THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 | January 1978 | Number 1

A Study of Network Performance and Customer Behavior During Direct-Distance-Dialing Call Attempts in the U.S.A.

By F. P. DUFFY and R. A. MERCER

(Manuscript received April 22, 1977)

A survey was conducted throughout the Bell System in October 1974 to gather detailed information about Direct-Distance-Dialing call attempts. The dispositions, setup times, and customer abandonment times associated with DDD attempts are discussed in detail in this article to provide network performance and customer behavior characteristics to network planners and administrators and to designers of equipment and systems which use, and interact with, the telephone network. It is shown that both network performance and customer behavior affect the call dispositions and the total call setup time; however, customer-dependent failures to complete account for 85 percent of all failures, and customer-determined components of the call setup time make up 71 percent of the total setup time. It is found that traffic composition in terms of the relative mix of business and residential originations exerts a strong influence on call dispositions. Network performance affects the probability of equipment blockages and failures and the interval from end of dialing to receipt of a network response. These are both found to depend on calling distance, while the latter is also affected by the types of originating and terminating local switching.

I. INTRODUCTION

A complex sequence of interactions and reactions is initiated each time a person or machine attempts to call another person or machine via the switched public telephone network. In the case of local calls, the setup process involves station equipment, subscriber loops, at least one local switching office (end office) with its multitude of equipment, and perhaps interoffice trunks and local tandem offices with possible local alternate routing capabilities. Several local switching arrangements are illustrated in Fig. 1a and b. In the case of long distance (toll) calls, the switching arrangements are more complex because of a five-level switching hierarchy, and they are more flexible because of the extensive use of alternate routing.[1] A standard toll switching arrangement is illustrated in Fig. 1c. At one extreme of this switching arrangement, a toll call may encounter ten switching offices interconnected by seven final intertoll trunks and two toll connecting trunks; at the other extreme, a toll call may encounter two end offices connected by a single direct intertoll trunk. In between these extremes, a toll call may be established through several switching offices interconnected by a combination of toll connecting, final intertoll, and high-usage intertoll trunks.

In general, a toll call may be established in several different ways between the same two originating and terminating stations. In Fig. 1c, there are four separate routes which may be traversed. The toll switching algorithm establishes calls on a "link-by-link" basis. At a given office in the switching hierarchy, precedence is given to the trunk group which provides the most direct route to the terminating end office. If the trunk group associated with that route is busy, an alternate route is sought by assigning precedence to the next most direct trunk group. This process is repeated at each consecutive office in the switching path until a route is established or until the setup process cannot proceed due to a system blockage.

A telephone customer making a call attempt has no perception of this process as it occurs, but is aware of the results: namely, the "system setup time" required for the network to provide an identifiable response after completion of dialing, and the probability that the calls are routed to the number dialed ("system completion"). Beyond these network effects, calling and called customer characteristics such as the time required to dial a number, the abandonment behavior of calling customers during the call setup process, and the time required for the called customer to answer impact strongly upon the overall completion probability and the call setup time experienced by telephone customers.

The combined influence of network performance and customer behavior tends to mold customer attitudes about the telephone network and plays an important role in determining the amount of equipment needed to serve customer requests and the revenues which are realized

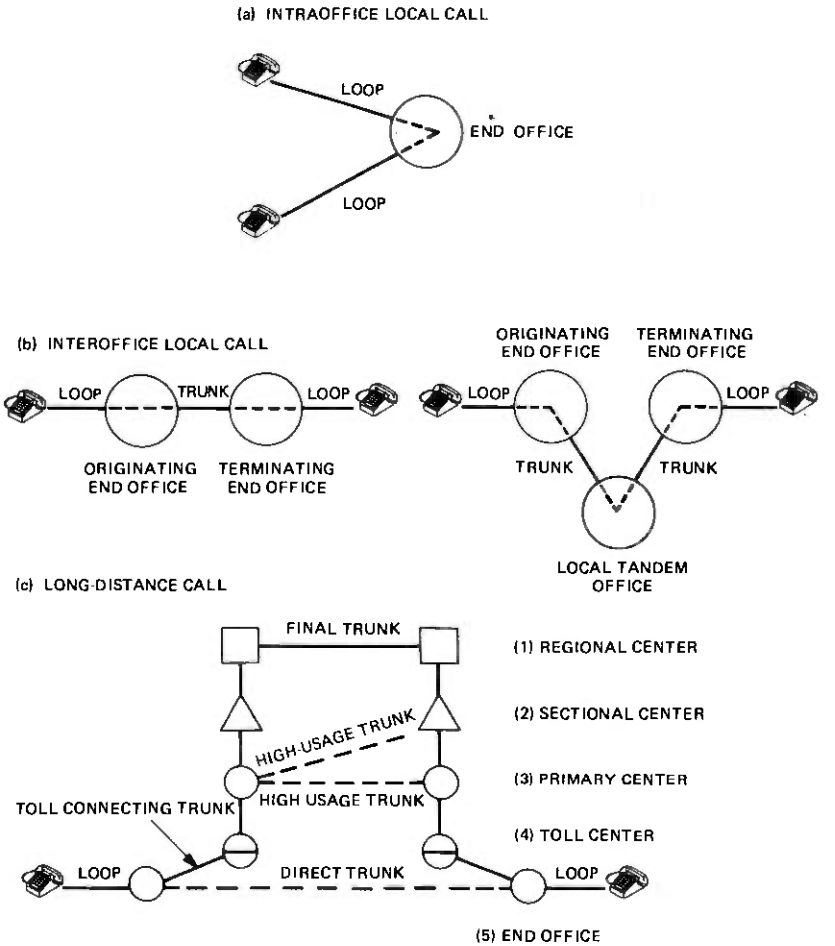


Fig. 1—Local and long-distance switching arrangements.

from successfully completed calls. As a result, the Bell System operating telephone companies continually collect detailed data on network performance and, to a lesser extent, customer behavior to identify and correct network problems and to be able to engineer the network properly to meet specified performance criteria.

Bell Laboratories has gathered a sample of these data using a sampling plan that provides a statistically valid systemwide characterization of the call setup process as seen by the calling customer. This article presents an analysis of the sampled data in the expectation that the information on overall network call setup performance and on customer behavioral characteristics which affect the call setup process will be useful to the planners and administrators of telephone networks and

to designers of equipment which utilizes and systems which interact with the telephone network. For instance, the probability of a fast retrial following the occurrence of a no circuit/reorder signal or announcement, as presented in Table III, can be an important parameter in trunk group size engineering.

Several previous studies by Bell Laboratories have dealt with the call setup process.[2–5] All but the first of these are of such different emphasis that the points of comparison with the current study are limited to a few statements on customer behavior. The Larsen and McGill study[2] is the most similar of the studies, and like the current one, is based on service observing data. However, the data were obtained from a number of sites chosen arbitrarily and not according to scientific sampling procedures. Hence, the representativeness of its results is unknown; and in particular, the confidence interval calculations do not properly account for the complexity of the service observing process (see Section II). That study emphasizes time of day and day of week trends in the completion percentage and call setup time results; and while a direct comparison with the current study is difficult, the qualitative statements on these trends are borne out by the present work. The overall completion percentage in the Larsen and McGill study is slightly over 1 percent lower and the call setup time about 1 second higher than currently, but the confidence intervals in the current study overlap these results.

The discussion in this article pertains to Direct-Distance-Dialing (DDD) toll call attempts. Those attempts which are directly dialed by the calling party represent approximately 88 percent of all the long-distance calls established within the Bell System. The sample design used to gather the data is described in Section II. The network and customer call attempt characteristics derived from this survey are presented in Sections III and IV.

II. SAMPLE DESIGN

The term "sample design" denotes the entire process of planning a sample survey. That process includes (*i*) the definition of the population to be studied, (*ii*) the adoption of a sampling plan to collect the data, and (*iii*) the derivation or selection of appropriate statistical estimation formulas.[6] The Bell System call attempt survey described in this article was designed to provide a systemwide characterization of the call setup process utilizing data collected through an existing Bell System procedure. Therefore, considerable care was required in the design of the sample survey to properly account for the complex sampling scheme inherent in the established procedure. This design is discussed in detail in terms of the three features state above in the following subsections in order to demonstrate the statistical validity of the results.

2.1 Population

A distinction must be made between the target population and the sampled population. The target population is the collection of elements about which information is desired. In some situations it may be unattractive or even impossible to sample from the target population. In those cases, the population definition is modified for sampling purposes, and the modified population is referred to as the sampled population. The modifications should be carefully applied so that the sampled population does not deviate too greatly from the target population.

The target population for the Bell System Call Attempt Survey was defined to be all DDD call attempts originated by Bell System subscribers including attempts to toll directory assistance operators. This population definition was modified to take advantage of an existing data source, Dial Line Service Observing (DLSO). The sampled population is restricted to call attempts which originate from local switching offices which serve more than 3000 subscriber lines and consists basically of calls which are directed to a terminating switching office more than 25 miles from the originating switching office or to a different Numbering Plan Area (NPA). Exceptions to this general description are as follows: (*i*) calls to a different NPA which are dailed as a seven-digit number (permissible by protecting codes) are excluded unless the calling distance exceeds 25 miles, and (*ii*) calls to Inward Wide Area Telephone Service (INWATS) subscribers, or calls to toll directory assistance, which are generally dialed as NPA-555-1212, are included without regard to any milage restrictions. Operator-dialed toll call attempts and customer-dialed, operator-serviced toll call attempts are not included in this study. In general, partially dialed attempts are also excluded from this study because of the classification problems posed by insufficient or incorrect digits.

2.2 Sampling plan

The sampling plan defines the method used to gather the survey data. It was intended that the call attempt data include end-to-end information about the setup process and the final result of that process so as to enable the characterization of calling and called customer behavior and telephone network responses. Although test calling procedures and human factors tests can provide valid information about some aspects of network responses and customer behavior respectively, only "live" telephone traffic can provide real-life information of this type.

Dial Line Service Observing (DLSO) is a function performed daily at 425 bureaus geographically located throughout the Bell System to obtain nationwide information about the quality of the call setup process for telephone traffic. An ongoing sample of actual call attempts is observed in over 3500 local switching offices, and the data are summarized peri-

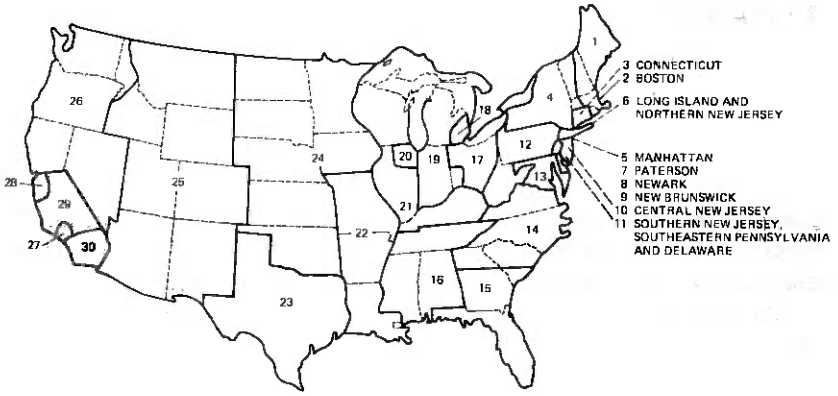


Fig. 2—Geographical stratification of the contiguous 48 states.

odically. Each observation begins at the instant the call attempt is originated and ends when the called party answers or when the calling party abandons. In the course of an observation, the call disposition and the timing sequence of call setup and abandonment events relative to the start of the attempt are recorded. These detailed records satisfy the survey criterion of end-to-end information, and the general application of DLSO satisfies the requirement for systemwide information. In addition, the DLSO definition of DDD call attempts agrees reasonably well with the target population definition for this survey. Therefore the sampled population definition in the previous subsection was intentionally designed to agree closely with the DLSO-DDD population in order to adopt DLSO as the data source for the Bell System call attempt survey.

The sampling plan can be briefly described as a four-stage selection process with stratifications imposed prior to selecting the first- and second-stage sample units. The first-stage units are DLSO bureaus. Prior to selecting the bureaus, the area of the contiguous 48 states was partitioned into 30 mutually exclusive geographical areas which are called primary strata. This stratification is illustrated in Fig. 2. Each primary stratum was designed to consist of a continuous geographical area and to be approximately the same size as the other primary strata with respect to the number of toll-call attempts originated from the area annually. The primary strata are much larger geographically in sparsely populated areas than in densely populated areas due to the second design criterion. This stratification guaranteed a dispersion of the sampled DLSO bureaus throughout the Bell System.

Five primary strata contained only one DLSO bureau. Those bureaus were selected as first-stage or primary sample units with a probability of one. Two DLSO bureaus were selected with replacement within each

Table I — Bell System call attempt survey primary or first-stage sample

Primary sample unit DLSO bureau	Primary stratum	Primary sample unit DLSO bureau	Primary stratum
Salem, Ma.	1	Cleveland, Oh.	17
Providence, R.I.	1	East Liverpool, Oh.	17
Boston, Ma.	2	Plymouth, Mi.	18
Bridgeport, Ct.	3	Pontiac, Mi.	18
New Haven, Ct.	3	Grand Rapids, Mi.	19
Glens Falls, N.Y.	4	Springfield, Oh.	19
White Plains, N.Y.	4	Blue Island, Il.	20
Manhattan, N.Y.	5	East Chicago, Il.	20
Brooklyn, N.Y.	6	Louisville, Ky.	21
Morristown, N.J.	6	Oshkosh, Wi.	21
Paterson, N.J.	7	Monroe, La.	22
Newark, N.J.	8	Saint Louis, Mo.	22
New Brunswick, N.J.	9	Fort Worth, Tx.	23
Asbury Park, N.J.	10	Houston, Tx.	23
Camden, N.J.	10	Salina, Ks.	24
Lansdowne, Pa.	11	Minneapolis, Mn.	24
Norristown, Pa.	11	Phoenix, Az.	25
Pittsburgh, Pa.	12	Las Cruces, N.M.	25
Fairmont, W.V.	12	Spokane, Wa.	26
Baltimore, Md.	13	Spokane, Wa.	26
Baltimore, Md.	13	Compton, Ca.	27
Athens, Ga.	14	Los Angeles, Ca.	27
Charoltte, N.C.	14	Oakland, Ca.	28
Delray Beach, Fl.	15	Oakland, Ca.	28
Melbourne, Fl.	15	Santa Ana, Ca.	29
Columbus, Ms.	16	Santa Ana, Ca.	29
Memphis, Tn.	16	Sacramento, Ca.	30
		Santa Cruz, Ca.	30

of the remaining 25 primary strata. Those selections were made with probabilities proportional to the sizes of the bureaus as measured by the number of annual toll call attempts which originated in the entities observed by the bureaus. Because sampling with replacement was employed to simplify the estimation formulas, four bureaus were selected twice. Two separate bodies of data were collected from each of these four bureaus. As a result of the first stage of sampling, the first-stage sample consists of 55 primary units which correspond to 51 distinct DLSO bureaus. The first-stage sample is listed in Table I. The analysis presented in this paper is based on 11,146 DDD call attempts observed by these bureaus during the study period.

Each DLSO bureau had the responsibility of observing between one and 124 local switching entities at the time the data were collected in October of 1974. A total of 3521 entities was being observed by the 425 DLSO bureaus in existence at that time for an average of approximately eight entities per bureau. Since the larger bureaus had a greater chance to be selected because the probabilities of selection were proportional to the sizes of the bureaus, there is an average of approximately 23 entities per DLSO bureau selected for the first-stage sample.

Somewhere between 20 and 100 service-observing loops were randomly assigned to groups of subscriber line appearances within each of these local switching entities. This assignment represents the second stage of sampling. The second-stage sample unit is a line group which is simply a collection of physically adjacent subscriber line appearances within a local switching entity. A line group varies in size from 50 to 600 subscriber line appearances depending upon the type of switching machine. Since the selection of line groups was made independently in each entity and since the entities associated with a DLSO bureau represented a partitioning of the bureau into mutually exclusive strata, the entities were treated as substrata in the sampling plan. The term "substratification" denotes a partitioning imposed within the first-stage sample units prior to selecting the second-stage sample. The second-stage sample units were randomly selected without replacement. In each line group belonging to the second-stage sample, the assigned service-observing loop was connected to a randomly selected subscriber line appearance. This selection of third-stage sample units generally was repeated within each local switching entity on a weekly basis during the month-long study. Observations of actual DDD call attempts on the selected subscriber lines formed the fourth-stage sample. These observations are referred to as sample elements since they comprise the last-stage sample and as such are members of the sampled population.

2.3 Estimation procedures

Because the Bell System call attempt survey data are not self-weighting, individual data items contribute differently to a system estimate. The complex sampling plan structure discussed in the previous subsection leads to varying probabilities of selection among the individual data items. The individual probabilities of selection are related to the amount of long distance telephone traffic a data item represents. Appropriate traffic weights, which are basically the inverse probabilities of selection for the data items, are used as sampling weights in all calculations to account for the individual contribution of a data item to a system estimate.

Ratio estimators of the form $r = x/y$, where r is an estimate of a population characteristic R and x and y are estimates of the population totals for two random variables X and Y, were used to calculate all system estimates. The distribution of such an estimate is approximately normal for sufficiently large samples. This assumption was used to calculate confidence intervals for the estimates presented in this study in the form $r \pm t\sigma_r$, where r is the ratio estimate, σ_r is the standard error, and t is the appropriate normal deviate. Since 90 percent confidence intervals are used consistently throughout this study, t has the value 1.65. Con-

fidence intervals are not given when the sample size is too small for the normality assumption.

The confidence intervals indicate the precision associated with the population estimates. Precision is a measurement of the probable sampling error associated with a statistical estimate of a population parameter. It gives a measure of the reliability of an estimate with respect to sampling errors; it does not take into consideration nonsampling errors. The 90 percent confidence intervals in this article mean that if the study were repeated 100 times, then 90 out of the 100 confidence intervals which could be calculated for the 100 separate estimates of a given population parameter would be expected to contain the true value of the population parameter.

III. NETWORK RESULTS

Call attempt performance is characterized through the presentation of disposition probabilities, setup and abandonment time distributions, and some aspects of customer retrial behavior. The call disposition results summarize the outcomes of DDD attempts in terms of the relative frequencies of completion and various reasons for failure to complete. The setup time results show the speed of network response during the provision of system signals and intercepts. They also illustrate customer behavior while dialing and answering telephone calls. The abandonment time results illustrate customer behavior in terms of releasing telephone facilities prior to call completion. In some cases, the facilities are released after an explicit system indication of failure to complete, while in other cases, the facilities are released following a subjective judgment of failure to complete or a desire to terminate the attempt prematurely. The customer retrial results characterize retrials which are initiated within 60 seconds of a failure to complete. Attention is focused on each of these attempt characteristics individually in Sections 3.1 through 3.3, respectively.

3.1 Call disposition

A summary of DDD call attempt dispositions is given in Fig. 3. An estimated 70.7 percent of the attempts are completed satisfactorily. Approximately one out of every four attempts are incomplete due to a called customer did not answer (DA) condition, called customer line busy (BY) condition, calling customer dialing irregularity (CDI), or called station number change or disconnect (C/D) situation. These customer-dependent failures to complete account for 24.8 percent of all DDD attempts and 84.6 percent of all DDD failures to complete. The second largest category of failures is manifested by abandonments without system responses (NR—no response) after dialing valid telephone numbers; they account for 2.6 percent of all DDD attempts and 8.9 percent of all DDD failures

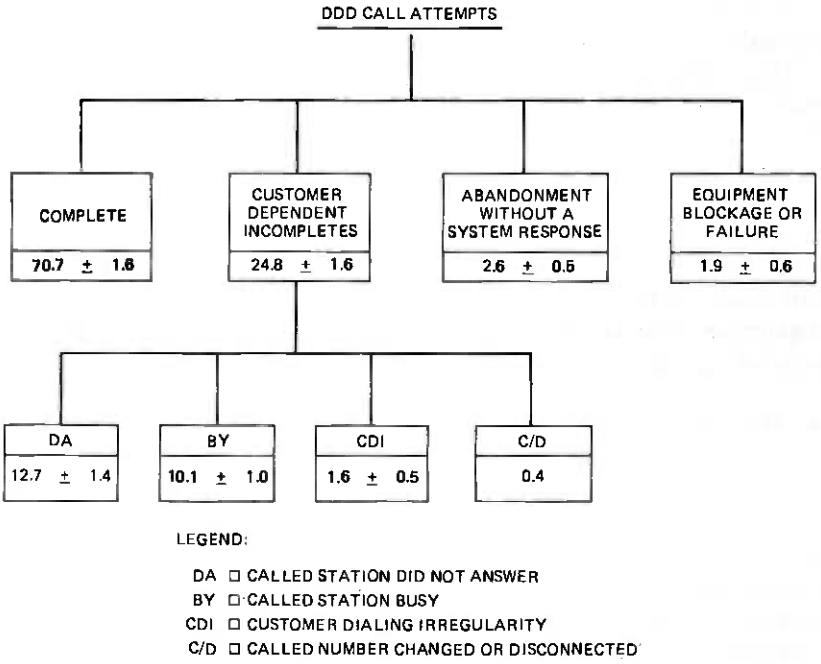


Fig. 3—DDD call attempt disposition percentage estimates with 90 percent confidence intervals.

to complete. These abandonments occur because the calling customer aborts prematurely or the network fails to respond properly to the request. The third and final category of incomplete attempts consists of equipment blockages and failures (EB & F). Since equipment blockages occur because of insufficient equipment to service the request and equipment failures are symptomatic of faulty equipment, the responsibility for these EB & Fs clearly belongs to the telephone network. These network failures account for 1.9 percent of all DDD attempts and 6.5 percent of all DDD failures to complete.

Complete attempts are calls properly processed by the switching and signaling network. They represent the ideal in call disposition since they fulfill the calling party service requests and generate revenues for the telephone companies. Customer-dependent failures (DA, BY, CDI, C/D) also are properly processed attempts from a switching and signaling viewpoint. While these attempts cannot be completed for reasons beyond network control, the calling party is informed about the nature of the failure. Therefore, the results in Fig. 3 show that at least 95.5 percent of all DDD attempts are properly processed by the telephone network. This percentage is a lower bound because a substantial number of abandonments without a system response are in fact premature disconnects

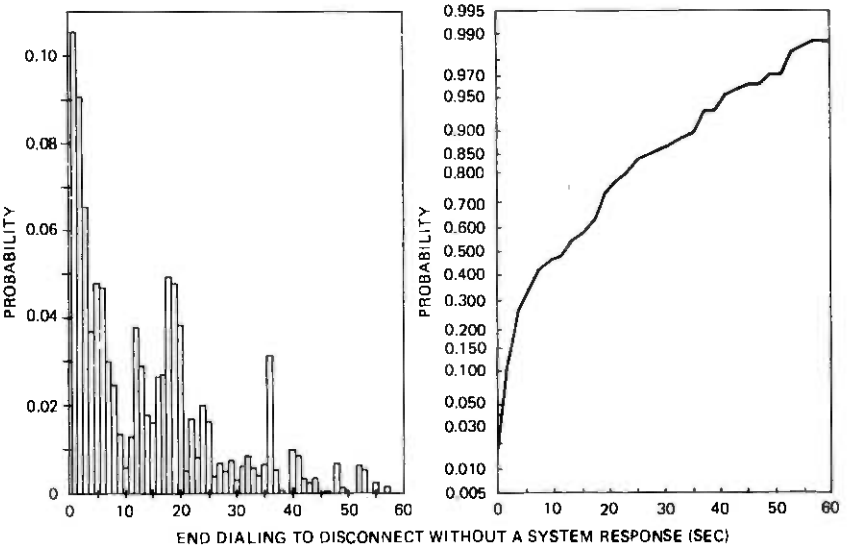


Fig. 4—Distribution of the time from end of dialing to abandonment without a system response.

on the part of the calling customers. A comparison of the distribution of time from end of dialing to abandonment without a system response in Fig. 4 with the distribution of time from end of dialing to the first system response in Fig. 5 illustrates that attempts abandoned within five seconds after dialing probably were intentionally prematurely disconnected by the calling parties, since the customers could have little expectation that a response should have been received during that interval. Only a few system responses are received within that time interval. This part of the distribution represents 30 percent of the NR conditions. At the other extreme of the distribution, it appears that about 14 percent of the NR conditions, or about 0.4 percent of all attempts, may be reasonably classified as network failures since very few system responses occur after 30 seconds. The remaining 56 percent of the NR distribution overlaps the system response distribution, and those abandonments cannot be easily classified as premature disconnects or network "high and dry" conditions.

While at least 95.5 percent of all DDD attempts are properly processed from a switching and signaling viewpoint, the disposition results also show that 1.9 percent fail to complete due to system blockages or equipment problems. This EB & F estimate is a lower bound for network failures since some of the abandonments without a system response, which were discussed above, represent network "high and dry" conditions. Individual estimates for blockages and equipment failures cannot be calculated because a common network signal, which is a tone inter-

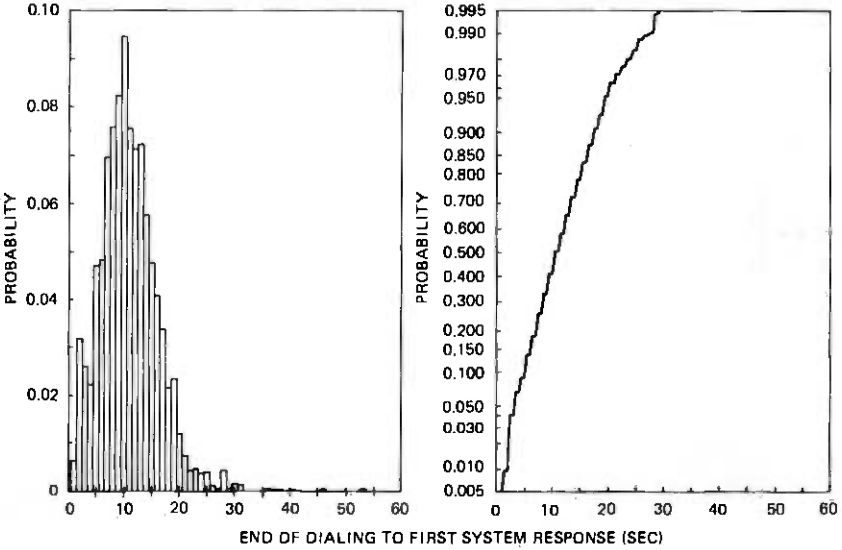


Fig. 5—Distribution of the time from end of dialing to the first system response after dialing a valid telephone number.

rupted 120 times per minute, is often issued for both failures. The telephone companies are responsible for providing and maintaining sufficient equipment to handle service requests within reasonable blockage and failure bounds. Since traffic loads are statistical in nature and equipment failures are random in nature, the system is continuously monitored and tuned to assure reasonable service within proper economic constraints. The EB & F estimate is an indication of the overall impact of system procedures for the provisioning and maintenance of equipment upon the DDD network response to customer service requests.

The reciprocal of the completion ratio gives the average number of attempts per successfully completed call. Using the DDD completion ratio of 0.707, this average is computed to be 1.4 attempts per successful completion. Substantial improvement in this average can only come about by a reduction in failures which are dependent upon calling and called customer behavior, since such failures account for about 85 percent of all DDD failures to complete. The occurrence of such failures can be reduced through the mutual cooperation of telephone companies and telephone customers. Reasonable service agreements in terms of the types and amounts of equipment required for individual customer communication requirements in conjunction with the application of new services made possible through the advent of Electronic Switching Systems (ESS) such as call forwarding and call waiting can produce a substantial impact.

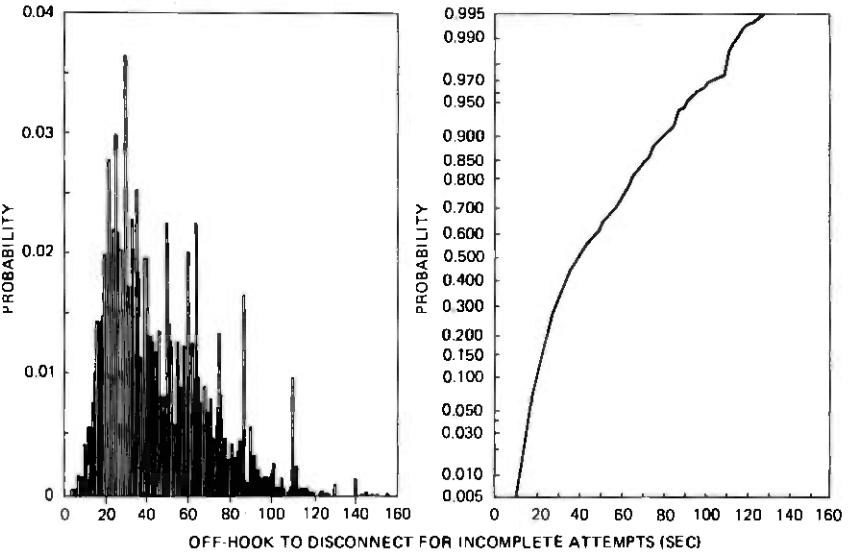


Fig. 6—Call attempt time for incomplete DDD attempts.

Complete eradication of all network failures would only decrease the average number of attempts per completed call by about 5 percent. Reduction of network failures is an area of continuing interest, and the historical record of EB & F reduction shows continuous improvement. Whenever practical, new technological advances continue to reduce network failures. For example, modern ESS offices will make a second attempt automatically if they are able to detect an irregularity in establishing the next link in a call.

3.2 Call attempt time

Call attempt time is defined as the nonconversational period which commences with calling station off-hook and terminates with called station answer for completed calls and with calling station return to on-hook for failures to complete. This usage of network facilities during the call setup process is absorbed as network overhead since revenues are not directly obtained. The average call attempt time with accompanying 90 percent confidence interval is 45.1 ± 1.8 seconds for incomplete attempts and 32.8 ± 0.7 seconds for completed calls. Call attempt times for incomplete attempts are not only longer on the average than those for completed calls, they are also more variable as is seen by the standard deviations of 24.7 and 11.0 seconds respectively. The distributions of call attempt time for incomplete attempts and completed calls are presented in Figs. 6 and 7 respectively. The overhead represented by these call attempt times amounts to an average of 50.8 seconds per

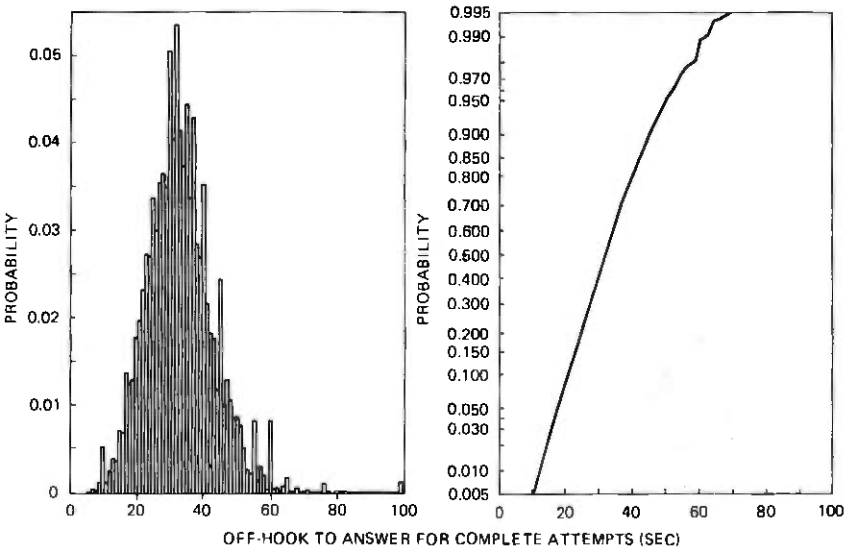


Fig. 7—Call attempt time (nonconversational) for successfully completed DDD attempts.

completed call, since 1.4 attempts are initiated for each successful attempt ($32.8 + 0.4 \times 45.1$). If the time from off-hook to start of dialing plus start of dialing to end of dialing (13.8 sec) is excluded on the basis that it impacts mainly upon local facilities, the toll network overhead is 31.5 seconds per completed call ($50.8 - 1.4 \times 13.8$).

Further insight into call attempt time is gained by considering the customer-controlled and network-controlled components. The components of the call attempt time interval are schematically illustrated in Fig. 8, which also contains the average holding time associated with each component. A more detailed statistical characterization, which includes the mean, standard deviation, and 10, 50, and 90 percent points of the cumulative distribution function, is given for each component in Table II. The calling customer components of call attempt time include the intervals from off-hook to start of dialing and from start of dialing to end of dialing, which are common to all attempts, and the intervals from the beginning of a system response to disconnect, from the beginning of an answer at the wrong station to disconnect, and from the end of dialing to abandonment without a system response for incomplete attempts. Called customer behavior is directly illustrated in two call attempt time components; namely, the intervals from the start of ringing to answer and from the end of dialing to answer with no audible ringing. These customer controlled components of call attempt time are discussed in Section 3.2.1 below. The network controlled components include the intervals from end of dialing to any of the various network signals, an-

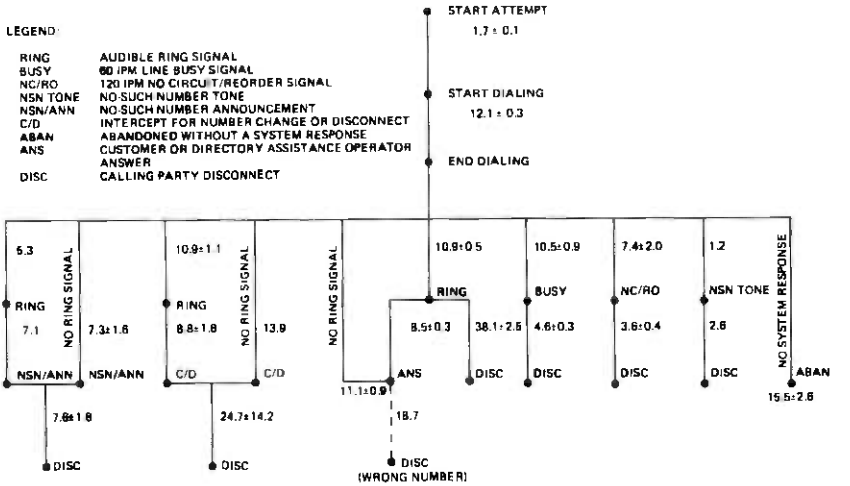


Fig. 8—Average DDD call attempt setup and abandonment times (seconds) with 90 percent confidence intervals.

nouncements, or manual intercepts, and they are discussed in Section 3.2.2 below.

3.2.1 Customer-controlled components of call attempt time

The intervals from off-hook to start of dialing and from start of dialing to end of dialing constitute the service request period of a call attempt. On the average, a request consumes 13.8 seconds of the overhead time per attempt. The type of dialing equipment at the originating location substantially impacts upon this service request time, since the average dialing time per DDD attempt is 13.7 seconds at rotary dial stations and 7.0 seconds at *TOUCH-TONE*® dialing stations. The interval from off-hook to start of dialing is also slightly longer for rotary dial customers than for *TOUCH-TONE* customers. The end result of these differences due to dialing equipment are rotary-dial service request times on the average of 15.5 seconds which are almost twice as long as the *TOUCH-TONE* service request times, which average 8.6 seconds.

In addition to influencing the service request phase of the call attempt process, calling customers also influence the speed of abandonment associated with incomplete or incorrectly completed attempts. Customers generally abandon faster when there is a positive network indication of failure to complete than when a subjective judgment of failure is required. This finding is borne out by a comparison of the abandonment times following network indications of line busy (BY), no circuit or recorder (NC/RO), and no-such-number (NSN) conditions with the abandonment times associated with disconnects after receipt of audible

Table II—DDD call setup and abandonment time statistics

Call setup or abandonment interval	Mean (sec)	Std. dev. (sec)	Cumulative distribution percentage points (sec)		
			10%	50%	90%
Off-hook to start of dialing	1.7 ± 0.1	2.0	1	2	4
Start of dialing to end of dialing	12.1 ± 0.3	5.1	6	12	18
End of dialing to ring before answer or disconnect	10.9 ± 0.5	5.0	5	11	17
End of dialing to answer without a ring signal	11.1 ± 0.9	5.3	5	10	17
Start of ringing to answer	8.5 ± 0.3	7.2	3	6	15
Start of ringing to disconnect without an answer	38.1 ± 2.5	20.7	18	33	59
Answer to disconnect for wrong numbers	18.7	8.9	—	—	—
End of dialing to busy (60 IPM) signal	10.5 ± 0.9	5.2	5	10	17
Start of busy to disconnect	4.6 ± 0.3	3.7	2	3	8
End of dialing to no circuit/reorder (120 IPM) signal	7.4 ± 2.0	7.4	0	6	15
Start of no circuit/reorder to disconnect	3.6 ± 0.4	2.3	2	3	6
End of dialing to no-such-number (NSN) tone	1.2	1.4	—	—	—
Start of NSN tone to disconnect	2.6	1.7	—	—	—
End of dialing to ring prior to NSN Announcement	5.3	4.1	—	—	—
Start of ringing to NSN announcement	7.1	4.9	—	—	—
End of dialing to NSN announcement without a ring signal	7.3 ± 1.6	5.4	1	6	11
Start of NSN announcement to disconnect	7.6 ± 1.8	10.4	2	5	13
End of dialing to ring prior to intercept for number change or disconnect (INT)	10.9 ± 1.1	4.6	6	10	19
Start of ringing to INT	8.8 ± 1.8	6.3	3	9	15
End of dialing to INT without a ring signal	13.9	4.4	—	—	—
Start of INT to disconnect	24.7 ± 14.2	24.6	4	15	81
End of dialing to customer abandonment without a system response	15.5 ± 2.8	17.8	1	12	35

ringing for DA conditions and with disconnects without a system response for network "high and dry" conditions. The reasons for the relatively slow abandonment times associated with intercepts for number changes or disconnects (C/D) will be explained later. The BY, NC/RO, and NSN indications immediately signal failure and encourage the calling party to abandon; while in other cases, the calling party abandons only after judging that the called party had sufficient time to answer or that the network has had sufficient time to respond.

Calling customers also appear to differentiate between the various definitive network indications for failure. The NSN tone, which continuously varies in frequency, is quite different from the BY and NC/RO signals. It consistently encourages rather fast abandonments as can be seen by the estimates for the mean and standard deviation in Table II. Either the characteristics of the NSN tone itself or the unfamiliarity of the tone may cause this behavior. More surprisingly, however, customers also seem to distinguish hetween BY and NC/RO signals since both the mean and the standard deviations are smaller for abandonment times after NC/RO signals compared with those after BY signals. The main distinction between these signals is that the BY tone is interrupted 60 times per minute and the NC/RO tone is interrupted 120 times per minute.

