THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

Total Network Data System:

Theoretical and Engineering Foundations

By W. S. HAYWARD* and J. P. MORELAND*

(Manuscript received July 14, 1982)

The Total Network Data System (TNDS) is a coordinated family of computer-based systems that collect and process network measurements to aid the engineers, administrators, and managers of the Bell System network in efficiently meeting service objectives. This paper describes these service objectives, the nature of telephone traffic and traffic measurements, and the theories and engineering assumptions underlying the use of these measurements in the design and administration of the trunk network and switching systems.

I. INTRODUCTION

The Total Network Data System (TNDS) is a coordinated family of computer-based systems that collect and process network measurements to aid the engineers, administrators, and managers of the Bell System network in efficiently meeting service objectives. In this paper we describe these service objectives, the traffic measurements used to monitor and design the network, and the theories underlying the use of these measurements in the various TNDS systems.

It is important to realize that, as a result of the continuing changes in the switching systems and methods of routing traffic in the Bell

* Bell Laboratories.

System network, new mathematical models and new service criteria are likely to be used as time goes by. Therefore what we are reporting here is a snapshot of what is going on today. We fully expect that there will be significant changes in the years to come.

The TNDS is described in the other articles of this issue. Our presentation of the theoretical foundation of the activities that the TNDS supports is divided into two parts, trunking and central office. We start with trunking because its simplest model is widely known and serves as an introduction.

II. TRUNKING

2.1 Service objectives

Consider the simple two-node network shown in Fig. 1. The trunk group joining end offices A and B provides the only route for calls between A and B and, as such, is called an only-route trunk group. Call attempts which arrive when all trunks are busy are said to be blocked and the customer is requested (by a reorder tone) to hang up and try the call at a later time.

The grade-of-service for an only-route group is defined to be the unweighted average blocking observed in the time-consistent busy hour of the busy season (defined below). For a given hour, the blocking is defined to be the ratio of the total number of blocked calls (overflows) to the total number of call attempts (peg count). When the hourly blockings are averaged over the time-consistent busy hour of the busy season, the resulting number is the observed grade-of-service for the group. The service objective is 0.01 average blocking.

The measure of telephone traffic used to define the time-consistent busy hour of the busy season is called the trunk-group offered load. Offered load is measured in units called erlangs, and is equal to the average number of busy trunks for a situation in which there is no blocking. Of course, in practice, offered load cannot be measured directly since calls are blocked. It can, however, be estimated from measurements of the carried load (average number of busy trunks) and blocking, as explained in Section 2.3.

When the hourly offered loads are averaged over the same hour for 20 consecutive business days, the maximum of these averages defines the time-consistent busy-hour load for the 20-day period.* The busy season is then defined to be the 20-day period for which the time-consistent busy-hour load is a maximum.

Formally, if PC_j and O_j denote, respectively, the number of call

* A different definition of busy hour is used for trunking than for central office equipment. The difference has grown up over the years because of fundamental differences between trunks and central offices, particularly with regard to forecasting capability, available measurements, and ability to respond to unforeseen shifts of traffic.

Fig. 1—Only-route trunk group.

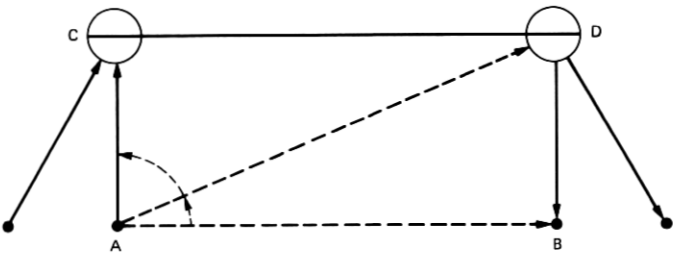

Fig. 2—Alternate-routing network.

attempts and blocked attempts during the jth time-consistent busy hour of the busy season, the observed blocking, b_j, during hour j, is defined to be

$$b_j = \begin{cases} O_j/PC_j, & \text{if } PC_j > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and the observed grade-of-service is defined to be

$$b = \frac{1}{n} \sum_{j=1}^{n} b_j, \tag{2}$$

where n is typically 20 consecutive business days. As explained in Section 6.3, the statistic b is used to detect service problems (blocking significantly greater than the 0.01 objective) and, if necessary, trigger corrective actions (e.g., trunk augments).

Figure 2 shows a more complex traffic network consisting of high-usage trunk groups (dashed lines) and final trunk groups (solid lines). Calls originating at end ofice A and destined for end office B are first offered to the primary high-usage group AB. If, however, all AB trunks are busy, the call is alternate routed to the intermediate high-usage group between A and the tandem switch D or, if all AD trunks are busy, to the final group AC. If the call arrives at D it is offered to the final group DB; if the call arrives at C, it is offered to the final group CD. Whenever a call is offered to a final group with all trunks busy, the customer receives a reorder tone or recorded announcement.

The grade-of-service for final trunk groups is defined in the same way as described above for only-route groups; the service objective is again 0.01 average blocking.

Service objectives are not, however, specified for high-usage groups. As explained in Section 2.4.1, they are sized so as to minimize the cost of the trunk network.

2.2 Traffic models

Modern trunking theory shows that three traffic parameters are required to predict busy-hour busy-season average blocking[1]: (1) average offered load during the time-consistent busy hour of the busy season, (2) peakedness (the ratio of the variance to the mean of the number of busy trunks for a trunk group so large that there is no blocking), and (3) the day-to-day variation of the busy-hour loads (variance of the daily offered loads in the time-consistent busy hour of the busy season).

2.2.1 Poisson/Erlang formulas

Prior to 1970, the Poisson formula,

$$P(c, a) = e^{-a} \sum_{n=c}^{\infty} \frac{a^n}{n!}, \tag{3}$$

where c is the number of trunks and a is the offered load in erlangs, was the standard formula for sizing trunk groups in the Bell System. It gives the blocking probability when Poisson traffic (i.e., traffic with random call arrivals) is offered to a group of c trunks and blocked calls are held. That is, calls are assumed to remain in the system—either waiting or holding a trunk if one is available—for intervals that are independent of whether they were initially blocked.

The Erlang loss formula,

$$B(c, a) = \frac{a^c}{c!} \Big/ \sum_{n=0}^{c} \frac{a^n}{n!}, \tag{4}$$

published in 1917 by A. K. Erlang of the Copenhagen Telephone Company, also assumes Poisson arrivals but assumes that blocked calls are cleared; i.e., blocked calls are assumed to leave immediately and to have no further effect on the telephone system. Note that both the Poisson and Erlang formulas are independent of the distribution of holding times.

