3) is the average number of calls on the trunk group connecting machines i and j. The left side of inequality (4.5) is the completion rate for calls originating at machine i and destined for machine j.

The dual variables for inequalities (4.2) and (4.3) are related to the incremental increase in the optimal carried load which could be achieved by adding one sender or marker or one trunk in the indicated place. The dual variables for inequality (4.5) give the incremental increase in the carried load which could be achieved by relaxing the minimum service constraint for a particular source-destination pair, i-j.

The amount of slackness in inequalities (4.2) and (4.3) gives information on how efficiently the toll centers and trunks are being utilized in carrying the given traffic.

XII. CONCLUSIONS

A steady-state, mean-flow-rate model of an overload telephone network with trunk and switching congestion was set up and optimized. The expected number of calls carried by this model is an upper bound for the number of calls which can be carried in the real network. The model takes into account the progression of calls along their routes as they are being set up and the amount of trunk space and switching machine capacity used by the calls which are in the process of being set up. The form of the model was such that the optimization could be carried out using conventional linear programming.

The calls carried by the optimized model can be used as a standard against which the calls carried by various network management control systems can be compared. The form of the controls of the optimized model should provide insight for network management. Computations indicate that there can be a significant difference between the number of calls carried by an unmanaged system and the optimally managed system. Some information which may be valuable in assessing the need for additional facilities in the network is available from peripheral information supplied by the linear program.

XIII. ACKNOWLEDGMENTS

The authors would like to express their appreciation for helpful discussions with Richard Ellis, Jack Holtzman, Sheldon Horing, Edwin Messerli, Patrick Spagon, Irvin Yavelberg, and Roger Wets. We also thank John Kohut and Pamela Vaughn for the use of programs they have written, and Terry Leventhal for finding the network response with No. 4 ESS controls used in Fig. 7.

REFERENCES

1. Bader, J. A., unpublished work.
2. Burke, P. J., unpublished work.
3. Burke, P. J., "Automatic Overload Controls in a Circuit-Switched Communications Network," Proc. Nat. Elec. Conf., *24* (December 1968), pp. 667–672.
4. Laude, J. A., "Local Network Management Overload Administration of Metropolitan Switched Networks," Proc. Nat. Elec. Conf., *24* (December 1968), pp. 673–678.
5. Weber, J. H., "A Simulation Study of Routing and Control in Communications Networks," B.S.T.J., *43*, No. 6 (November 1964), pp. 2639–2676.
6. Beneš, V. E., "Optimal Routing in Connecting Networks Over Finite Time Intervals," B.S.T.J., *46*, No. 10 (December 1967), pp. 2341–2352.
7. Kushner, H. J., *Introduction to Stochastic Control*, New York: Holt, Rinehart & Winston, 1971.
8. Ross, S. M., *Applied Probability Models with Optimization Applications*, New York: Holden Day, 1970.
9. Rishel, R. W., "Necessary and Sufficient Dynamic Programming Conditions for Continuous Time Stochastic Optimal Control," SIAM J. Cont., *8*, No. 4 (1970), pp. 559–572.
10. Gamkrelidze, R. D., "On Sliding Optimal States," Sov. Math.—Dokl. (1962), pp. 559–562.
11. Warga, J., "Relaxed Variational Problems," J. Math. Anal. & Applic., *4* (1962), pp. 111–128.
12. Little, J. D. C., "A Proof of the Queueing Formula: $L = \lambda W$," Oper. Res., *9*, (1961), pp. 383–387.
13. Jewell, W. S., "A Simple Proof of: $L = \lambda W$," Oper. Res., *15* (1967), pp. 1109–1116.
14. Wilkinson, R. I., "Some Characteristics of Telephoning Behavior," Advanced Tel. Traffic Eng. Conf., Michigan State University (July 23, 1970).
15. Franks, R. L., and Rishel, R. W., "Overload Model of Telephone Network Operation," to be published in November 1973 B.S.T.J.