Abandonments which follow definitive network signals for failure to complete occur faster than those which follow recorded, automatically composed (computer-generated) or manual intercepts. NSN announcements and C/D intercepts convey verbal messages, and this transfer of information requires additional time. While all NSN announcements were automated, one third of the C/D intercepts were manual. This partially explains the slower C/D abandonment time, since the average abandonment time for the manual C/D intercepts is approximately 10 seconds slower than for the automated C/D intercepts. In addition, the average abandonment time following automated C/D intercepts is almost twice as slow as the average for automated NSN announcements. In some instances an automated C/D intercept is followed by an "operator intercept cut-through" to provide the calling party with an opportunity to discuss the problem with a person. In such cases the time to abandon could well be even greater than for an immediate manual intercept. The occurrence of such cut-throughs are not recorded in the data base; however, if it occurred during a significant number of the automated C/D intercepts, it might well account for slower abandonment times. That some such cut-throughs occurred is suggested by the presence of several outlying values in the automated C/D abandonment time data. These outlying values along with several long holding times for the manual C/D intercepts inflate the confidence interval substantially and result in an average overall C/D abandonment time of

24.7 seconds which is quite larger than the median of 15 seconds. Excluding the automated C/D intercepts which appear to be followed by operator intercept cut-throughs, the average abandonment time for automated C/D intercepts is still longer than for NSN announcements. Since the standard messages which are conveyed are similar in length, this difference must be attributed to customer reaction.

While some type of system response is encountered for 97.4 percent of all DDD attempts, the remaining 2.6 percent are abandoned after dialing without a system response. They are abandoned because the calling party no longer desires to complete the call, is interrupted by some event, or perceives that the network has failed to respond properly to the dialed digits. Abandonments without a system response were discussed in Section 3.1 above. The conclusions were (*i*) about 30 percent clearly appear to abort prematurely, (*ii*) about 14 percent clearly appear to be network failures, and (*iii*) about 56 percent of the distribution for these abandonment times overlaps the distribution for the times from end of dialing to first system response.

The final calling customer abandonment characteristic to be discussed in this article pertains to the 19 wrong numbers which were observed during the survey. The average time from answer at a wrong station to abandonment is 18.7 seconds; however, this estimate is based upon a very small number of observations. The confidence interval for this estimate is omitted in Fig. 8 and Table II because the small number of observations negates the meaningfulness of the confidence interval calculation.

The large differences in average abandonment time among the various reasons for abandoning cause the high variability seen earlier for the call attempt times associated with incomplete attempts. The longest average call attempt time is associated with the single most frequent reason for failure to complete, called customer did not answer. The average call attempt time for DA conditions is 62.8 seconds, of which 38.1 seconds are consumed by the calling party listening to ringing. An average of between six and seven rings are received before the calling party abandons. Since over 99 percent of all DDD answers occur prior to the 38.1 second average abandonment time, it is meaningful to investigate possible benefits associated with an effort to change customer behavior through educational programs. Not only may it be beneficial to discourage very long holding times, it may also be beneficial to discourage very short holding times to give called customers more opportunity to respond.

At the other end of the abandonment time spectrum, abandonments following a busy signal occur rather fast. However, 30 percent of the customers who encounter a BY signal remain on the line over 5 seconds following the initial receipt of the tone. Since 10.1 percent of all DDD

attempts result in busies, this excessive holding time represents a drain on network facilities without any possibility of completion.

Focusing attention at the terminating end of a call, one aspect of called customer behavior is provided by the following answer time results: (*i*) the average time from end of dialing to called station answer with audible ringing observed at the calling station is 19.4 seconds, and (*ii*) the average time from end of dialing to called station answer without the receipt of audible ringing at the calling station is 11.1 seconds. The substantial difference between estimates is mainly attributed to toll directory assistance calls. The first category, which represents 96.1 percent of all DDD calls (messages), contains relatively few calls to toll directory assistance (5 percent), while slightly over two-thirds of the calls in the second category are to toll directory assistance. The average time to answer for toll directory assistance operators is 7.4 seconds when ringing is heard at the calling station and 1.1 seconds when it is not. This suggests that in the former case the operators are busy when the call appears, and that in the latter case an idle operator answers almost immediately; therefore, ringing is not observed at the calling station. Customer answers take an average of 8.5 seconds when ringing is observed at the calling station and 4.2 seconds when it is not. These results along with the composition of the two categories described above account for about two-thirds of the difference between the 19.4 and 11.1 second estimates. The remaining third is explained by differences in connect times for toll directory assistance calls and other DDD calls. The average time to establish a connection from the originating local switching office to the called station is 3.8 seconds shorter for toll directory assistance calls than for other DDD calls.

About 71 percent of the average call attempt time is under customer control during the service request or answer and abandonment phases of the call attempt process. The service request phase represents 33 percent of the call attempt time, and the answer and abandonment phase represents 38 percent. The remaining 29 percent is under network control for the purpose of establishing a physical connection between the calling and called stations.

3.2.2 Network controlled components of call attempt time

Upon completion of dialing, the network attempts to route a call to the called station by establishing transmission and signaling paths between a number of local and toll switching offices. A successful connection may be established, whereupon audible ringing is normally returned to the calling station, or the network may be unable to reach the called station, whereupon a specific network indication of failure to complete is normally returned to the calling station. The first system response is defined as the first network signal, network intercept, or called station

answer without audible ringing at the calling station which follows the end of dialing. The distribution of these response times is given in Fig. 5 and has an average time to respond of 10.9 ± 0.5 seconds with a standard deviation of 5.3 seconds. In Fig. 8 and Table II, summary statistics are listed for the individual system responses which include answer without audible ringing at the calling station, audible ringing, busy signal, no circuit or reorder signal, no-such-number tone or announcement, and intercept for number change or disconnect.

The average times from end of dialing to ring prior to answer or disconnect, busy, NC/RO, and NSN tone are 10.9, 10.5, 7.4, and 1.2 seconds respectively. The averages for ring and busy are relatively long because a connection must be established to the far-end local switching office and the called customer line must be tested before a ring or busy signal is returned to the calling station. A NC/RO signal is returned to the calling station when an equipment blockage or failure occurs. The average time from end of dialing to receipt of the NC/RO signal is shorter than the ring and busy averages because the signal is applied at the office in which the EB & F is encountered. That office may be the originating local switching office, an intermediate toll switching office, or the terminating local switching office. The average time from end of dialing to a NSN tone is the shortest of the four because the tone generally can be applied at the originating local switching office or the first toll switching office encountered in the switching path.

The two remaining responses which qualify as a first system response after dialing are recorded, composed, or manual intercepts for no-such-number conditions and number change or disconnect conditions. The NSN intercept is encountered near the originating end of a connection in a manner similar to the NSN tone. The C/D intercept occurs at the far-end of a connection. As a result of these distinctions, it is clear why the average time from end of dialing to C/D intercepts is almost twice as long as the average for NSN intercepts. Both types of intercept often are preceded by audible ringing at the calling station. Note that the average time from end of dialing to ringing prior to a NSN intercept is only 5.3 seconds. This clearly reflects the fact that NSN intercepts occur towards the near end of a connection. The average time from end of dialing to ringing prior to a C/D intercept is 10.9 seconds which is and should be in agreement with the average for ringing prior to answer or disconnect.

3.3 Fast retrials

In the event that an attempt is not successfully completed, the calling party must decide to permanently or temporarily abandon or to try again. The retrial probability and time to retry are completely determined by calling customer behavior under the influence of the various

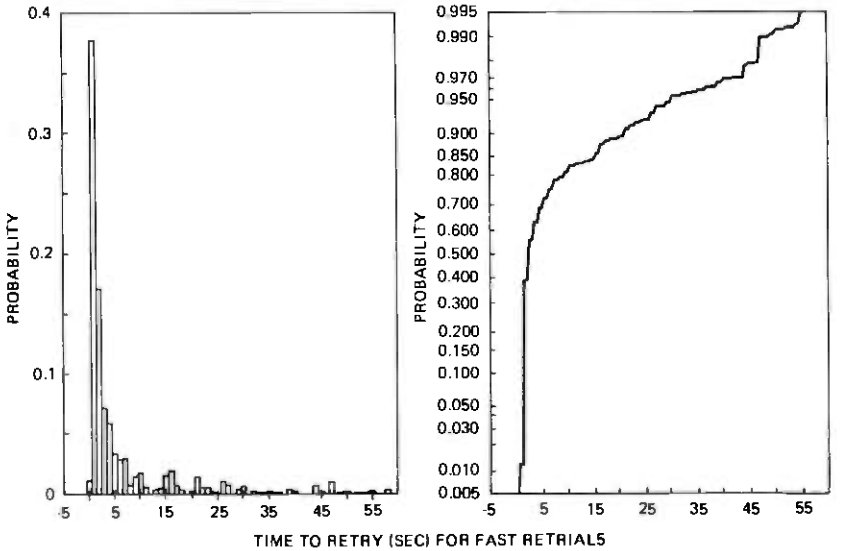


Fig. 9—Time to retry for retrials which occur within 60 seconds of a DDD failure to complete.

network indications for failure to complete. Retrial disposition is a function of both calling and called customer behavior, and to a lesser degree network behavior such as the persistence of NC/RO conditions.

An estimated 19.1 ± 2.5 percent of all DDD attempts which fail to complete are retried within 60 seconds of the failure, and 34.3 ± 5.4 percent of those retrials are successful. Over 50 percent of these fast retrials are initiated within two seconds of the abandonment time for the preceding failure. This time to retry is characterized by the distribution given in Fig. 9 for which the average is 6.7 ± 1.3 seconds and the standard deviation is 10.7 seconds. The substantial difference between the average and median times is caused by the high degree of positive skewness associated with the distribution.

Fast retrial statistics are summarized in Table III for the most frequent types of failures to complete. The fast retrial rates and the contribution to the overall fast retrial phenomenon for individual types of failures to complete are listed in the top part of the table. The major conditional disposition probabilities associated with retrials after failure to complete are given in the lower part of the table. Since these probabilities were derived from rather small samples, they are only intended to provide an indication of what happens to fast retrails, and are quoted without confidence limits. The sample size for the DA, BY, abandoned without a system response, NC/RO, and NSN results are 95, 183, 166, 66, and 85 attempts respectively.

Almost one-third of all DDD fast retrials follow a failure due to called

Table III — Retrial characteristics for retrials occurring with 60 seconds of an unsuccessful attempt

Unsuccessful call attempt dispositions	Percent which retry* (within 60 seconds)	Percent of retrials* (within 60 seconds)
Called party did not answer (DA)	5.2 ± 1.6	11.7 ± 3.9
Called party busy (BY)	17.3 ± 3.8	30.9 ± 5.1
Calling party abandoned without a system response (ABAN)	53.4 ± 6.5	24.4 ± 4.2
No circuit/reorder condition (NC/RO)	41.7 ± 13.8	11.2 ± 3.5
No-Such-Number condition (NSN)	55.5 ± 8.8	14.2 ± 4.7

Unsuccessful call attempt disposition	Conditional disposition probabilities (60 second retrials)
Called party did not answer (DA)	Complete: 0.23 DA: 0.13 No order: 0.57
Called party busy (BY)	Complete: 0.20 BY: 0.61
Calling party abandoned without a system response (ABAN)	Complete: 0.46 No order: 0.19
No circuit/reorder condition (NC/RO)	Complete: 0.49 NC/RO: 0.40
No-such-number condition (NSN)	Complete: 0.31 NSN: 0.23 No order: 0.24

Percent which retry—The percent of call attempts which fail to complete and are tried again within 60 seconds for the specific failures listed.

Percent of retrials—The percent of all retrials occurring within 60 seconds of a failure to complete which are attributable to the specific failures listed.

customer busy. These retrials are very likely to result in a second BY condition as is evidenced by the conditional probability for a BY following a BY of 0.61. The conditional probability for successful completion after a BY condition is only 0.20. While the fast retrial rate for called customer did not answer conditions is relatively low, such retrials account for 11.7 percent of all fast retrials because DA conditions occur relatively frequently on first attempts. Over half of these retrials which follow DA conditions are no orders, i.e., the calling party goes off-hook and back on-hook without dialing any digits or dials a partial number and returns to the on-hook state within 10 seconds after the end of dialing. The conditional probability for successful completion following a DA conditions is 0.23.

Fast retrial rates are relatively high for attempts which are abandoned without a system response or encounter NC/RO indications. The conditional probabilities for successful completion are also relatively high at 0.46 and 0.49 respectively. While less than 10 percent of the retrials after an abandonment without a system response also are abandoned without a system response, the reoccurrence of NC/RO having already received a NC/RO indication is very likely as can be seen by the conditional probability of 0.40 in Table III.

The fast retrial rate is also relatively high for attempts encountering NSN intercepts. The conditional disposition probabilities for fast retrial

following NSN intercepts present a mixed picture. These probabilities are 0.31, 0.23, and 0.24 for successful completion, NSN intercepts, and no orders respectively. The latter two probabilities indicate substantial confusion on the part of the calling party, since the first suggests the customer again dials the incorrect number while the second indicates the customer is too uncertain to continue dialing.

Fast retrials comprise only a small segment of the complete retrial process, since hours and even days may elapse between a failure to complete and a legitimate retrial. The data collection scheme for the DLSO call attempt survey was not designed to capture information on other than immediate retrials; however, a systemwide survey to characterize the complete retrial process is currently in progress at Bell Laboratories.

IV. CORRELATION OF NETWORK RESULTS WITH SEVERAL ATTEMPT CHARACTERISTICS

Call disposition is strongly influenced by certain attempt characteristics such as class of subscriber service and calling distance. Call setup time also is influenced by these characteristics and by the type of switching which is encountered during the setup process. While the class of subscriber service effects are totally dependent upon customer influences, the calling distance and switching machine effects are also dependent upon network influence. Call disposition characteristics vary by day of week and time of day; however, the major causative influence for these variations is the relative traffic composition by class of subscriber service. The roles which class of subscriber service, calling distance, and type of switching play in the call attempt process, are explored in the following subsections.

4.1 Class of subscriber service

Service observing is restricted to originating attempts. While the originating class of service is known for most observations, the terminating class of service is unknown for all observations except for calls to directory assistance and INWATS calls. The survey results discussed in this subsection are predicated on originating class of service with the main service distinction being residential versus business. Substantial differences in customer behavior are clearly shown by the DDD completion rates of 76.6 ±2.2 and 66.9 ±2.5 percent for business and residential customers respectively. The primary cause of this 10 percentage point difference is attributed to called customer behavior in terms of did not answers and busies, since the business DA rate of 8.9 ±1.6 percent is approximately half the residential rate of 16.4 ±2.3 percent, and the business BY rate of 8.6 ±1.4 percent is approximately three-fourths the

residential rate of 11.3 ±2.1 percent. Of the remaining dispositions, only customer abandonments without a system response border upon differing significantly. These estimates are 2.7 ±0.7 and 1.7 ±0.5 percent for business and residential attempts. Although the business abandonment rate is higher, business customers are as patient as residential customers in the sense that the median holding time prior to abandoning without a system response is 12 seconds in both cases. In addition, the distributions for the time from end of dialing to first system response are basically the same for business and residential traffic. These systemwide abandonment time and response time similarities for business and residential traffic seem to contradict the difference between the business and residential abandonment rates. The cause of this apparent contradiction is not known. It may be due to some characteristic such as routing for which the information is not available, or it may be due to sampling error.

Traffic composition in terms of the relative percentages of business and residential originated attempts exerts a strong influence upon overall DDD dispostion characteristics. In the following paragraphs it is shown that variations in the mixture of business and residential originations are responsible for many day of week, time of day, and calling distance trends associated with DDD call dispositions. In general, completion increases as the percentage of business originations increase because the occurrence of called station did not answer and busy conditions decline.

Weekend traffic is primarily residential in nature. Weekday traffic is almost evenly split between business and residential originations. As a result of this difference in traffic composition, the weekend completion rate of 69.1 ± 4.8 percent is lower than the weekday completion rate of 71.8 ± 1.8 percent. A high rate of DA conditions, which is typical of residential traffic, is the primary cause of the lower weekend completion rate.

In addition to this weekday versus weekend variation, traffic composition varies by hour of day. The percentage of business-originated DDD attempts by hour of day for weekday traffic is illustrated in Fig. 10 for the hours of 8:00 A.M. to 10:00 P.M. The business-residential composition is fairly constant during the *morning* hours of 8:00 A.M. to 12:00 noon. During the *noon* hour the business originations drop by about 10 percentage points. In the *afternoon* business originations initially jump to about 58 percent of the total traffic and then gradually decrease to about 49 percent of the traffic by 4:00 P.M. At this time a *transitional* period begins during which the business-residential mixture changes dramatically. By 7:00 P.M. the percentage of business originations is down to about 16 percent of the total traffic. The *evening* traffic from 7:00 to 10:00 P.M. consists of almost completely residential originated

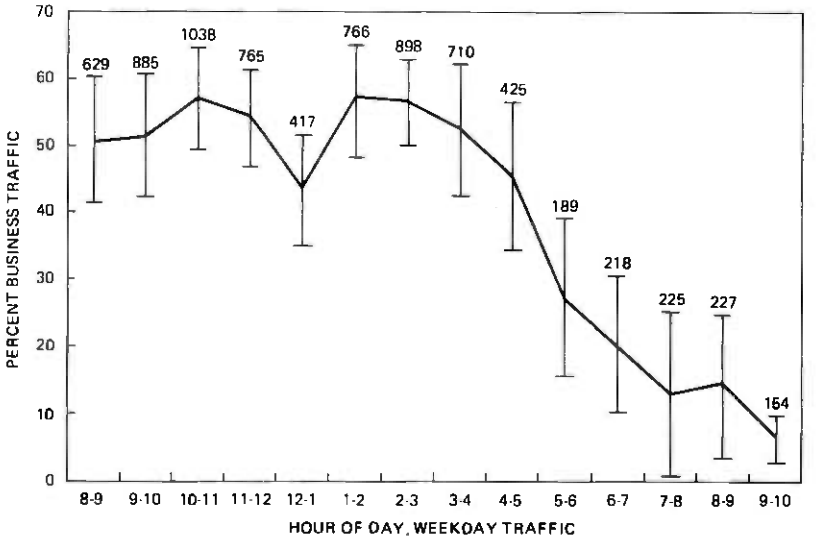


Fig. 10—Percent business-originated attempts with 90 percent confidence interval by hour of day.

attempts.

The variations in business-residential traffic composition are clearly evident in the disposition results for the five periods of day in italics above. The estimates for percent complete, which are listed in Table IV, increase and decrease along with the percentage estimates for business-originated attempts which are also listed in Table IV. These changes in completion are accompanied by corresponding changes in the opposite direction for the composite estimates of DA + BY conditions. While completion increases when the percent of business-originated attempts increases due to fewer DA and BY conditions, the increase in completion is not as great as the decrease in DA + BY conditions because the calling customer abandonment rate also increases with business-originated attempts. Considering the previous description of business and residential call disposition characteristics, it is clear that the time of day variations in call disposition are essentially caused by variations in the business-residential traffic composition throughout the day.

Business-residential traffic composition also varies with calling distance, which is defined as the airline distance between the originating and terminating local switching offices. While the majority (56 percent) of DDD attempts with calling distances between 26 and 400 miles originates from residential stations, the majority (also 56 percent) of the attempts with longer calling distances originates from business stations. This change in traffic composition contributes to several call disposition trends related to calling distances. Positive correlation of completion

Table IV — DDD call disposition estimates with 90 percent confidence intervals for five periods of the day for weekday traffic

Disposition	Percentage estimates by period of day				
	8–12 A.M.	12–1 P.M.	1–4 P.M.	4–7 P.M.	7–10 P.M.
Complete to desired station	74.1 ± 2.3	67.5 ± 8.3	72.0 ± 2.9	66.5 ± 5.0	59.1 ± 8.6
Called station did not answer	8.7 ± 1.5	15.2 ± 6.3	11.6 ± 2.1	13.6 ± 2.6	25.3 ± 9.0
Called station busy	9.6 ± 1.4	11.3 ± 6.2	9.4 ± 1.6	15.1 ± 5.5	11.5 ± 3.9
Calling customer abandoned without a system response	3.4 ± 0.8	1.7	2.3 ± 0.7	1.4	1.2
Percent business traffic	53.6 ± 6.2	43.5 ± 8.3	55.1 ± 6.2	33.2 ± 7.1	13.3 ± 5.8

with calling distance comes about because the business completion rate is higher than the residential rate and the percentage of business originations increases as calling distance gets longer. In a similar manner, negative correlation of the DA + BY rate with calling distance comes about because the business DA + BY rate is lower than the residential rate and the percentage of business originations increases as calling distance gets longer. The abandonment rate without a system response also is positively correlated with calling distance. In this case, the business abandonment rate of 2.7 ±0.7 percent appears to be invariant with calling distance, while the residential abandonment rates of 1.5 ±0.6 percent for attempts with calling distances between 26 and 400 miles and 3.0 ±2.1 percent for those with calling distances greater than 400 miles appear to be positively correlated with calling distance. These trends, along with additional call disposition trends which are related to calling distances but are not influenced by class of subscriber service, are discussed more thoroughly in Section 4.2.

Class of subscriber service plays only a minor role in determining call setup and abandonment times, since statistically, call setup and abandonment time results for business- and residential-originated traffic are similar with the exception of customer dialing times. The average residential rotary dialing time is 1.5 seconds longer than the average time for business customers. The average residential *TOUCH-TONE* dialing time is 1.3 seconds longer than the average time for business customers. Residential and business customers originate almost the same percentage of Home company, Home Numbering Plan Area (H-HNPA) attempts. Therefore, the differences in dialing times cannot be attributed to calling patterns in which one group is required to dial NPA codes more frequently than the other. Most likely, business customers have shorter dialing times because they use the telephone more frequently or are more familiar with the numbers they call. The only clear correlations of call setup and abandonment times with day of week or time of day once again is for customer dialing time for which the average dialing time on weekends is longer than the average for weekdays, and the average dialing time in the evening is longer than the average during the remainder of the day. These findings are in total agreement with the differences between business and residential dialing times and the traffic composition characteristics which were discussed previously. Namely, dialing times tend to be longer when the traffic is dominated by residential originations and shorter when the traffic is dominated by business originations.

4.2 Calling distance

Beyond the customer influences discussed above, there are additional network effects upon the call attempt process that appear in the calling

Table V—DDD call disposition percentage estimates with 90 percent confidence intervals as a function of calling distance

Disposition	26–400 miles	>400 miles
Complete to desired station	69.8 ± 2.0	72.0 ± 3.9
Called station did not answer	13.9 ± 1.6	10.9 ± 2.8
Called station busy	9.8 ± 1.3	9.0 ± 2.3
Calling customer abandoned without a system response	2.3 ± 0.5	4.0 ± 2.1
No circuit/reorder condition	1.4	2.5
No-such-number condition	1.7 ± 0.7	0.5
Other failures to complete	1.2 ± 0.5	1.1

distance discussion in this section and in the type of switch discussion in Section 4.3 below.

As stated previously, calling distance is defined as the airline distance between the originating and terminating local switching offices. Generally, as this distance increases, the call setup process becomes more complex due to additional toll switching, intraoffice processing, and interoffice trunking. This added setup complexity has considerable impact upon call disposition and setup times. It does not have much impact upon customer abandonment times.

Direct-Distance-Dialing attempts with calling distances of 25 miles or less are excluded from the following analyses because many such attempts do not belong to the sampled population for the DLSO call attempt survey (see Section 2.1). INWATS attempts are also excluded from the following analyses because the calling distances are not known for the intrastate, INWATS attempts. All remaining attempts to both customer stations and toll directory assistance operators are included in the calling distance analyses which follow.

Disposition results for attempts with calling distances between 26 and 400 miles and for attempts with calling distances greater than 400 miles are listed in Table V. As stated in the previous subsection, the completion and abandonment without a system response rates both increase and the DA + BY rate decreases as the calling distance becomes longer. The completion and DA + BY trends are primarily attributed to changes in business-residential traffic composition. The abandonment without a system response trend is attributed to a dependence of residential abandonments upon calling distance.

The occurrence of no circuit/reorder conditions increases as the calling distance gets longer. This trend is caused by the added complexity to the call setup process. The increase in toll switching creates more opportunities for failure to obtain the required facilities and for equipment irregularities.

The occurrence of no-such-number conditions decreases as calling distance increases due to the substantial difference between the 3.3 ±1.7

percent intra-NPA NSN rate and the 0.7 ±0.3 inter-NPA rate. While over three-fourths of the intra-NPA NSN conditions are due to the omission of the toll access code 1, which is required in many areas of the country, only about one-third of the inter-NPA NSN conditions are due to that reason. Apparently the presence of an NPA code in the called number reminds the calling party to dial the access code.

The impact of calling distance upon setup time is shown in Table VI. Average dialing times are greater for attempts with long calling distances because such attempts more frequently require an NPA code to establish a call. The interval from end of dialing to a ring or busy signal also increases as calling distances get longer because additional switching tends to slow down the setup time. The estimates for end of dialing to a NC/RO signal do not indicate any dependence upon calling distance. Because NC/RO conditions can occur at any intermediate office within the switching path and ring or busy are always returned from the terminating local switching office, the average time from end of dialing to NC/RO is shorter than to ring or busy, while the standard deviation is wider.

The last four intervals listed in Table VI represent customer abandonment times. The second and third of these intervals are for abandonments after positive indications of failure to complete; namely, busy and NC/RO signals. The results show that there is no correlation between calling distance and the time to disconnect after hearing a NC/RO signal. The results also indicate that there may be a positive correlation between calling distance and the time to disconnect after hearing a busy signal; however, the evidence for such a correlation is rather weak. The two remaining abandonment times for disconnect after receipt of audible ringing with no answer and disconnect without a system response are dependent upon calling customer judgments that the attempts would not complete satisfactorily. The time to disconnect after receipt of audible ringing without an answer is not correlated with calling distance. The time to disconnect without a system response may be positively correlated with calling distance; however, once again, the evidence for such a correlation is rather weak.

4.3 Local switching and call setup time

The average setup time from end of dialing to ring or busy varies substantially for attempts classified by the types of local switching. The average setup times with accompanying 90 percent confidence intervals and standard deviations are listed in Table VII for originating and terminating local switching classifications. Estimates for attempts which originate from rotary dial and *TOUCH-TONE* dialing stations are given separately for the originating local switching classifications.

Call attempts which originate from *TOUCH-TONE* dialing stations served by step-by-step (SXS) switching machines have a significantly

Table VI—DDD call attempt setup and abandonment time statistics as a function of calling distance

Setup or abandonment interval	26–400 miles		>400 miles	
	Mean (sec)	Std. dev. (sec)	Mean (sec)	Std. dev. (sec)
Rotary dialing time	13.1 ± 0.4	4.6	14.4 ± 0.3	4.3
TOUCH-TONE® dialing time	6.9 ± 0.4	3.3	8.5 ± 1.0	4.4
End dialing to ring or busy	11.1 ± 0.6	5.2	12.5 ± 0.4	4.2
End dialing to no circuit/reorder	8.6 ± 3.4	8.7	6.9 ± 3.1	5.6
Start of audible ring to disconnect without an answer	39.9 ± 3.3	20.7	39.3 ± 4.2	21.1
Start of busy to disconnect	4.4 ± 0.4	3.5	5.8 ± 1.6	4.6
Start of no circuit/reorder to disconnect	3.8 ± 0.4	2.5	3.7 ± 0.9	1.7
End dialing to abandonment without a system response	14.8 ± 3.1	13.8	20.2 ± 7.3	27.9

Table VII—DDD call attempt statistics for the time from end of dialing to ring or busy as a function of originating and terminating local switching

Originating switch	Rotary dial customers		*TOUCH-TONE*® customers	
	Mean (sec)	Std. dev. (sec)	Mean (sec)	Std. dev. (sec)
PAN	13.6 ± 3.0	7.4	—	—
SXS	12.6 ± 1.6	5.8	16.5 ± 1.7	4.7
5XB	10.1 ± 0.4	4.4	10.8 ± 0.6	4.6
1XB	10.8 ± 0.8	5.8	12.3 ± 1.9	4.6
ESS	9.2 ± 0.6	4.6	9.5 ± 0.9	3.6

Terminating switch	All customers	
	Mean (sec)	Std. dev. (sec)
SXS	12.7 ± 0.6	4.6
CDO	12.8 ± 1.0	4.7
5XB	9.7 ± 0.7	5.0
1XB	9.5 ± 0.6	3.8
ESS	7.9 ± 0.5	3.7

Legend:

PAN	Panel
SXS	Step-by-step
5XB	No. 5 crossbar
1XB	No. 1 crossbar
ESS	Electronic Switching System
CDO	Community Dial Office

longer delay between the end of dialing and the beginning of audible ring or busy than attempts which originate from rotary dial stations served by SXS switching machines. *TOUCH-TONE* dialing signals, which are generated at a subscriber station, are stored and converted to dial pulses at a SXS office before the telephone number is processed. On the other hand, dial pulses, which are generated at a rotary dial station, directly drive a SXS switch during the dialing interval itself; therefore, the telephone number is almost completely processed at the end of dialing. These operational procedures account for much of the difference between the SXS estimates; however, unknown differences in calling distances, routing patterns, and toll switching may also contribute to the 3.9 second difference in average setup times. There are no significant differences between rotary dial and *TOUCH-TONE* results for the other types of switching.

The average setup time for attempts originating from customer stations served by SXS machines is significantly higher than the averages for those originating through No. 5 crossbar (5XB), No. 1 crossbar (1XB), and ESS machines. In the previous paragraph it was stated that the dial pulses generated from a rotary dial station directly drive a SXS switch, and hence, at the end of dialing the dialed number is almost completely processed by the originating SXS machine. This would tend to shorten the interval from end of dialing to ring or busy. However, the same SXS procedure occurring at the terminating switching office tends to lengthen the setup interval as is illustrated by the SXS and CDO estimates in the lower portion of Table VII. An estimated 50 percent of all DLSO-observed call attempts which originate through SXS machines also terminate through SXS machines. This terminating percentage is 20 to 30 per-

centage points higher than for the other types of switching machines. In addition to this influence, which tends to increase the length of the call setup interval, there is a second influence which comes into play: the calling distance distribution for DDD attempts originating from customers served by SXS machines is significantly biased towards longer calling distances relative to the distributions for attempts originating from customers served by the other types of switching machines. In the previous subsection it was shown that call setup time increases as the calling distance gets longer. Thus, the long setup times for rotary dialed calls handled by SXS machines are not caused by the originating SXS machines. They come about because of call destination and calling distance characteristics associated with the attempts. However, SXS machines at the terminating ends of calls do cause significantly longer setups.

In addition to the SXS findings above, the results in Table VII indicate that calls placed by customers served by panel machines have rather slow setup times and calls placed or received by customers served by ESS machines have rather fast setup times. Terminating ESS machines generally provide immediate ringing when a desired line is not busy, while other machines generally do not. This difference in procedures tends to shorten setup times for calls terminating through ESS machines. Other causes which underly these characteristics have not been uncovered. The precaution of not totally attributing these findings to the panel or ESS machines is exercised because of the absence of routing and toll switching information.

V. CONCLUSIONS

This survey has confirmed that both network performance and customer behavior play a role in determining the overall completion probability and call setup time experienced by telephone customers. In the area of completion performance, the percentage of equipment blockages and failures has been steadily declining in recent years due to extensive efforts throughout the Bell System to provide adequate facilities to achieve low blocking and to detect and correct equipment failures. As a result, customer-dependent failures now account for at least 85 percent of all failures to complete. Thus the greatest potential for improved completion performance lies in reducing such failures. This may be brought about, for instance, through the offering of new services made possible by the advent of Electronic Switching Systems which will serve to reduce the occurrence of called customer did not answer and called line busy failures.

One manifestation of the dominance of customer behavior on completions is the important influence of class of subscriber service. Many trends by day of week, time of day, and calling distance can be directly

attributed to changes in the relative proportion of business and residence traffic.

The total call attempt time is made up of several components, most of which are customer-controlled, such as dialing time and time from start of ringing to answer by the called customer. The one network-dependent component, time from end of dialing to network response, represents about 29 percent of the total call attempt time. Calling distance and type of local switching have both been shown to affect this interval. The advent of common channel interoffice signaling is expected to reduce this component from its current value of 10.9 seconds to 2 or 3 seconds, which will reduce the overall call setup time by 20–25 percent.

Customer holding time after the onset of ringing and after a positive indication of failure to complete has a significant effect on total call attempt time and hence on network load. Programs directed toward encouraging faster customer abandonment or toward releasing network facilities more promptly after receipt of busy of no circuit/reorder signals would reduce this load on network facilities.

VI. ACKNOWLEDGMENTS

The authors wish to thank Mrs. Carol Hassol who faithfully assisted us during the first half of this study. We also thank Anna Marron of AT&T and Bertha Salo formerly of AT&T for coordinating the interactions between Bell Laboratories and the individual operating telephone companies. Finally, we recognize that this survey would not have been possible without the individual efforts of many Bell System employees in every operating company.

REFERENCES

1. Notes on Distance Dialing, Section 3, New York: American Telephone and Telegraph Company, 1975.
2. W. A. Larsen and R. McGill, "Attempt Ratios and Operating Times for Long Distance Toll Traffic," unpublished work, 1971.
3. G. W. Riesz, "Factors Influencing the Call Completion Ratio," 8th International Teletraffic Congress (ITC), 1976.
4. W. S. Hayward, Jr., and R. I. Wilkinson, "Human Factors in Telephone Systems and Their Influence on Traffic Theory, Especially with Regard to Future Facilities," 6th ITC, 1970.
5. R. I. Wilkinson and R. C. Radnick, "The Character and Effect of Customer Retrials in Intertoll Circuit Operation," 5th ITC, 1967.
6. M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, Vols. I and II, New York: Wiley, 1953.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Physical and Transmission Characteristics of Customer Loop Plant

By L. M. MANHIRE

(Manuscript received May 25, 1977)

A recent Bell System customer loop survey was jointly conducted by AT&T and Bell Laboratories. Bell Labs developed the sampling plan and performed the data analysis while AT&T assumed responsibility for the sample selection, data collection, and all intermediate data processing prior to analysis. This paper presents the principal physical composition and calculated transmission statistical characterization of customer loop plant as defined by that survey. Comparisons of data obtained from this survey and a similar survey made in 1964 are also presented.

I. INTRODUCTION

A recent Bell System customer loop survey was jointly conducted by AT&T and Bell Laboratories. Bell Labs developed the sampling plan and performed the data analysis while AT&T assumed responsibility for the sample selection, data collection, and all intermediate data processing prior to analysis.

This paper presents the principal results of this Bell System customer loop survey. This survey provides a statistical characterization of both the physical composition and the calculated transmission characteristics of customer loop plant. Comparisons of data obtained from the survey and a similar survey made in 1964 are also presented.

The characterization of the Bell System customer loop plant as defined by the 1964 loop survey was published in 1969 by P. A. Gresh.[1]

The current Bell System survey consisted of a simple random sample of 1100 main stations selected from the population of all Bell System main stations. Lines serving official telephones, foreign exchanges, dial teletypewriter exchanges (TWX), and special services were, however, omitted from the survey, as it was felt their design would not be representative of customer loop plant.

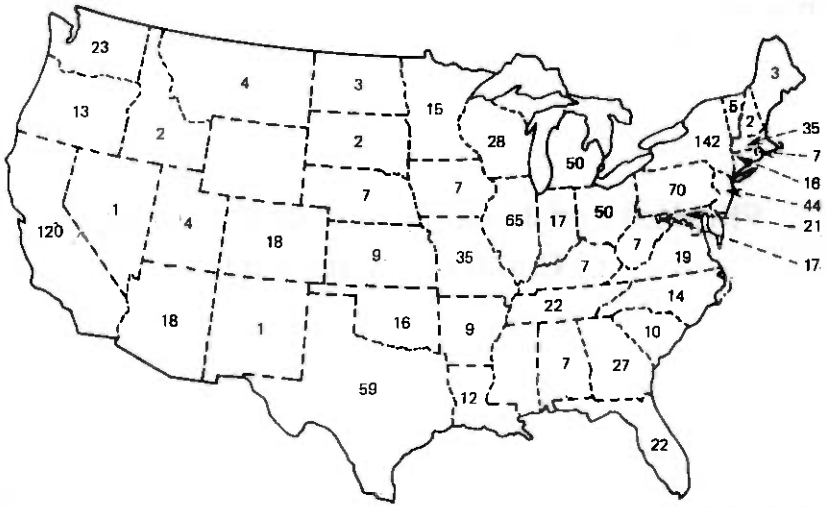


Fig. 1—Geographic distribution of sampled loops, 1973 customer loop survey.

Analysis of the 1100 sampled main stations consisted of encoding and keypunching the raw data to enable computer processing. A number of computer programs were then accessed which in addition to tabulating the physical composition of subscriber loops also calculated and evaluated the transmission performance of the customer loop plant.

II. DESIGN OF THE SURVEY

Definition of the sampling population, sampling plan, and sample size were based on the statistical model used in the 1964 loop survey. The population for the current survey was defined as all 57,293,521 Bell System main stations as of January 1, 1973. A simple random sample was drawn with the size determined so as to provide precision equal to that obtained in the 1964 survey. The precision was measured by the width of the 90% confidence interval for the estimated mean working length. The desired confidence interval (at 90 percent confidence level) of ±450 feet on the average cable distance to the sampled main stations dictated a sample size of 1100 randomly selected main stations. The actual sample produced a ±490 foot confidence interval. The survey had a wide geographical dispersion with every state except Wyoming contributing to the sample and heavier contributions from the metropolitan areas (Fig. 1).

III. LOOP SURVEY RESULTS—PHYSICAL COMPOSITION

Data obtained in the 1973 loop survey included a detailed physical makeup diagram of each of the sampled loops. Distributions of such

Table I — 1973 Loop Survey: Summary of main station characteristics

Main station quantity	Mean 1964	Mean 1973	90% Confidence limits on mean (±) 1964	90% Confidence limits on mean (±) 1973	Significant level for difference of means, %
Working length	10,613 ft	11,413 ft	476 ft	490 ft	99
Total bridged tap	2,478 ft	1,821 ft	172 ft	113 ft	99+
Working bridged tap	228 ft	121 ft	74 ft	53 ft	95
Airline distance	7,758 ft	8,410 ft	386 ft	395 ft	95
Working length airline distance	1.50	1.51	.03	.05	<80

Note: Drop wire excluded except when individual lengths exceed 400 feet.

items as type of cable construction, composition by gauge, pair size of cables and grade of service (individual or multiparty service) distribution were generated in addition to such physical quantities as working length to main station and total bridged tap. These quantities are defined in the illustration on the next page. Similar data was gathered in 1964 and a comparison of the two surveys had been made to detect changes in

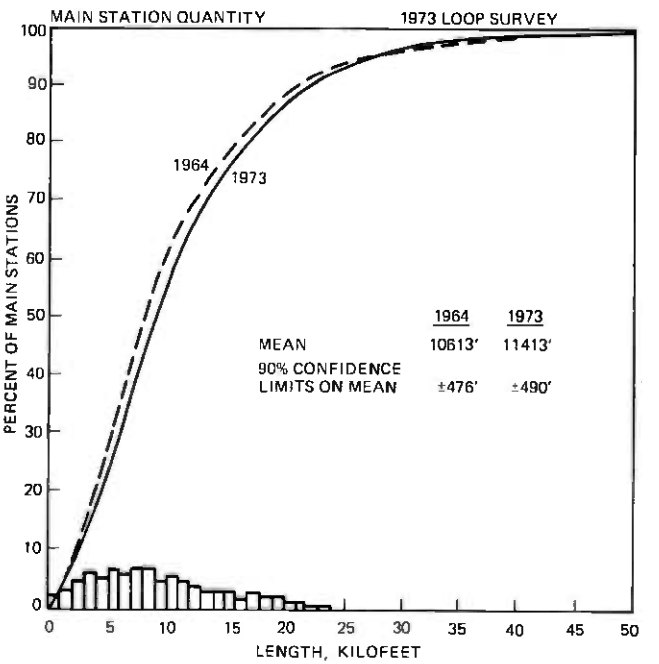


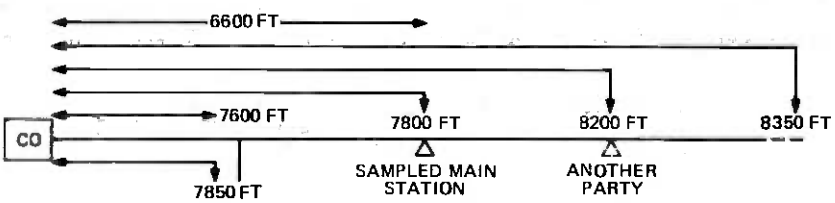
Fig. 2—Working length to main station.

plant composition over the intervening 9 years between the two surveys.