While the Erlang formula was derived from an apparently more realistic assumption of blocked call behavior, it, in practice, underestimated the measured blocking corresponding to an average busy season load a. Accordingly, after considerable discussion during the 1920s, AT&T decided to use the Poisson formula, which predicts a higher probability of blocking than the Erlang formula. It was generally thought that blocked call behavior, while not the same as assumed for the Poisson formula, was the prime reason for the Poisson formula's superiority. Much later it was found that day-to-day load variation was the major underlying cause.

2.2.2 Wilkinson's Equivalent Random method

The assumption of Poisson (random) call arrivals accurately models most first-offered traffic, but does not adequately describe overflow traffic from high-usage trunk groups. That is, overflow traffic is more variable than Poisson traffic because it arrives in bunches; consequently, more trunks are required than the Poisson formula predicts when final groups receive overflow traffic. Accordingly, in 1956, shortly after the introduction of automatic alternate routing of customer-dialed calls, R. I. Wilkinson developed the Equivalent Random method to size final trunk groups in an alternate routing network.[2]

The basis of Wilkinson's method, as illustrated in Fig. 3, is the assumption that the superposition of the individual overflows offered to a final trunk group with c trunks can be represented by a single overflow from a (fictitious) group of s trunks with Poisson offered load a^*. The parameters a^* and s^* are chosen so that the resulting overflow has the same mean, α, and variance, v, as the actual traffic offered to the final group. With this approximation, the Erlang loss formula, with $c + s^*$ trunks and offered load a^*, can be used to estimate the blocking on the final group. That is, the Equivalent Random approximation to blocking probability, which simulation studies have shown to be remarkably accurate, is given by

$$B(c, \alpha, z) = \frac{a^*}{\alpha} B(c + s^*, a^*), \tag{5}$$

where $z = v/\alpha$ is called the peakedness of the traffic offered to the final. Formulas for computing z, a^*, and s^* are given in Ref. 1.

2.2.3 Day-to-day variation

Wilkinson showed that day-to-day load variation causes trunk group average blocking to be significantly higher than that predicted under the assumption of a constant offered load.[3] Furthermore, he showed that the distribution of measured daily offered loads is well approximated by a gamma distribution, $\Gamma(\alpha \mid \bar{a}, v_d)$, with mean $\bar{a}$ and variance v related to the mean $\bar{a}$ by

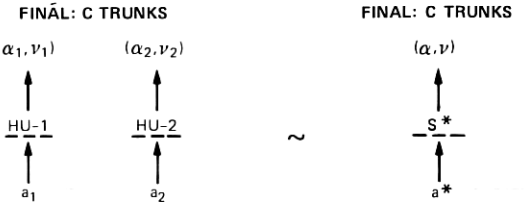

Fig. 3—Wilkinson's Equivalent Random method.

$$v_d = 0.13\bar{a}^{\phi}, \tag{6}$$

where ϕ is a parameter that describes the level of day-to-day variation. Wilkinson's studies—which showed that v_d is relatively larger for overflow traffic than for Poisson traffic—led to the use of three values of ϕ for engineering applications: $\phi = 1.5$, $\phi = 1.7$, and $\phi = 1.84$, which are referred to respectively as low, medium, and high day-to-day variation.

Using this model for day-to-day variation, Wilkinson's formula for predicting trunk group average blocking is given by

$$\bar{B}(c, \bar{a}, z) = \int_0^{\infty} B(c, \alpha, z) d\Gamma(\alpha | \bar{a}, v_d). \tag{7}$$

Since it was not practicable to apply Wilkinson's method without the use of a computer, his now famous $\bar{B}$ capacity tables were not available in the Bell System until about 1970.

2.2.4 Neal-Wilkinson tables

In 1976, Hill and Neal refined Wilkinson's $\bar{B}$ tables by developing mathematical models that account for the effects of the finite (one-hour) measurement interval.[1] That is, the service objective is defined in terms of the expected single-hour blocking, $E(O/PC)$. Thus, since eq. (7) provides an estimate of the probability of blocking, $E(O)/E(PC)$, it must be modified to remove the assumption that the measurement interval is infinite. Furthermore, Hill and Neal modified the formula for day-to-day load variation to account for the fact that part of the measured variation in daily offered loads is due to the finite measurement interval and must, therefore, be subtracted from the observed variation. Their work led to the new Neal-Wilkinson trunk group capacity tables, which are now the Bell System standard for sizing final and only-route trunk groups.

2.3 Traffic measurements and data transformations

This section describes the traffic measurements used to monitor network service, and the conversion of these measurements into estimates of the traffic parameters used to design the network.

2.3.1 Measurements

The Trunk Servicing System (TSS) derives estimates of average blocking on final groups. These estimates are used to detect service problems that trigger the demand servicing action described in Section 2.5. In addition, TSS derives estimates of trunk-group offered load, peakedness, and day-to-day variation. These estimates are used in demand servicing to correct existing service problems and in planned

servicing to forecast future trunk requirements. These estimates are derived from three traffic measurements: peg count, overflow, and usage (i.e., carried load).

In electromechanical offices, usage is measured in units called CCS (hundred call seconds) by a traffic usage recorder (TUR) which scans the trunk group every 100 seconds and increments a register for each busy trunk. Thus, the usage or carried load in erlangs (average number of busy trunks) is obtained by dividing hourly register count by 36; i.e., one erlang equals 36 CCS. In electronic switching system offices, a similar process is used to estimate usage, but there is no need for specialized equipment since traffic data are collected by the switching system's central processor.

Because of the discrete scanning (i.e., once every 100 seconds) there is, of course, some statistical sampling error associated with the usage estimate. However, our studies have shown that this error is negligible compared with the unavoidable statistical errors associated with the finite measurement interval.

2.3.2 Data transformations

The process used to derive estimates of blocking, offered load, peakedness, and day-to-day variation consists of three basic steps: data validation, computation of traffic parameters, and data substitution. Data validation is a mechanized process used to determine whether there are unusual traffic conditions (indicated by unusually high blocking due, for example, to a snowstorm) or problems in the data collection process (indicated by inconsistent measurements such as overflow exceeding peg count or usage exceeding 36 CCS per trunk). Such data must be detected and deleted; otherwise they would propagate through the trunk provisioning process and possibly cause unnecessary trunk augments to be made.

Under normal conditions, when peg count (PC_i), overflow (O_i), and usage (U_i) are available for each day in the study period (normally twenty consecutive business days), the required traffic parameters are estimated as follows:

1. The study period average blocking, b, is estimated as given by eqs. (1) and (2).
2. The study period offered load a (in erlangs) is estimated by

$$a = \frac{1}{n} \sum_{i=1}^{n} a_i, \tag{8}$$

where

$$a_i = \frac{u_i/36}{i - b_i} \tag{9}$$

is an estimate of the daily offered load.

3. The observed variance, v, of the daily offered loads is estimated by

$$v = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - a)^2. \tag{10}$$

Peakedness is not directly measured. Instead, to reduce data collection costs, a value of z is inferred from the relation

$$b = B(c, a, z), \tag{11}$$

where c is the number of trunks in the group and $B(c, a, z)$ is the equivalent random blocking formula.

Since there are many cases when complete UPCO (usage, peg count, and overflow) data are not recorded or are invalidated, TSS includes procedures to estimate traffic parameters with less than complete UPCO data. For example, if PCO is available but U is not, and if $b_i \neq 0$, an estimate of the daily offered load a_i is obtained by solving the equation

$$b_i = B(c, a_i, z), \tag{12}$$

where a typical value for z is assumed. Alternatively, if U is available, but PCO is not, the daily blocking is estimated by solving the equation

$$u_i/36 = a_i[1 - B(c, a_i, z)] \tag{13}$$

for a_i; then

$$b_i = 1 - \frac{u_i/36}{a_i}. \tag{14}$$

To minimize the impact of sampling error, TSS requires a minimum of three days per week of both peg count and overflow or three days per week of usage measurements. If less than this amount of data is available, the traffic parameters are estimated by using the most recent historical values; this process is called data substitution.

2.4 Trunk forecasting

The Trunk Forecasting System (TFS) is used by most Bell operating telephone companies to provide forecasts of interoffice message-trunk requirements for each of the next five years. The process consists of two basic steps: the estimation of future busy-season traffic loads and the design of a traffic network that minimizes the cost of the trunks required to satisfy these anticipated demands.

Since the network design process determines the traffic loads that must be forecast, we first discuss the concepts underlying the design

of minimum cost traffic networks and then describe the load forecasting process.

2.4.1 Network design

The objective of the network design process is to determine the number of trunks in each high-usage and final trunk group that minimizes the cost of the trunks provided subject to the constraint that the average blocking on final groups does not exceed 0.01 in any hour.

To illustrate the basic concepts, we first consider the simple case of engineering the network shown in Fig. 4 for a single hour. The problem is to determine the value of n, the number of trunks in the high-usage group, which minimizes the cost function

$$\text{COST} = nC + N_a C_a, \tag{15}$$

where C is the cost per trunk on the high-usage route, C_a is the cost per circuit on the alternate route (which, in reality, consists of at least two trunk groups and a tandem switch), and N_a is the number of alternate route circuits required to meet the service objective. The load offered to the high-usage group is a; the traffic offered to the alternate route, A, consists of overflow traffic from the high-usage group under consideration plus "background traffic" consisting, in general, of overflow from other high-usage groups and first-routed traffic.

Qualitatively, the trade-off is between the less expensive high-usage trunks (i.e., the direct route has no switching cost and is usually physically shorter) and the more efficient alternate route circuits (i.e., the total number of circuits, $n + Na$, is minimized by providing all circuits in a common group where they can be shared by all traffic).

Since N_a depends upon both the mean A and the peakedness Z of the traffic offered to the alternate route, we should account for the change in both A and Z as n is varied. However, the procedure is considerably simplified without significant loss in accuracy by ignoring the variation of Z with n and by assuming that the rate of change of A with respect to N_a is a constant, γ. Thus, with these approximations, the condition for minimum cost may be written as

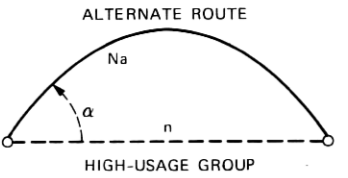

Fig. 4—ECCS engineering.

$$-\frac{dA}{dn} = \frac{\gamma}{C_R},$$

where $C_R = Ca/C$ is called the cost ratio.

Since only the overflow portion of A varies with n, and since the overflow load is $aB(n, a)$, where $B(n, a)$ is the Erlang-B blocking formula, the value of n that minimizes the COST is determined by

$$-\frac{d}{dn}[aB(n, a)] = \gamma/C_R. \tag{16}$$

The quantity of the left is approximately $a[B(n-1, a) - B(n, a)]$ and is called the load carried on the last trunk. The quantity on the right is called the economic load on the last trunk or, when the load is expressed in CCS, the ECCS; in fact, a more careful derivation, which recognizes that n is restricted to integer values, shows that the optimum value of n is the largest value for which the load on the last trunk is greater than or equal to the ECCS.

If the time-consistent busy hour for each group in the network were the same, that hour would be the only one needed in the sizing process. However, since such complete coincidence seldom occurs and since the existing algorithms are designed to use only a single-hour load, the question arises as to which hour to use to size each group.

At first glance, the solution may appear simple: since the service objective must be satisfied in all hours, final-trunk groups must be sized for their busy-hour loads, and high-usage groups should be sized for their offered loads in the alternate route's busy hour. However, there are two problems here. First, the various legs of the alternate route may have different busy hours and second, these busy-hours may depend upon the size of the subtending high-usage groups.[4] Thus, in theory, it is not possible to preselect a single hour to use in designing a minimum-cost network that satisfies the service objectives in all hours.

In practice, however, a heuristic, called the significant-hour method, has been found to produce networks that do not differ significantly from those obtained by using all hourly data. To illustrate this method, consider the two-level network shown in Fig. 2. For group AB, for example, two significant hours are considered: the A-office originating cluster busy hour (the hour for which the total load offered to the set of high-usage groups and the final originating at A is maximum) and the B-office terminating cluster busy hour (the hour for which the total load terminating at B is a maximum). Of these two hours, which are preselected in TSS from current traffic measurements, the one for which the AB load is larger is called the control hour and the group is sized using its control-hour load. A more complete discussion of this method is given in Ref. 4.