Table I gives the means and confidence limits for the principal physical properties of both the 1964 and 1973 surveys and the level of significance for the observed differences. Cumulative distributions of these quantities are presented in Fig. 2 through 5. The distribution of working bridged tap is not given, since 93 percent* of the sampled main stations were served by loops having zero working bridged tap and consequently the distribution is not particularly useful.

As Table I indicates, the estimated average working length from the central office to the main station is 11.4 kilofeet with 90 percent probability that the true mean value lies within ±490 feet of this estimate. Notice that this is 800 feet longer than the average in 1964 and that this observed increase can statistically be said to be a true increase.

Loop Sketch Showing Main Station Quantities



Main Station Quantities	
Working length to sample (WL)	7800 ft
Total bridged tap (TBT)	800 ft
Working bridged tap (WBT)	400 ft
Airline distance to sample (AL)	6600 ft
Working distance/airline distance (ratio)	1.238

Analyses of principal loading characteristics of loop plant are given in Table II. As indicated, the percent of loaded loops has increased from 16.4 to 22.9 percent. The majority of this 6.5 percentage point increase is attributable to an increase in loops requiring loading (main stations served by loops over 18 kilofeet in length). A second contribution to the increase is attributed to the fact that greater care has been taken to assure that loops in excess of 18 kilofeet are loaded (long nonloaded loops have decreased from 1.5 percent to 1.0 percent). The final contribution

* 92 percent of the sampled main stations were individual lines and 14 percent of the 8% party-line stations were working alone.

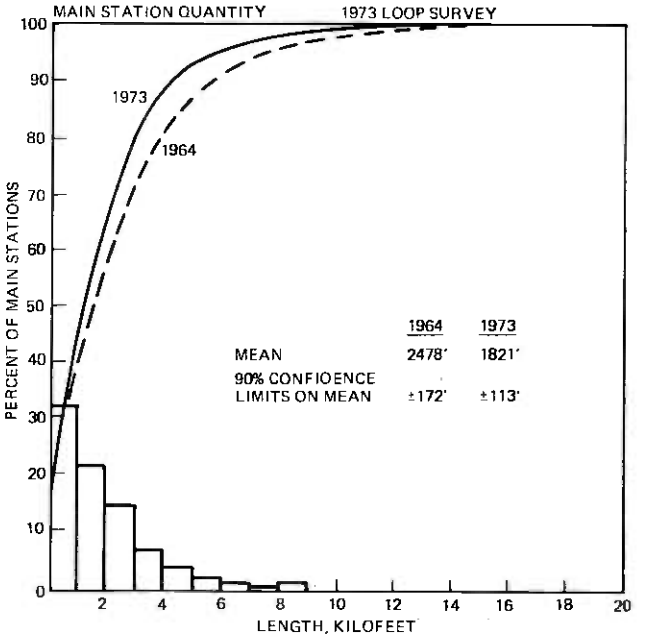


Fig. 3—Total bridged tap.

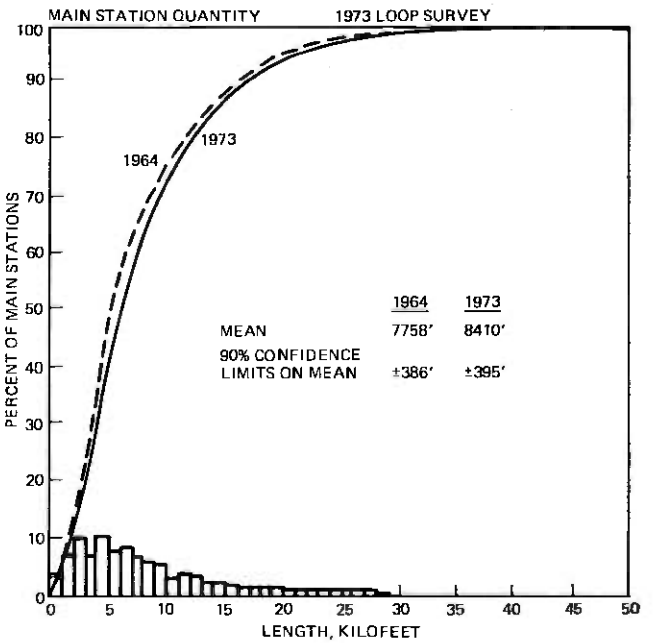


Fig. 4—Airline distance to main station.

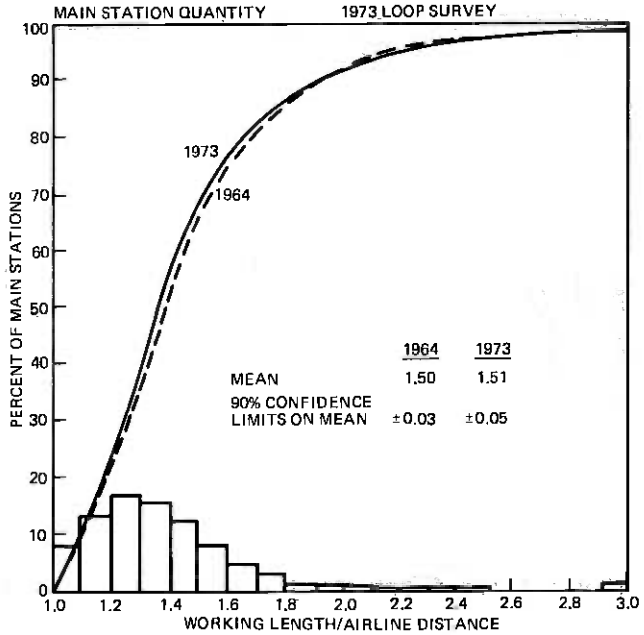


Fig. 5—Ratio of working length to airline distance.

to the increase in loaded loops is the fact that the percentage of main stations served by short loaded loops has increased (from 4.6 percent to 6.7 percent). This increase in short loaded loops can not be accounted

Table II — 1973 Loop survey: Summary of principal loading information

Main stations served by	1964 % MS	1973 % MS	1973 90% Confidence Interval (±%)	1973 Change Since 1964 Level of Significance
Loaded loops	16.4	22.9	2.1	99+
With H88	15.5	22.6	2.1	99+
With loading other than H8	0.9	0.3	0.2	93
Less than 18 kf central office to main station and loaded*	4.6	6.7	1.2	97
Having load spacing deviations				
exceeding ±500 feet	5.3	8.6	1.4	99+
exceeding ±120 feet†	NA	18.5	1.9	—
Nonloaded loops	83.6	77.1	2.1	99+
Long nonloaded loops (>18 kilofeet)	1.5	1.0	0.5	<80

† The normal confidence intervals for percentages close to 0 are not meaningful and so are indicated by (—).

* Based on distance from CO. to sampled main station and does not necessarily indicate improper design.

for by an increase in PBX/centrex services (PBX/centrex service has increased from 1.8 percent in 1964 to 2.5 percent in 1973).

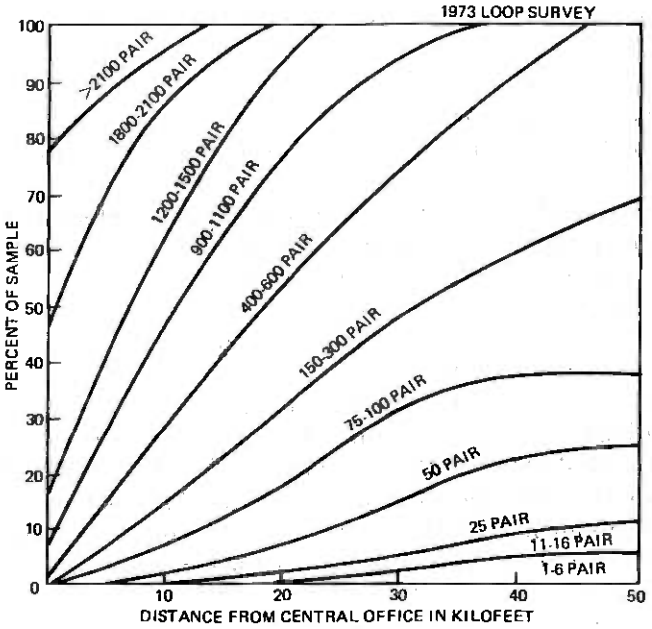
Survey data analysis also provides information on loop plant composition by gauge, pair size, type of construction (aerial, buried, or underground), and grade of service as a function of distance to the sampled main station.

The sampled loops were examined at 1 kilofoot intervals, beginning at the central office, to obtain the various distributions presented in Figs. 6–9. The gauge distribution at 10 kilofeet, for example (Fig. 7a), was derived by determining the gauge of each loop sample at the 10 kilofoot point of all 494 loops which extend beyond 10 kilofeet. A number of conclusions can be drawn from these analyses. First, inspection of pair sizes as a function of distance from the central office (Fig. 6a) reveals decreasing pair size with distance. But also notice that there has been a shift toward the use of larger pair size cables for longer distances as compared to 1964 (Fig. 6b). For example, at 30 kilofeet only 47 percent of the loops are in pair sizes of 100 pairs or less in 1973 while in 1964, at the same distance, nearly 75 percent of the loops were served by cables of 100 pairs or less. Similarly, examination of the gauge distribution (Fig. 7a) shows a transition to coarse gauge with increasing distance (at 20 kilofeet only 30 percent of the sampled loops consist of finer than 22 gauge cable). It can also be seen (Fig. 7b) that there has been a shift since 1964 toward use of the finer gauge cables at further distances from the central office. For example at 10 kilofeet, 28 percent of the loops sampled in 1973 were contained in 26 gauge cables while in 1964 only 15 percent of the loops were contained in 26 gauge cables at this distance.

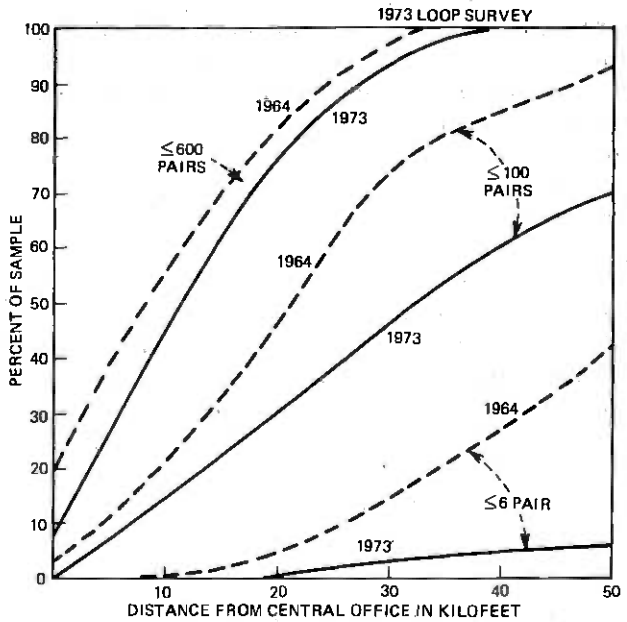
As shown in Fig. 8a, longer loops are predominately either aerial or buried (at 35 kilofeet, 52 percent of the sampled loops were constructed with buried facilities and 48 percent with aerial). Figs. 8b–d show that in the intervening 9 years between the two loop surveys there has been an increase in buried cable coupled with decreasing use of aerial facilities with greater distance from the central office. For example, in the 1973 survey at the distance of 20 kilofeet, 36 percent and 49 percent of all plant was buried and aerial respectively, while in 1964 the values were 12 percent for buried and 72 percent for aerial (aerial facilities consist of both wire and cable). Noteworthy also is the grade of service distribution (Figs. 9a–c), which shows with greater distance from the central office the increasing use of individual party service coupled with more uniform distribution of two-, four-, and eight-party service than was prevalent in 1964.

IV. LOOP SURVEY RESULTS—TRANSMISSION PERFORMANCE

The 1973 loop survey data also provided sufficient information to derive the equivalent "T" network for each loop which was then used

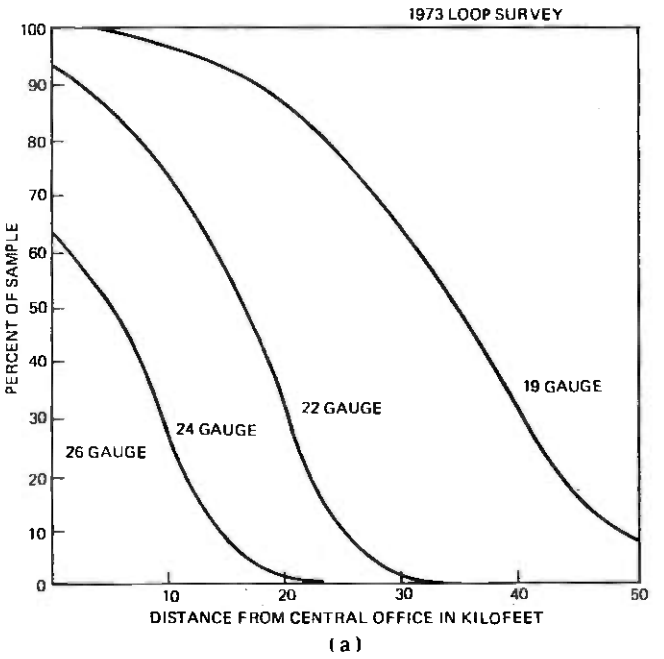


(a)

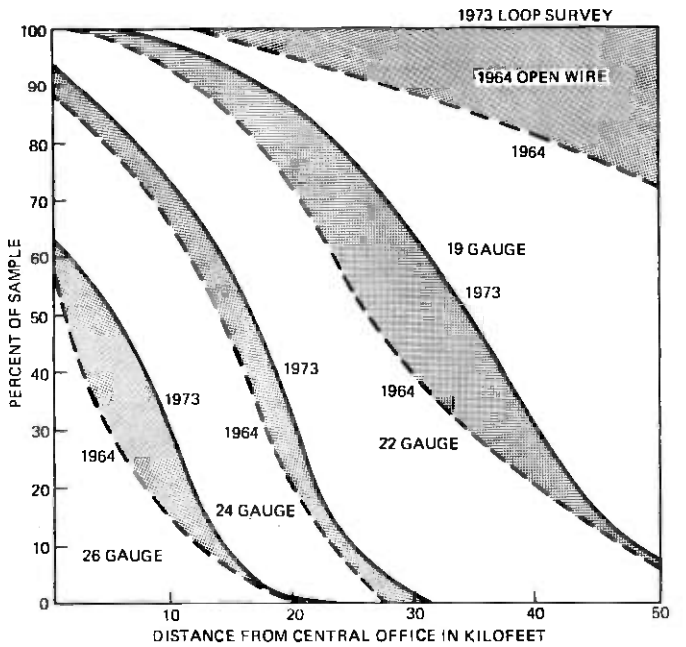


(b)

Fig. 6—(a) Pair size distribution. (b) Pair size distributions, 1973 vs. 1964 surveys.



(a)



(b)

Fig. 7—(a) Gauge distribution. (b) Gauge distributions, 1973 vs 1964 surveys.

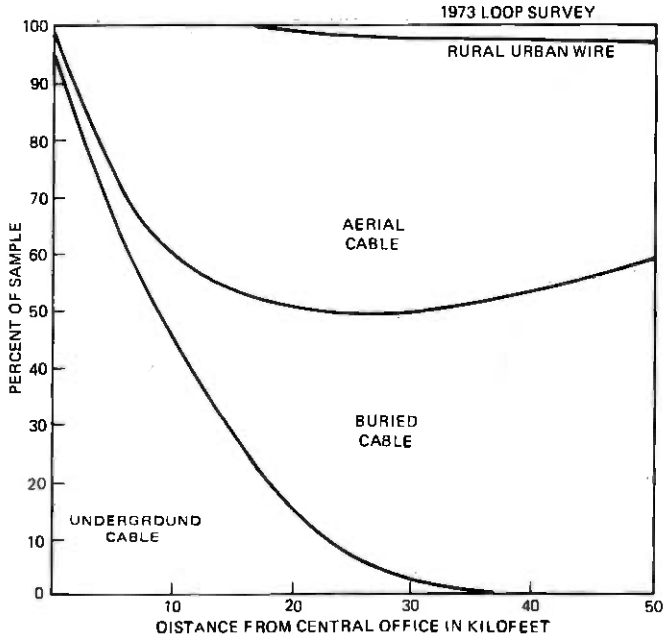
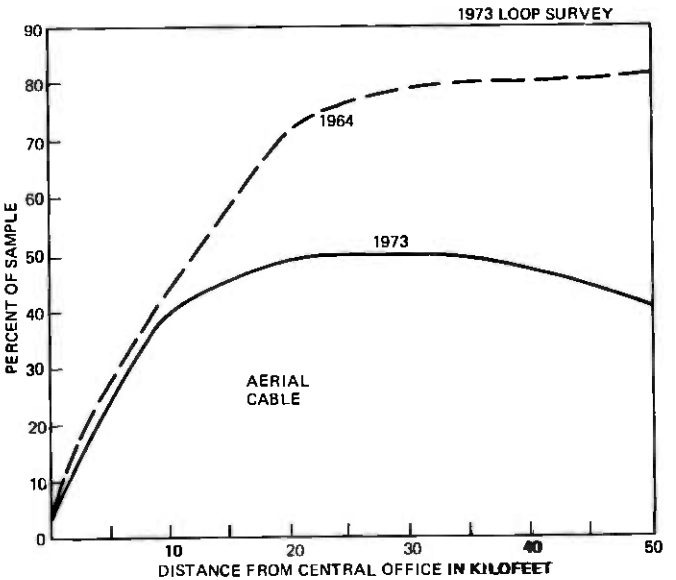
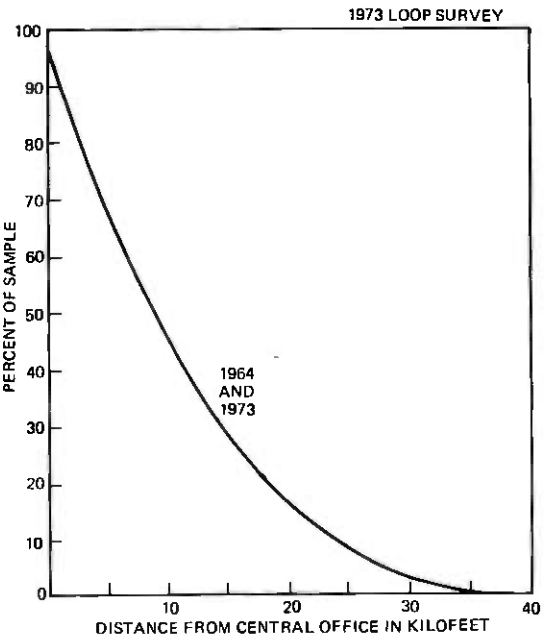


Fig. 8—Type of construction. (a) Distribution. (b) Buried cable.

to derive the calculated transmission loss at nine voice-band frequencies. In the 1964 loop survey, the loss of each sampled loop was determined by two methods: theoretical calculations as described above, and actual measurements. Comparison of the calculated and measured transmission losses showed that calculated losses based on the physical composition

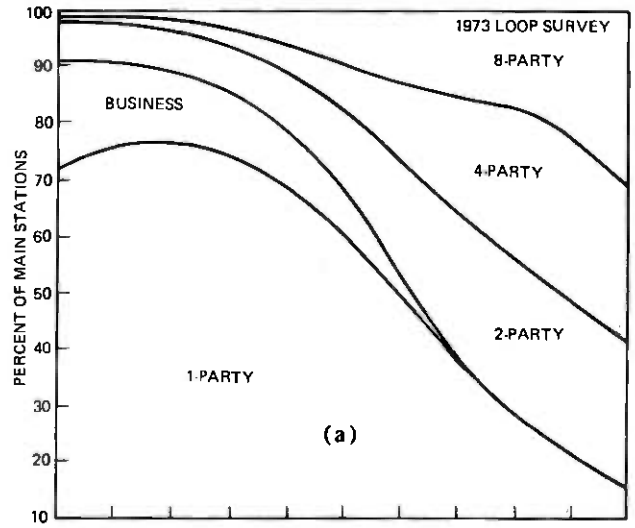


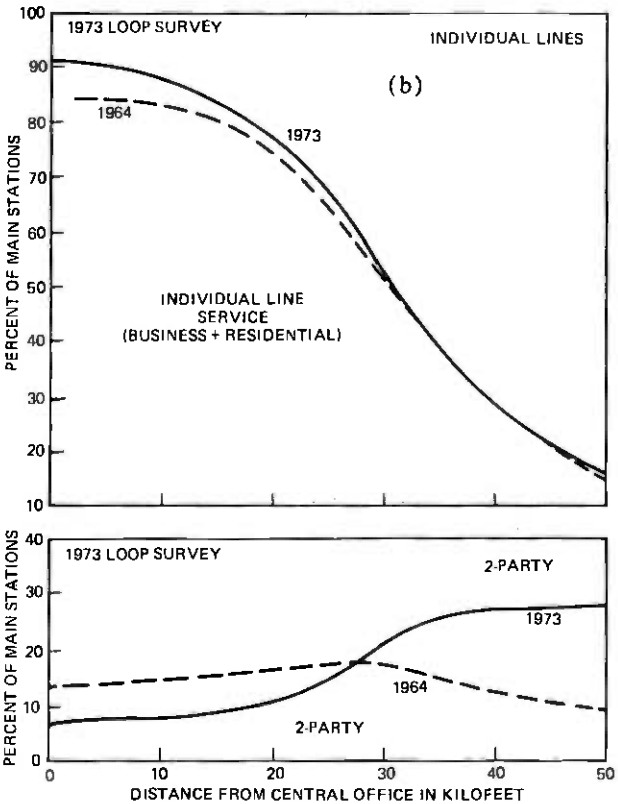
(c)

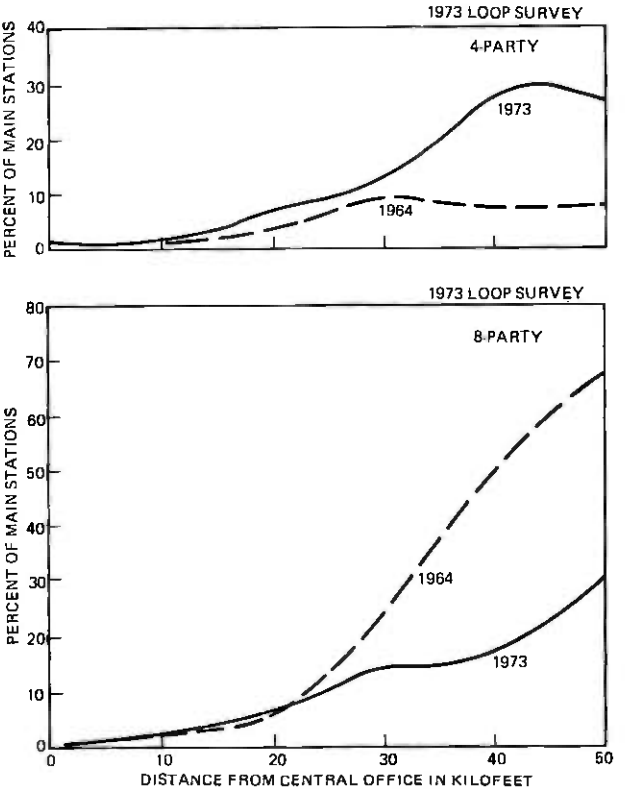


(d)

Fig. 8—(c) Aerial (cable and wire). (d) Underground cable.


1973 LOOP SURVEY
8-PARTY
BUSINESS
4-PARTY
2-PARTY
1-PARTY
(a)
PERCENT OF MAIN STATIONS
100
90
80
70
60
50
40
30
20
10


1973 LOOP SURVEY
INDIVIDUAL LINES
(b)
1964
1973
INDIVIDUAL LINE
SERVICE
(BUSINESS + RESIDENTIAL)
PERCENT OF MAIN STATIONS
1973 LOOP SURVEY
2-PARTY
1973
1964
2-PARTY
PERCENT OF MAIN STATIONS
0
10
20
30
40
50
DISTANCE FROM CENTRAL OFFICE IN KILOFEET



(c)

Fig. 9—Type of service. (a) Distribution. (b) Individual lines and two-party. (c) Four-party and eight-party. One-, two-, four-, and eight-party include residence service only. Business includes PBX, centrex, and coin.

of the loops were sufficiently close to the measured transmission losses. Consequently, actual transmission measurements were not made on the sampled loops in the 1973 survey. The distribution of calculated insertion losses at 1, 2, and 3 kHz, a scatter diagram of the 1 kHz insertion losses, return loss at 3 kHz, echo return loss (equal weighting over the 500 to 2500 Hz band), and dc resistance are presented in Figs. 10–13.

The cumulative distributions of calculated 1, 2, and 3 kHz insertion losses for customer loop plant is presented in Fig. 10. Note that approximately 98 percent of all Bell System main stations are served by loops having 1 kHz insertion loss less than or equal to 8 dB with a calculated mean loss of 3.7 dB. Similarly, at 3 kHz the calculated mean loss is 7.5 dB with 95 percent of all main stations served by loops having less than 15 dB insertion loss. A scatter diagram of 1 kHz calculated insertion

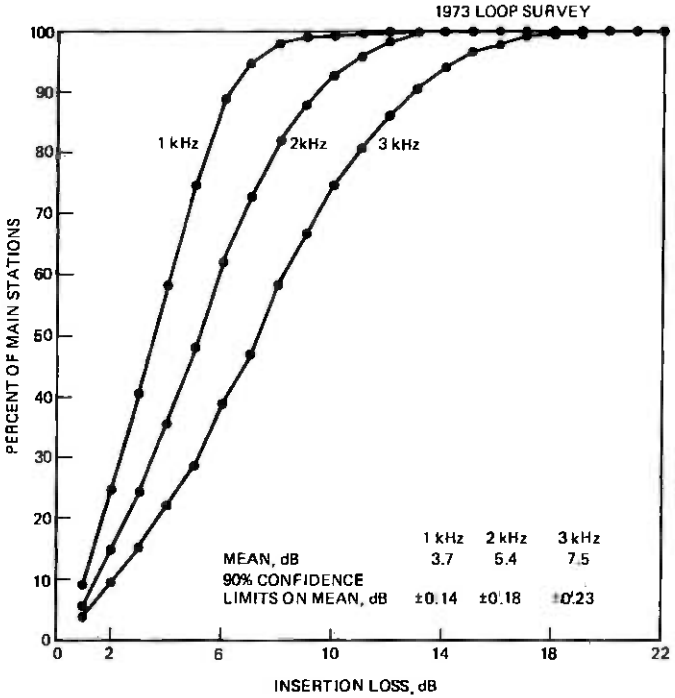


Fig. 10—Insertion loss.

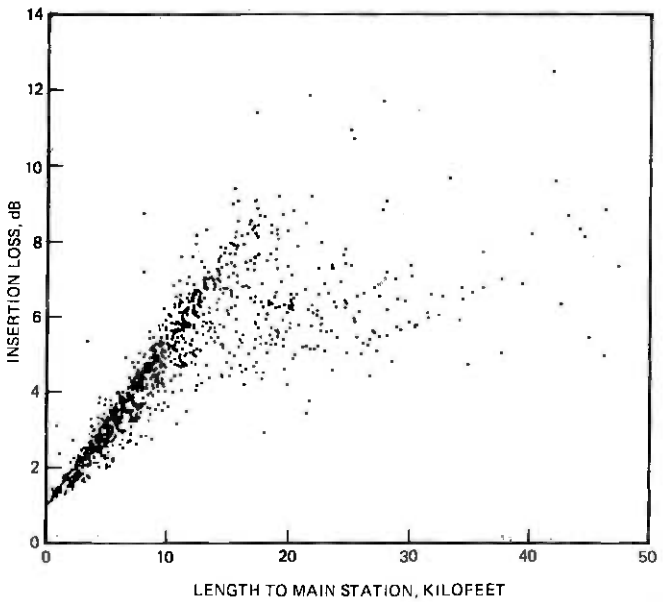


Fig. 11—Scatter diagram of calculated 1-kHz insertion losses.

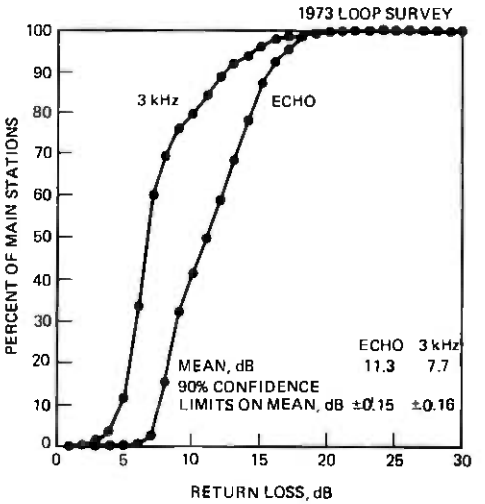


Fig. 12—Return loss.

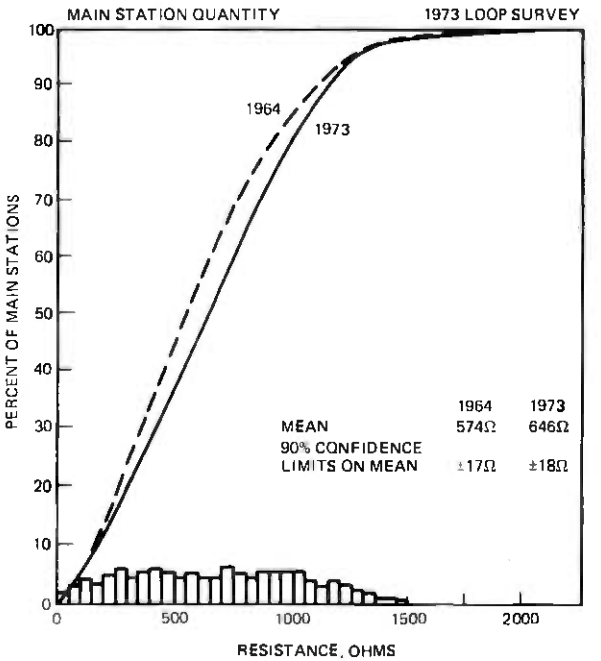


Fig. 13—Calculated resistance to main station.

loss as a function of loop length is shown in Fig. 11. Table IV presents calculated insertion loss at nine voiceband frequencies. Note that mean insertion losses across the voice band have increased, principally because

Table IV — Insertion loss (calculated with 900-ohm terminations at both ends)

Freq., Hz	1964 Mean, dB	1964 90% Confidence Interval, dB (±)	1973 Mean, dB	1973 90% Confidence Interval, dB (±)	St. Dev., dB	Difference 1964–1973 Significance Level
200	2.4	0.06	2.9	0.16	3.2	99+
300	2.5	0.07	2.8	0.10	2.1	99+
500	2.7	0.08	3.0	0.11	2.3	99+
1000	3.5	0.10	3.7	0.14	2.8	94
1500	4.3	0.13	4.6	0.16	3.2	98
2000	5.3	0.16	5.4	0.18	3.6	<80
2500	6.2	0.18	6.4	0.21	4.1	<80
3000	7.3	0.21	7.5	0.23	4.7	<80
3200	7.9	0.23	8.3	0.26	5.2	93

of an increase in the use of finer gauge cable and the resulting increase in average loop resistance (Fig. 13).

Calculated return loss performance is presented in Table V for nine frequencies and the cumulative distributions for the 3 kHz return loss and echo return loss is shown in Fig. 12. Note that it has remained essentially unchanged since 1964.

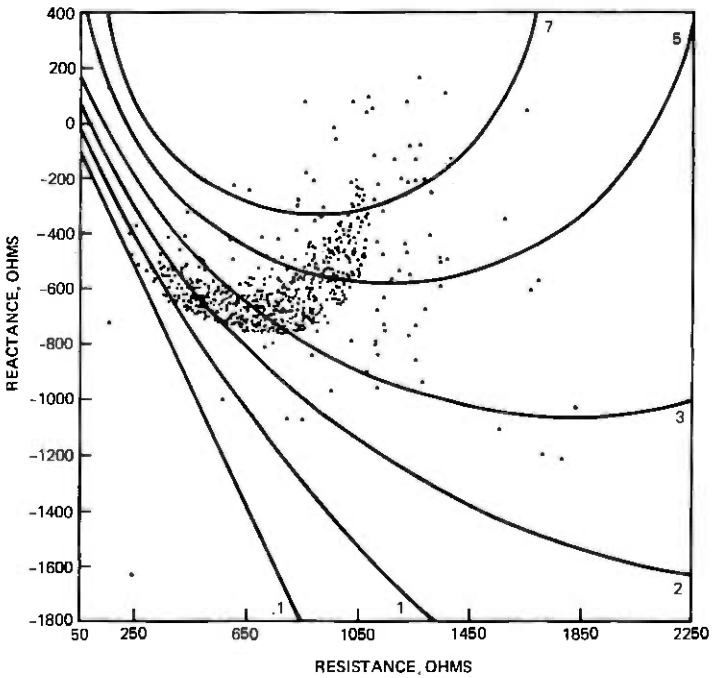


Fig. 14—Nonloaded loop input impedances at 1 kHz. *RL* circles based on 500-type telephone set impedance (22-gauge H88 cable termination at the central office).

Table V — Return loss (calculated using a 900 ohm + 2μF termination)

Freq., Hz	1964 Mean, dB	1964 90% Confidence Interval, dB (±)	1973 Mean, dB	1973 90% Confidence Interval, (dB (±)	St. Dev., dB	Difference 1964–1973 Significance Level
200	8.0	0.11	8.3	0.10	2.0	99+
300	10.2	0.12	10.3	0.11	2.2	<80
500	13.4	0.17	13.1	0.16	3.2	97
1000	15.4	0.30	15.3	0.31	6.2	<80
1500	13.1	0.27	13.3	0.28	5.6	<80
2000	10.9	0.25	11.2	0.27	5.3	82
2500	9.1	0.20	9.1	0.20	4.0	<80
3000	7.7	0.16	7.7	0.16	3.2	<60
3200	7.1	0.15	7.0	0.15	2.9	<60
Echo	11.2	0.15	11.3	0.15	3.1	<80

The final transmission characteristic to be presented is the loop input impedance at the station set. The input impedance of each loop was computed by using the equivalent "T" network for each loop.

Central office terminations equivalent to four-wire trunks and two-wire trunks were represented by 900 ohms plus 2.16 μF and midsection

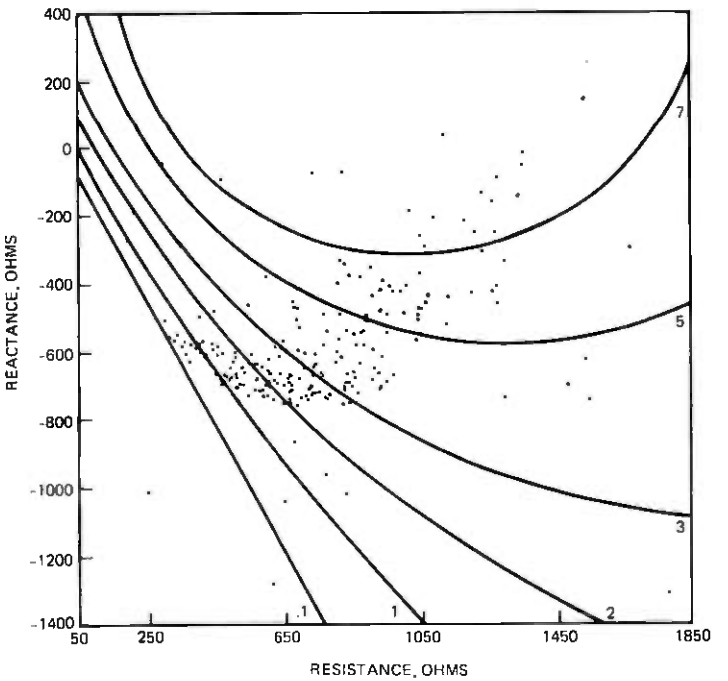


Fig. 15—Loaded loop input impedances at 1 kHz. *RL* circles based on 500-type telephone set impedance (22-gauge H88 cable termination at the central office).

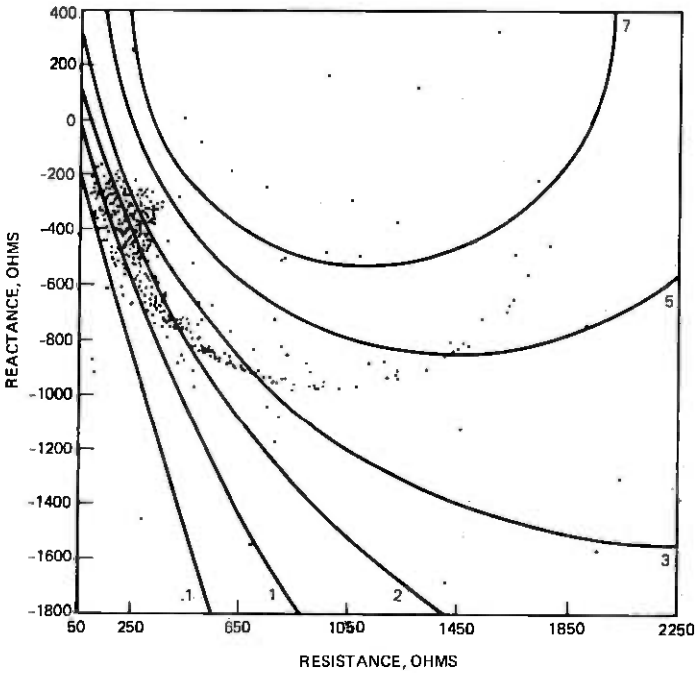


Fig. 16—Nonloaded loop impedances at 3 kHz. *RL* circles based on 500-type telephone set (22-gauge H88 cable termination at the central office).

input impedance of 22-gauge H88 loaded cable, respectively. Intraoffice calls were simulated based on a Monte Carlo technique of selecting 500 pairs of loops. Using a random number generator a loop was chosen from the 1100 loops in the 1973 loop survey data base. The central office termination was the input impedance as measured at the central office of another randomly chosen loop.