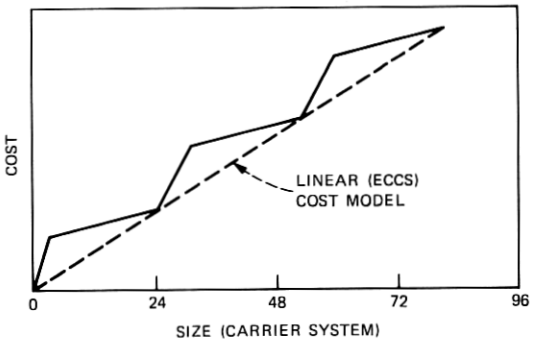

Fig. 5—Modular engineering.

The trunk quantities provided by the ECCS method must be modified to account for several factors. First, the administrative cost for a high-usage group is not included in the ECCS model. Since this cost is independent of group size, its inclusion leads to a prove-in or minimum group size; for larger groups it plays no role in the determination of the optimum size. While the exact prove-in size depends upon the specific costs, in practice we use a three-trunk minimum for local networks and a six-trunk minimum for toll networks.

Second, the ECCS model assumes a linear relationship between the cost and size of a high-usage group. However, for carrier systems, such as the T1 system, which provides the capacity for 24 time-division multiplexed channels on two pairs of wires, the cost is the step-like function illustrated in Fig. 5. Consequently, the ECCS solution is rounded to the nearest module of 24 for two-way groups or 12 for one-way groups. This rounding procedure is called modular engineering.

Finally, the ECCS method does not address the dynamic aspects of the trunk forecasting process, specifically, the time-varying nature of demands (which may call for the removal of trunks in one year only to have them added back in a future year) and forecast uncertainty. Currently, these factors are accounted for by manual adjustments to smooth the trunk requirements and avoid uneconomical disconnects and to introduce some reserve capacity to limit the amount of activity required in demand servicing.

2.4.2 Load forecasting

The objective of the load forecasting process is to estimate future busy season (control hour) first-route loads for each trunk group. First-route load, i.e., total offered load minus overflows from subtending high-usage groups, is projected since future offered load depends, of course, on the future sizes of subtending high-usage groups.

The standard load-forecasting algorithms currently in use in the

Bell System obtain estimates of future busy-season first-route loads by multiplying the most recent measurement of busy-season first-route load by an aggregate growth factor, for example, the average of the growth factors obtained by trending the total office loads at each end of the trunk group. Descriptions and comparisons of the various algorithms currently in use are given in Ref. 5.

As we explained in Ref. 5, the existing algorithms have two significant sources of error: First, on account of the finite amount of data, measured loads can have large statistical errors; standard deviations fall in the range of 5 to 10 percent for trunk-group data. Second, individual trunk-group growth factors can differ from the aggregate growth factor. These errors are significant since they lead to an increase in the reserve capacity required to satisfy anticipated customer demands.[6]

To reduce forecast error, and hence reserve trunking capacity, a new algorithm, called the Sequential Projection Algorithm (SPA), has been developed to forecast busy-season traffic loads within the Bell System.[5,7] SPA is based on a linear Kalman filter model, which establishes a linear trend for individual trunk group loads, together with logic for detecting and responding to outlier measurements. A complete description of SPA is given in Ref. 7.

2.5 Trunk demand servicing

To compensate for the effects of forecast error, demand servicing determines which trunk groups should be augmented to satisfy current busy season demands. The process uses current traffic measurements to detect service problems on finals (blocking significantly greater than objective) and to determine which high-usage and/or final groups should be augmented to correct these problems in a cost-effective manner. In theory, the demand-servicing process could also direct the removal of trunks in groups providing significantly better than objective service. In practice, however, the decision to remove trunks is, generally, part of the trunk-forecasting process described in Section 2.4.

The first step in demand servicing is to determine when action should be taken. This decision is based upon the observed average blocking b defined by eq. (2). Specifically, when b exceeds a threshold b_u, the final group is declared to be overloaded and corrective action is taken; otherwise the measured blocking is assumed to be acceptable.

Because of the statistical nature of the demands, the finite number of days in the study period, and the finite (one-hour) measurement interval each day, the measured blocking can be smaller or larger than the underlying true mean blocking.[8] For example, Fig. 6 shows a

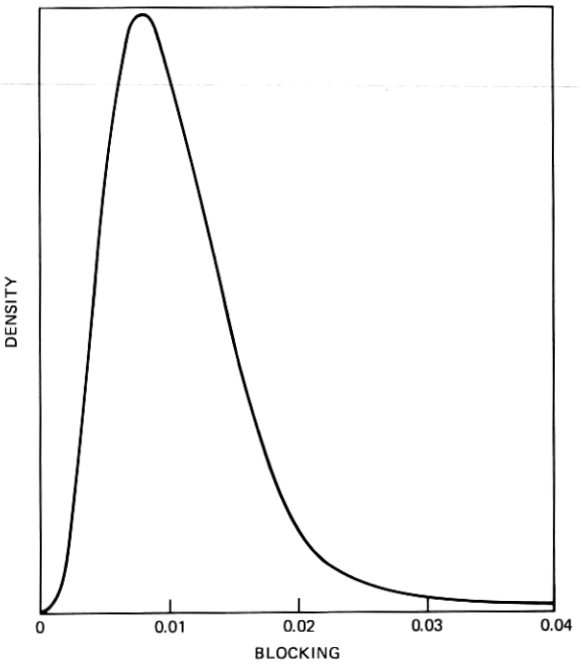

Fig. 6—Distribution (density) of measured blocking.

typical distribution of observed blocking for a correctly sized final group.

Accordingly, the threshold b has been selected to achieve a reasonable balance between two types of servicing errors: false alarms, or Type I errors, occur when b exceeds the threshold but the group is not overloaded; misses, or Type II errors, occur when a service problem is not detected; i.e., when the measured blocking falls below the threshold but the group is underprovided. The threshold u has been selected to give a false alarm probability of less than 2.5 percent; in most cases this threshold corresponds to a miss probability of less than 10 percent when the group is overloaded to at least a 0.05 blocking level.[9] For the standard 20-day study period, the threshold b_u falls in range of 0.07 (small groups) to 0.02 (large groups).

If a service problem has been detected, the next step is to determine which groups to augment. At present, this decision is based on the servicer's judgment, but a new trunk-demand servicing policy, described below, has been developed for possible inclusion in TSS.[9]

Although the simplest procedure would be to augment the final group, this procedure would tend to drive the network away from the optimal balance between high-usage and final trunks. Consequently, the demand servicing policy should attempt to resolve a service prob-

lem at its source. That is, the problem should be corrected by augmenting those groups—starting at the lowest level in the hierarchy—that are significantly underprovided (based upon the ECCS design) and are contributing to the service problem on the final group.