Scatter diagrams of loop input impedance at the station set are presented for 1 and 3 kHz voiceband frequencies. There is wide variation in the input impedance of loaded and nonloaded loops; consequently the two populations (849 nonloaded and 251 loaded loops) were separated for this study. Return loss circles at 0.1, 1, 2, 3, 5, and 7 dB are also superimposed on the scatter diagrams. The return loss circles were based on the 500-type telephone set as the reference impedance since 84.3 percent of the sampled loops in the 1973 loop survey were terminated with this type of telephone set. Further, since loaded loops have nearly 15 mA lower loop current than nonloaded loops, the mean loop currents of the individual populations were used.

The scatter diagrams for loops with simulated two-wire trunk terminations (22-gauge H88 loaded cable) are presented in Figs. 14 through 17. The dispersion of points can be partially attributed to the use of

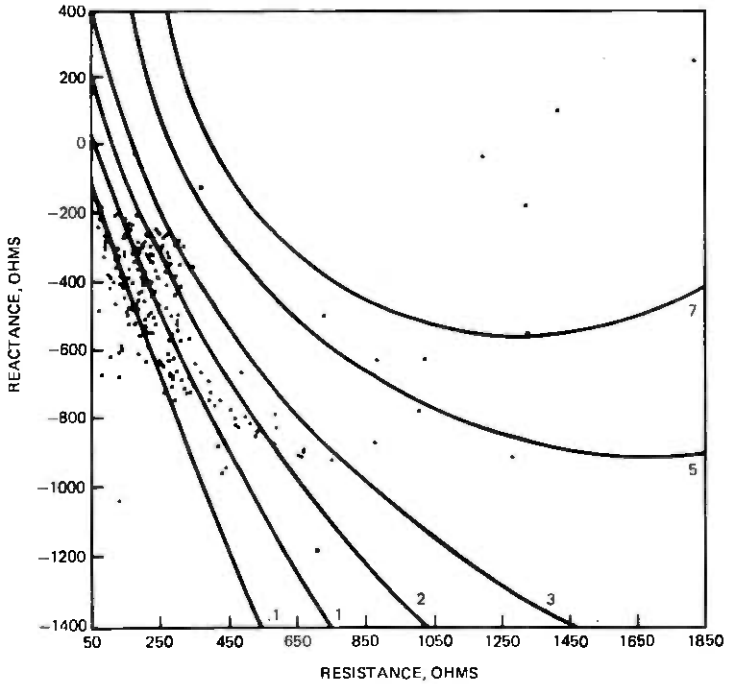


Fig. 17—Loaded loop input impedances at 3 kHz. *RL* circles based on 500-type telephone set impedance (22-gauge H88 cable termination at the central office).

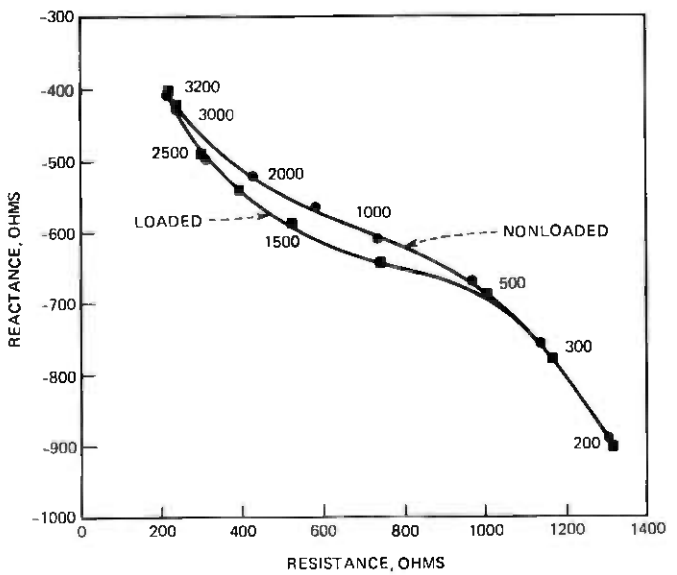


Fig. 18—Median input impedances at all nine frequencies for loaded and nonloaded loops (22-gauge H88 cable termination at the central office).

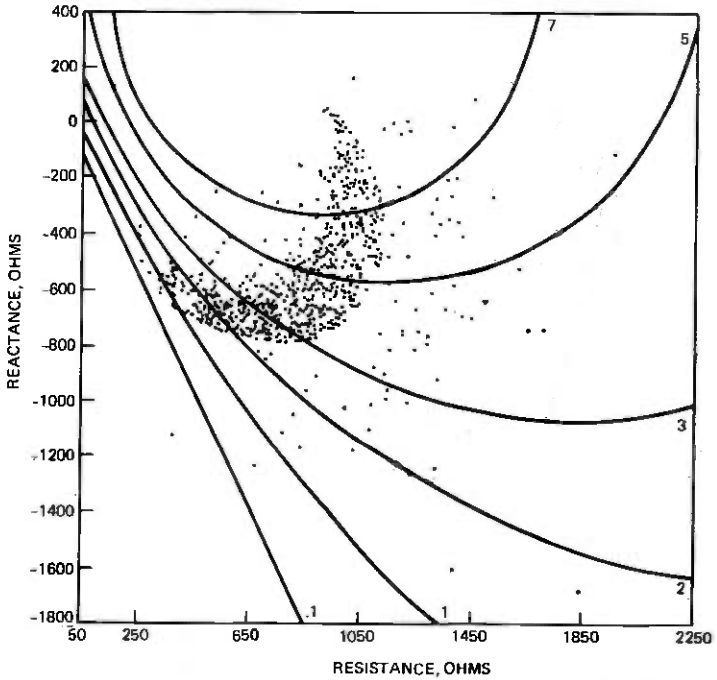


Fig. 19—Nonloaded loop input impedances at 1 kHz. *RL* circles based on 500-type telephone set impedances (900 ohm + 2.16 μF cable termination at the central office).

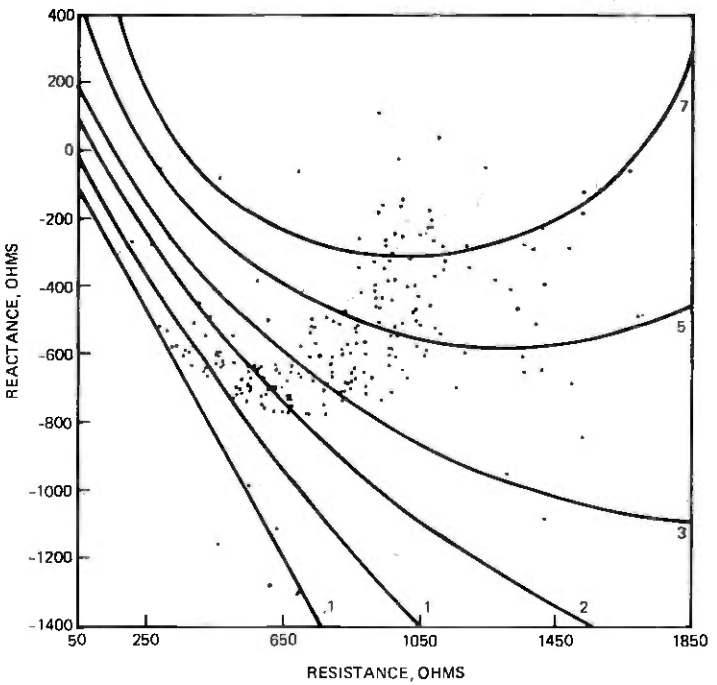


Fig. 20—Loaded loop input impedances at 1 kHz. *RL* circles based on 500-type telephone set impedances (900 ohm + 2.16 μF cable termination at the central office).

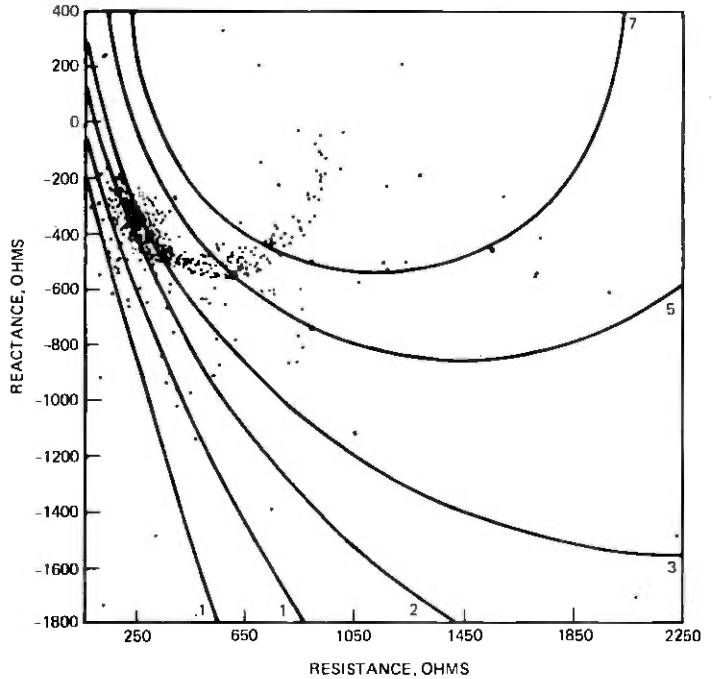


Fig. 21—Nonloaded loop input impedances at 3 kHz. *RL* circles based on 500-type telephone set impedances (900 ohm + 2.16 μF cable termination at the central office).

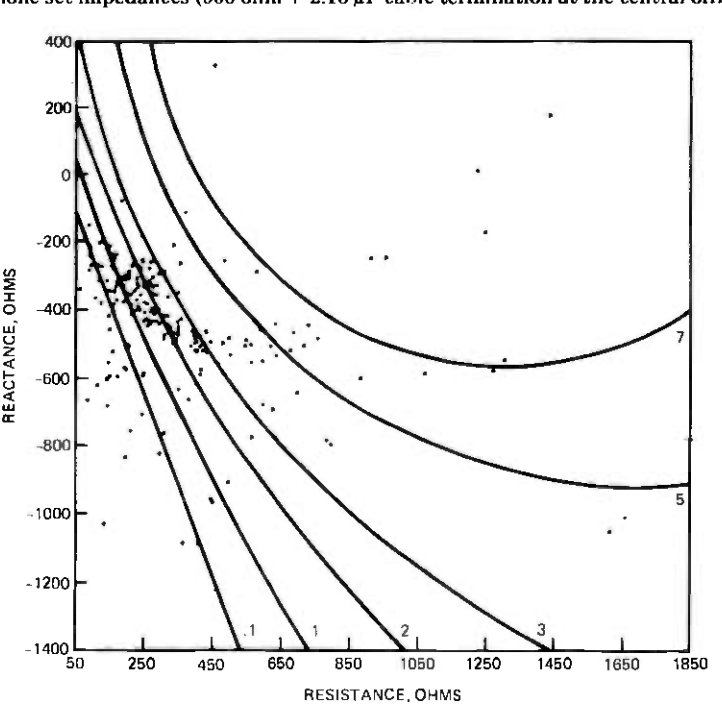


Fig. 22—Loaded loop input impedances at 3 kHz. *RL* circles based on 500-type telephone set impedances (900 ohm + 2.16 μF cable termination at the central office).

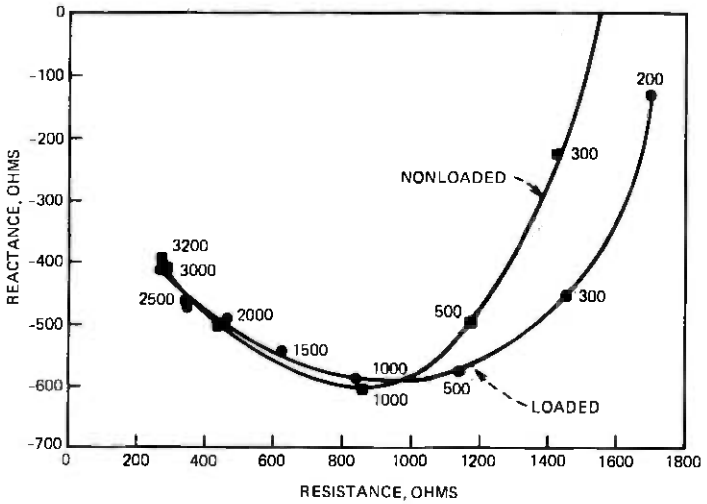


Fig. 23—Median input impedances at all frequencies for loaded and nonloaded loops (900 ohm + 2.16 μF cable termination at the central office).

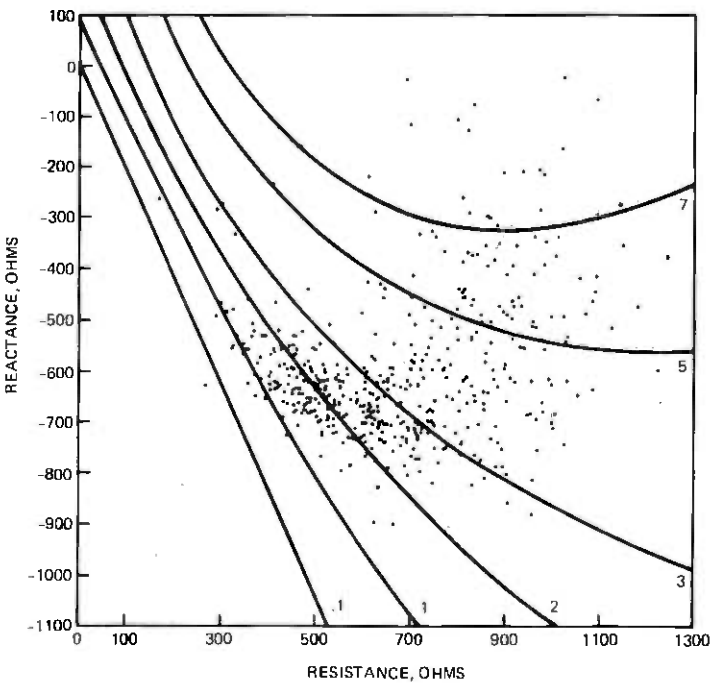


Fig. 24—Intraoffice simulation input impedances at 1 kHz. *RL* circles based on 500-type telephone set impedance.

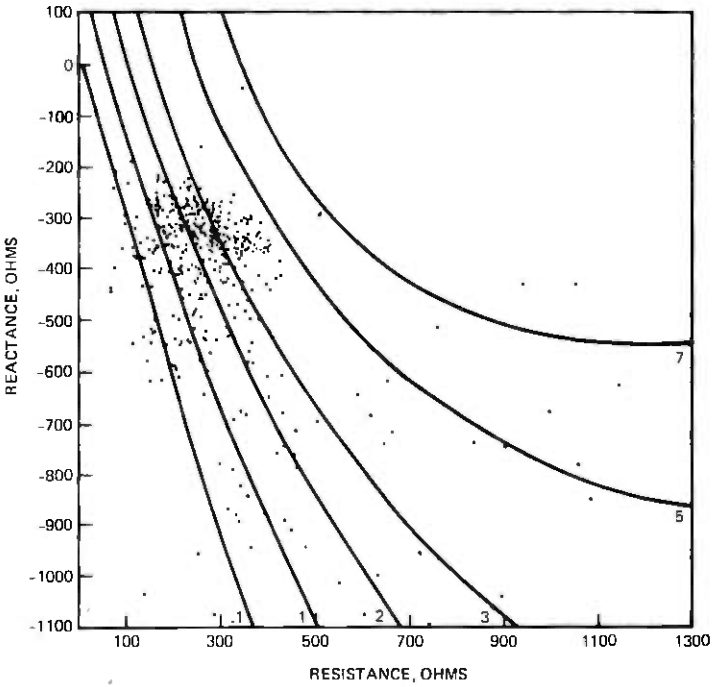


Fig. 25—Intraoffice simulation input impedances at 3 kHz. *RL* circles based on 500-type telephone set impedance.

coarser gauge cable than is required to meet the resistance design objective of 1300 ohms. Notice that loaded loops (Figs. 15 and 17) are more scattered due to their vulnerability to load coil spacing irregularities than are the nonloaded loops (Figs. 14 and 16). Median impedances at all nine voiceband frequencies for simulated two-wire trunk terminations are shown in Fig. 18. The scatter diagrams for loops with simulated four-wire trunk terminations (900 ohms plus 2.16 μF) are shown in Figs. 19 through 22. Similarly, the scatter plots indicate more irregularities present in loaded loops than nonloaded ones. Curves of the median input impedances at all nine voiceband frequencies of loaded and nonloaded loops with a four-wire trunk termination at the central office are presented in Figure 23. The intraoffice call simulation results can be seen in the scatter diagrams presented in Figs. 24 and 25. These reveal the effect of connecting together two random loops, one terminated by a station set and the other's impedance calculated at the station set. The median input impedances for intraoffice calls are presented in Fig. 26 along with median impedances of simulated two-wire and four-wire trunk terminations of nonloaded loops.

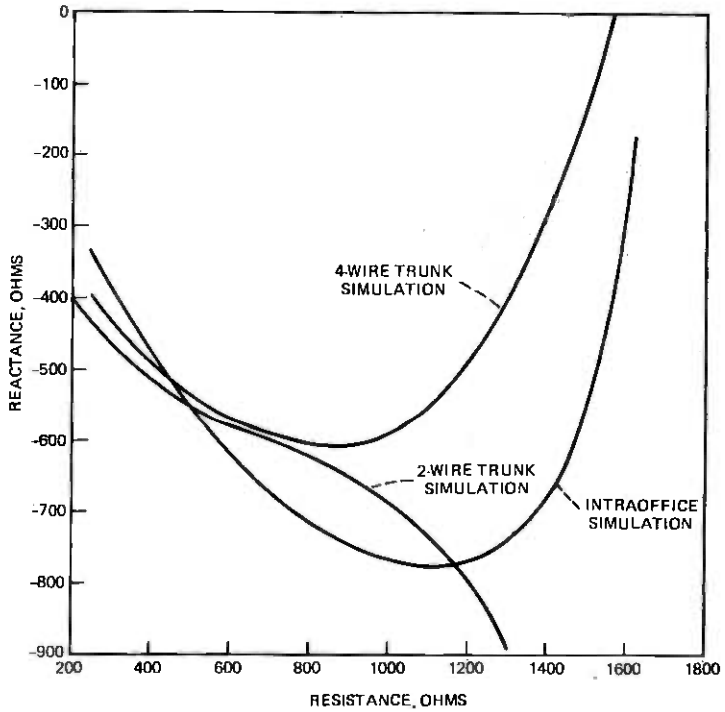


Fig. 26—Median input impedances at all nine frequencies for simulated intraoffice calls, simulated two-wire trunk terminations (22-gauge H88 loaded cable), and simulated four-wire trunk terminations (900 ohm + 2.16 μF cable) of nonloaded loops.

V. CONCLUSIONS

The analyses of the 1973 loop survey can be summarized in the following four principal results.

(*i*) The average customer loop length is currently 11.4 kilofeet with only 4 percent of the main stations located beyond 30 kilofeet from their serving office. The length distributions show a statistically significant (99 percent level of significance) increase in average loop length since 1964, with the average loop length increasing by 800 feet.

(*ii*) The average calculated 1 kHz insertion loss of the Bell System loop plant is currently 3.7 dB* with 95 percent of all main stations being served by loops having a calculated 1 kHz loss of less than 7.5 dB. At 3 kHz the average calculated loss is 7.5 dB with 95 percent of all main stations served by loops having less than 15 dB insertion loss.

* Anomalies in plant design and record errors will tend to make actual losses slightly higher than calculated losses. For example, the 1964 loop survey, which included actual loss measurements as well as calculated losses, showed that average measured loss at 1 kHz was approximately 0.3 dB higher than the calculated value.

(*iii*) The percentage of main stations served by loaded loops has increased since 1964 from 16.4 to 22.9 percent. Sixty percent of the increase is a direct result of the increase in the percent of customer loops requiring loading (loops longer than 18 kilofeet). The remainder is attributed to an increase in the use of loading for loops shorter than 18 kilofeet and a reduction in the percentage of nonloaded loops longer than 18 kilofeet.

(*iv*) Average dc resistance to the sampled main station has increased from 574 to 646 ohms since 1964. This increase is attributed to the trend towards the use of finer gauge cable in the loop plant combined with an increase in the average customer loop length.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge N. Reina and C. Schroeder of AT&T who selected the sample, collected the data, and verified its accuracy. Without their contributions this paper could not have been written.

REFERENCE

1. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3385.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Level Reassignment: A Technique for Bit-Rate Reduction

By B. PRASADA, F. W. MOUNTS, and A. N. NETRAVALI

(Manuscript received June 17, 1977)

Three techniques are described for reducing the entropy of a digitally coded monochrome picture signal by dynamically changing the input-output relationships of a single quantizer. This is done by reassigning the input levels of the quantizer to different representative output levels in such a way as to reduce the entropy of the quantized output, while keeping the visibility of the resultant quantization noise below a certain specified threshold. These techniques are simulated both in the pel and the transform domain for a simple differential pulse code modulation system and Hadamard transform coding system, respectively. It is found that they reduce the entropy by about 20 to 25 percent in the pel domain without noticeably sacrificing the picture quality. In the Hadamard transform domain, the entropy of the first transform coefficient can be decreased by about 23 percent with little change in picture quality.

I. INTRODUCTION

It is well known that the statistics of picture signals are nonstationary and that the required fidelity of reproduction demanded by the human eye varies from picture element (pel) to picture element. Consequently, for efficient digital representation of pictures, it is desirable to adapt coding strategies to those local properties of the picture signal which determine the visual sensitivity to quantization noise. In this study, we make use of the spatial masking properties of the human observer to adapt the quantization strategies for encoding the picture signal. We define spatial masking as the reduction in the ability of a person to visually discriminate amplitude errors which occur at or in the neighborhood of significant spatial changes in the luminance. To this end, we borrow, from our earlier work, measures of luminance spatial activity in both the transform[1,2] and the pel[3] domain, and the relationships (called the visibility functions) between the amplitude accuracy and

these measures of spatial detail. We use these visibility functions to change dynamically the input-output mapping of a single quantizer to reduce the bit rates. This is done by reassigning the input of the quantizer to a different representative level than normal in such a way as to reduce the entropy of the quantized output, while keeping the visibility of quantization noise below a certain threshold. We demonstrate three different algorithms for level reassignment and evaluate their potential by measuring the entropy of the quantized output for a given picture quality. It is worth pointing out that there are other methods of adapting the quantizer characteristics. Some of these are discussed in Refs. 3, 5, and 6.

1.1 Summary of the approach

To illustrate the coding strategy, let e_{in} and e_{out} be the input and output to a quantizer, whose decision levels are denoted by X_i, $i = 1, \ldots, N + 1$ and representative levels by Y_i, $i = 1, \ldots, N$. The quantizer input-output mapping in the absence of adaptation is:

$$\text{If } X_i < e_{in} \leq X_{i+1}, \text{ then } e_{out} = Y_i, \quad i = 1, \ldots N \tag{1}$$

In the adaptive quantization, instead of representing the input e_{in} as Y_i, we change the representation to Y_j ($j \neq i$) provided the visibility of the quantization noise ($e_{in} - Y_j$) is below a given threshold. The changed representation Y_j is chosen, in general, to decrease the entropy* of the quantizer output. This is done in several different ways. In one case, we increase the frequency of occurrence of the inner levels† of the DPCM quantizer. Here we make use of the fact that the frequency of occurrence of the inner levels of a DPCM quantizer is generally high, and, therefore, making it higher decreases the entropy. In another case, for alternate levels Y_i, Y_j is taken to be Y_{i-1} or Y_{i+1}, depending on the noise visibility. This has the effect of decreasing the frequency of occurrence of alternate quantizer levels, and increasing the frequency of occurrence of the other levels, and thus reducing the entropy. We refer to such a change in the use of the quantizer levels as level reassignment. Spatial detail surrounding the pel and the corresponding visibility function are used in determining whether quantization noise ($e_{in} - Y_j$) is visible. We have studied three different algorithms which differ only in detail from the approach outlined above. They are described in Section III.

---

* We define the entropy of the quantizer output as

$$\epsilon = -\sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_j} \log_2 \left[\frac{n_i}{\sum_{j=1}^{N} n_j} \right]$$

where n_i is the frequency of occurrence of level Y_i.

† Inner levels are those which are close to level zero.

1.2 Summary of results

We simulated the level reassignment algorithms on a digital computer. The efficiency of an algorithm was measured by evaluating the difference in entropy between a nonadaptively coded picture and an adaptively coded (using level reassignment) picture having the same quality. In the pel domain, reassigning all the quantizer levels to the lowest level permitted by the quantization noise visibility leads to degradation at edges in the form of slope overload and, therefore, allows only about 9 percent reduction in entropy before picture quality is degraded. To avoid degradation of these edges, only a few inner levels (three inner levels on both sides of zero of a 15-level quantizer) were reassigned. This allows a higher threshold for the quantization noise before it is visible and leads to about 14 percent decrease in entropy. Alternate reassignment, in which certain alternate levels are reassigned to one of their two surrounding levels depending on the visibility of quantization noise, also allows a higher threshold for quantization noise and results in a 25 percent decrease of entropy. Delayed reassignment, in which reassignment is done only if the reassigning of the present sample to a lower level does not "adversely" affect the coding of the next sample, gives a 22 percent decrease in entropy. Thus alternate reassignment and delayed reassignment appear promising.

In the transform domain, we considered differential PCM coding of the first Hadamard coefficient. Using alternate reassignment, it was possible to reduce the entropy of the coded first transform coefficient by about 23 percent. We also give results on different methods of evaluating the visibility of the quantization noise and the effect of changing the picture content.

II. MASKING AND VISIBILITY FUNCTIONS[3]

We have constructed a simple measure of luminance spatial activity. We call this measure the masking function. In the pel domain, the masking function at a pel is taken as the weighted sum of the luminance slopes at the pel under consideration and at the neighboring pels. Thus, at point (i, j) the two-dimensional masking function $M_{i,j}$, using a 3×3 neighborhood of slopes is given by

$$M_{i,j} = \sum_{n=i-1}^{i+1} \sum_{t=j-1}^{j+1} \alpha^{\|(n,t)-(i,j)\|} \cdot \tfrac{1}{2} \cdot [|m_{n,t}^{H}| + |m_{n,t}^{V}|] \qquad (2)$$

where $\|(n, t) - (i, j)\|$ is the Euclidean distance between points (n, t) and (i, j) normalized by the distance between horizontally adjacent pels; $m_{n,t}^{H}$, $m_{n,t}^{V}$ are the horizontal and vertical slopes of the image intensity at point (n, t); α is a constant taken to be 0.35 based on the tests given in Ref. 3.

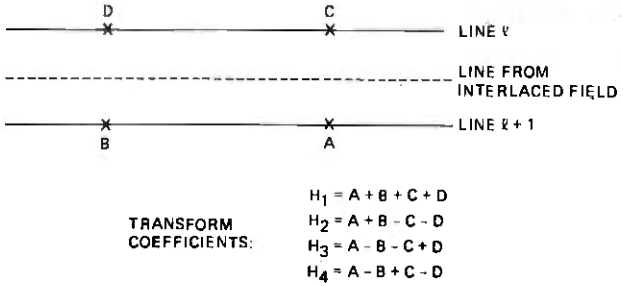


Fig. 1—Definition of Hadamard coefficients. *A, B, C, D* are the pel positions, and H_1, H_2, H_3, H_4 are the Hadamard coefficients.

In this framework, many masking functions are possible and some of these are discussed in Ref. 3.

In the Hadamard transform domain[1,2] using a 2×2 transform of a block of pels in the same field defined as in Fig. 1, the measure of spatial activity in a block is taken as

$$H = \max(|H_2|, |H_4|) \tag{3}$$

It is clear from Fig. 1 that H is the maximum magnitude of the average line or element difference of pels within the block.

We hypothesized that the masking of the quantization noise is related to the spatial detail as measured by the masking functions (M or H). The precise relationship was obtained through subjective tests.* In these tests, the test picture was obtained by adding varying amounts of noise (to simulate the quantization noise) to all pels (or blocks in the case of transform domain) where the measure of spatial detail (M or H) has a given value. Such a picture was then compared in an A/B test with an unimpaired picture to which the subjects added controlled amounts of white noise everywhere in the picture until they found both pictures to be subjectively equivalent.

Visibility functions were derived from such subjective equivalence. They are shown in Fig. 2. Figure 2a shows the visibility of noise added to a pel as a function of masking function M, at that pel, whereas Fig. 2b shows the visibility of noise added to Hadamard transform coefficient H_1 as a function of $\max(|H_2|, |H_4|)$. Both these visibility functions decrease with respect to their argument, implying that at higher values of spatial detail the visibility of quantization noise is lower. Details of the subjective tests and visibility functions are given in Refs. 1–3. From this work we know that the visibility of noise at a point is proportional to the power of noise and the proportionality constant is given by the value of

* These tests are similar to the ones described by Candy and Bosworth.[4]

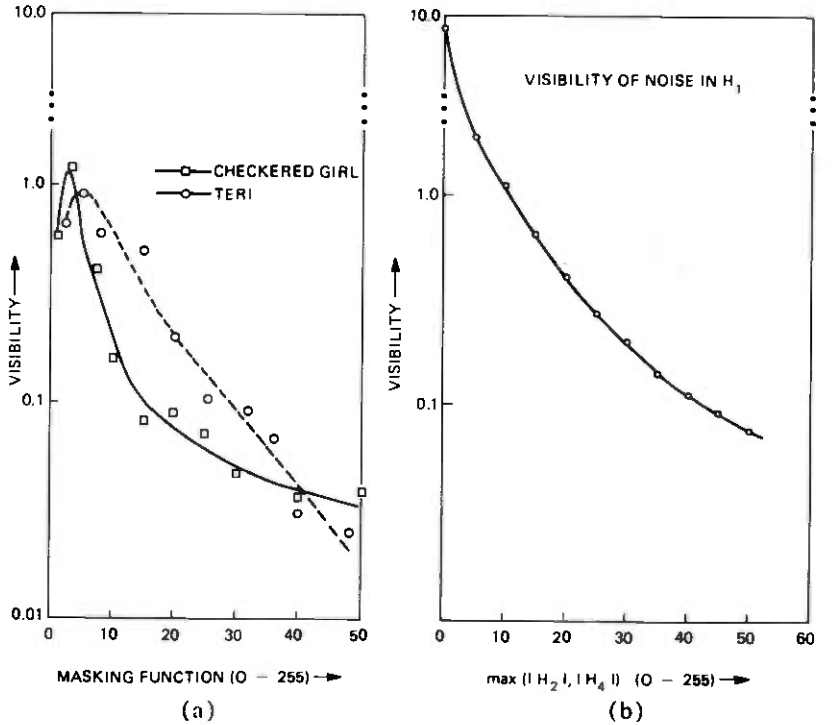


Fig. 2—(a) Visibility function for Checkered Girl and Teri using a two-dimensional one neighbor masking function. (b) Visibility function for noise added to H_1 as a function of max ($|H_2|$, $|H_4|$).

the visibility function evaluated using the spatial detail at that pel. Thus visibility functions allow us to judge the visibility of the quantization noise.

III. DETAILS OF THE TECHNIQUES

The entropy-reducing techniques described briefly in Section I have been simulated on a computer. In this section we give details of the technique as well as the results of the simulation. We have taken two different head and shoulders views (Fig. 3a, called Checkered Girl, Fig. 3b, called Teri) and processed them using our algorithms. We are dealing with a still picture, but since the quantization noise in a real television system varies from frame to frame even with still pictures, we coded three frames of the same picture to which a small amount of random noise (to simulate camera noise) was added. Due to the randomness of this noise from frame to frame, the noise appeared to move when the three frames were displayed in a repetitive sequence.

(a) (b)

Fig. 3—Original pictures of (a) Checkered Girl and (b) Teri. These are 256 × 256 arrays with signal having a bandwidth of 1 MHz; each sample is linearly quantized with 8-bit PCM.

In the pel domain, we started with a DPCM coding system which had a 15-level quantizer optimized for minimum mean square quantization error. Using this quantizer, the coded pictures had noticeable defects, but were considered to be of acceptable quality by the authors. The entropy of the quantized output for Checkered Girl is 2.93 bits/pel and for Teri 2.64 bits/pel. In our experiments, the parameters of the coder were varied to decrease the entropy while keeping the quality of the coded picture similar to that of the 15-level DPCM coded picture. Thus the efficiency of the adaptive techniques described in the next section was judged by the difference between the entropy of the output of the adaptive coder and the 2.93 bits/pel (for Checkered Girl) and 2.64 bits/pel (for Teri) required for the nonadaptive coder. In the transform domain, we applied alternate reassignment to the DPCM coding of the first Hadamard coefficient using the first Hadamard coefficient of the horizontally adjacent block as the predictor. We used a 37-level quantizer which resulted in a picture of acceptable quality as judged by the authors and entropy of 3.75 bits/block for Checkered Girl and 3.76 bits/block for Teri for coding of H_1.

3.1 Reassignment to the lowest level

The 15-level DPCM quantizer characteristic that was used in our simulations is shown in Fig. 4. The level numbers are also marked in the

	0	3	8	15	24	33	42	58
LEVEL	0 1	2 5	6 11	12 19	20 28	29 37	38 49	50 255
NO.	0	1	2	3	4	5	6	7

Fig. 4—A 15-level symmetric quantizer used for DPCM processing in the pel domain. The symbol

$$\overset{z}{\overbrace{}}$$
$$x\ \ y$$

means that inputs between x and y including x and y are represented by z.

Table I — Change in quantizer level distribution from nonadaptive to adaptive coder which reassigns all levels to the lowest possible level for Checkered Girl. Only non-negative levels are shown.

			ith level of adaptive coder							
			0	1	2	3	4	5	6	7
		Total pel count per level	26566	10624	3029	1874	1206	678	363	590
jth level of nonadaptive coder	0	16825	16825	0	0	0	0	0	0	0
	1	14465	5465	9000	0	0	0	0	0	0
	2	6161	3424	1286	1451	0	0	0	0	0
	3	3307	532	302	1474	999	0	0	0	0
	4	1804	180	22	84	817	701	0	0	0
	5	965	78	10	13	43	480	341	0	0
	6	663	27	3	5	12	21	325	270	0
	7	740	35	1	2	3	4	12	93	590

figure. Let the decision levels for the ith level be X_i, X_{i+1} and the representative levels be Y_i. The algorithm for positive i proceeds as follows:

If e_{in} is the input to the quantizer and $X_i < e_{in} \leq X_{i+1}$, i.e., without adaptation e_{in} would be quantized as Y_i, then,

(i) Evaluate the quantization error in representing e_{in} by the next lower level, i.e., error = $e_{in} - Y_{i-1}$.

(ii) Compute the visibility of this error by

$$e_{vis} = |e_{in} - Y_{i-1}|^{\gamma} \cdot f(M) \tag{4}$$

where M is the value of the masking function at the pel, $f(\cdot)$ is the visibility function, and γ is a constant.

(iii) If the visible error is less than a preassigned threshold then the input e_{in} can be represented as level Y_{i-1}. The above three steps are then repeated until the lowest level, at which either the visible error is above the threshold or the zero level is reached.

Similar steps can be followed when i is negative. Thus the lowest level (in magnitude) to which the input can be reassigned such that the visibility of the quantization is below a given threshold is taken as its representation. In the simulations γ was set to 1, and the threshold was varied until the lowest entropy was obtained for our picture quality.

The results of this simulation, for Checkered Girl, are shown in Table I. In this table the (j, i)th entry is the number of pels which would be quantized by $\pm j$th level in a nonadaptive coder but are quantized by $\pm i$th level in the adaptive coder. The first column gives the total number of pels quantized by a level in the nonadaptive coder and the first row gives the total number of pels quantized by a level in the adaptive coder. Thus,

Table II — Change in quantizer level distribution from nonadaptive to adaptive coder which reassigns the inner three levels to the lowest possible level for Checkered Girl. Only nonnegative levels are shown.

			*i*th level of adaptive coder							
			0	1	2	3	4	5	6	7
		Total pel count per level	29526	10190	2453	460	1663	827	611	685
*j*th level of nonadaptive coder	0	16950	16950	0	0	0	0	0	0	0
	1	14533	5697	8836	0	0	0	0	0	0
	2	7229	5483	605	1141	0	0	0	0	0
	3	3917	1396	749	1312	460	0	0	0	0
	4	1663	0	0	0	0	1663	0	0	0
	5	827	0	0	0	0	0	827	0	0
	6	611	0	0	0	0	0	0	611	0
	7	685	0	0	0	0	0	0	0	685

levels ±2 had 6161 pels in the nonadaptive coder out of which 3424 went to level 0, 1286 went to level ±1, and 1451 remained in level ±1, in the adaptive coder. Clearly the distribution of pels using different quantizer levels in the adaptive coder is more peaked than the corresponding distribution in the nonadaptive coder. This causes a decrease in the entropy of the coder output. The decrease in the entropy without any noticeable change of picture quality is about 0.27 bits/pel which is approximately 9.2 percent.

One of the reasons for such a small decrease of entropy is the feedback process inherent in a DPCM coder. If, in quantization of a pel, the adaptive quantizer changes its occupancy from level $\pm j$ to $\pm k$ $(k < j)$, then the differential signal to be quantized for the next pel generally increases in magnitude, and there is a good chance that it will occupy a higher quantization level than in the case of the nonadaptive coder. This, to some extent, hampers our original objective of making the distribution of pels in quantizer levels highly peaked. This feedback affects the very high levels more severely. Quantizing a "higher level sample" by a lower level, results in spreading of the picture edges, which is subjectively very annoying.

To obviate this, and knowing that the higher quantization levels have little effect on the entropy, we used the above algorithm to reassign only the inner ±1, ±2, and ±3 levels. The results of this simulation are shown in Table II. The 0th level contains more pels than before. This reduces the entropy to 2.51 bits/pel which is a reduction of about 14 percent without any noticeable sacrifice of picture quality.

Using the inner ±1, ±2, ±3 level reassignment algorithm, we studied the effects of the variation of γ in eq. (4). Small γ results in good repro-

duction of flat areas, but results in severe edge effects similar to the edge busyness. A value of γ larger than 2, on the other hand, results in better edge reproduction but shows granular noise in flat areas. We used three different values of γ, namely 0.5, 1.0, and 2.0, and varied the threshold T such that for each case we got the lowest entropy without changing the picture quality. The entropy for $\gamma = 0.5$ was 2.62 bits/pel, for $\gamma = 1.0$, it was 2.51 bits/pel and for $\gamma = 2.0$, it was 2.44 bits/pel. Thus, $\gamma = 2.0$ was clearly superior to either 0.5 or 1.0. We used $\gamma = 2.0$ for all our subsequent simulations.