Since the determination of when a high-usage group is underprovided is subject to the same sources of statistical error as blocking estimates, thresholds that complement those for final groups have been developed for high-usage groups.[9] Thus, when the difference between the required number of trunks and the number of trunks in service exceeds the threshold, the group is a potential candidate for servicing. However, since only those groups contributing to the blocking problem are considered as candidates for servicing, an overloaded primary high-usage group that subtends an acceptably loaded intermediate high-usage group is not selected as a candidate.

Thus, as explained in Ref. 9, the proposed demand-servicing process is an iterative one that starts at the top of the hierarchy (the final) to identify those undertrunked groups that subtend undertrunked groups at the next higher level; such groups are selected as candidates for servicing. The lowest-level candidate, whose servicing will contribute the maximum reduction in load offered to the final, is then augmented. The blocking on the final is then recomputed. If a problem still exists, the trunk requirements and candidacy of all groups is then reevaluated. The algorithm is terminated when the blocking on the final has been reduced to an acceptable level.

III. SWITCHING SYSTEMS

3.1 Service objectives

Switching systems typically divide the process of switching calls into functional parts. Each part may require specialized equipment, such as customer receivers to receive dial pulses, or dedicated areas of memory in which to store information relative to the function, such as the number dialed. Equipment and memory of this kind are generally provided only as required to handle the traffic of each installation—a traffic engineering job requiring data from TNDS. The central processor which time-shares the control of most functions has a finite capacity, so it too must be traffic engineered. Because of the different kinds of traffic and the different function of a switching system in serving customers, the service criteria and hours of measurement differ, not only from those applied to trunks but even among switching systems according to their place in the network. This and the next section discuss service objectives and measurement timing; the remaining sections discuss application.

In setting service criteria to balance customer service against the cost of providing it, a comprehensive consideration has to be given to

what blocking or delaying calls means to a customer. In toll switching and in trunking, any particular subscriber is likely to place only a small proportion of calls through a particular trunk group or switch. If average blocking is achieved here, the customer is not likely to build up a feeling of frustration when a call is blocked; the next call is very likely to go through. In a local office, however, an overloaded switching system can affect all calls placed by the customer, regardless of destination, and blocking may remain high for a considerable time. The feeling of frustration can be much higher therefore, and service criteria must be chosen for closer control of service than can be achieved by considering averages alone. Day-to-day variation is, obviously, most important. Another important factor is balance; during the high-traffic hours, a switching system can give excellent *average* service while giving consistently poor service to a subset of customers.

A further difference applies to administration. As point-to-point loads vary with time, the components from which trunks are constructed can be reconfigured to meet changing demands. Very little of this kind of administration of equipment can be done with a switching system. Additions can normally be made only at the end of an engineering period, which is usually two years.

In addition to criteria of call blocking, which are expressed in terms of fraction of calls not completed because of congestion, delay criteria are used for engineering and administering switching systems. Delay criteria, which were not discussed in the section on trunking, are expressed in terms of the probability of exceeding an objective delay time. They require measurement methods different from those of load or loss. Ideally, delays would be measured separately for each call; however, a very busy period is a poor one for the switching system to divert its activity from call processing to service measurement. Instead, the usual delay measurement is made by placing test calls, at fixed intervals either within the system or "externally" by an attached test-call originator. In this way, a fixed amount of time is devoted to the measurement which is independent of the load on the system. The total number of test calls and the number of test calls that are delayed over an objective time are reported to the data acquisition system of TNDS.

3.2 Measurement timing

As with trunking, most traffic measurements of switching systems are based on a concept of a busy hour, that is, a single clock hour of the day such as 9:00 to 10:00 am, which is identified as the hour of the day in which, on the average, more traffic is carried than in any other. It has long been recognized that the busiest hour of a particular day may not coincide with the "busy hours" as defined above. A newly

developed engineering concept to take this into account is described at the end of this section.

As we mentioned previously, definitions involved in traffic measurement differ from those for trunks. Various "engineering hours" used in switching are described in the following sections.

3.2.1 Average busy-season busy hour

Through special studies, often made at the start of the busy season (which is itself usually well known through previous data) by means of special programs available in TNDS, the hour of the day that exhibits the highest average load, when averaged over the selected days of the study period, is identified as the busy hour. The load may be number of calls, usage, or both, depending on the type of switching system component being measured. Different components may have different busy hours. Measurements are taken in the busy hours on all days throughout the year. The traffic measurements of those three months that have the highest average busy-hour traffic are averaged to get the average busy-season busy-hour traffic. (Weekends and holidays are usually excluded from this calculation.) Provision is made within TNDS to measure loads in more than one candidate busy hour, to handle cases where different switch components have different busy hours, or when two hours are so close together in load that a clear decision between them cannot be made ahead of time.

3.2.2 Average 10-high-day busy hour

As the name implies this measurement period still uses the busy hour; however, the average usage or call volumes of the ten highest busy hours of all the months of the year are used rather than just those of the busiest three months.

3.2.3 High-day busy hour

This is the highest of the 10-high-day busy hours for a given year. Provision is made for deleting measurements made on days of a very unusual type that are not expected to recur. Service impairment on days with such unusual traffic extremes is generally accepted by the public because the cause is obvious: blizzard, flood, tornado, local or national disaster. Provision of sufficient equipment to handle such extremes without service impairment is uneconomical; network management serves under such circumstances to help the network complete as many calls as possible.

To assist the traffic engineer in evaluating the 10-high-day and high-day busy-hour values, a comparison is made by the TNDS Central Office Equipment Report (COER) system based on a theoretical model of traffic variation. Studies have shown that the probability distribu-

tion of the busy hour traffic carried by central office components can often be approximated by a gamma distribution. The gamma distribution is determined by its first two moments, so that a fitting curve can be generated from the first two moments of the observed data. COER makes such a calculation for the current year of data being collected and prints out in parallel columns the values of the observed highest 15 measurements and the values computed for a gamma distribution with the same mean and standard deviation as all of the busy-hour measurements taken to date. A quick visual check will give the engineer some idea of whether this traffic is typical. In particular, an impression can be given of whether there is more or less volatility in the traffic under study. Also, if high-day engineering is required, the difference between actual high-day load and the gamma projection will help the administrator judge whether the high-day load may have been a data outlier.