3.1.1 Alternate reassignment

Alternate reassignment changes alternate levels (positive or negative) to one of the two adjacent levels depending on the quantization error visibility. This has the effect of reducing the level occupancy of alternate levels, thereby reducing the entropy. Also the highest level (positive or negative) was not reassigned for better edge reproduction. In our simulation of this scheme only alternate levels i.e., ± 1, ± 3, ± 5 were reassigned. Level 1, for example, was reassigned to either level 0 or level 2 depending upon both of the following conditions:

(*i*) Reassign to level 0 if

$$|e_{\text{in}} - Y_0|^{\gamma} \leq |e_{\text{in}} - Y_2|^{\gamma} \tag{5}$$

otherwise reassign to level 2; and

(*ii*)

$$|e_{\text{in}} - Y_k|^{\gamma} f(M) \leq T \tag{6}$$

$k = 0$ or 2, whichever satisfies condition 1.

Level 1 was not reassigned if both the above conditions were not met. Here again, both the positive and negative levels were reassigned similarly. This resulted in further improvement in the entropy of coded output. We tried alternate reassignment on levels ± 1 and ± 3, and ± 5 with results as shown in Table III. Note that as expected, the number of pels in levels ± 1, ± 3, and ± 5 have decreased significantly. The entropy of the quantized output for this simulation was 2.19 bits/pel, which is a decrease of about 25 percent.

3.1.2 Delayed reassignment

Another way to counter the feedback in the DPCM coder is to wait until the next sample for the coding of the present sample and reassign the present sample only if the next sample is not "adversely" affected by reassigning the present sample. The reassignment of the kth sample

Table III — Change in quantizer level distribution from nonadaptive to adaptive coder which does alternate reassignment of levels ±1, ±3, ±5, for Checkered Girl. Only nonnegative levels are shown.

			*i*th level of adaptive coder							
			0	1	2	3	4	5	6	7
		Total pel count per level	34802	1845	7868	90	2642	1	852	621
*j*th level of nonadaptive coder	0	16274	16274	0	0	0	0	0	0	0
	1	22192	18528	1845	1819	0	0	0	0	0
	2	4322	0	0	4322	0	0	0	0	0
	3	2561	0	0	1727	90	744	0	0	0
	4	1408	0	0	0	0	1408	0	0	0
	5	799	0	0	0	0	490	1	308	0
	6	544	0	0	0	0	0	0	544	0
	7	621	0	0	0	0	0	0	0	621

to the lowest possible level is made only

(*i*) If the visibility of error, i.e., $|e_{in} - Y_j|^{\gamma} \cdot f(M)$, is less than a threshold T_1; and

(*ii*) If the prediction error for the next sample does not change by more than "T_2", a specified value.

In our simulation, we set the threshold T_1 at a high value and varied T_2 between 1 and 15. $T_2 = 5$, which implies that reassigning the present sample should not change the prediction error of the next sample by more than 5, gave the best results. The results of this simulation for

Table IV — Change in quantizer level distribution from nonadaptive to adaptive coder which does delayed reassignment of all the levels for Checkered Girl. Only nonnegative levels are shown.

			*i*th level of adaptive coder							
			0	1	2	3	4	5	6	7
		Total pel count per level	37659	5235	3181	1742	1209	710	504	637
*j*th level of nonadaptive coder	0	13946	13946	0	0	0	0	0	0	0
	1	23171	21219	1952	0	0	0	0	0	0
	2	7826	2020	3170	2636	0	0	0	0	0
	3	2454	347	97	446	15640	0	0	0	
	4	1405	90	16	90	129	1080	0	0	0
	5	804	19	0	8	45	97	635	0	0
	6	591	9	0	1	4	29	70	478	0
	7	680	9	0	0	0	3	5	26	637

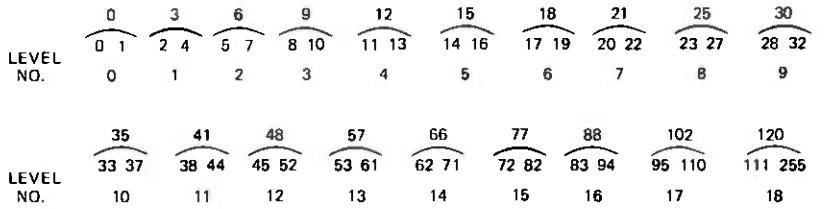


Fig. 5—A 37-level symmetric quantizer characteristic used for DPCM processing of the first Hadamard coefficient. (Only the positive portion is shown.)

Checkered Girl are shown in Table IV. It is seen that the highest level occupancy did not change significantly in the adaptive coder, but occupancy in the lower levels changed rather significantly. The resulting entropy was 2.28 bits/pel, a decrease of about 22 percent over the nonadaptive coder. It appears that delayed reassignment and alternate reassignment perform somewhat similarly in their bit-rate reduction capabilities.

3.2 Reassignment in the Hadamard domain

Alternate reassignment in the Hadamard domain was simulated for the first Hadamard coefficient "H_1" of a 2×2 block. As mentioned in our earlier work,[1] we did DPCM coding of the first coefficient to make use of some of the statistical redundancy not utilized due to small block size. It was found that, when all the other coefficients, i.e., H_2, H_3, and H_4, were not quantized, the 37-level quantizer whose characteristics are shown in Fig. 5, gave a satisfactory picture without any adaptation. Using this quantizer, alternate reassignment was simulated for levels ± 1, ± 3, ± 5, ± 7, ± 9, ± 11. The entropy of the quantized output was 2.90 bits/block for coding of H_1, which amounts to a 23 percent decrease over that obtainable from the nonadaptive coder.

3.3 Sensitivity to picture variation

In order to study the sensitivity of the reassignment algorithms to picture variation, we coded Teri using all the above techniques. The visibility functions used were those obtained for Teri. The results are shown in Table V. It is clear from this table that the percentage decrease in entropy using reassignment is a little higher for Teri than for Checkered Girl. For example, using alternate reassignment there is about 30 percent decrease in entropy. We also coded Teri and Checkered Girl with each other's visibility functions and found relatively insignificant changes in entropy. However, the thresholds for the least entropy had to be different for the two pictures. Taking the lower value of the

Table V — Entropy results with level reassignment for Checkered Girl and Teri

Algorithm	Entropy (bits/pel) Checkered Girl	Teri
Reassigning all levels to lowest level	2.66	2.31
Reassigning inner ±3 levels to lowest level	2.44	2.23
Alternate reassignment on ±1, ±3, ±5 levels	2.19	1.87
Delayed reassignment	2.28	1.91
Hadamard transform coding, alternate* reassignment on ±1, ±3, ±5, ±7, ±9, ±11 levels	2.90	2.84

* Without level reassignment, the entropy of the coded H_1 coefficient was 3.77 bits/block for Checkered girl and 3.76 bits/block for Teri.

threshold and using it for both pictures increased the entropy for one of the pictures by about 5 percent.

IV. SUMMARY

We have described several coding algorithms to change dynamically the input-output relationships of a single quantizer in order to achieve bit-rate reduction. The algorithms reassign the input to the quantizer to a different representative level in such a way as to reduce the entropy of the output, while keeping the visibility of the quantization below a specified threshold. The techniques were demonstrated for DPCM coding systems both in the pel and the Hadamard transform domain. Although we do not discuss it here, it is possible to extend these techniques to PCM quantizers. The inherent feedback, which exists in a DPCM system, counters the reassignment strategy to some extent. Methods were devised to decrease the effect of feedback on the bit-rate reduction. It is possible to reduce the entropy by about 20 to 30 percent by our adaptive techniques, without changing the picture quality.

V. ACKNOWLEDGMENTS

Our work was greatly simplified by the use of the picture-processing computer system developed mainly by Jean-David Beyer and Ralph C. Brainard. We are grateful to Alan S. Kobran for programming assistance and to Barry G. Haskell who developed many of the general-purpose programs which we used.

REFERENCES

1. F. W. Mounts, A. N. Netravali, and B. Prasada, "Design of Quantizers for Real Time Hadamard Transform Coding of Pictures," B.S.T.J., *56*, No. 1 (January 1977), pp. 21–48.
2. A. N. Netravali, B. Prasada, and F. W. Mounts, "Some Experiments in Adaptive and Predictive Hadamard Transform Coding of Pictures," B.S.T.J., *56*, No. 8 (October 1977), pp. 1531–1547.

3. A. N. Netravali and B. Prasada, "Adaptive Quantization of Picture Signals Using Spatial Masking," Proceedings of IEEE, April 1977, pp. 536–548.
4. J. C. Candy and R. H. Bosworth, "Methods for Designing Differential Quantizers Based on Subjective Evaluations of Edge Busyness," B.S.T.J., *51*, No. 7 (September 1972), pp. 1495–1516.
5. E. R. Kretzmer, "Reduced-Alphabet Representation of Television Signals," IRE Nat. Conv. Rec., *4*, Part 4, 1956, pp. 140–147.
6. J. O. Limb, "Adaptive Encoding of Picture Signals," Symposium on Picture Bandwidth Compression, MIT, Cambridge, April 1969.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Fiber-Optic Array Splicing with Etched Silicon Chips

By C. M. MILLER

(Manuscript received August 8, 1976)

The two-dimensional array splicing concept using aluminum chips was reported earlier as a technique that is potentially suitable for field-splicing an optical cable containing linear arrays of optical fibers. Two arrays are stacked, epoxied, polished, and positioned to form a butt-joint splice. This paper reports the use of preferentially etched silicon chips to fabricate higher precision arrays which deviate from perfect uniformity by only 2.5 μm (0.0001 inch) on the average. Average losses for splices assembled with these arrays have been in the range 0.16 dB to 0.32 dB with a yield of 98.8 percent. These average losses are close to the anticipated values based on the measured mean offset between corresponding fiber axes in the spliced arrays. After several combinations of arrays were assembled and loss measurements made, the original array configuration was measured to test for array deterioration, loss measurement repeatability and final alignment repeatability. Only one fiber position showed evidence of contamination, and the mean splice loss was repeatable to within 0.02 dB.

I. INTRODUCTION

A two-dimensional array splicing approach for connecting fiber optic cable has been reported previously.[1] This mass splicing technique has given encouraging results and was used in the Atlanta Fiberguide System Experiment. The method of fabrication of the connector halves involves the following operations:[1,2]

(*i*) Ribbons containing linear arrays of fibers are prepared by removing all ribbon and coating material from the fibers at the end of the ribbons.

(*ii*) A stack is formed by interleaving chips and layers of fibers to form a two-dimensional array (see Fig. 1).

(*iii*) The array is potted to maintain the geometry.

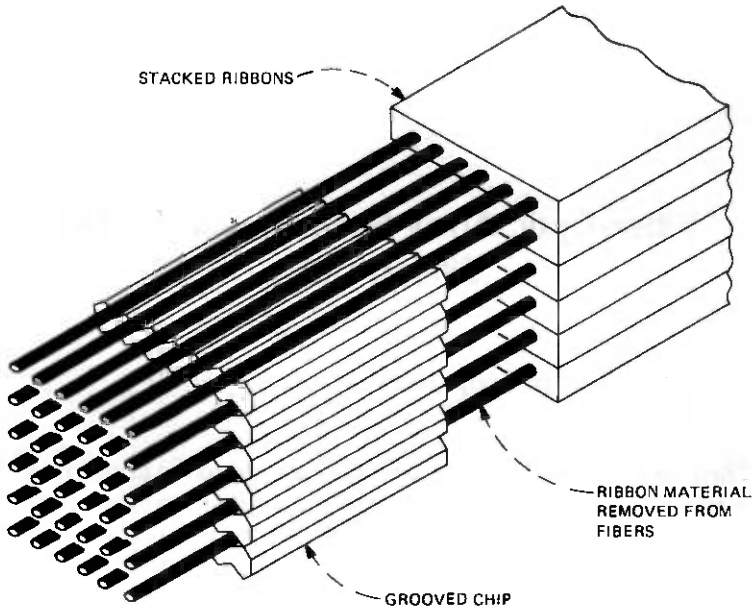


Fig. 1—Stacking fiber ribbons.

(*iv*) "Good ends" on all fibers in the array are obtained by grinding and polishing the array.

(*v*) An array splice is made by butting two arrays, prepared as previously described; the arrays are held in position using the unoccupied grooves on top and bottom of the array for final alignment.

The purpose of this paper is to give a progress report on the array splicing method for connecting groups of optical fibers. Improvements in the chips and in the fabrication of the arrays are described. A method is given for determining a "figure of merit" for a two-dimensional array. Four array-splice loss tests are described and correlation of loss and offset is presented.

II. IMPROVEMENTS

Several improvements have been incorporated in the process for fabrication of two-dimensional arrays. These include improved tools and fixtures for assembly and end preparation. A stack of 12 prepared ribbons interleaved between 13 chips can be assembled without microscopes or micromanipulators in approximately 10 minutes. An improved, simplified polishing fixture has replaced the previous version.[2] This fixture uses a negative chip mounted to the inside wall of the holding fixture to accurately position the array.

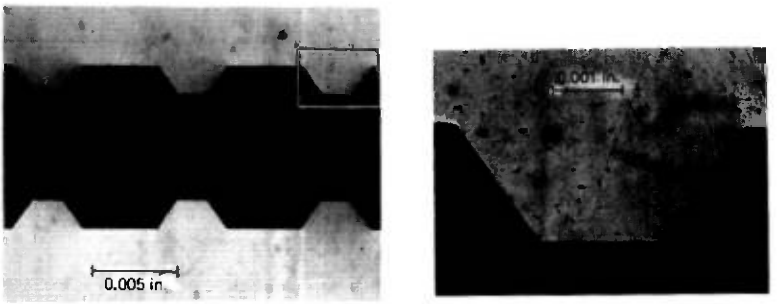


Fig. 2—Silicon chip cross section.

The most significant improvement in the array splicing technique has been the introduction of preferentially etched silicon chips.[3] These chips were developed and produced by C. M. Schroeder[6] at Western Electric Engineering Research Center using photolithographic techniques applied to (100) oriented silicon wafers. Oxide masks are generated on both surfaces of the silicon wafer and the wafer is submerged in a basic solution such as potassium hydroxide. The etch rate in the unmasked region in the (100) direction is much greater than in the (111) direction. Groove angles are dependent only on the accuracy of slicing the wafer relative to the crystal axis, which is on the order of 1°. Groove opening is dependent only on the mask accuracy which is approximately ±1 micron. Alignment between top and bottom grooves is accomplished by using a dual exposure channel with near collimated ultraviolet light. Finished wafers are laser scribed and broken into individual chips. Figure 2 consists of two photographs of the cross section of a typical silicon chip. A companion paper[6] describes the chip fabrication process in greater detail. Occasionally groove bridges occur (perhaps in 1 percent of the chips) which must be detected by visual inspection and discarded. Figure 3 shows a bridged groove on a silicon chip.

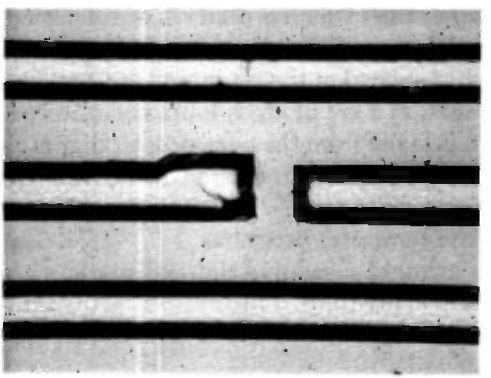


Fig. 3—Bridged groove on a silicon chip.

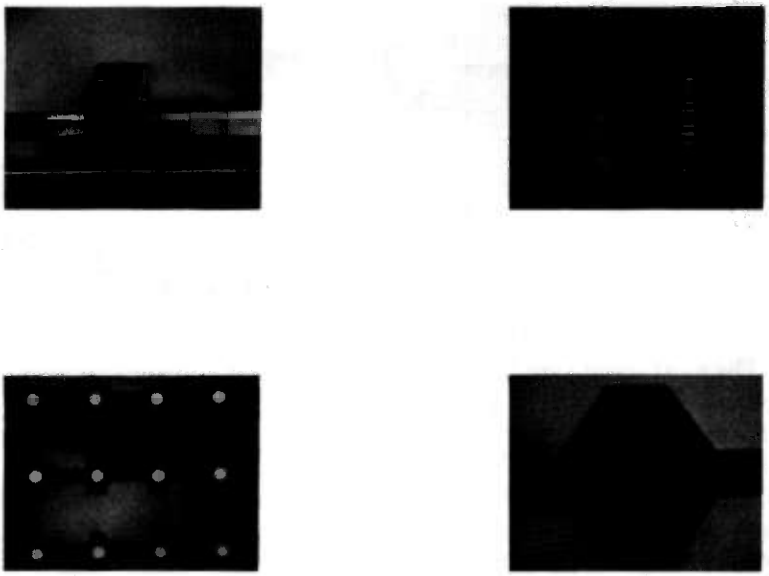


Fig. 4—Array splicing.

Connectors assembled with silicon chips have given greatly improved array alignment. Figure 4 shows three views of a connector end at different magnifications. The completed splice shown in Fig. 4 is one of the splices used in the Atlanta Fiberguide System Experiment.

III. ARRAY ALIGNMENT MEASUREMENTS

An accurate method has been developed for determining an alignment "figure of merit" which characterizes the degree of uniformity of a two-dimensional array. A Leitz-Leica Microscope with a micrometer stage is used to measure the coordinates (x,y) of each fiber in the array relative to a corner fiber which is used as the reference origin (0,0). The accuracy and resolution of these measurement is ± 1.25 micron. A small dark spot which is located at the center of the fiber core* (as seen in Fig. 4) is used to position a set of cross-hairs accurately and quickly. The coordinate data is taken from the microscope's digital readout and typed into a computer data file for processing. Iterative programs are used to generate a uniform array of "best fit." The flow chart of Fig. 5 maps the main loops in the computer programs.

Consider an array of measured data points (x,y) as compared to a perfectly uniform array (u,v). The uniform array is defined by a value for the uniform spacing of points along both coordinate axes (x_s,y_s), the

* The dark spot is thought to be due to loss of germania or the inside surface of the preform which produces a dip in the index of refraction.

origin relative to the origin of the measured data (x_o,y_o) and the angle between the u-axis of the uniform array and the x-axis of the measured data (θ).

The measured data is transformed by translation and rotation into the coordinate system of the uniform array by the following transformations.

$$x_t = a_{11}x + a_{12}y + b_1$$

$$y_t = a_{21}x + a_{22} + b_2$$

where

$$a_{11} = a_{22} = \cos\theta$$

$$a_{12} = \cos\left(\frac{\pi}{2} - \theta\right)$$

$$a_{21} = \cos\left(\frac{\pi}{2} + \theta\right)$$

and

$$b_1 = -x_o \cos\theta - y_o \sin\theta$$

$$b_2 = x_o \sin\theta - y_o \cos\theta.$$

The magnitude and angle of the vector, q_i, joining the ith point of the uniform array with the corresponding ith transformed data point is calculated as simply

$$|q_i| = (x_{t_i} - u_i)^2 + (y_{t_i} - v_i)^2$$

and

$$\angle q_i = \arctan\frac{y_{t_i} - v_i}{x_{t_i} - u_i}$$

The parameters of the uniform array, θ, x_o, y_o, x_s and y_s are now changed systematically according to the flow charts in Fig. 5 by using two separate computer programs. When this "number-crunching" is finished, the output gives the values of θ, x_o, y_o, x_s and y_s that define a uniform array of "best fit." That is, the mean value of $|q_i|$ is a minimum for this uniform array. This minimum mean value of $|q_i|$ is used as a "figure of merit" for the connector being characterized. The x and y components of q_i (the offsets) are calculated and printed out along with the standard deviation of the minimum $|q_i|$ and the standard deviation of the x and y components of q_i.

Several additional calculations are made as a check on the measured data to eliminate gross errors. A histogram is calculated which gives the number of offsets, $|q_i|$, between o to σ, σ to 2σ, 2σ to 3σ, and greater than

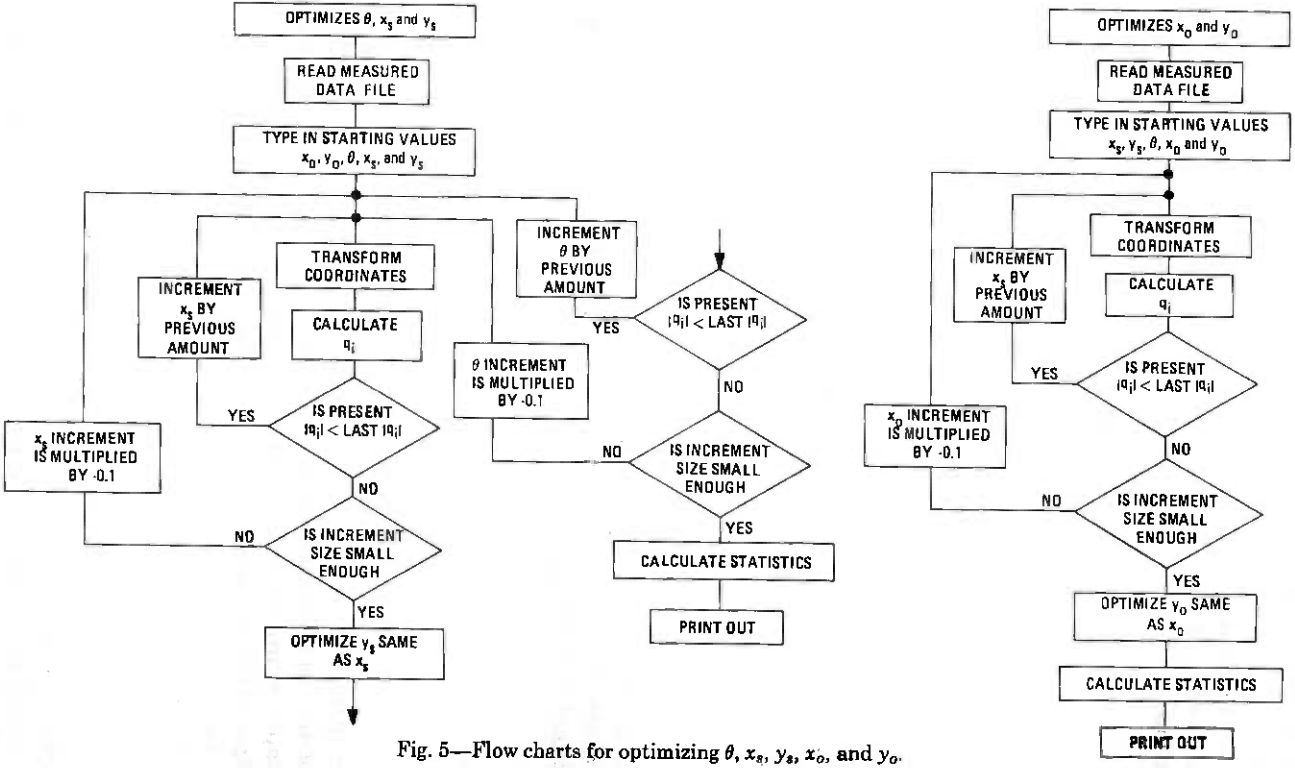


Fig. 5—Flow charts for optimizing θ, x_s, y_s, x_o, and y_o.

3σ. Typically, a few offsets will be in the 2σ to 3σ interval and seldom any greater than 3σ. Also the angle of the offset, $\angle q_i$, places it in one of four quadrants. The distribution of offsets among the four quadrants is printed out as a check to see if the offset vectors are more or less uniformly distributed among the four quadrants.

The primary use of these alignment calculations is for use as a "figure of merit" for an individual connector half. From the outset, it was realized that assembly and measurement of splice loss for a 144-fiber array splice would be tedious and time-consuming. This method of alignment characterization provides a basis for alignment comparison of connector halves, while requiring only one connector half to be assembled and without requiring splices to be assembled or losses to be measured.

IV. ARRAY ALIGNMENT RESULTS

Table I lists the alignment results for selected connector halves. A further description of these connector halves is needed. Connectors 34 and 35 were the last 12 × 12 aluminum chip arrays[1] assembled. The 0.3 mil to 0.4 mil value for the minimum mean offset is typical for 12 × 12 aluminum arrays. (All offset data are in mils, 0.001 inch.)

Silicon chip arrays have from the outset exhibited greatly reduced minimum mean offsets. Connectors 38 through 41 had mean offset values close to 0.1 mil. Connectors 40 and 41 were the input and output fanout connectors for the Atlanta Fiberguide System Experiment. Connectors 45 and 46 were the connector halves on the fiberguide experiment cable. These connectors mated with fanout connectors 40 and 41 to provide access to the cable, and performed satisfactorily. Splice loss data for these specific splices are unavailable due to the fact that they appeared in series with other splices in the system configuration.

V. SPLICE LOSS TEST CONNECTORS

Connectors 47 through 50 were fabricated on the ends of two 0.8 meter long ribbon stacks for the purpose of measuring splice loss for these improved silicon chip arrays. Fiber diameter variations were measured to be approximately ±2.5 percent for the fiber used in these connectors. Connectors 47 and 49 were epoxied at the same time in the holding vise. The stack height for connector 47 was less than that for connector 49 by 0.4 mils and this "slop" may account for the slightly higher minimum mean offset. Connectors 48 and 50 were fabricated separately and gave improved offset characteristics.

VI. LOSS MEASUREMENT PROCEDURE

Array splice losses were measured using the setup shown in Fig. 6. The beam from a Spectra Physics He-Ne laser was expanded and focused with

Table I — Connector data

Connector number	Array size	Material	Min. mean offset, mils	Var. of offset, mils	x_s, mils	y_s, mils	x_o, mils	y_o, mils	θ, millirads
34	12 × 12	Al	0.33	0.20	9.03	9.51	0	−0.19	−1.4
35	12 × 12	Al	0.39	0.18	9.02	9.80	−0.05	−0.17	−2.5
38	12 × 12	Si	0.094	0.05	9.02	11.40	0.01	−0.06	−0.62
39	12 × 12	Si	0.104	0.07	9.015	11.38	−0.01	−0.04	−0.53
40	12 × 12	Si	0.105	0.08	9.011	11.87	0.03	0.03	1.55
41	12 × 12	Si	0.104	0.05	9.021	11.85	0.09	0.04	0.70
45	12 × 12	Si	0.129	0.06	9.021	11.75	0.05	−0.14	0.44
46	12 × 12	Si	0.114	0.075	9.018	11.804	−0.06	0.024	1.79
47	12 × 12	Si	0.179	0.072	9.001	11.604	−0.13	−0.05	−1.39
48	12 × 12	Si	0.135	0.061	9.019	11.658	0.03	−0.03	−1.12
49	12 × 12	Si	0.106	0.051	9.017	11.667	−0.01	−0.13	1.89
50	12 × 12	Si	0.095	0.052	9.012	11.591	0.02	−0.01	0.68

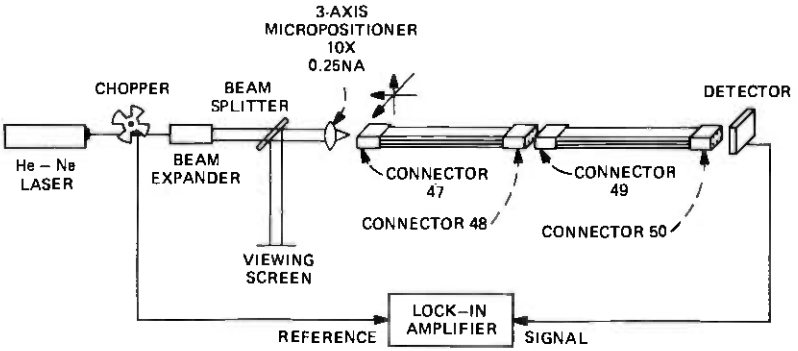


Fig. 6—Loss measurement setup.

a 10X, 0.25 NA, lens onto the core of a fiber in the array. The reflected image from the fiber core was observed on the viewing screen (Fig. 6) and the laser spot positioned above the center of the core approximately midway to the core-clad boundary. The exact spot in this region was not critical since input power to the splice was found to be close to maximum as long as the spot was about midway between core center and cladding.

Input power measurements were obtained by exciting each fiber of connector 47 and measuring the power exiting connector 48. The splice was then assembled by placing connector 49 (or connector 50) in position against connector 48. Again, each fiber of connector 47 was excited by positioning the spot as previously described and the power exiting connector 50 (or connector 49) was measured. Once again the same vicinity on the fiber core almost always corresponded to maximum transmission through the splices. This technique, by contrast with the technique of maximizing input and output power, allowed a larger time constant (1 or 3 seconds) to be used on the lock-in amplifier, thereby reducing noise fluctuations. The time required for a full 12 × 12 array loss measurement was reduced from 4 hours with the previous technique to less than 2 hours with the new technique. The fact that only two negative losses occurred (−0.03 dB and −0.06 dB) compared to 130 positive losses between 0 and +0.1 dB is indicative of measurement accuracy and repeatability substantially better than 0.1 dB.

VII. LOSS MEASUREMENT RESULTS

7.1 Case 1—Connector 48 to connector 49

Figure 9 is a splice loss histogram for case 1. In this case, the ribbons were organized so that a particular fiber was spliced to its mating section. This splicing of "identical" fibers was done to minimize the effect of mismatch due to core diameter variations.

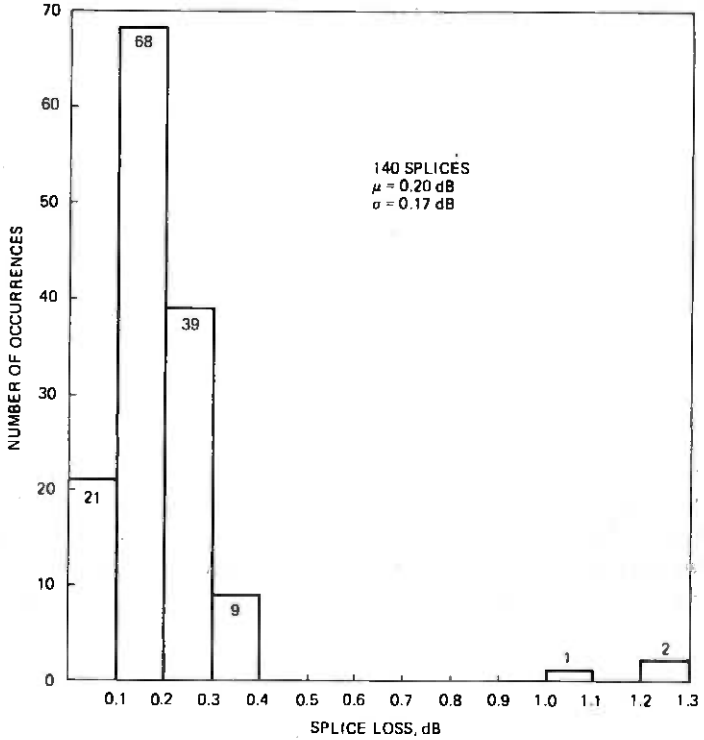


Fig. 7—12 × 12 array splice, connector 48 to connector 49.

Three fibers were broken within connectors 47 or 48 as indicated by a low input power to the splice and one fiber was broken in connector 49 or 50 as indicated by a large loss. These four fibers were not included in Fig. 7 due to definite scattering, indicating breaks. Two fibers also had low input power indicating a break in connector 47 or 48 although the losses for these splices were 1.05 dB and 1.29 dB respectively. It is suspected that another break occurred (1.26 dB loss) in either connector 49 or 50 as this was the only other loss above 0.4 dB.

These seven high losses are due to breaks and if they are excluded from the data, a mean loss of 0.18 dB results with a variance of 0.09 dB. Seven breaks in fabricating four connectors involving 576 fibers corresponds to 98.8 percent yield, and, although this is high, it is believed that improvements in the yield can be made.

7.2 Case 2—Connector 48 to rotated connector 49

Figure 8 is a splice loss histogram for the case of rotating connector 49 by 180° before joining it with connector 48. Fiber diameter mismatches occur in this case and have contributed to the higher mean loss

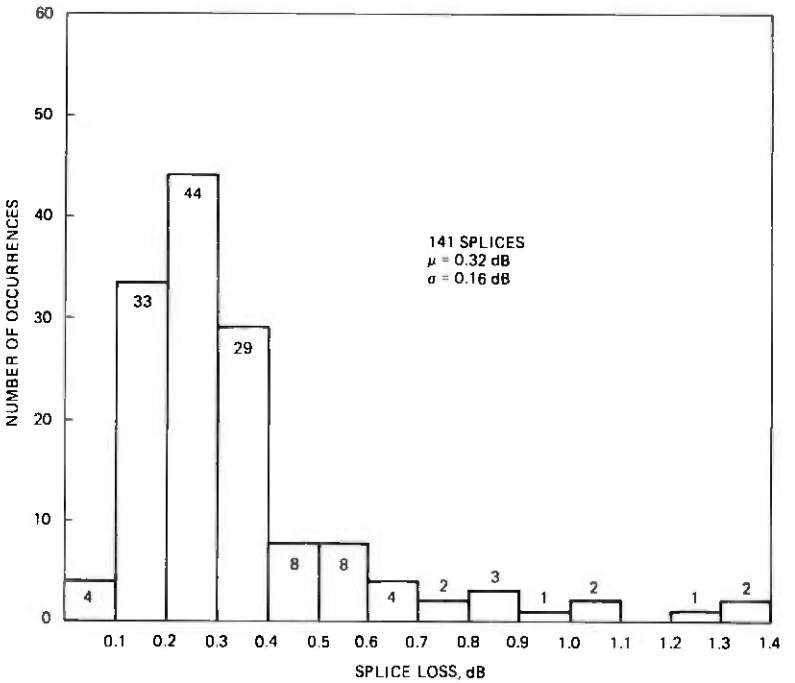


Fig. 8—12 × 12 array splice, connector 48 to inverted connector 49.

and variance. Of the five losses above 1 dB in Fig. 8, three can be attributed to breaks in fibers, leaving two losses over 1 dB unexplained thus far.

Considering the possible effects of ±2.5 percent core diameter variation—if splice loss increases with the square of the ratio of the core diameters when the transmitting core is larger,[4] then a worst-case effect would be 0.42 dB. This worst-case effect may account for the two losses over 1 dB mentioned earlier. Since no core diameter mismatch effect occurs when the transmitting core is the smaller, the mean effect is probably less than half the worst case or perhaps on the order of 0.1 dB. Thus the core diameter mismatch could account for most of the increased mean loss.

7.3 Case 3—Connector 48 to connector 50

Figure 9 is the splice loss histogram for this case. All losses greater than 1 dB have been previously identified as breaks. In this case, improved alignment accuracy appears to be present, since a significant portion of the mean loss could be due to diameter mismatch effects. It is unclear, at this time, how mismatch losses and alignment losses combine. From Table I it is seen that connector 50 has the least value of minimum mean offset and the least θ; therefore, improved splice loss is expected.

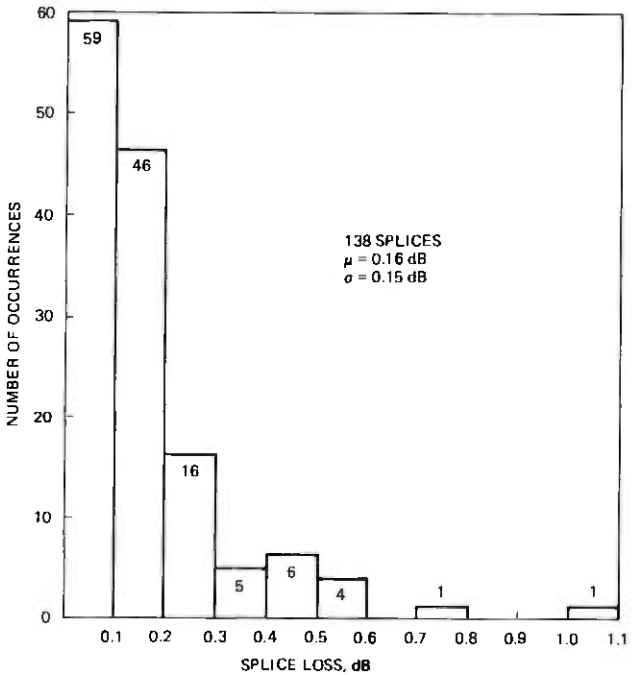


Fig. 9—12 × 12 array splice, connector 48 to connector 50.

It is also seen in Table I that y_s for connector 50 is significantly less than y_s for connector 48. This difference accumulates throughout the array and a connector height measurement shows that connector 48 is 0.52 mils larger than connector 50. When the connectors are placed in the final alignment fixture, a greater force is applied to the thicker connector; however, it was unclear how much deformation occurred and to what extent connector height differences affect array splice losses. This case shows that 0.52 mils can be accommodated with little or no degradation in the splice.

7.4 Case 4—Connector 48 to connector 49

This case is a repeat of case 1. The connectors had been handled, assembled and disassembled several times since the first set of measurements. This test was to see if any noticeable deterioration had taken place and to check the repeatability of the final alignment procedure.

Figure 10 shows that the mean loss is reduced slightly from case 1. All losses greater than 0.9 dB occur in positions which have been identified as broken fibers. It is interesting to note that the three losses between 1 dB and 1.3 dB in case 1 are between 0.7 dB and 1.0 dB in this case. One fiber position measured 0.15 dB in case 1 and increased to 0.89 dB in case 4 possibly due to contamination or end damage.

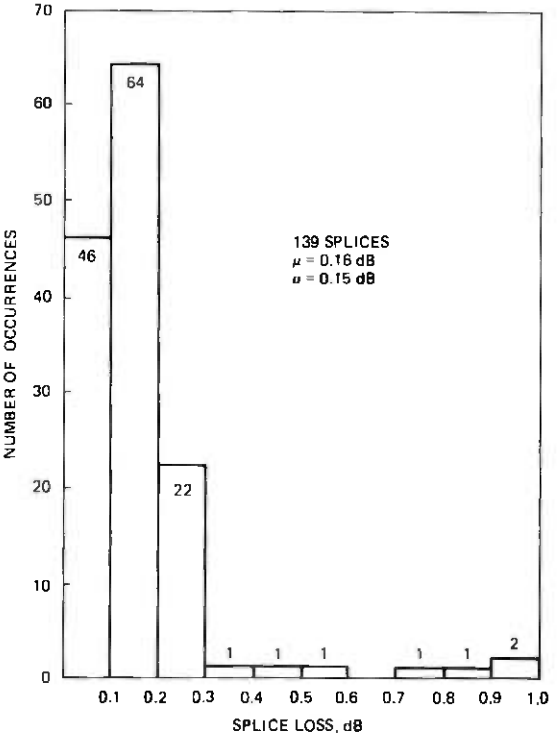


Fig. 10—12 × 12 array splice, connector 48 to connector 49.