3.2.4 Extreme-value measurement period

This type of measurement period differs markedly from the three previous ones. As the costs of collecting data have changed, it has been possible to do a limited amount of processing at the collection point so that long-term storage of many measurements can be eliminated. The hour with the highest measured value for the day can be retained and all others discarded. The resulting measurement can be used to obtain reliable estimates of the extremes of traffic and service to be encountered. A different criterion of service is possible now; instead of average or high-day service, the objective can be set in terms of the expected number of hours or days in which service fails to meet a given criterion. For example, one objective now used is that on the average no more than one day a month will contain an hour in which dial tone delay exceeds 8 percent over 3 seconds. Because the criterion applies to only one day of the month, the probability of delay can be made significantly higher than the average 1.5 percent over 3 seconds used for average busy-season busy hour. In this example the criterion was picked so that in an average office there would be no noticeable change in service on changing from busy-hour engineering to extreme-value engineering.

Extreme-value methods are still being incorporated into the TNDS. Studies of the statistics of extremes indicate that extreme-value methods result in much more accurate projection of peak period loads than do current methods.[10,11] This leads to more accurate forecasting and engineering.

3.3 Types of equipment

Measurement periods and the accompanying criteria of service vary

among the three general types of central office components that are described next. Also in this section, we discuss both load (traffic intensity) and call volume. For example, the common control elements of a system are affected primarily by the number of calls processed, while the switching network is affected primarily by the average number of simultaneous conversations in progress.

3.3.1 Switching network

The capacity of the switching network portion of a switching system is almost always a function of the conversation load. Load capacity is specified in units of erlangs or CCS. (See Section 2.3.1.)

Electromechanical switching networks in general react with rather slowly rising blocking to increasing load, so that a criterion of loss averaged over the busy hours is used similar to the trunking criterion. On the other hand, time-division networks generally display very low blocking until high load is reached where blocking rises quickly.[12] In the latter case a high-day criterion is used.

The theoretical reasons for this can be seen by considering the effect of the number of parallel paths on the load-loss characteristics. Blocking in a typical three-stage network tends to follow the simple formula:[13]

$$B = [1 - (1 - a)^3]^n, \tag{17}$$

where:

B = blocking (loss)
a = average load in erlangs carried by a link connecting two stages
n = number of parallel paths.

This formula is plotted in Fig. 7 for 10, 30, and 128 parallel paths to show the effects that different blocking characteristics may have upon the selection of network service criteria. For illustration purposes, the ratios of high-day to 10-high-day to average busy season, busy hour are selected at 1.15:1.10:1.00. If the objectives are $B = 0.02$ for the average busy-season busy hour, $B = 0.05$ for the 10-high day average, and $B = 0.10$ for the high-day, three different networks might require three different engineering criteria.

The curves show the result of having selected the appropriate criterion. For the network with ten parallel paths the *ABS* criterion is appropriate; the 10-high-day and high-day service will be better than required. For a network with 30 parallel paths, however, the 10-high-day criterion is picked; the high-day and *ABS* service will be better than required. Finally, for the network with 128 parallel paths only the high-day is of concern; essentially no calls will be blocked on any other day. In fact, for this network it is likely that a service criterion will not be used directly, but that a maximum occupancy for the high-

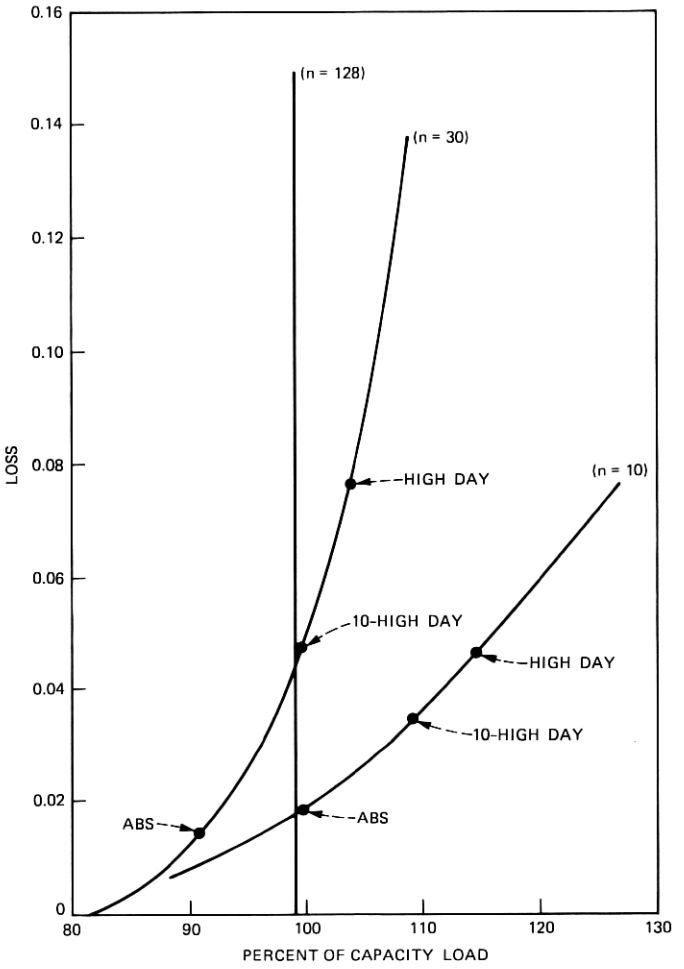

Fig. 7—Effect of parallel paths on service criterion—based on Loss = $[1 - (1 - a)^3]^n$.

day load will be adopted in order to leave room for error in predicting the high-day load.

3.3.2 Service circuits

Service circuits are the circuits that are provided in switching systems to give tones or to record customer dialing, to accept in-band signals from other offices, or to transmit such signals to other offices. Depending on the size of the office, these are provided usually in small groups ranging from five to ten, but they may occur in large groups of well over 100. Because of the way in which service circuits are used by the switching systems, service circuits delay rather than block cus-

tomers' calls during periods of congestion. The delay criterion of service used here is expressed in percent of calls delayed over a specified number of seconds. Depending on the group's size and holding time, the average, 10-high-day or high-day criterion may be used. The choice of criteria depends to a large extent on the judgment of the traffic engineer. Customer reaction, cost of equipment, stage in the progress of a call, traffic volatility, and interaction with other parts of the switching system all play a part.

The capacities of many service circuit groups are determined from delay calculations using the erlang delay formula when service times are highly variable or using the Crommelin-Pollaczek delay curves when holding times are nearly constant.

3.3.3 Central processors and controls

The control components of switching systems that are mainly used in setting up a call generally have very short holding times and are provided in small quantities, often only one per system. The probability of delay is very high, but, because of the very low holding time per call, delays are usually unnoticeable. Criteria are based on delays on the highest day, because when delays start to become noticeable the processor is usually close to saturation, where no further increase in load can be handled. In such cases if additional load is offered, customer calls will be blocked. Customers try again when blocked, so the number of attempts will rise faster than the carried load. Eventually, so much processing time is used for uncompleted attempts that the number of good calls completed by the switching system is reduced.

3.4 Traffic engineering models

The engineering of most central office equipment is based on assumptions of random call originations with constant source loads during the busy hour but with day-to-day variation. In general, peakedness is assumed at unity, although it is known that there are phenomena, such as customer retrials, that can produce some of the effects of peakedness. Also, toll offices high in the hierarchy will handle a large quantity of overflow traffic, and thus peaked traffic. The effect of this peakedness may seriously affect capacity in the periods of overload. In spite of the differences from the model underlying the Poisson formula, that is, constant traffic intensity and random call arrivals, the formula has been the work horse for much of central office engineering because it was the best choice in predicting actual system performance.

The Poisson formula is not applicable to internal switching networks. Formulae of the general form of eq. (17) give remarkably close approximation to the performance of this type of network. Observa-

tions show that a particular network will display a load-loss relationship that parallels such a theoretical curve. Differences may be attributed not only to the simplified assumptions of the formula but also to imbalance effects, day-to-day variation, systematic load variation within the hour, or minor variations in the network structure from that assumed in the model. The TNDS supplies data from which comparisons of actual with theoretical load-loss relationships can readily be made.

Common control equipment, including call processors, present a different traffic modeling problem from what has been considered so far. Loss or delays are insignificant most of the time. The event of call processor overload is so infrequent that measurements are usually too few for making useful load-loss curves to compare with theory.

With electromechanical switching systems, the proportion of time that a processor, such as a marker, is busy can be measured with little complication. A plot of occupancy against calls gives a clear picture of the number of calls that will fully occupy the processor. Capacity is usually picked at a lower number, say 95 percent of full occupancy, and applied to high-day busy-hour traffic.

A stored program controlled processor, however, is busy all of the time; when it is not processing calls, it is performing maintenance or audit functions or looking for work. The control program is designed to reduce the time spent in non-call processing activities as the call processing load increases. In many systems, this is done by organizing the processing into a fixed order of tasks. Processing jobs that must be done with little delay are given priority over those with less stringent time requirements. So that some maintenance and audit work is done even in periods of heavy load, the search for work is made in a cyclical fashion over the various priorities with many repeat searches for high-priority work. Thus the time taken between visits to the lowest priority level becomes longer with increased processing load. The capacity of the processor, in terms of high-day busy-hour calls, is then found by analyzing observations of offered originating-plus-incoming calls with the corresponding observations of basic processor cycles, i.e., visits to the lowest priority level. A first-order least-squares fit to these points is made and a 90-percent confidence line is computed and drawn below the fitting line (90 percent of the observations are expected to fall above this line). Next a horizontal line is drawn at the minimum number of visits that allow acceptable service as determined by simulations and field studies of many systems. The intercept of the 90-percent line and the minimum-number-of-visits line gives the estimated call capacity for the high-day busy hour. This is illustrated in Fig. 8.

The choice of 90-percent confidence is, of course, a judgment choice.

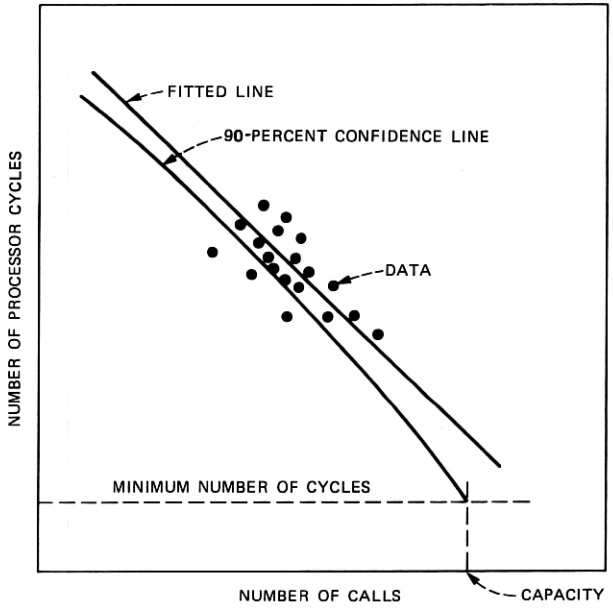

Fig. 8—Use of measurements to determine central processor capacity.

It is based on experience gained from many switching systems. A load greater than capacity will often cause no problem; on the other hand, a load less than capacity may rarely overload the processor.

3.5 Load balance

The switching networks of most switching systems require balancing of offered traffic over the network switches and frames. Unbalance arises from the nature of traffic sources—some lines or trunks generate more traffic than others. Correction of unbalance is particularly needed in the concentrating stages of the network, where the number of input lines exceeds the number of output links. While the first objective of load balancing is to equalize service among customers, it also achieves a second objective of improving the load-carrying capacity of the network. The latter effect of load balancing arises from the nonlinear load-loss curve that exists for most networks. As Fig. 9 illustrates, the higher loss in a section of network loaded above average is not completely compensated for by the lower loss in a section loaded, by the same amount, below average. Therefore, reducing the spread of the distributions of loads around the mean will reduce average blocking at a given load or enable increased loading at a given service objective.