VIII. CORRELATION OF LOSS AND OFFSET

A computer program has been written that calculates this mean value of offset between two sets of fiber position data. Two fiber positions that are well separated within the array are selected on the basis of low splice loss at these points. One position is used as the origin, the other is used to fix the position of the x-axis. This operation defines a coordinate system common to both arrays. Each set of fiber position data is transformed by the previously described set of transformations, and the mean offset is calculated as before. After generation of a computer data file containing the measured splice loss data, a correlation coefficient, ρ, is calculated.

$$\rho = \frac{\sum_{i=1}^{n} |q_i| \cdot l_i - n|\overline{q_i}| \cdot \bar{l}_i}{\left(\sum_{i=1}^{n} |q_i|^2 - n|\overline{q_i}|^2\right)\left(\sum_{i=1}^{n} l_i^2 - n\bar{l}_i^2\right)}$$

where l_i = splice loss data and $\bar{l}_i$ = mean splice loss.

Table II — Splice data

Splices	Probable $\lvert\bar{q}_i\rvert$, mils	$\bar{l}_i$, dB	ρ	Loss limit, dB
1. 48 to 49	0.16–0.18	0.18	<0.2	0.5
2. 48 to inv (49)	0.18–0.20	0.32	<0.15	1.0
3. 48 to 50	0.32	0.16	0	0.5
4. 48 to 49	0.16–0.18	0.16	<0.2	0.5

To eliminate the predominating effect of high-loss fibers, a loss limit is set to exclude high losses caused by fiber breaks. Table II lists the results of the mean offset and correlation coefficient calculation. It is seen in Table II that the correlation coefficients are small, indicating little or no correlation between offset and loss on a fiber-by-fiber basis. This had not been the case with aluminum chip splices which exhibited correlation coefficients of typically 0.7. The reduced correlation coefficient for silicon chip arrays indicates that alignment accuracies and the resulting splice losses are approaching the limit of measurement resolution as applied on a fiber-by-fiber basis.

Another approach is to compare the mean offset to the mean loss. Figure 11 is a measured loss versus offset curve[5] for a single fiber similar in index profile and diameter to the fibers used in these arrays. Normalizing a mean offset of 0.17 mils (from Table II for connector 48 spliced to 49) with respect to fiber core radius yields a mean offset of approximately 0.16 core radius. From Fig. 11, this corresponds to a transmission of 96 percent or 0.18 dB loss. The measured mean loss was 0.18 dB (excluding fiber breaks), in agreement with the value predicted from the measured mean offset.

For connector 48 spliced to inverted connector 49, the mean offset was 0.19 mils (0.176 core radius) which corresponds to 95 percent transmission or 0.28 dB. The 0.04 dB by which the mean measured loss exceeded 0.28 dB is probably attributable to core diameter mismatch.

Connector 48 spliced to 50 does not follow the same trends as the previous two splices, due to the 0.52 mil overall connector height difference and the supposed unequal deformation during final alignment. The 0.32 mil (0.30 core radius) mean offset should cause a 0.75 dB mean splice loss, from Fig. 11. The measured mean loss of 0.16 dB suggests that significantly unequal array deformation is taking place in this splice as previously supposed.

IX. CONCLUSIONS

Etched silicon chips have enabled uniform arrays to be fabricated with an average error of 0.1 mil without complex tools or microscopes. Exceptionally low mean splice loss data have been obtained which correlated well with mean offset.

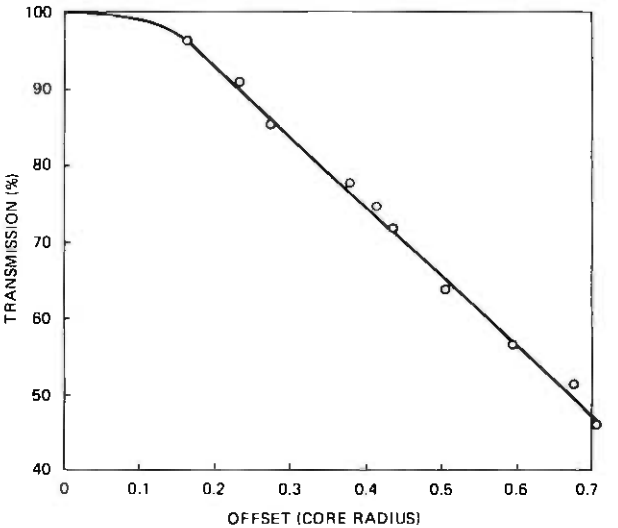


Fig. 11—Measured transmission vs. transverse offset.

Fiber diameter variations of ±2.5 percent caused an increase in the mean loss on the order of 0.1 dB; however, it is not known to what extent array uniformity was affected by these fiber diameter variations. Two arrays with an overall height difference of 0.52 mil had very low mean splice loss, indicating that this height difference was equalized within the splice holding fixture. After several combinations of arrays were assembled and loss measurements made, the original array configuration was measured to test for array deterioration, loss measurement repeatability and final alignment repeatability. Only one fiber position showed evidence of contamination, and repeatability was within 0.02 dB for the mean.

X. ACKNOWLEDGMENTS

The author appreciates and thanks C. M. Schroeder for the work he did to produce the excellent silicon chips reported in this paper. D. N. Ridgway made fiber position measurements, ran computer programs and assisted in measuring splice losses for which the author is indeed grateful.

REFERENCES

1. C. M. Miller, "A Fiber-Optic-Cable Connector," B.S.T.J., *54*, No. 9 (November 1975), pp. 1547–55.
2. C. M. Miller, "Array Splicing-Progress Report," unpublished work.
3. E. Z. DeRossett and C. M. Schroeder, "Optical Fiber Connector Alignment Chips Preferentially etched from (100) Oriented Silicon Wafers," unpublished work.

4. F. L. Thiel, "Utilizing Optical Fibers in Communications Systems," International Conference on Communications, Conference Record, *II*, Session 32, June 16–18, 1975.
5. C. M. Miller, "Transmission vs. Transverse Offset for Parabolic-Profile Splices with Unequal Core Diameter," B.S.T.J., *56*, No. 7 (September 1976), pp 917–927.
6. C. M. Schroeder, "Preferentially Etched Alignment Chips from (100) Oriented Silicon Wafers," B.S.T.J., this issue.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Accurate Silicon Spacer Chips for an Optical-Fiber Cable Connector

By C. M. SCHROEDER

(Manuscript submitted July 1, 1977)

Assembled silicon array connectors have been fabricated with low splice loss. The silicon chips can be manufactured with submicrometer repeatability and accuracy. Improvements in chip thickness variations are anticipated which in turn should reduce the overall groove opening variations to a point where individual chip selection is not required.

I. INTRODUCTION

Low-loss multifiber splices have been obtained with the use of accurate silicon spacer chips. These spacer chips are used in a unique splicing technique developed by Bell Laboratories[1] and used in the Atlanta Fiberguide experiment. This multifiber cable splice is a stacked array consisting of two properly prepared cable connectors butted end to end.

A cable array connector is a laminate sandwich of silicon alignment chips with "V" grooves on both top and bottom surfaces interleaved with optical fibers epoxied to form a two-dimensional array (Fig. 1). A completed connector may consist of up to 144 optical fibers positioned by up to 13 alignment chips. Each fiber in the array must be positioned accurately to its counterpart in a mating array to obtain low splice loss. The transverse misalignment of the mating fibers should be no greater than one-tenth the core radius or typically 2.5 micrometers.[2,3] This can only be obtained with alignment chips having high dimensional accuracy with respect to thickness, groove geometry, and position.

The feasibility of the multifiber splice was shown with alignment chips manufactured of aluminum. The aluminum chip, however, could not be manufactured repeatedly with the high dimensional accuracies required. An improved chip was developed, that met the above goals, using (100) oriented silicon[4] and photolithographic techniques. This improved chip has demonstrated a significant reduction in the overall array splice loss.

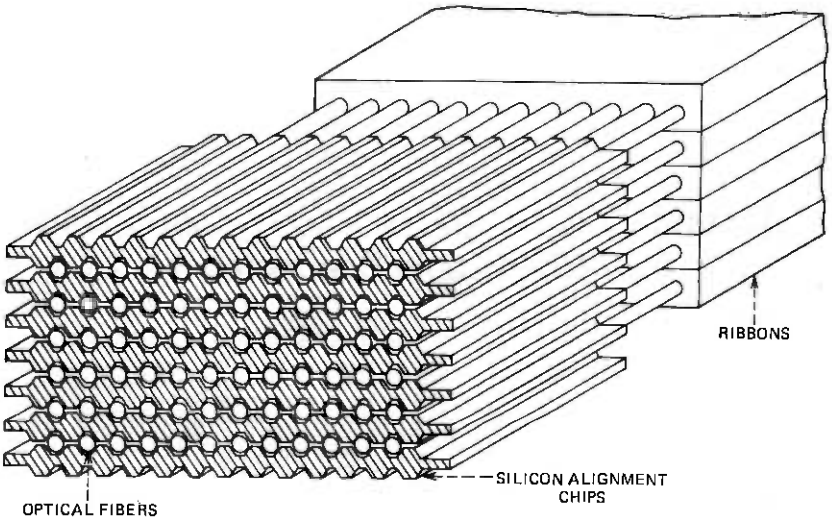


Fig. 1—Schematic of a multifiber array connector.

II. PHOTOLITHOGRAPHIC AND PROCESSING TECHNIQUES

The use of silicon for alignment chips offers several advantages. The material is readily available, it can be handled and processed relatively easily, and it can produce precise V grooves to photomask accuracy.

When a (100) surface slice of silicon covered by an oxide mask is submerged in a basic solution, the etch rate is much greater in the ⟨100⟩ direction than in the ⟨111⟩ direction (Fig. 2). This anisotropic etching of the silicon results in a precise V groove to an angle of 70.53° with the reaction self-stopping at the point where the (111) planes intersect. The groove opening is determined by the opening in the oxide mask. The (110) plane is used as a reference to align the photomask features parallel to the intersection of the (100), (111) planes.

The processing of the silicon first starts with the selection of a 5 cm diameter (100) surface oriented wafer polished on both surfaces to a thickness of 0.25 mm. Thickness variations and surface defects should

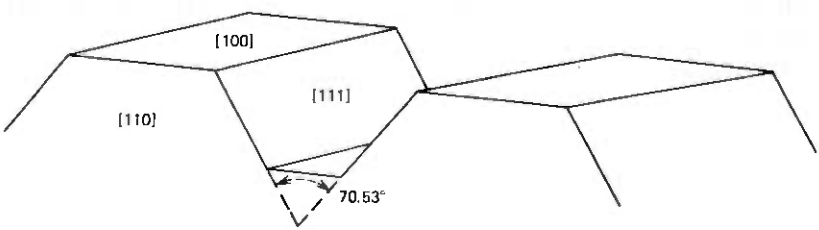


Fig. 2—Crystallographic planes used to manufacture alignment chips.

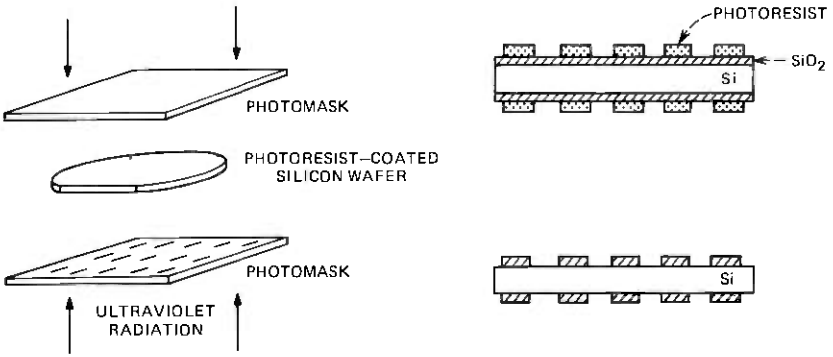


Fig. 3—Processing techniques.

be minimized for best results. The wafer is cleaned, then placed into a tube furnace for a 1-micrometer oxide growth. The wafer is then coated on both surfaces with a positive photoresist, placed between two prealigned photomasks and exposed to ultraviolet radiation (Fig. 3). The exposed wafer is removed, then developed leaving open windows of silicon oxide which are in turn etched in a buffered hydrofluoric acid solution. The remaining photoresist is removed leaving a silicon wafer covered by an oxide mask which is then placed into a potassium hydroxide solution for preferential etching[5]. A cross section of a typical silicon alignment chip is shown in Fig. 4.

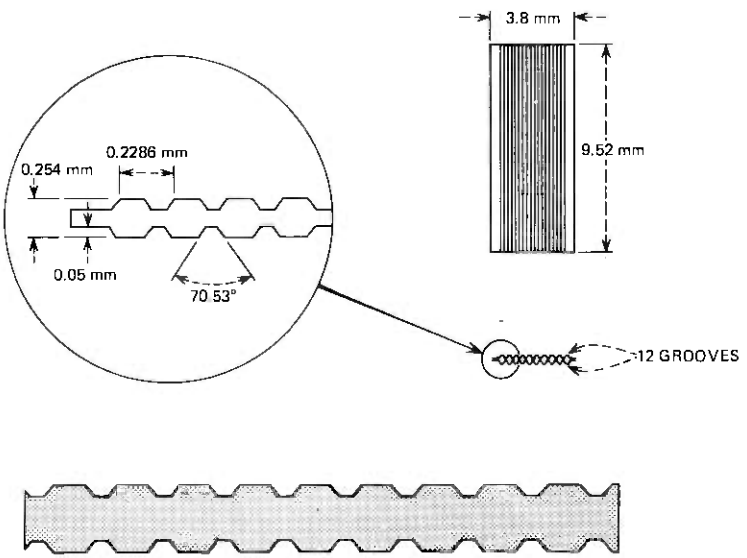


Fig. 4—Cross section of typical alignment chip with dimensions.

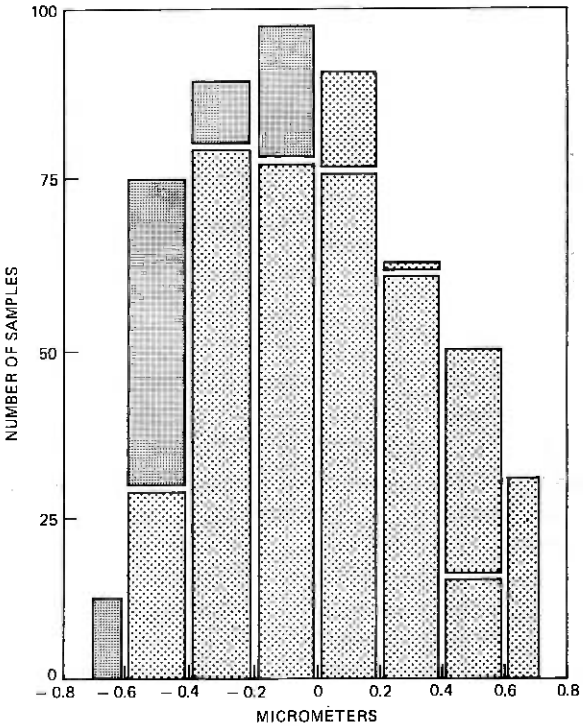


Fig. 5—Data on groove opening variations.

III. ALIGNMENT CHIP EVALUATION

A listing of variations that can cause splice loss in a multifiber connector in their present order of importance is as follows:

(*i*) Alignment chip thickness variations
(*ii*) Fiber diameter variations
(*iii*) Groove opening variations
(*iv*) Transverse misalignment of the top and bottom grooves
(*v*) Aperiodic grooves

Excluding fiber diameter variations, the remaining list must be controlled by alignment chip accuracy.

Measurements on the photomask features and the chip profile were made with the use of a Hewlett Packard model 5526A laser interferometer. The object whose features are to be measured is placed onto a traversing stage under the cross hair of a microscope. A reflecting cube attached to the stage monitors the movement of one of the two paths of the interferometer. This movement is compared to a fixed length of the second path with the changes displayed on a monitor. The repeatability

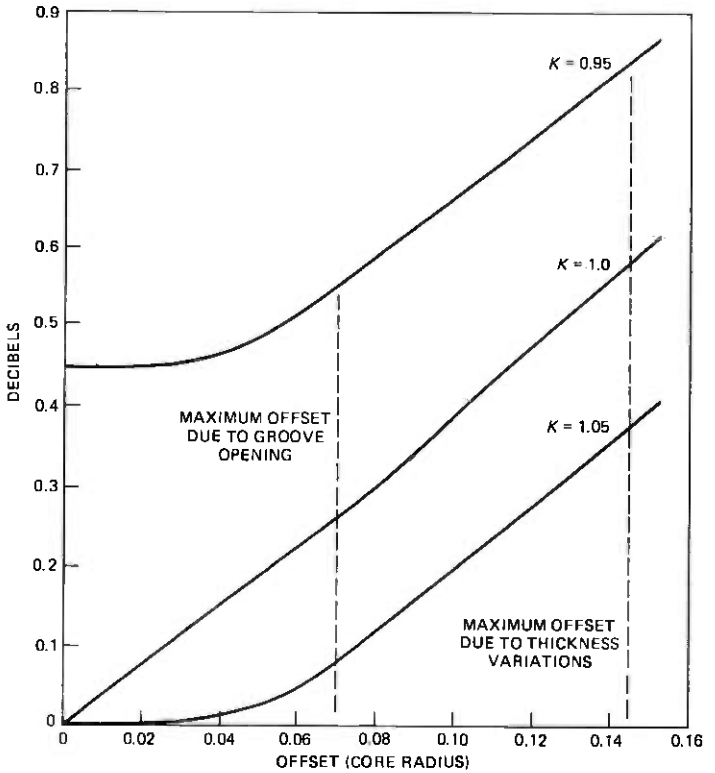


Fig. 6—Splice loss as a function of offset.

of the data was found to be 0.2 micrometer, limited chiefly by the vernier acuity of the eye.

Measurements on the photomask features indicate the required dimensions were well within the 1 micrometer stated accuracy for the masks with line-width variations not measurable across the mask face. Profile measurements on an etched silicon wafer indicate the groove periodicity of 228.6 micrometers was maintained within the measurement error. The transverse misalignment of the top and bottom grooves was held to 1 micrometer or less and the groove depth maintained at 50 micrometers. Variations of this type do not significantly add to the overall splice loss. The major contributors to array splice loss other than fiber diameter variations, was found to be chip thickness variations and groove opening variations.

The groove opening directly affects the vertical position of the optical fiber. The groove angle is constant but relatively steep with 1 unit of groove opening corresponding to 1.4 units of vertical drop in the optical fiber. The data on groove opening variations are shown in Fig. 5. These

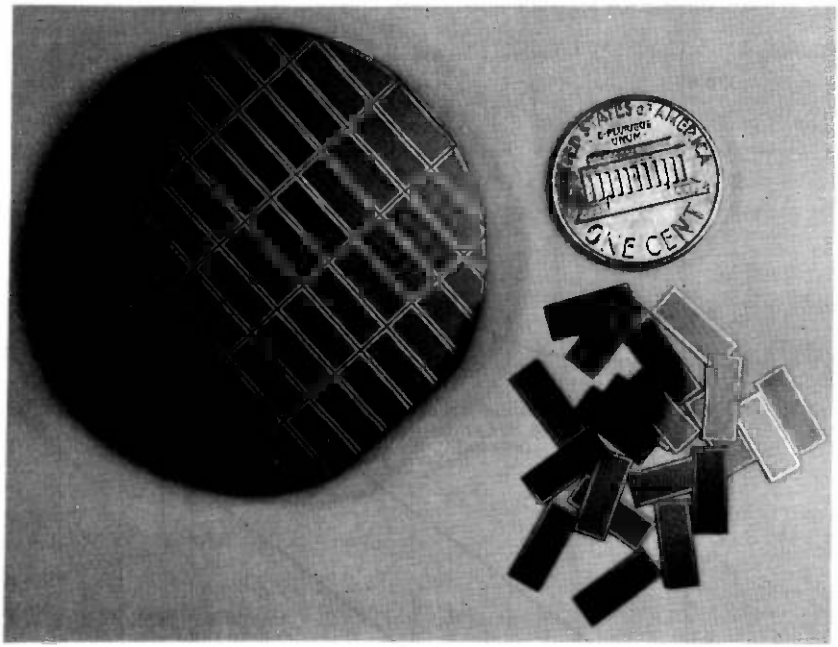


Fig. 7—Laser-scribed silicon wafer.

data were taken on both surfaces of one silicon wafer with the cross-hatched region being data from surface 1 superimposed onto data from surface 2. An overall shift in data of 0.2 micrometers can be observed when comparing the two surfaces. This shift was caused primarily by unequal exposure times. Maximum groove width variations were found to be ±0.7 micrometers with a standard deviation of 0.23 micrometers. The maximum difference in groove width of 1.4 micrometers would cause vertical positioning errors of the optical fibers of 1.96 micrometers for a worst-case situation. While this deviation does not exceed the 0.1 core radius offset, it does point to a need to minimize groove width variations, which are probably caused by variations in wafer thickness with Fresnel diffraction playing a part in undercutting at the areas where the wafer and photomask did not contact. Thickness variations are at present the largest potential contributor to splice loss with total variations of 4 micrometers encountered on some wafers.

IV. LOSS DUE TO TRANSVERSE OFFSETS

Splice loss as a function of offset for small offsets of parabolic index fibers is shown in Fig. 6.[3] K is the ratio of the radius of the receiving fiber to the radius of the transmitting fiber. The maximum offsets due to groove opening and thickness variations are displayed to indicate the

probable loss that could be expected in the worst-case situation. To avoid variations of this type an individual chip selection is required at present.

Splice loss data for 144-fiber array connectors using silicon alignment chips have been made with the results presented in a companion report by C. M. Miller.

V. SUMMARY AND CONCLUSION

A laser scribed silicon wafer is shown in Fig. 7. Each 5-cm diameter wafer yields 36 usable alignment chips 3.8 × 9.52 mm in size.

Assembled silicon array connectors have been fabricated with low splice loss. The silicon chips can be manufactured with submicrometer repeatability and accuracy. Improvements in chip thickness variations are anticipated which in turn should reduce the overall groove opening variations to a point where individual chip selection would not be required.

REFERENCES

1. C. M. Miller, "A Fiber-Optic Cable Connector," B.S.T.J., *54*, No. 9 (November 1975), pp. 1547–1555.
2. D. L. Bisbee, "Measurements of Loss to Offsets," B.S.T.J., *50*, No. 10 (December 1971), pp. 3159–3168.
3. C. M. Miller, "Loss vs. Transverse Offset for Parabolic Fiber Splices," B.S.T.J., *55*, No. 7 (September 1976), pp. 917–927.
4. L. P. Boivin, "Thin-Film Laser-to-Fiber Coupler," Appl. Opt., *13*, No. 2 (February 1974), pp. 391–395.
5. R. C. Kragness et al., "Etchant for Precision Etching of Semiconductors," Patent No. 3,506,509.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

On the Phase of the Modulation Transfer Function of a Multimode Optical-Fiber Guide

By I. W. SANDBERG, I. P. KAMINOW, L. G. COHEN, and W. L. MAMMEL

(Manuscript received June 1, 1977)

We consider the range of validity of a Hilbert-transform approach in which the measured magnitude of the modulation-transfer-function of an optical fiber is used to compute the fiber's impulse response. It is argued that a key "minimum-phase assumption" can fail to be satisfied in important cases, and a few closely related experimental and analytical results are presented.

I. INTRODUCTION

Pulse dispersion in an optical fiber transmission line limits its information-carrying capacity by limiting the temporal spacing of input pulses that can be resolved at the output. The impulse response $g(t)$, by which we mean the output power of a fiber excited by a unit impulse of optical power, provides the necessary information concerning the distortion of the pulse by modal and material dispersion. For a strictly monochromatic pulse source, only modal dispersion contributes to the distortion. However, with regard to the corresponding measurement problems, it is difficult to obtain sufficiently short (<0.5 nsec) and monochromatic (< 10 Å) input pulses to accurately study fibers with very low modal dispersion. Another difficulty is the lack of availability of suitable sources that are tunable over a wide range of wavelengths, including wavelengths longer than 1 μm which are of interest for practical fiber systems.

One can obtain $g(t)$ from the modulation frequency transfer function (MTF) $G(\omega)$, which is the envelope response of the fiber to an incoherent optical signal sinusoidally modulated in amplitude at angular frequency ω. Personick[1] has shown that to the extent that certain reasonable approximations hold, $g(t)$ and $G(\omega)$ are a Fourier transform pair. In principle, the MTF can be determined experimentally. The method employed in Refs. 2 and 3 uses a xenon lamp and monochromator as a tunable

source that can be sinusoidally modulated to high frequency (>1 GHz) by an electrooptic modulator. It is straightforward to measure the magnitude $|G(\omega)|$ of the transfer function

$$G(\omega) = |G(\omega)|e^{i\theta(\omega)} \tag{1}$$

using available components, including an RF spectrum analyzer. However, because the fibers must be long (~1 km) in order to obtain good measurement precision, the measurement of $\theta(\omega)$ appears to be formidable. As $f = \omega/2\pi$ varies from zero to 1 GHz, $\theta(\omega)$ varies nearly linearly with f from zero to $10^4\pi$ radians for a 1 km long fiber. However, the contribution of $\theta(\omega)$ to pulse dispersion is due to a nonlinear deviation $\Delta\theta(\omega)$, on the order of 2π, from the much larger linear phase shift $\theta_0(\omega)$; i.e.,

$$\theta(\omega) = \theta_0(\omega) + \Delta\theta(\omega) \tag{2}$$

$$\theta_0(\omega) = \omega L/v \tag{3}$$

where L is the fiber length and v is an effective envelope velocity taken to be independent of ω. Hence, direct measurement of phase distortion in the presence of the large frequency-dependent $\theta_0(\omega)$ could be subject to large error as v varies with temperature or other environmental factors. One is therefore led[2,3] to consider methods of mathematical computation of $\theta(\omega)$ from $|G(\omega)|$ using Hilbert-transform theory.

The main purpose of this note is to report on results which indicate that unfortunately the Hilbert-transform approach described in Refs. 2 and 3 is, in general, not a helpful one for the particular problem at issue, even though early experimental results suggested otherwise.†

Roughly speaking, it is known that by using a Hilbert-transform relation the phase can be obtained from the magnitude of a transfer function provided that the transfer function is "minimum phase." A standard condition (which is by no means sufficient) for "minimum-phase behavior" is that the Laplace transform of the impulse response that corresponds to the transfer function have no zeros in the closed right half-plane. This is in accord with the observation that for an ideal waveguide with constant positive delay τ_0 and transfer function $G_0 = e^{-i\omega\tau_0}$, the phase cannot be determined from a knowledge of the function $|G_0(\omega)|$ alone, but a transfer function that has the same magnitude as that of the waveguide is $G_1(\omega) = 1$ for which the phase is zero for all ω.††

† The statement on page 1518 of Ref. 2 concerning the possible lack of "approximate minimum phase behavior" was motivated by the results of the joint work described here.

†† The mathematical reason that G_0 is not a "minimum-phase" function (even though $e^{-z\tau_0}$ has no zeros for $\mathrm{Re}(z) \geq 0$) is that $\ln|e^{-z\tau_0}|$ fails to satisfy a sufficiently strong growth condition in the half-plane $\mathrm{Re}(z) \geq 0$. See Sections 5.1 and 5.3 for related material.

For our purposes, the difference between $G_0(\omega)$ and $G_1(\omega)$ is unimportant, because we are willing to ignore a constant time delay that can easily be estimated. More generally, it is reasonable to ask if $e^{i\omega\tau_0}G(\omega)$ is a minimum-phase function, where τ_0 denotes the linear part of the delay. An early impulse response measurement on an actual fiber suggested[2,3] that this might indeed be the case. However, the further analytical and experimental study reported on here shows that nonminimum-phase behavior is likely to arise, and can arise, in important actual cases. Some additional closely related material is also presented.

Methods for circumventing the difficulties described in this note are under study, and it is expected that they will be described in a later paper.

II. SOME ANALYTICAL PROPERTIES OF THE MULTIMODE TRANSFER FUNCTION

In the general case, the transfer function of a fiber can be written in the form

$$G(\omega) = \int_{T_a}^{T_b} e^{-i\omega\tau} da(\tau) \tag{4}$$

in which the integrator $a(\tau)$ is a real-valued monotonically nondecreasing function of τ,† and T_a and T_b, which are fixed by the refractive indices of core and cladding, are the smallest and largest modal delays, respectively.†† Often $a(\tau)$ is normalized so that

$$\int_{T_a}^{T_b} da(\tau) = 1$$

For a fiber that can propagate n discrete modes without mode mixing, (4) becomes

$$G(\omega) = \sum_{j=1}^{n} d_j e^{-i\omega\tau_j} \tag{5}$$

in which each d_j is a positive constant that represents the initial excitation of the jth mode. Typically, $n > 100$. Most of our discussion is concerned with the important particular case in which (5) holds, and, in order to avoid a lack of continuity of the presentation, proofs of the results discussed are given in a separate section. We assume that the τ_j are ordered so that $\tau_1 < \tau_2 < \ldots < \tau_n$.

In (5), $G(\omega)$ is the generalized Fourier transform of a finite train of

† Thus, roughly speaking, $da(\tau)$ in (4) can be replaced with $b(\tau)d\tau$ in which the function $b(\tau)$ is nonnegative and may contain impulses corresponding to discrete modes. See Refs. 1 and 2 for the relevant background material.

†† We are of course assuming that material dispersion can be neglected.

not-necessarily-equally-spaced impulses. Let $H(z)$ denote the corresponding Laplace transform. That is, let $H(z)$ denote

$$\sum_{j=1}^{n} d_j e^{-z\tau_j}$$

for all complex z. Of course $G(\omega) = H(i\omega)$ for all ω.

A standard Hilbert-transform method[4,2,3] for determining the phase $\hat{\theta}(\omega)$ of a transfer function $\hat{G}(\omega)$, from the function $|\hat{G}(\omega)|$, when it is possible to do so, is to use the formula†

$$\hat{\theta}(\omega) = \frac{2\omega}{\pi} \int_0^\infty \frac{\ln|\hat{G}(y)|}{y^2 - \omega^2}\, dy \tag{6}$$

which, roughly speaking, amounts to a direct application of the Hilbert transform

$$\mathrm{Im}[f(i\omega)] = \frac{2\omega}{\pi} \int_0^\infty \frac{\mathrm{Re}[f(iy)]}{y^2 - \omega^2}\, dy \tag{7}$$

in which $f(z)$ is any complex-valued function of z that satisfies certain conditions (such as those described in Section 4.1††) which include the condition that $f(z)$ is analytic for $\mathrm{Re}(z) \geq 0$.

The way in which (7) is used to obtain (6) is of course to let $f(z)$ be a suitably defined single-valued variant of $\ln[\hat{H}(z)]$, in which $\hat{H}(z)$ is the Laplace transform of the time function whose Fourier transform is $\hat{G}(\omega)$. That is what gives rise to the well-known requirement that $\hat{H}(z)$ be zero-free in the closed right half-plane.

2.1 The zeros of H(z), and related material

With regard to the location of the zeros of $H(z)$, according to Proposition 2 of Section IV: $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ if

$$d_1 > \sum_{j=2}^{n} d_j \tag{8}$$

and if (8) is not satisfied, then given an arbitrarily small positive number ϵ, and *any* set of n numbers $t_1 < t_2 < \ldots < t_n$, we have $H(z) = 0$ for some z in the closed right half-plane for a choice of the τ_j such that $|\tau_j - t_j| \leq \epsilon$ for all j. In particular, and roughly speaking, unless (8) is satisfied, given any set of n delays, there is a set of delays arbitrarily close to those

† The integrals in (6) and (7) are to be interpreted as Cauchy principal values.

†† Section 4.1 contains an outline of a proof that (7) holds under certain specific conditions. The derivation given, for instance in Ref. 4, lacks rigor in that, for example, the point s in Ref. 4 is initially assumed to be a point internal to a certain contour, while subsequently an expression based on that assumption is evaluated for s *on* the contour. The main reason for including the basically tutorial material of Section 4.1 is that it is used to prove another result described in this note.

delays such that $H(z)$ is not "minimum phase." Notice that it is not claimed that $H(z)$ has a zero in the closed right half-plane whenever (8) is violated.† However, Proposition 2 does imply that whenever (8) is not satisfied it is incorrect to assert that $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ when the τ_j are known only to within some positive tolerance ϵ, no matter how small ϵ is. Therefore, $H(z)$ is zero-free in the closed right half-plane, and that property is structurally stable in the sense indicated, if and only if (8) is met. This result suggests that it would not be surprising to encounter "nonminimum-phase" behavior with fibers for which the total power in a sufficient number of the modes corresponding to the delays $\tau_2, \tau_3, \ldots, \tau_n$ is considerable.

An idealized example in which a somewhat analogous conclusion is reached is as follows. Consider an n-mode fiber without mode mixing for which the modal delays are equally spaced by δ sec, so that $\tau_j = \tau'_0 + j\delta$ for each positive integer j. Assume that the fractional power into the jth mode is given by $d_j = ce^{\gamma j}$ for all j, in which γ and c are constants with $c > 0$. Then

$$H(z) = ce^{-z\tau'_0}\frac{e^{(\gamma - z\delta)n} - 1}{1 - e^{-(\gamma - z\delta)}}$$

and the condition that $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ will be satisfied if and only if $\gamma < 0$.†† Of course $\gamma < 0$ means that modes with larger delays have smaller excitation. A similar conclusion is reached for the continuum mode-mixing case‡ in which the integrator $a(\tau)$ of (4) has a continuous derivative that is proportional to $e^{\gamma\tau}$.

While the discussion in the preceding paragraphs suggests‡‡ that there are important cases in which (6) cannot be used, it certainly does not rule out the possibility that there is some other method for determining $\theta(\omega)$ from $|G(\omega)|$ (which, for example, might possibly exploit the fact that the d_j are positive).

In this connection, consider (5). In order to avoid the necessity of introducing a function equal to $e^{i\omega\tau_1}G(\omega)$, assume throughout the remainder of this section that $\tau_1 = 0$ (which of course simply provides a normalization††).

Suppose, for example, that $n = 2$ and that $d_2 = (1 - d_1)$ with $0 < d_1 < 1$ and $d_1 \neq 1/2$. Then $|G(\omega)| = d_1^2 + (1 - d_1)^2 + 2d_1(1 - d_1)\cos(\omega\tau_2)$

† In fact, we show that that claim would be false.

†† Since (8) is violated when γ is negative and sufficiently close to zero, we see that a given specific $H(z)$ can be zero-free in the closed right half-plane when (8) is violated.

‡ This example was suggested by H. E. Rowe.

‡‡ Little information is available concerning how to accurately specify $G(\omega)$ for an actual fiber using purely analytical methods. Very small geometrical perturbations can have significant effects on the impulse response of graded-index fibers.[5]

†† It is easy to see that without some such normalization, it is is not possible to determine $\theta(\omega)$ from $|G(\omega)|$.

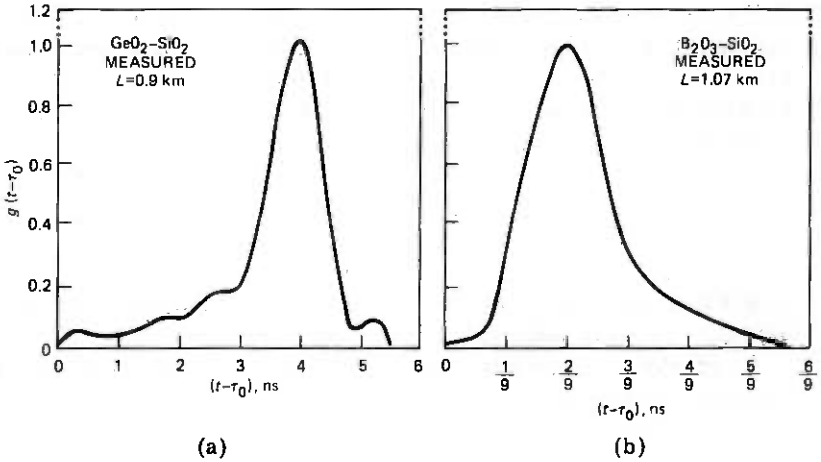


Fig. 1—Measured (deconvolved) impulse responses for (a) a GeO_2–SiO_2 fiber with $\alpha < \alpha_{opt}$; (b) a B_2O_3–SiO_2 fiber with $\alpha \approx \alpha_{opt}$.

which clearly is unchanged if d_1 is replaced with $(1 - d_1)$. Hence given only $|G(\omega)|$ for all ω, and that (5) with $\tau_1 = 0$ holds, it is not possible to find a unique $\theta(\omega)$.†

With regard to results in the opposite direction, when (5) holds and (8) is satisfied, it is true that the phase function $\theta(\omega)$ can be obtained from $|G(\omega)|$ by using (6). This is proved in Section 5.3.

III. EXPERIMENT

For each of two fibers A and B the impulse response was measured by injecting 0.4 nsec pulses (2σ width) from a GaAs laser ($\lambda = 0.9\ \mu$m) and observing the pulse distortions after propagation through the fibers. Fiber A was 1 km long and had a graded index GeO_2–SiO_2 core with $\alpha \approx 1.9$ and $\alpha_{opt}(0.9\ \mu m) \approx 2.0$. Fiber B was 1 km long and had a graded index B_2O_3–SiO_2 core with $\alpha \approx \alpha_{opt}(0.9\ \mu m) \approx 1.8$. The measured impulse responses are shown in Figs. 1a and b. These impulse-response functions were Fourier transformed to obtain $|G(\omega)|$ for each case. Then each $|G(\omega)|$ together with its phase *calculated* from piecewise-linear formulas[4] based on (6) was used to calculate a corresponding impulse response. The plots are shown in Figs. 2a and b for fibers A and B, respectively. It may be seen that the calculated and measured responses agree quite well for fiber B with a suitable normalization and translation to bring them into

† Analogous examples can be given which hold for all $n \geq 2$. For instance, let $H(z, \gamma)$ denote the expression for $H(z)$ for the idealized exponential-excitation case mentioned above, with c chosen to depend on γ so that $H(0, \gamma) = 1$. It is not difficult to verify that we have $|H(i\omega, \gamma)| = |H(i\omega, -\gamma)|$ for all ω for each γ.

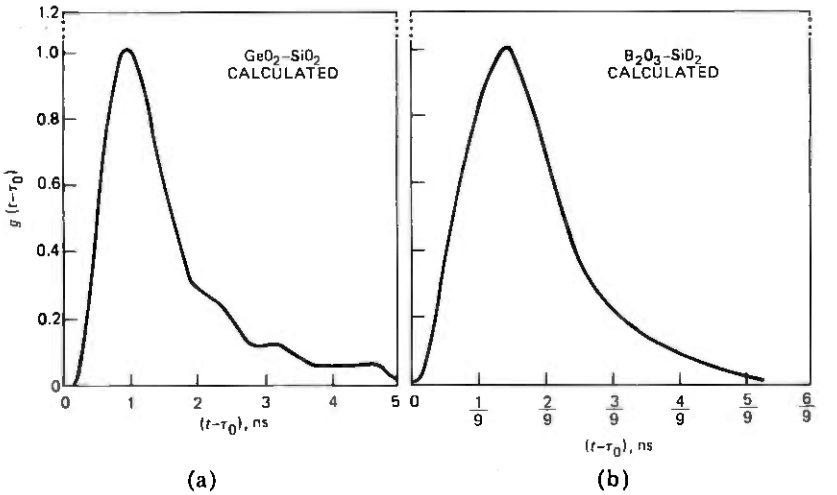


Fig. 2—Impulse responses calculated from the magnitudes of the fourier transforms of the measured impulse responses in Figs. 1(a) and (b).

coincidence. On the other hand, the measured and calculated responses for fiber A do not agree.†

The fact that the functions of Figs. 1a and 2a appear to be approximate mirror images of one another suggests that all of the dominant zeros associated with the transfer function of fiber A might lie in the *right* half-plane. In this connection, we note that a general function $H(z)$ of the form defined in Section II can have zeros in *both* half-planes. For example, with $H(z, \gamma)$ as defined in a preceding footnote, the product $H(z, 1)H(z, -1)$, which can be written in the same form as $H(z)$, has zeros in both half planes.