It is an objective of line-load balancing to control the assignment of

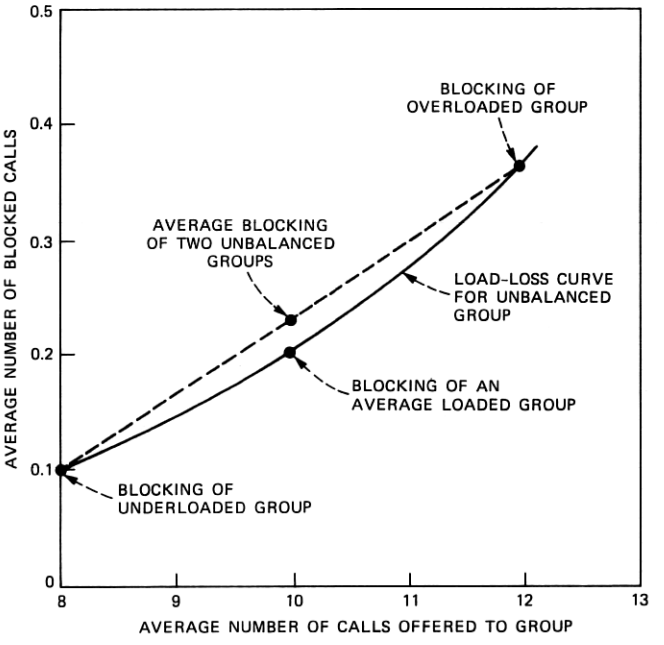

Fig. 9—Effect of unbalance on blocking.

new lines in such a fashion that all units stay in traffic balance. In a 55,000-line 1A *ESS** switching equipment for example, this implies keeping track of a number of separately tracked "loading modules" that may approach a thousand. Even in smaller offices the problem is far from trivial. The measurement and processing systems of TNDS provide the means of controlling load balance—without excessive manual effort—by accumulating line-group usage for all of the designated hours of the week. Also, measurements of individual line usage are taken for guidance in selecting lines when extreme unbalance requires shifting lines from one particular group to another.

The first step in balancing concentrator loads is to assign lines in such a fashion that every concentrator has the same number of lines of each identifiable traffic or class of service. Unfortunately, this is not enough to ensure balance.

The first systematic application of traffic measurements to load balance was reported in a study by D. H. Barnes in 1958.[14] Barnes proposed a method of statistical control that would separate effects of chance load variations from those of systematic load balances. It was implemented on the computers of that day as well as in a manual

* Trademark of Western Electric.

procedure (the "score method") that would permit the calculations to be made with simple mathematical operations.

The type of solution proposed by Barnes has been refined over the years as more computing power has become available and more measurement refinement has become possible. The operational problem to be solved remains the same however: to separate the random variations from the systematic ones out of the large variations that occur in the hourly measured traffic on a concentrated group of lines. This is the same problem faced in the design of a statistical quality control procedure. It involves estimating load variance and measurement error, and displaying the results in a form that will enable managers to take appropriate action without moving lines needlessly or letting service deteriorate. The TNDS Load Balance system takes over the computation of balance indices and processing of the balance data.

IV. NETWORK MANAGEMENT

The background of network management is covered in a companion article to this one and in Ref. 15. The main difference in the requirements placed by network management on TNDS from those of other traffic data applications is the very fast response required between traffic events and the reporting of them. Because of the long-established need to keep network managers informed within seconds of a major control loss or within minutes of serious call completion problems, the traffic-gathering parts of TNDS must have—and do have—the capability of passing data quickly and accurately to the network management computer and relaying control orders quickly back to the measured switching systems.

V. CONCLUSION

Throughout the evolution of TNDS there has been a continuous increase in the accuracy and availability of traffic data. This in turn has made possible more accurate traffic models and has spurred activity for better traffic theory and better network operating practices. What has been reported here is the state of the art at this time. The coming years may be expected to bring more changes and improvements.

REFERENCES

1. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability-Engineered Trunk Group," B.S.T.J., *55*, No. 7 (September 1976), pp. 831–2.
2. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
3. R. I. Wilkinson, "Some Comparisons of Load and Loss Data with Current Teletraffic Theory," B.S.T.J., *50*, No. 9 (October 1971), pp. 2807–34.

4. M. Eisenberg, "Engineering Traffic Networks for More Than One Busy Hour," B.S.T.J., *56,* No. 1 (January 1977), pp. 1–15.
5. A. J. David and C. D. Pack, "The Sequential Projection Algorithm: A New and Improved Traffic Forecasting Procedure," 9th International Teletraffic Congress, Torremolinos, Spain, 1979.
6. R. L. Franks et al., "A Model Relating Measurement and Forecast Errors to the Provisioning of Direct Final Trunk Groups," B.S.T.J., *58,* No. 2 (February 1979), pp. 351–78.
7. J. P. Moreland, private communication.
8. S. R. Neal, "Blocking Distributions for Trunk Network Administration," B.S.T.J., *59,* No. 6 (July 1980), pp. 829–44.
9. C. R. Szelag, "Trunk Demand Servicing in the Presence of Measurement Uncertainty," B.S.T.J., *59,* No. 6 (July 1980), pp. 845–60.
10. D. H. Barnes, "Extreme Value Engineering of Small Switching Offices," 8th International Teletraffic Congress, Melbourne, Australia, 1976.
11. K. A. Friedman, "Extreme Value Analysis Techniques," 9th International Teletraffic Congress, Torremolinos, Spain, 1979.
12. J. F. Huttenhoff et al., "No. 4 ESS: Peripheral System," B.S.T.J., *56,* No. 7 (September 1977), pp. 1029–55.
13. C. Y. Lee, "Analysis of Switching Networks," B.S.T.J., *34,* No. 6 (November 1955), pp. 1287–1315.
14. D. H. Barnes, "Statistical Methods for the Administration of Dial Offices," 2nd International Teletraffic Congress, The Hague, 1958.
15. D. G. Haenschke, D. A. Kettler, and E. Oberer, "Network Management and Congestion in the U. S. Telecommunications Network," IEEE Trans. Commun., *COM-29,* No. 4 (April 1981), pp. 376–85.

AUTHORS

Walter S. Hayward, Jr., A.B., 1943, S.M. (Electrical Engineering), 1947, Harvard University; Bell Laboratories, 1947—. Mr. Hayward has worked in the field of telephone traffic and switching systems engineering. In 1961, he was appointed Head, Electronic Switching Studies Department. In 1964, he was appointed Director of the Traffic Studies Center. He is now Consultant, SPC Studies Center. Member, IEEE, ORSA, and ACM.

James P. Moreland, B.S.E.E., 1964; M.S.E.E., 1964; Ph.D. (E.E.), 1967, Ohio State University; Research Associate, Electroscience Laboratory, 1964–1968, Instructor, Electrical Engineering, 1967–1968, Ohio State University; Bell Laboratories, 1968—. At Ohio State, Mr. Moreland worked on studies of scattering theory and optical heterodyne detection. At Bell Laboratories, he has been concerned with clock-synchronization schemes for digital communications networks, optical-fiber transmission studies, and traffic and facility network planning. He is presently a Supervisor in the Trunk Traffic Engineering Department. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.