IV. CONCLUSIONS

With regard to the overall problem of determining the impulse response $g(t)$, direct measurement of the phase appears to be difficult and the general use of the "minimum phase" assumption to calculate the phase does not appear to be justified.

Methods for circumventing the difficulties described are under study, and it is expected that they will be described in a later paper.

V. APPENDIX: PROOFS

5.1 Outline of a derivation of a well-known formula

Let z be a complex variable with real and imaginary parts x and y, respectively, and let z^* denote the complex conjugate of z. Let f be a

† An early computational error led to a reversal of the sign of the time scale for the phase calculations as reported in Refs. 2 and 3. Thus, it was erroneously reported that $g(t)$ for measured and calculated responses matched for fiber A.

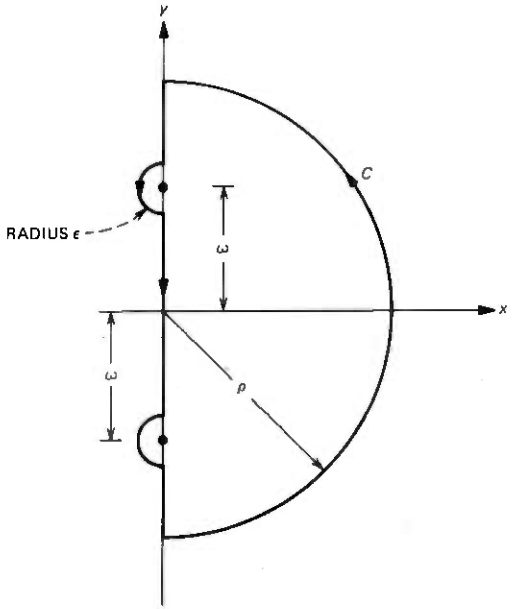


Fig. 3—Contour in the (x, y) plane.

complex-valued function of z defined throughout an open subset S of the (x, y)-plane that contains the half-plane $x \geq 0$ such that: $f(z^*) = f(z)^*$, $\mathrm{Re}[f(iy)]$ is an even function of y, and $\mathrm{Im}[f(iy)]$ is an odd function of y.

Proposition 1. Suppose that f is analytic on S, and that $|f(z)/z| \to 0$ as $|z| \to \infty$ in the half-plane $x \geq 0$. Then

$$\mathrm{Im}[f(i\omega)] = \frac{2\omega}{\pi} P \int_0^\infty \frac{\mathrm{Re}[f(iy)]}{y^2 - \omega^2}\, dy$$

for each real $\omega > 0$, in which P denotes a Cauchy principal value.

Outline of a Proof. Assume that the hypotheses of Proposition 1 are satisfied. Let $\omega > 0$ be given, and choose $\epsilon > 0$ such that $\epsilon < \omega$ and f is analytic for $|z - i\omega| < 2\epsilon$. With ρ any positive number such that $\rho > \omega + \epsilon$, let C denote the contour shown in Fig. 3 which consists of a semicircular arc of radius ρ, two semicircular arcs of radius ϵ, and a portion of the y-axis.

By Cauchy's integral theorem,

$$f(i\omega) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z - i\omega}\, dz$$

Therefore,

$$\mathrm{Im}[f(i\omega)] = \frac{\omega}{2\pi i}\int_C \frac{f(z)}{(z - i\omega)(z + i\omega)}\,dz \tag{10}$$

Consider separately the following contributions to the right side of (10): I_1 the integral along the y-axis from $z = i\rho$ to $z = -i\rho$ excluding the two gaps due to the semicircular ϵ arcs, the sum I_2 of the integrals along the ϵ arcs, and the integral I_3 along the remaining arc of radius ρ.

We find at once that

$$I_1 = \frac{\omega}{\pi}\left[\int_0^{\omega-\epsilon} + \int_{\omega+\epsilon}^{\rho}\right]\frac{\mathrm{Re}[f(iy)]}{y^2 - \omega^2}\,dy$$

Using the fact that the integral of $(z - i\omega)^{-1}$ over the upper ϵ arc is $i\pi$, it follows that $I_2 = \frac{1}{2}\,\mathrm{Im}[f(i\omega)] + \delta(\epsilon)$ in which $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Also, $I_3 = \beta(\rho)$ in which $\beta(\rho) \to 0$ as $\rho \to \infty$.

Since $\mathrm{Im}[f(i\omega)] = I_1 + I_2 + I_3$, we have

$$\mathrm{Im}[f(i\omega)] = 2I_1 + 2\delta(\epsilon) + 2\beta(\rho)$$

from which it is clear that the limit

$$\lim_{\rho\to\infty}\lim_{\epsilon\to 0} 2I_1$$

exists and is equal to $\mathrm{Im}[f(i\omega)]$. This completes the outline of the proof.†

5.2 Result concerning the zeros of H(z)

Let $H(z)$ denote††

$$\sum_{j=1}^{n} d_j e^{-z\tau_j}$$

for all complex z, in which $n \geq 2$, $d_j > 0$ for all j, and the τ_j are real numbers such that $\tau_1 < \tau_2 < \ldots < \tau_n$.

Let $t_1, t_2 \ldots, t_n$ denote *any* set of n real numbers with the property that $t_1 < t_2 < \ldots < t_n$.

Consider the condition

$$d_1 > \sum_{j=2}^{n} d_j \tag{11}$$

Proposition 2. We have $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ if (11) holds. If (11) is not satisfied, then for any positive ϵ there is a choice of the τ_j such that $|\tau_j - t_j| \leq \epsilon$ for all j and $H(z) = 0$ for some z with $\mathrm{Re}(z) \geq 0$.

† For further results concerning Hilbert transforms, see, for example, Ref. 6.
†† For the reader's convenience the definition of $H(z)$ is repeated here.

Proof. If (11) holds and $z = x + iy$ with $x \geq 0$,

$$|H(z)| = \left| e^{-z\tau_1} \left[d_1 + \sum_{j=2}^{n} d_j e^{-z(\tau_j - \tau_1)} \right] \right| \geq e^{-x\tau_1} \left(d_1 - \sum_{j=2}^{n} d_j \right) > 0$$

Suppose now that

$$d_1 \leq \sum_{j=2}^{n} d_j \tag{12}$$

and let ϵ be given. Let $z = x + iy$ with $y = \pi/\epsilon$. Choose $\tau_1 = t_1 - \epsilon$, and for each $j = 2, 3, \ldots, n$, choose τ_j such that $e^{-iy(\tau_j - \tau_1)} = -1$ and $|\tau_j - t_j| \leq \epsilon$. We have

$$H(z) = e^{-z\tau_1} \left[d_1 - \sum_{j=2}^{n} d_j e^{-x(\tau_j - \tau_1)} \right] \tag{13}$$

in which $\tau_j - \tau_1 > 0$ for $j \geq 2$. Using (12) and (13), it is clear that there is an $x \geq 0$ such that $H(z) = 0$, which completes the proof.

5.3 A corollary of proposition 1

Concerning the function H defined in Section 5.2 consider the following hypothesis.

Hypothesis: $\tau_1 = 0$ and (11) is satisfied.

We notice that if the hypothesis holds, then, for each real ω, $\mathrm{Re}[H(i\omega)] > 0$ and we have $H(i\omega) = |H(i\omega)|e^{i\phi(i\omega)}$, in which $\phi(i\omega)$ denotes the principal value of $\tan^{-1}\{\mathrm{Im}[H(i\omega)]/\mathrm{Re}[H(i\omega)]\}$.

Proposition 3. If the hypothesis holds, then

$$\phi(i\omega) = \frac{2\omega}{\pi} P \int_0^\infty \frac{\ln|H(i\omega)|}{y^2 - \omega^2}\, dy$$

for $\omega > 0$ in which P denotes a Cauchy principal value.

Proof: Assume that the hypothesis is satisifed. We shall show that Proposition 1 can be used.

Let x_0 be a negative number such that

$$d_1 > \sum_{j=2}^{n} d_j e^{-x_0\tau_j}$$

We see that there is a positive constant δ such that $\mathrm{Re}[H(z)] > \delta$ for $\mathrm{Re}(z) > x_0$. Let $\phi(z)$ denote the principal value of $\tan^{-1}\{\mathrm{Im}[H(z)]/\mathrm{Re}[H(z)]\}$ for $\mathrm{Re}(z) > x_0$.

For each complex z such that $z \neq 0$, let $p(z)$ denote the principal value of $\ln z$. Thus, $p[H(z)] = \ln|H(z)| + i\phi(z)$ for $\mathrm{Re}(z) > x_0$, and $p[H(z)]$ is analytic in z for $\mathrm{Re}(z) > x_0$.

REFERENCES

1. S. D. Personick, "Baseband Linearity and Equalization in Fiber Optic Digital Communication Systems," B.S.T.J., *52*, No. 7 (September 1973), pp. 1175–1194.
2. L. G. Cohen, H. W. Astle, and I. P. Kaminow, "Wavelength Dependence of Frequency-Response Measurements in Multimode Optical Fibers," B.S.T.J., *55*, No. 10 (December 1976), pp. 1509–1524.
3. L. G. Cohen, H. W. Astle, and I. P. Kaminow, "Frequency Domain Measurements of Dispersion in Multimode Optical Fibers", Appl. Phys. Lett., *30* (1977), pp. 17–19.
4. E. A. Guillemin, *Theory of Linear Physical Systems*, New York: Wiley, 1963, pp. 252, 275, 536, and 552–555.
5. D. Gloge, "Principles of Optical Fiber Transmission," Conference of Optical Fiber Communication, September 1975, London, IEEE Conference Publication 132.
6. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford Press, 1948, pp. 119–151.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Reduction of Network States Under Symmetries

By V. E. BENEŠ

(Manuscript received May 24, 1977)

It is a folk-theorem of traffic theory that if all sources have the same stochastic behavior, then symmetries of a telephone connecting network can be used to lump together equivalent states and to reduce the number of equations to be solved for the state probabilities. The structural and algebraic bases of this idea, and its connections to stochastic models, are studied here by means of concepts from lattice theory, group theory, and combinatorics generally. It is shown that when offered traffic is homogeneous and routing is structurally consistent, the state equations for certain natural Markov processes (representing operating telephone networks) can be substantially simplified by restricting attention to "macrostates," defined as the structural equivalence classes of states, of which there are typically many fewer than of states. Reduced state equations are then obtained for general networks under simple Markovian traffic assumptions.

I. INTRODUCTION

Most telephone connecting networks are built in stages of identical units, arranged symmetrically in frames, and joined by symmetric cross-connect fields (Fig. 1). As a result, their structure has so much symmetry that it is possible quickly to identify at least some network states that are "essentially" equivalent in that their combinatorial structure is the same, and that they differ only in point of renaming terminals, switches, links, customers, etc. It has long been known in the informal lore of traffic theory that if the traffic sources at the terminals have the same stochastic behavior, then these symmetries of the network could be used as a basis for lumping together equivalent states and reducing the number of equations to be solved for the state probabilities.

Such a line of thought was first pursued in a formal way by S. P. Lloyd[1] in unpublished work dated September 19, 1955, and we follow it further

here. But whereas Lloyd right away considered probability transition rate matrices (for Markov processes) which he assumed admitted symmetry operations, we shall instead first relate the relevant symmetries to the network graph and the semilattice of states, without reference to a probabilistic model, and only later consider a natural traffic model. Our approach remains combinatorial as long as possible, and allows us eventually to include the effects of routing decisions, to connect the reduction ideas with *optimal* routing, and to find that certain natural transition matrices *necessarily* admit the network symmetries. In particular, we show that a traffic model used in previous work[2] has a transition rate matrix that admits the symmetries of the network graph, provided only that the routing matrix used also admits these symmetries.

An additional practical incentive for the present study is the fact that for small networks, such as concentrators, it is often possible to press the advantages gained by reducing the states to their equivalence classes, and to solve completely the problem of optimal routing. One can then devise a circuit or a finite-state machine to mechanize the optimal routing policy, as has been done by A. F. Bulfer for the RTA concentrator structure.[3,4] Indeed, historically, it was our attempts to *prove* some of Bulfer's surprising empirical results that led to a realization that a thorough study of structural equivalence of states was valuable and necessary.[5,6]

II. SUMMARY

Various preliminaries are in Sections III and IV: a model for discussing connecting networks, with an account of the role of symmetry. Prior results of S. P. Lloyd on the use of groups to reduce state equations are described in Section V, and there is a heuristic discussion of some necessary conditions on such groups in Section VI: they should be groups of automorphisms of the semilattice of states that preserve the relation of having the same calls up. The symmetry group G_n of the network graph appears in Section VII; it is used in Section VIII to define structural equivalence of states and its associated group G_ν, and in Sections IX and X to define a natural homomorphic image G_η of G_n into G_ν which usually gives a more economical description of equivalence in terms of a group than does G_ν. The semilattice of states induces a partial ordering of the reduced states, described in Section XI.

The next four sections are devoted to detailed calculations of symmetry groups G_n, G_ν, and G_η for various well-known networks: individual 2×2 switches in Section XII, a random slip concentrator in Section XIII, a small three-stage network in Section XIV, and general crossbar networks in Section XV, including frames, cascades of stages, and the No. 5 crossbar type network. The final four sections consider a simple sto-

chastic model for traffic in a network with routing decisions, and show how the reduced state equations are derived in this setting.

It can be concluded that symmetry groups afford a precise definition of structural equivalence for network states; this equivalence in turn allows a substantial reduction in the number of equilibrium equations for state probabilities in suitable stochastic models, thus extending the range of computable examples. For modest networks the reduction method can be used to perform optimal routing in explicit ways.[5,6]

III. PRELIMINARIES

We shall use a model for the structural and combinatorial aspects of a connecting network. This model arises by considering the network structure to be given by a graph G whose vertices are the terminals of the network, and whose edges represent crosspoints between terminals by pairs, with some of the terminals designated as inlets or outlets. Calls in the network are described by paths on G from an inlet to an outlet. Thus a connecting network ν is a quadruple $\nu = (G,I,\Omega,S)$ where G is a graph depicting network structure, I is the set of vertices of G which are inlets, Ω is the set of outlets, and S is the set of permitted or physically meaningful states. It is possible that $I = \Omega$ (one-sided network), that $I \cap \Omega = \phi$ (two-sided network), or that some intermediate condition obtain, depending on the "community of interest" aspects of the network ν. Variables w, x, y, and z at the end of the alphabet denote states, while u and v denote a typical inlet and a typical outlet, respectively.

A possible state x can be thought of as a set of disjoint chains on G, each joining I to Ω. Not every such set of chains need represent a state in S: wastefully circuitous chains may be excluded from S. The set S is partially ordered by inclusions $\leq$, where $x \leq y$ means that state x can be obtained from state y by removing zero or more calls. It is reasonable that if y is a state and x results from y by removal of some chains then x should be a state too; i.e., S should be closed under "hangups." It can be seen from this requirement that the set S of permitted states has the structure of a semilattice, that is, a partially ordered system whose order relation is definable in terms of a binary operation $\cap$ that is idempotent, commutative, and associative, by the formula $x \leq y$ iff $x = x \cap y$. Here for $x \cap y$ we can simply use literal set intersection: $x \cap y$ is exactly the state consisting of those calls and their respective routes which are common to x and y.

An *assignment* is a specification of what inlets are to be connected to what outlets. The set A of assignments can be represented as the set of all fixed-point-free correspondences from subsets of I to Ω. The assignments form a semilattice in the same way that the states do, and A is related to S as follows: call two states x,y in S equivalent as to assignment, written $x \sim y$, iff all and only those inlets $u \epsilon I$ are connected

in x to outlets $v \epsilon \Omega$ which are connected to the same v in y, though possibly by different routes; the realizable assignments can then be identified with the equivalence classes of states under $\sim$, and there is a natural map γ: $S \rightarrow A$, the projection that carries each state x into the assignment $\gamma(x)$ it realizes, i.e., the equivalence class it belongs to under $\sim$.

With x and y states such that $x \geq y$, it is convenient to use x-y to mean the state resulting from x by removing from x all the calls in y. Similarly, with a and b assignments such that $a \geq b$, we use a-b to mean the assignment resulting from a by dropping all the connections intended in b. Note that here x-y, a-b have their usual set-theoretic meaning.

It can now be seen that the map γ is a semilattice homomorphism of S into A, with the properties:

$$x \geq y \Rightarrow \gamma(x) \geq \gamma(y)$$

$$x \geq y \Rightarrow \gamma(x - y) = \gamma(x) - \gamma(y)$$

$$\gamma(x \cap y) \leq \gamma(x) \cap \gamma(y)$$

$$\gamma(x) = \phi \Rightarrow x = 0 = \text{zero state, with no calls up}$$

Not every assignment need be realizable by some state of S. Indeed, it is common for practical networks to realize only a vanishing fraction of the possible assignments, and the networks that do realize every assignment, the so-called *rearrangeable* networks, have been the objects of substantial theoretical study. Thus the image set $\gamma(S)$ of realizable assignments is typically much smaller than the set A it is embedded in. A *unit* assignment is, naturally, one that assigns exactly one outlet to some inlet, and it corresponds to having just one call in progress. It is convenient to identify calls c and unit assignments, and to write $\gamma(x) \cup c$ for the larger assignment consisting of $\gamma(x)$ and the call c together, with the understanding of course that c is "new in x" in the sense that neither of its terminals is busy in x.

We denote by A_x the set of states that are immediately above x in the partial ordering $\leq$ of S, and by B_x the set of those that are immediately below. Thus

A_x = {states reachable from x by adding a call}

B_x = {states reachable from x by a hangup}

For c new in x, let $A_{cx} = A_x \cap \gamma^{-1}[\gamma(x) \cup c]$; A_{cx} is the subset of states of A_x that could result from x by putting up the call c, because $\gamma^{-1}\gamma(y)$ is precisely the equivalence class of y under $\sim$. If A_{cx} is empty then we say c is *blocked* in x: there is no $y \epsilon A_x$ which realizes the larger assignment $\gamma(x) \cup c$. It can be seen that with F_x the set of new calls of x that are not

blocked, the family $\{A_{cx}, c \epsilon F_x\}$ forms the partition of A_x induced by the equivalence $\sim$.

IV. HOW DOES SYMMETRY HELP?

Symmetry in the structure of a connecting network, together with the theoretically convenient (but in practice false) assumption of homogeneous or interchangeable traffic sources, leads to simplifications in calculating state probabilities, loss, carried load, and other traffic quantities. In most cases the simplification occurs because network states that have the same structure to within renaming of customers, links, and switches are in a definite sense equivalent, and because if traffic sources are interchangeable, such equivalent states can turn out to have the same equilibrium or transition probabilities. Whether they do or not depends on the rest of the traffic model, especially on the rule used to make routing decisions: roughly speaking, if the rule is *consistent* in that it opts for analogous routes for analogous calls in equivalent states, then equivalent states will (or at least can) have the same probabilities. In such cases the state probabilities can be calculated from those of the equivalence classes, of which there are usually many fewer, by considering, in place of the original microscopic stochastic process on the set of states, a macroscopic one taking values on the equivalence classes.

Our problem in this paper is to make all these notions, especially that of "equivalent" states, as precise as possible in the general network setting. In view of the central role of symmetry it is natural to expect that the equivalence idea be expressed mathematically by means of group theory, and particularly, in terms of the symmetry group of the graph G depicting network structure. Applications to optimal routing in networks and concentrators will appear in later work.[5,6] These applications are considerably complicated by the following "problem of refusals": It turns out that analytical methods for finding optimal routing rules are greatly simplified if, as operator of the network, the telephone company is allowed the option of refusing to complete an unblocked call if it thinks that this denial of service will improve performance according to some criterion of interest; with this added option the task of finding out when to decline unblocked calls is part of the routing problem, a part which it turns out is usually much harder than actually choosing the best route if the call is to go in; however, it is often possible to solve the routing problem *up to refusals,* that is, to specify optimal routes for calls that might go in without ruling on whether they go in or not.

V. PRIOR RESULTS OF S. P. LLOYD

The relevance of group theory was well understood in 1955 by S. P. Lloyd, whose unpublished work[1] is now sketched. He used a standard

method of identifying an equivalence relation on a set with a group of bijections of the set into itself. This method considers a group G_v of correspondences of S onto itself, and describes "equivalent" states as follows: G_v is said to be *transitive* on a subset X of S iff

(*i*) $x \epsilon X, g \epsilon G_v \Rightarrow gx \epsilon X$

(*ii*) $x,y \epsilon X \Rightarrow \exists g \epsilon G_v$ such that $gx = y$

X is then called a transitive set. With $|X|$ the cardinality of X, it can be shown that $|X|$ divides the order of G_v, and that each member of X appears exactly

$$\frac{|G_v|}{|X|}$$

times in the array gx, $g \epsilon G_v$, where x is any element of X. For each $x \epsilon S$ define $\pi x = \{gx : g \epsilon G_v\}$. It can be seen that each πx is a transitive set, and that for any $x,y \epsilon S$ we have either $\pi x = \pi y$ or $\pi x \cap \pi y = \phi$. Thus the πx form a partition of S, and so G_v induces a corresponding equivalence relation $\equiv$ according to $x \equiv y$ iff there is a $g \epsilon G_v$ such that $gx = y$. Conversely, given an equivalence relation $\equiv$ on S, the set of all bijections $g:S \leftrightarrow S$ with $gx \equiv x$ forms a group G_v under composition which induces $\equiv$ in the sense above, and we have the following "summation formula:" With α an equivalence class of $\equiv$, x any member of α, and f a real valued function on S,

$$\sum_{y \epsilon \alpha} f(y) = \frac{|\alpha|}{|G_v|} \sum_{g \epsilon G_v} f(gx)$$

Now let x_t be a continuous-parameter Markov process taking values on S, with a stationary transition rate matrix $Q = (q_{xy})$, assumed to be ergodic. There is then a unique probability vector $p = \{p_x, x \epsilon S\}$ such that p solves the "statistical equilibrium" equations

$$\sum_{x \epsilon S} p_x q_{xy} = 0, \qquad y \epsilon S \tag{1}$$

p is the stationary probability distribution for x_t. The group G_v and the relation $\equiv$ become relevant to eq. (1) when the matrix Q of rates is unaffected by the permutations (of its rows and columns) corresponding to $g \epsilon G_v$; indeed, if G_v and $\equiv$ express what we mean by saying that equivalent states differ only in respect of renaming customers, switches, etc., this is the precise way that symmetry affects the problem of calculating state probabilities. This relevance is recognized in the following

Def. 1: *Q admits* G_v iff $g \epsilon G_v$, $x,y \epsilon S \Rightarrow$

$$q_{xy} = q_{(gx)(gy)}$$

which leads quickly to Lloyd's basic 1955 results:

Theorem 1: Let G_ν be a group of bijections of S onto itself such that Q admits G_ν, and let $E = \{\pi x, x \epsilon S\}$ be the set of equivalence classes induced by G_ν. Then

(*i*) The projection map $\pi: x \to \pi x$ defines a "macroscopic" Markov process πx_t on E, with transition rates

$$q_{\alpha\beta} = \sum_{y \epsilon \beta} q_{xy} \text{ for } x \epsilon \alpha \text{ and } \alpha \epsilon E$$

and stationary probabilities $\{p_\alpha, \alpha \epsilon E\}$ satisfying

$$\sum_{\alpha \epsilon E} p_\alpha q_{\alpha\beta} = 0, \beta \epsilon E$$

(*ii*) For each $x \epsilon \alpha \epsilon E$,

$$p_x = \frac{1}{|\alpha|} p_\alpha$$

Thus if Q admits G_ν, then equivalent states have *the same* stationary probabilities, and these can be computed from a *reduced state equation* of lower order.

Proof of Theorem 1: Everything follows from the fact that if $\beta \epsilon E$, then

$$\sum_{y \epsilon \beta} q_{xy}$$

has the same value for all $x \epsilon \alpha \epsilon E$; so we prove this first. If $z \epsilon \alpha$ then by transitivity of α there is a $g \epsilon G_\nu$ with $x = gz$, so that

$$\sum_{y \epsilon \beta} q_{xy} = \sum_{y \epsilon \beta} q_{(gz)y} = \sum_{y \epsilon \beta} q_{(gz)(gy)} = \sum_{y \epsilon \beta} q_{zy}$$

The second equality arises from $g^{-1}\beta = \beta$, the third from Q's admitting G_ν. Since x and z were arbitrary elements of α, the result is proved. It implies, by results[7] of M. Rosenblatt and C. J. Burke, that πx_t is a Markov process with transition rates

$$q_{\alpha\beta} = \sum_{y \epsilon \beta} q_{xy}, \text{ any } x \epsilon \alpha$$

and πx_t is ergodic if x_t is. Hence it has a unique stationary probability distribution $\{p_\alpha, \alpha \epsilon E\}$ satisfying

$$\sum_{\alpha \epsilon E} p_\alpha q_{\alpha\beta} = 0, \beta \epsilon E$$

Thus (i) is proved; to prove (ii) we show first that

$$p_\alpha = \sum_{x \epsilon \alpha} p_x$$

where $\{p_x, x \epsilon S\}$ is the stationary distribution of x_t. We find

$$\sum_{\alpha \epsilon E} \sum_{x \epsilon \alpha} p_x q_{\alpha\beta} = \sum_{x \epsilon S} p_x \sum_{y \epsilon \beta} q_{xy} = \sum_{y \epsilon \beta} \sum_{x \epsilon S} p_x q_{xy} = 0$$

since the inner sum in the last term is always zero. Thus

$$\left\{ \sum_{x \epsilon \alpha} p_x, \ \alpha \epsilon E \right\}$$

satisfy the (reduced) equilibrium equations for πx_t, and so by the uniqueness of its (probability vector) solution,

$$p_\alpha = \sum_{x \epsilon \alpha} p_x$$

Now define q by $q_x = |\pi x|^{-1} p_{\pi x}$ and consider that

$$\begin{aligned} 0 = \sum_{\alpha \epsilon E} p_\alpha q_{\alpha\beta} &= \sum_{x \epsilon S} |\pi x|^{-1} p_{\pi x} q_{(\pi x)\beta} \\ &= \sum_{x \epsilon S} q_x \sum_{y \epsilon \beta} q_{xy} = \sum_{y \epsilon \beta} \sum_{x \epsilon S} q_x q_{xy} \end{aligned}$$

However if $z \epsilon \beta$, there is $g \epsilon G_\nu$ with $y = gz$, and (since q is constant on equivalence classes) $q_x = q_{(g^{-1}x)}$ so that

$$\sum_{x \epsilon S} q_x q_{xy} = \sum_{x \epsilon S} q_x q_{x(gz)} = \sum_{x \epsilon S} q_{(g^{-1}x)} q_{(g^{-1}x)z} = \sum_{x \epsilon S} q_x q_{xz}$$

Thus

$$\sum_{x \epsilon S} q_x q_{xy}$$

is constant over equivalence classes and so it must be zero. Hence q is a probability vector solution of the equilibrium equation for x_t, so $q = p$, i.e., $p_x = |\pi x|^{-1} p_{\pi x}$.

Lloyd's theorem accurately captured the relevance of the rate matrix Q's admitting the group G_ν, and he gave examples of the application of his result to small networks, such as individual switches and partial access concentrators, but he did not elaborate on the groups G_ν to be considered. However, in applying such a result to traffic in connecting networks we want to be sure that the groups G_ν we use reflect the intuitive notions of invariance of structure under renaming of terminals, switches, etc. The theorem thus leaves us with these important questions: What groups G_ν are appropriate or available for describing equivalence

relations ≡ useful in applications to traffic in networks? What traffic models give rise to rate matrices Q that admit these groups?

Furthermore, since we expect the applications we make of Lloyd's theorem to networks to depend on both network symmetry and customer interchangeability, not to mention routing, it would be well to have a formulation in which these items are clearly separated, as they are not in Theorem 1. What we need is a more specific definition of "equivalence" of states, one independent of probabilistic models, and peculiar to the network applications we intend, and one that reflects the idea of invariance of structure under renaming. These requirements will be met by constructing some groups that are appropriate from the symmetry group of the network graph G; Lloyd entertained[8] such an idea but did not describe it in Ref. 1. We shall first argue that certain natural necessary conditions, to be met by groups considered "appropriate," imply that they should be automorphism groups of $(S, \leq)$ whose elements preserve $\sim$; then we show how such groups arise directly from the symmetry group of the network graph.

VI. TWO INTUITIVELY NECESSARY CONDITIONS

We need properties and concepts that help make precise the notion of "structurally equivalent" states. Some of these will now be arrived at quickly and intuitively. Consider therefore two states x and y that differ only in point of renaming terminals and links, but otherwise have the same structure. In such a situation we expect to be able to make a correspondence between the calls in progress in x and those in y, because structural equivalence requires that each call c in progress in x have at least one analog in progress in y, playing the role of c in the structure of y. This being so, we see in the same way that the elements of B_x (obtainahle from x by a hangup) have a natural correspondence to those of B_y, going beyond the fact that $|B_x| = |B_y| = |x| = |y|$: namely, to $z \epsilon B_x$ we assign a state in B_y obtained by hanging up an analog of the call hung up in x to get z, i.e., an analog of $\gamma(x - z)$.

An exactly similar situation holds for the sets A_x and A_y of states which are respectively above x and y; actually, more is true, since for every call c free and not blocked in x there will be an analog (possibly more than 1), call it c', free and not blocked in y, and to every way of putting up c in x, i.e., for every $z \epsilon A_{cx}$, there will be a corresponding way of putting up an analog c' in y, and thus a natural correspondence of A_{cx} with $A_{c'y}$. It is intuitively clear that $A_{c'y}$ cannot have either more or fewer states than A_{cx}, else there would have to be something structurally different about x and y to account for the discrepancy.

Let now g denote the partial natural correspondence we have so far, with the properties $gx = y$, $g(B_x) = B_{gx}$, and $g(A_{\gamma(z-x)x}) = A_{\gamma(gz-gx)gx}$

for $z \epsilon A_x$. As we indicated, not even this much of g is unique. Nevertheless we suggest that g can be extended in a similar way, and to be defined on all of S in such a way as to satisfy the commutation $g(B_x) = B_{gx}$ and the distribution $g(A_{\gamma(z-x)x}) = A_{\gamma(gz-gx)gx}$ for $z \epsilon A_x$. It is not hard to see that these conditions are the same as

$$x \leq y \Rightarrow gx \leq gy$$

$$x \sim y \Rightarrow gx \sim gy$$

The map g then takes any state into one equivalent in structure, in such a way as to preserve order in $(S, \leq)$ and also equivalence in the other sense of having the same people talking. We can expect that $g(S) = S$, and that g is one-to-one.

Now an isomorphism between two partially ordered systems (POS) is precisely a bijection that preserves order both ways; in our case the two POS coincide, so g is called an *automorphism* of $(S, \leq)$. Since equivalence is transitive, it follows that if there are automorphisms g, h such that $y = gx$, $z = hy$, then there must be one f such that $z = fx$, namely the composition $f = hg$. Hence one is led naturally to consider groups of automorphisms that preserve $\sim$. Thus while *any* equivalence relation on S can be described by a group of bijections on S, we claim that to adequately express what is meant by structure invariant under renaming, the groups of interest for a theory of networks should be automorphism groups of $(S, \leq)$. Thus the first part of this study of equivalence $\equiv$ between states is the search, for a general network ν, for a group G_η of automorphisms of $(S, \leq)$ that preserve $\sim$ such that the "usual" definition of equivalence via a group, viz.,

$$x \equiv y \text{ iff } \exists g \epsilon G_\eta \ni x = gy$$

agrees with what we mean by structural equivalence.

VII. SYMMETRIES OF THE NETWORK GRAPH

Structural equivalence of states rests ultimately on properties of the network graph G that are independent of and prior to the choice of inlets I, outlets Ω, and states S that complete the description of a network $\nu = (G, I, \Omega, S)$. In an informal way, one might say that the equivalence of two states under renaming of terminals and links really depends on what it means for the network to "look the same" to distinct terminals. As an example consider two arbitrary distinct inputs on the left side of the standard No. 5 crossbar type network in Fig. 1. It is obvious intuitively that if the frames and switches are identical, and the connections within and between frames correspond to complete bipartite graphs in the usual way, then the network "looks the same" to two such inlets. The same is true of any two interframe junctors, or of any two links from the same

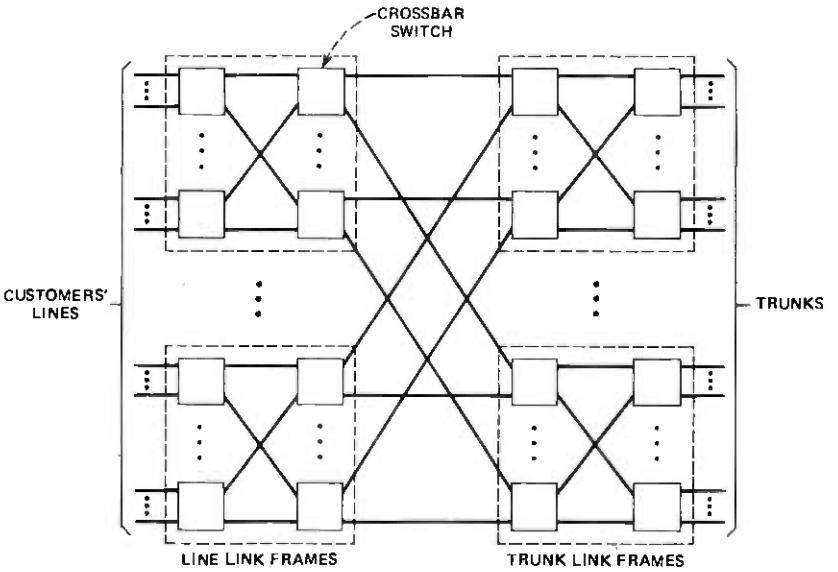


Fig. 1—Connecting network.

or from two distinct frames. We seek to clarify this informal notion of "looking the same," and to develop it into a precise definition of structural equivalence in terms of a natural symmetry group for the network graph.

Let us think of the terminals of ν as the vertices, and of the crosspoints as the edges, of the network graph G. It is clear that G is determined by giving a relation N on the set T of terminals such that for t,s in T, tNs iff there is a crosspoint or edge between t and s. N is the symmetric "nextness" or adjacency relation that completely depicts the network structure.

Now suppose that we rename the terminals in T according to some permutation τ. Most permutations would play havoc with the adjacency relation; that is, if t and s had a crosspoint between them, then τt and τs easily might not, and conversely. But there might be *some* permutations other than the identity which *preserved* N in the sense that for every $t,s \epsilon T$

$$tNs \text{ iff } (\tau t)N(\tau s)$$

In this case the permuted terminals have crosspoints between them in exactly the same pattern as the unpermuted. It is the existence of such an "N-preserving" permutation τ that we take as the precise meaning of "looking the same."

The network "looks the same" to two terminals t and s iff there is an N-preserving permutation τ of T into itself such that $s = \tau t$. It is ap-

parent that these N-preserving permutations form a group, which we call the symmetry group of the network graph G. We remark that this group may be trivial (if there are no N-preserving permutations except the identity), and that in any case it in no way depends on what terminals have been designated as inlets or outlets, or on what ways of closing the crosspoints are to be allowed as physically meaningful states.

In most telephone connecting networks the set I of inlets and that Ω of outlets are fixed sets of terminals, and one is not interested in whether the network "looks the same" to an inlet as it does to an intermediate link or junctor; in most cases it will not, in any case, because of their different functions in operation. So it makes sense to restrict the N-preserving bijections we think of as renamings of switches and terminals to those which either preserve both I and Ω, or else map each onto the other. This restriction defines a subgroup G_n, called the symmetry group of G for I and Ω, which will be used to define structural equivalence of states.

VIII. SYMMETRIES OF THE SET OF STATES

The set S of states of a network $\nu = (G,I,\Omega,S)$ represents all the ways of closing the crosspoints which we regard as physically sensible. It is closed under hangups, that is, under removal of a chain from a state; it need not be closed under adding new chains, nor even under adding new chains which by themselves already represent a state with one call in progress. It is convenient, however, to require that sets of chains in a structural equivalence class either all belong to S, or that none of them does. This requirement of course implicitly assumes that we know what structural equivalence is before we choose states for S; it will be seen that the definition of equivalence below applies to *arbitrary* sets of chains on G, so the requirement can be met as we choose such sets to belong to S. Specifying the set S represents definite choice of the ways in which the network with graph G is to be used.

We shall now use the symmetry group G_n of the network to define what we mean by two states' differing only in point of renaming links, terminals, etc. Indeed, it can be seen intuitively that the symmetry group G_n of the network induces a natural equivalence relation on sets of chains on G, and thus on whatever such sets we choose as states: two sets of chains on G are equivalent if there is some group element $\tau \epsilon G_n$ such that a terminal t is busy in the first iff τt is busy in the second. The incorporation of a map $\tau \epsilon G_n$ in the definition ensures that *simultaneously* the network looks the same to a terminal t and to its analog τt. For states, then, we define structural equivalence $\equiv$ by

Def. 2: $x \equiv y$ iff $\exists\, \tau \epsilon G_n \ni t_1, \ldots, t_m$ are the terminals busy in x iff $\tau t_1, \ldots, \tau t_m$ are those busy in y.

It is apparent that since G_n is a group, $\equiv$ is an equivalence relation on S, partitioning S into a set E of equivalence classes $\alpha,\beta,\ldots$, and inducing a natural projection map $\pi: S \to E$ such that

$$\pi x = \{y : y \equiv x\} \epsilon E$$

The elements α of E will be the *reduced network states*. As in Section V, we see that there is an associated group G_ν that provides an alternative description of $\equiv$; G_ν is the group of all bijections $g: S \leftrightarrow S$ which map each equivalence class α into itself. In fact, G_ν is the largest strictly imprimitive group of bijections of S onto itself whose sets of imprimitivity are precisely the $\alpha \epsilon E$:

Def. 3: $G_\nu = \{g : S \leftrightarrow S \ni g(\alpha) = \alpha \text{ for } \alpha \epsilon E\}$

Remark 1: As in Section V, we have $x \equiv y$ iff $\exists g \epsilon G_\nu \ni gx = y$.

Remark 2: Although G_ν is a group of bijections which does characterize $\equiv$, it is typically *not economical*. It turns out that a much smaller subgroup of G_ν, defined directly from G_n, suffices to characterize $\equiv$ as in Remark 1. These subgroups appear in Section 10, and they are the $\sim$-preserving automorphism groups desired in Section VI.

IX. ACTIONS OF $\tau \epsilon G_n$ ON A AND S

An assignment a is a correspondence or injection from a subset of I into Ω. Thus an element $\tau \epsilon G_n$ acts in a natural way on an assignment $a \epsilon A$ to produce a new assignment τa according to the rule that u and v correspond in a iff τu and τv correspond in τa. Similarly an element $\tau \epsilon G_n$ acts in a natural way on a set X of chains on the graph G to produce a new set τX of chains consisting of the τ-images of chains in X, thus: $t_1, \ldots, t_l$ is to be a chain of X iff $\tau t_1, \ldots, \tau t_l$ is a chain of τX. In particular a $\tau \epsilon G_n$ acts on a state $x \epsilon S$ to produce a set of chains τx, and it is reasonable to assume, as we have done here, that the choice of S is consistent with the symmetry group G_n in that S is closed under the action of any $\tau \epsilon G_n : x \epsilon S$, $\tau \epsilon G_n \Rightarrow \tau x \epsilon S$. This will ensure that either all the sets of chains in a structural equivalence class are states, or none of them is.

Remark 3: Since τ may map Ω onto I its action on $a \epsilon A$ may reverse the "usual" order of the pairs $(u,v) \epsilon I \times \Omega$ to $(\tau u, \tau v) \epsilon \Omega \times I$. Therefore we do not distinguish between an assignment

$$a = \{(u,v) \epsilon I \times \Omega : (u,v) \epsilon a\}$$

from its inverse

$$a^{-1} = \{(v,u) \epsilon \Omega \times I : (u,v) \epsilon a\}$$

or else we specify that when $\tau I = \Omega$ then the action is defined by

$$\tau a = \{(\tau v, \tau u) : (u,v) \epsilon a\}$$

X. HOMOMORPHISM OF G_n INTO G_ν

The network symmetry group G_n depends only on the basic network structure: the adjacency relation N of the graph G that depicts the network, and the choice of I and Ω, since $\tau \epsilon G_n$ are restricted to preserve inlets and outlets. The state symmetry group G_ν describing $\equiv$, however, also depends on what sets of chains on the graph G are chosen as physically interesting or important states. Since G_n is used in the definition of $\equiv$, and thus of G_ν, it is not surprising that G_n and G_ν happen to be algebraically related: there is natural homomorphic image G_η of G_n in G_ν, consisting entirely of $\sim$-preserving automorphisms of $(S, \leq)$, and describing the same equivalence relation $\equiv$. Thus for most purposes it is more convenient to use the "equivalent" subgroup G_η than the full symmetry group G_ν associated with $\equiv$ by the standard method. The big group G_ν induced by $\equiv$, incidentally, is not necessarily an automorphism group: consider, e.g., the 2×2 switch of Fig. 4 supra; clearly $5 \equiv 6$ and $1 \equiv 3$, but $1 \leq 5$ and $3 \not\leq 6$. Thus the map g defined by the permutation (13) (24) (56) belongs to G_ν but is not an automorphism.

Theorem 2: The action of $\tau \epsilon G_n$ on states x defines a homomorphism $\eta: G_n \to G_\nu$ according to the rule that $\eta(\tau)x = \tau x$; the image group $G_\eta = \eta(G_n)$ is an automorphism group that preserves $\sim$.

Proof: For $\tau \epsilon G_n$, define $\eta(\tau) \epsilon G_\nu$ by the condition $\eta(\tau)x = \tau x$ that $t_1, \ldots, t_l$ is a chain of $x \epsilon S$ iff $\tau t_1, \ldots, \tau t_l$ is a chain of $\eta(\tau)x (= \tau x)$. The subset $\eta(G_n)$ is closed under composition, so it is a subgroup G_η of G_ν. If $x \leq y$, then $\eta(\tau)x \leq \eta(\tau)y$ for each $\tau \epsilon G_n$, so that each $\eta(\tau)$ is an automorphism of $(S, \leq)$. It is easily verified that $\eta(\tau_1 \tau_2) = \eta(\tau_1)\eta(\tau_2)$; preservation of $\sim$ follows from the definition of action.

Theorem 3: If every terminal $t \epsilon T$ is busy in some state $x \epsilon S$, then the homomorphism η of Theorem 2 is an isomorphism, i.e., it is injective (one-to-one).

Proof: If $\eta(\tau_1) = \eta(\tau_2)$, let x be a state with one call up in which a terminal t is busy. Then x consists of a single chain $t_1, \ldots, t_l$ such that t is some t_i and $\tau_1 t_j = \tau_2 t_j$ for $1 \leq j \leq l$. Then $\tau_1 t = \tau_2 t$, and since t is arbitrary we have $\tau_1 = \tau_2$, and so η is injective.

Remark 4: The following conditions are all equivalent:

(*i*) $x \equiv y$
(*ii*) For some $\tau \epsilon G_n$, $x = \tau y = \eta(\tau)y$
(*iii*) For some $h \epsilon G_\eta$, $x = hy$
(*iv*) For some $g \epsilon G_\nu$, $x = gy$

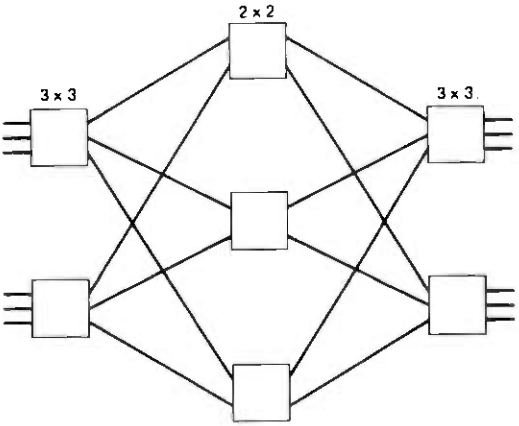


Fig. 2—Small 3-stage network.

XI. PARTIAL ORDERING OF REDUCED STATES

It is natural to try to partially order the set E of reduced states according to this idea: an equivalence class α is "above" another β if some element $x \epsilon \alpha$ can be reached from some member $y \epsilon \beta$ by adding more calls, i.e., if $x \geq y$ for some $x \epsilon \alpha$ and $y \epsilon \beta$. Formally, we make the

Def. 4: $\alpha \geq \beta$ iff $\exists x \epsilon \alpha, y \epsilon \beta \ni x \geq y$

For this to define a partial ordering it must be reflexive, antisymmetric, and transitive. The first is obvious; we prove the other two. Let then $\alpha \geq \beta$ and $\beta \geq \alpha$; to show $\alpha = \beta$ it is enough to show $\alpha \cap \beta \neq \phi$, because each is an equivalence class. There exist $x,z \epsilon \alpha$ and $y,w \epsilon \beta$ such that $x \geq y$ and $w \geq z$. Since $x \equiv z$ and $y \equiv w$, there exist *automorphisms* $\tau_1,\tau_2 \epsilon G_\eta$ such that $\tau_1 x = z$ and $\tau_2 y = w$. Hence $\tau_1 x \geq \tau_1 y$ and

$$\tau_2 y = w \geq z = \tau_1 x \geq \tau_1 y$$

so that $\tau_1^{-1}\tau_2 y \geq y$. But a state z that is above another $\tau_1 y$ and has the same number of calls in progress equals it: $z = \tau_1 y$. So $\alpha \cap \beta \neq \phi$.

To prove transitivity let $\alpha \geq \beta$ and $\beta \geq \gamma$, so that there exist $x \epsilon \alpha$, $y \epsilon \beta$, $z \epsilon \beta$, and $w \epsilon \gamma$ such that $x \geq y$, $y \equiv z$, and $z \geq w$. There is an automorphism $\tau \epsilon G_\eta$ such that $\tau y = z$, whence

$$\tau x \geq \tau y = z \geq w$$

Hence there is something in α which is above something in γ, i.e., $\alpha \geq \gamma$. Thus $\geq$ is a partial ordering on $E = \pi(S)$.

Remark 5: $x \geq y \Rightarrow \pi x \geq \pi y$

Remark 6: The partial order $\geq$ on E need not be a semilattice, and *a fortiori* the projection $\pi: S \rightarrow E$ need not be a semilattice homomorphism, as is $\gamma: S \rightarrow A$, if A is ordered by inclusion. Figures 2 and 3 provide a

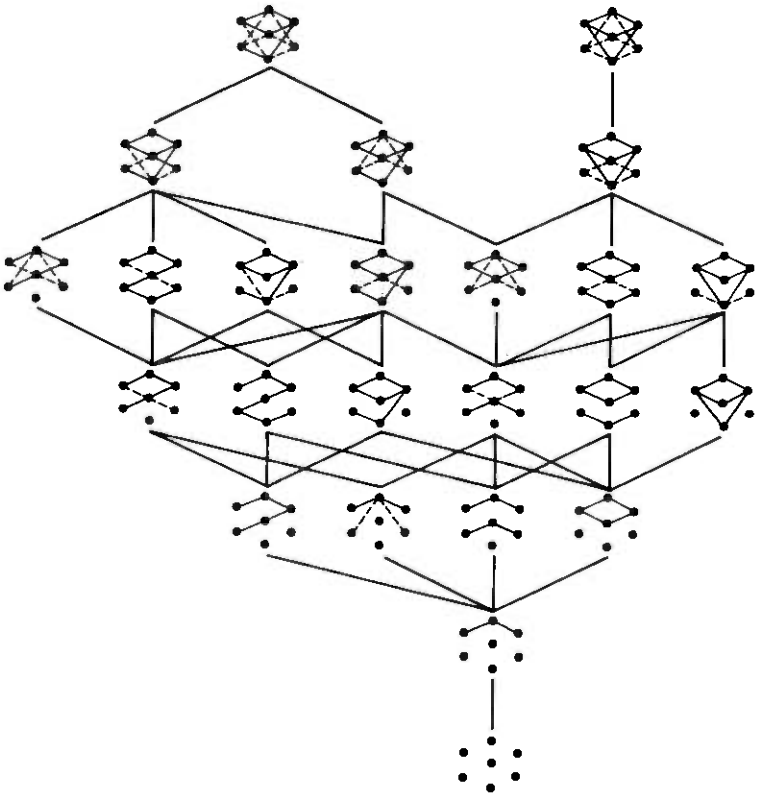


Fig. 3—Reduced state diagram for small 3-stage network of Fig. 2.

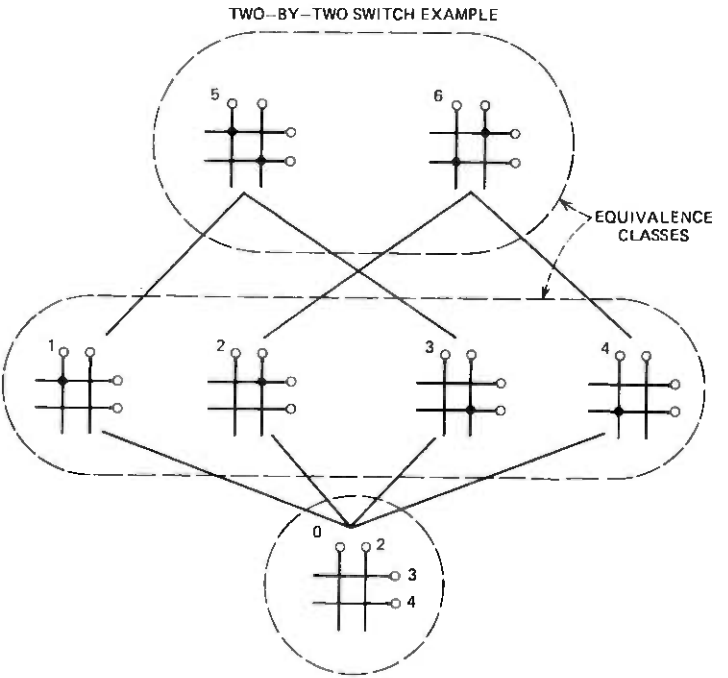
counterexample, as well as an illustration of the reduced states of a network.

XII. 2 × 2 SWITCH

The next four sections are devoted to increasingly complex examples that will illustrate the notions we have introduced to make precise the idea of structural equivalence. Our first and simplest example is the 2 × 2 switch shown in Fig. 4. We arbitrarily label the inlets 1 and 2 and the outlets 3 and 4. So with $I = \{1,2\}$ and $\Omega = \{3,4\}$ the network graph G is a square with the terminals of I on diagonally opposite vertices, and similarly for Ω. The adjacency relation N consists of exactly the pairs (1,3), (1,4), (2,3), and (2,4) together with the results of interchanging the first and second members of these pairs so as to make N symmetric.

The maps of the terminal set $T = \{1,2,3,4\}$ into itself which preserve N and either preserve both I and Ω, or carry one onto the other, are exactly the permutations

identity, (12), (34), (12)(34), (13)(24), (23)(14), (1423), (1324)



		T			
	N	1	2	3	4
T	1	0	0	1	1
	2	0	0	1	1
	3	1	1	0	0
	4	1	1	0	0

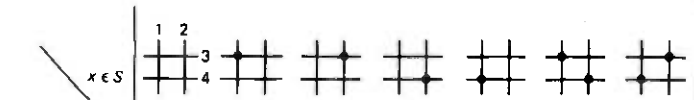
Fig. 4—Network graph, adjacency matrix, and states for 2 × 2 switch.

These 4-permutations form the symmetry group G_n of the 2 × 2 switch for $I = \{1,2\}$, $\Omega = \{3,4\}$. This group has a nice geometric meaning in terms of the network graph, the square in Fig. 4: it consists precisely of all the rotations and reflections of the square into itself. The multiplication table for this group is given in Table I, along with "generators" $A = (1423)$ and $B = (12)$ which, under the relations $A^4 = B^2 =$ identity, $BA = A^3B$, identify the group as (an isomorph of) the dihedral group D_4 of order 8. Every terminal is busy in some state, so Theorem 3 applies and we need only calculate G_n to define $\equiv$, instead of passing through G_ν.

The symmetry group G_n describes the equivalence of states through the action table in Fig. 5. Here the columns are indexed by states $x \epsilon S$ and the rows by group elements $\tau \epsilon G_n$, and the τ,x entry is the state τx defined by the action of τ on x, i.e., by replacing every chain $(t_1,t_2)\epsilon x$ by $(\tau t_1, \tau t_2)$.

Table I — Multiplication table of G_n for 2 × 2 switch

g \ f	I	B (12)	A^2B (34)	A^2 (12)(34)	AB (13)(24)	A^3B (14)(23)	A (1423)	A^3 (1324)
I	I	(12)	(34)	(12)(34)	(13)(24)	(14)(23)	(1423)	(1324)
(12)	(12)	I	(12)(34)	(34)	(1324)	(1423)	(14)(23)	(13)(24)
(34)	(34)	(12)(34)	I	(12)	(1423)	(1324)	(13)(24)	(14)(23)
(12)(34)	(12)(34)	(34)	(12)	I	(14)(23)	(13)(24)	(1324)	(1423)
(13)(24)	(13)(24)	(1423)	(1324)	(14)(23)	I	(12)(34)	(12)	(34)
(14)(23)	(14)(23)	(1324)	(1423)	(13)(24)	(12)(34)	I	(34)	(12)
(1423)	(1423)	(13)(24)	(14)(23)	(1324)	(34)	(12)	(12)(34)	I
(1324)	(1324)	(14)(23)	(13)(24)	(1423)	(12)	(34)	I	(12)(34)



$\tau \in G_n$ \ $x \in S$	0	1	2	3	4	5	6	$\eta(\tau)$
I	0	1	2	3	4	5	6	I
(12)	0	2	1	4	3	6	5	(12) (34) (56)
(34)	0	4	3	2	1	6	5	(14) (23) (56)
(12) (34)	0	3	4	1	2	5	6	(13) (24)
(13) (24)	0	1	4	3	2	5	6	(24)
(14) (23)	0	3	2	1	4	5	6	(13)
(1423)	0	2	3	4	1	6	5	(1234) (56)
(1324)	0	4	1	2	3	6	5	(1432) (56)
	$\tau x\ (= \eta(\tau) x)$							

Fig. 5—Action and isomorphism for the 2 × 2 switch.

Each row of the table defines a bijection of $S = \{0, \ldots, 6\}$ into itself that corresponds to the row index $\tau \epsilon G_n$, and thus an isomorphism $\eta: G_n \rightarrow S_7$ depicted by

τ	id	(12)	(34)	(12)(34)	(13)(24)	(23)(14)
$\eta(\tau)$	id	(12)(34)(56)	(14)(23)(56)	(13)(24)	(24)	(13)
					(1423)	(1324)
					(1234)(56)	(1432)(56)

It can be verified that $G_\eta = \eta(G_n)$ is an automorphism group of $(S, \leq)$,

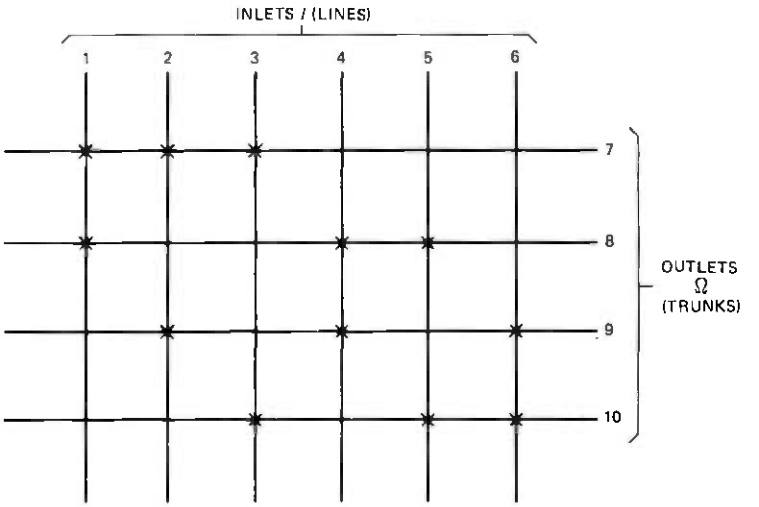


Fig. 6—Random slip concentrator.

isomorphic to D_4 and preserving $\sim$. It remains to check that $\equiv$ defined with the help of G_n is in fact the intuitive one that is indicated in Fig. 4: looking at the values of η in the two line table above we easily see that $5 \equiv 6$, $1 \equiv 2 \equiv 3 \equiv 4$, and no other distinct states are equivalent, so the equivalence classes are $\{0\}$, $\{1,2,3,4\}$, and $\{5,6\}$, as we had guessed. The group G_ν of *all* bijections which preserve these equivalence classes is isomorphic to $S_2 \times S_4$, a group of order 48; in contrast, the subgroup G_η is sufficient to characterize $\equiv$ and is only of order 8.

XIII. RANDOM SLIP CONCENTRATOR

The crosspoint assignment for a 6-line to 4-trunk random slip concentrator is shown in Fig. 6. The phrase "random slip" is old telephone terminology that clearly originated as a description of the even or regular way in which the incomplete access of lines to trunks is distributed over the network, something like the statisticians' balanced block designs. Figure 6 may also be depicted as the labelled tetrahedron of Fig. 7, with the interpretation that an edge (line) "has access" to the two vertices (trunks) that it connects. Figure 7 leads naturally to the network graph (Fig. 8) and the adjacency matrix (Table II): just add a vertex of degree 2 in the middle of each edge, with the same label as the edge. In this network every t is busy in some state, so Theorem 3 applies, $G_\eta \approx G_n$, and we need only calculate G_n. The reduced states are shown in a convenient representation as a partially ordered system in Fig. 9.

Thus we seek maps of $I = \{1,2,\ldots,6\}$ and $\Omega = \{7,8,9,10\}$ into themselves which preserve adjacency. It can be seen from Figs. 6 and 7 that every

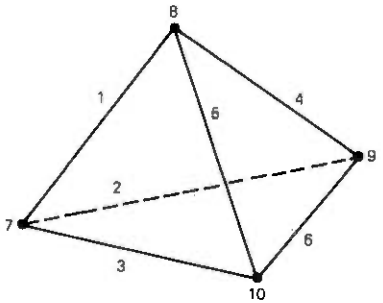


Fig. 7—Labeled tetrahedron.

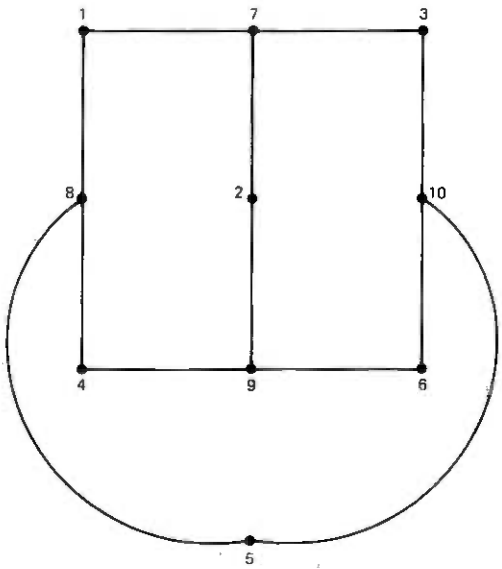


Fig. 8—Network graph for concentrator.

Table II — Adjacency matrix for concentrator

N	1	2	3	4	5	6	7	8	9	10
1							1	1	0	0
2							1	0	1	0
3			0				1	0	0	1
4							0	1	1	0
5							0	1	0	1
6							0	0	1	1
7	1	1	1	0	0	0				
8	1	0	0	1	1	0			0	
9	0	1	0	1	0	1				
10	0	0	1	0	1	1				

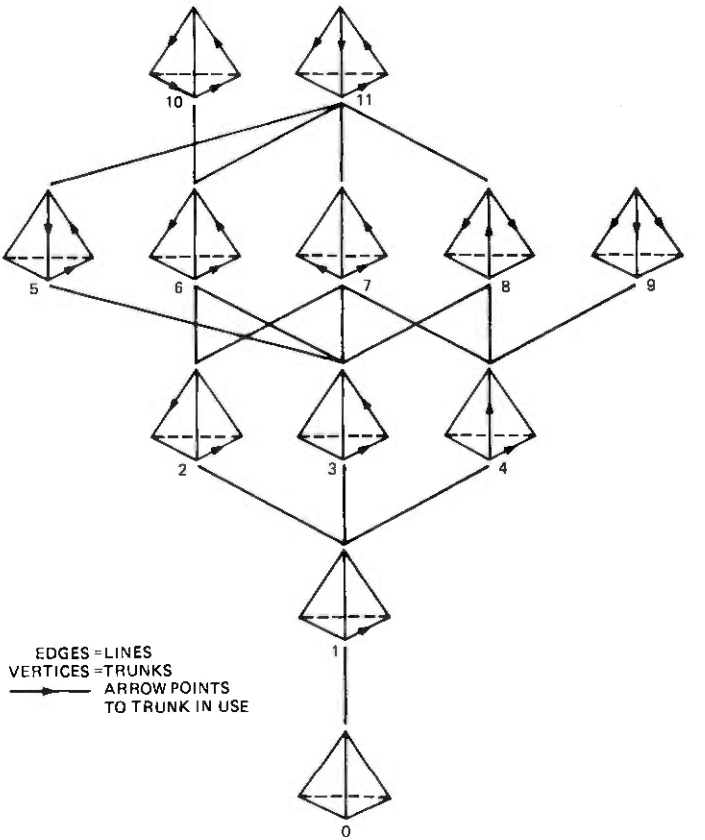


Fig. 9—Reduced states of random slip concentrator.

permutation of {7,8,9,10} *forces* a unique permutation of {1, . . . ,6} if adjacency is to be preserved; graphically, we see that all the maps we seek are rotations or reflections of the tetrahedron of Fig. 7 into itself, and that all permutations of {7,8,9,10} are allowed. More formally, every transposition of Ω is allowed and requires a unique pair of disjoint transpositions in I according to the table

(78) forces (24)(35)
(89) forces (12)(56)
(9 10) forces (23)(45)
(7 10) forces (15)(26)
(7 9) forces (14)(36)
(8 10) forces (13)(46)

Since the allowed maps form a group, and every permutation of Ω is a product of transpositions, every permutation of Ω forces a unique one

Table III — G_n for concentrator

id.	id.
(78)	(24)(35)
(89)	(12)(56)
(9 10)	(23)(45)
(7 10)	(15)(26)
(79)	(14)(36)
(8 10)	(13)(46)
(78)(9 10)	(25)(34)
(89)(7 10)	(16)(25)
(79)(8 10)	(16)(34)
(789)	(142)(356)
(798)	(124)(356)
(78 10)	(153)(246)
(7 10 8)	(135)(264)
(79 10)	(145)(263)
(7 10 9)	(154)(236)
(89 10)	(123)(465)
(8 10 9)	(132)(456)
(789 10)	(1463)(25)
(798 10)	(2453)(16)
(79 10 8)	(1265)(34)
(78 10 9)	(1562)(34)
(7 10 89)	(3542)(16)
(7 10 98)	(3641)(25)

of I, as claimed. It follows that G_n is isomorphic to S_4, and consists of the permutations shown in Table III. The "forced" permutations of I are of course a subgroup of S_6 isomorphic to S_4.

XIV. CLOS NETWORK EXAMPLE

The crosspoint structure and network graph for the simplest 3-stage Clos rearrangeable network are shown in Figs. 10–11. A planar form of the network graph is on Fig. 12. At the start of this paper we loosely described the invariances of structure to be studied as those associated with "renaming terminals, switches, and links." The little Clos network now under discussion gives us a specific example of what this means: it means, of course, permuting these entities while preserving the structure, e.g., interchanging switches in a stage while dragging along each switch's rat's

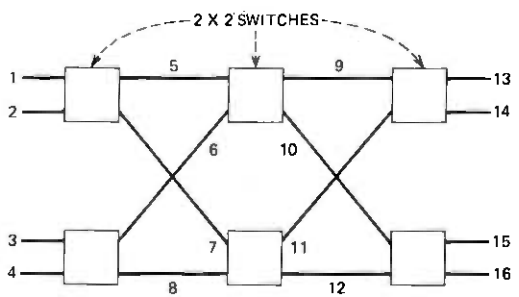


Fig. 10—Simple 3-stage Clos network.

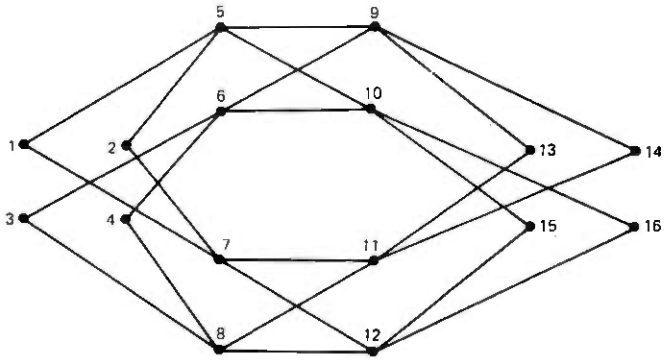


Fig. 11—Crosspoint structure and network graph for 3-stage Clos network made of 2 × 2 switches.

nest of links that connect to other stages. In particular, the organization of crosspoints into switches must also be preserved by these permutations of terminals, switches, and links. In the next section (XV) we shall describe how this constraint provides an approach to calculating the group G_n for general crossbar networks. For the present we show how the approach applies to the example under discussion.

Any 16-permutation that preserves the adjacency or crosspoint structure depicted in Fig. 11 must permute the inner terminals or links $\{5,6,\ldots,12\}$ among themselves only, and it can do so in exactly the 16 ways shown in Table IV. It is readily seen that (i) these 16 ways correspond to interchanging switches, and rotating or reflecting the network,

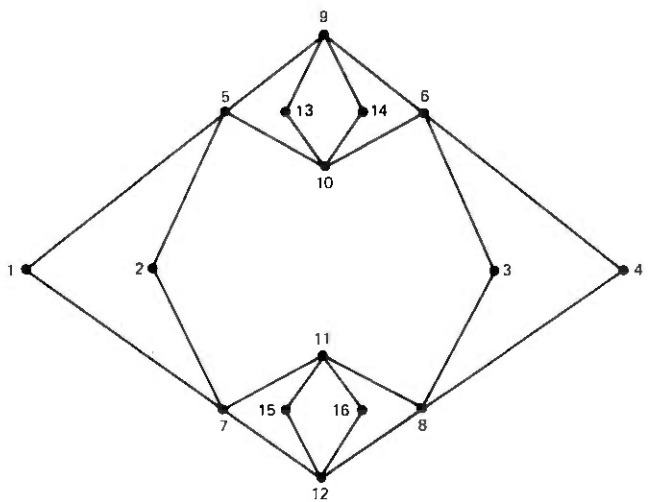


Fig. 12—Planar form of network graph of Fig. 11.

Table IV — Allowed permutations of inner terminals for 3-stage Clos network of 2 × 2 switches

No.	Map	Type	Action, if simple
1	id.	*ss*	move no switches
2	(57)(68)(9 11)(10 12)	*ss*	interchange m. sw.
3	(9 10)(11 12)	*sm*	interchange rt. sw.
4	(57)(68)(9 12)(10 11)	*sm*	interchange rt. and m. sw.
5	(56)(78)	*ms*	interchange l sw.
6	(58)(67)(9 11)(10 12)	*ms*	interchange l. and m. sw.
7	(56)(78)(9 10)(11 12)	*mm*	interchange rt. and l. sw.
8	(58)(67)(9 12)(10 11)	*mm*	rt., m., and l.
9	(59)(7 11)(6 10)(8 12)	*aa*	rotate about vertical axis
10	(5 11)(79)(6 12)(8 10)	*aa*	
11	(5 10 69)(7 12 8 11)	*ad*	
12	(5 12 6 11)(7 10 8 9)	*ad*	
13	(5 9 6 10)(7 11 8 12)	*da*	
14	(5 11 6 12)(7 9 8 10)	*da*	
15	(5 12)(6 11)(7 10)(89)	*dd*	
16	(5 10)(69)(7 12)(8 11)	*dd*	

and (*ii*) that they form a subgroup of S_{16}. Each of these ways can "go with" a number (here always 16) of permutations of outer terminals among *themselves*, to form an element of G_n. By writing down the possible ways this *matching* can be done (so as to preserve adjacency) we get a brute force way of calculating the group G_n for the 3-stage Clos network made of 2 × 2 switches.

The matching in question has a block structure: the permitted permutations of links can be partitioned in such a way that to each parititon element there corresponds a set of permitted permutations of outer terminals any one of which can "go with" each link permutation in the element. This block structure is indicated in Table IV by the type symbol; all allowed link permutations of the same type can match with all the same permutations of the outer terminals. Thus to present G_n

Table V — G_n for Clos network example

Type	*ss*	*sm*
	id.	(13 1)(14 16)
	(12)	(13 16)(14 15)
	(34)	(12)(13 15)(14 16)
	(13 14)	(12)(13 16)(14 15)
	(15 16)	(34)(13 15)(14 16)
	(12)(34)	(34)(13 16)(14 15)
	(12)(13 14)	(12)(34)(13 15)(14 16)
	(12)(15 16)	(12)(34)(13 16)(14 15)
	(34)(13 14)	(13 15 14 16)
	(34)(15 16)	(13 16 14 15)
	(13 14)(15 16)	(12)(13 15 14 16)
	(12)(34)(13 14)	(12)(13 16 14 15)
	(12)(34)(15 16)	(34)(13 15 14 16)
	(12)(13 14)(15 16)	(34)(13 16 14 15)
	(34)(13 14)(15 16)	(12)(34)(13 15 14 16)
	(12)(34)(13 14)(15 16)	(12)(34)(13 16 14 15)

Table V — (Continued)

Type	*ms*	*mm*
	(13)(24)	(1324)(13 15)(14 16)
	(14)(23)	(1324)(13 16)(14 15)
	(13)(24)(13 14)	(1423)(13 15)(14 16)
	(14)(23)(13 14)	(1423)(13 16)(14 15)
	(13)(24)(15 16)	(13)(24)(13 15)(14 16)
	(14)(23)(15 16)	(13)(24)(13 16)(14 15)
	(13)(24)(13 14)(15 16)	(14)(23)(13 15)(14 16)
	(14)(23)(13 14)(15 16)	(14)(23)(13 16)(14 15)
	(1324)	(1324)(13 15 14 16)
	(1324)(13 14)	(1423)(13 15 14 16)
	(1423)(13 14)	(1423)(13 16 14 15)
	(1324)(15 16)	(13)(24)(13 15 14 16)
	(1423)(15 16)	(13)(24)(13 16 14 15)
	(1324)(13 14)(15 16)	(14)(23)(13 15 14 16)
	(1423)(13 14)(15 16)	(14)(23)(13 16 14 15)

Type	*aa*	*ad*
	(1 13)(2 14)(3 15)(4 16)	(1 15 3 13)(2 16 4 14)
	(1 13)(2 14)(3 16)(4 15)	(1 16 3 13)(2 15 4 14)
	(1 14)(2 13)(3 15)(4 16)	(1 15 4 13)(2 16 3 14)
	(1 14)(2 13)(3 16)(4 15)	(1 16 4 13)(2 15 3 14)
	(1 13)(2 14)(3 15 4 16)	(1 15 3 14)(2 16 4 13)
	(1 13)(2 14)(3 16 4 15)	(1 16 3 14)(2 15 4 13)
	(1 14)(2 13)(3 15 4 16)	(1 15 4 14)(2 16 3 13)
	(1 14)(2 13)(3 16 4 15)	(1 16 4 14)(2 15 3 13)
	(1 13 2 14)(3 15)(4 16)	(1 15 3 13 2 16 4 14)
	(1 13 2 14)(3 16)(4 15)	(1 16 3 13 2 15 4 14)
	(1 14 2 13)(3 15)(4 16)	(1 15 4 13 2 16 3 14)
	(1 14 2 13)(3 16)(4 15)	(1 16 4 13 2 15 3 14)
	(1 13 2 14)(3 15 4 16)	(1 15 3 14 2 16 4 13)
	(1 13 2 14)(3 16 4 15)	(1 16 3 14 2 15 4 13)
	(1 14 2 13)(3 15 4 16)	(1 15 4 14 2 16 3 13)
	(1 14 2 13)(3 16 4 15)	(1 16 4 14 2 15 3 13)

Type		
	(1 13 4 15)(2 14 3 16)	(1 15)(2 16)(3 13)(4 14)
	(1 13 4 16)(2 14 3 15)	(1 15)(2 16)(3 14)(4 13)
	(1 13 3 15)(2 14 4 16)	(1 16)(2 15)(3 13)(4 14)
	(1 13 3 16)(2 14 4 15)	(1 16)(2 15)(3 14)(4 13)
	(1 14 4 15)(2 13 3 16)	(1 15)(2 16)(3 13 4 14)
	(1 14 4 16)(2 13 3 15)	(1 15)(2 16)(3 14 4 13)
	(1 14 3 15)(2 13 4 16)	(1 16)(2 15)(3 13 4 14)
	(1 14 3 16)(2 13 4 15)	(1 16)(2 15)(3 14 4 13)
	(1 13 3 15 2 14 4 16)	(1 15 2 16)(3 13)(4 14)
	(1 13 3 16 2 14 4 15)	(1 15 2 16)(3 14)(4 13)
	(1 13 4 15 2 14 3 16)	(1 16 2 15)(3 13)(4 14)
	(1 13 4 16 2 14 3 15)	(1 16 2 15)(3 14)(4 13)
	(1 14 3 15 2 13 4 16)	(1 15 2 16)(3 13 4 14)
	(1 14 3 16 2 13 4 15)	(1 15 2 16)(3 14 4 13)
	(1 14 4 15 2 13 3 16)	(1 16 2 15)(3 13 4 14)
	(1 14 4 16 2 13 3 15)	(1 16 2 15)(3 14 4 13)

it is enough to list the "outer" permutations that match each type of "inner"; this is done in Table V. The reduced states are in Fig. 13.

The multiplication table of the group of "inner" or switch permuta-

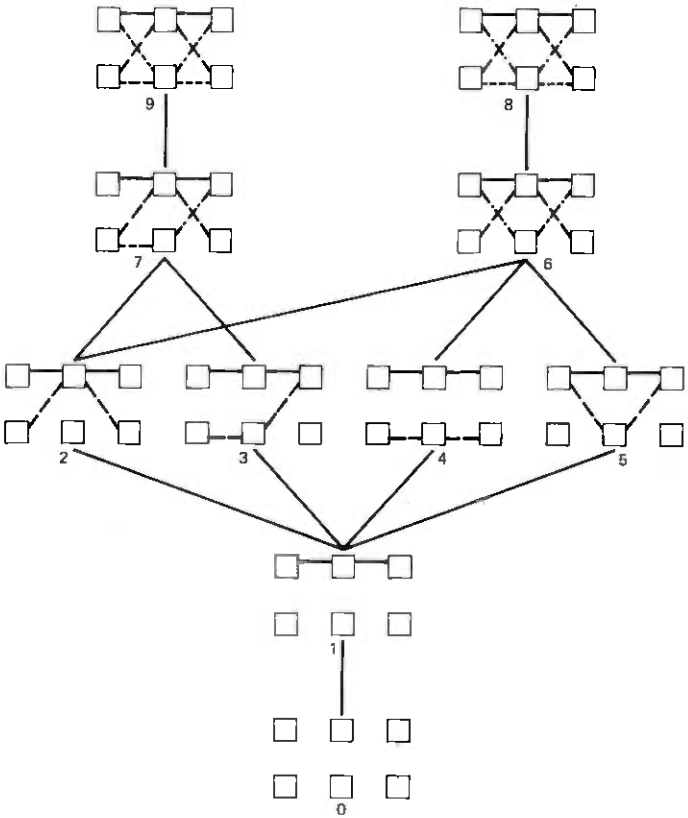


Fig. 13—Reduced states for 3-stage Clos network made of switches.

tions can be calculated from the cycle representations of Table IV; it is displayed in Table VI, from which it can be seen that elements $R_1 = 2$, $R_2 = 4$, and $R_3 = 15$ satisfy the relations

$$(R_1R_3)^2 = (R_2R_1)^4 = (R_2R_3)^4 = \text{identity}$$

which identify the group as (an isomorph of) $S_2 \times D_4$. This factorization is related to the fact that the type of a product is determined by the types of the factors in a way that is summarized in Table VII, in which the row is the type of the first factor, and the column that of the second. Indeed the type symbols themselves form an isomorph of D_4 under the "multiplication" Table VII. What is more, we can identify this group as the subgroup ($\cong D_4$) of the link or switch permutation group which restricts attention to the motion of the outer switches. Clearly the motion of the two middle switches is independent of that of the outer ones; this fact shows up in the feature that every type consists of just two permutations, and in the factor S_2 in the switch permutation group. The possible in-

Table VI — Multiplication table for switch permutation group

	id. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
id.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	2	1	4	3	6	5	8	7	10	9	12	11	14	13	16	15
	3	4	1	2	7	8	5	6	11	12	9	10	16	15	14	13
	4	3	2	1	8	7	6	5	12	11	10	9	15	16	13	14
	5	6	7	8	1	2	3	4	13	14	16	15	9	10	12	11
	6	5	8	7	2	1	4	3	14	13	15	16	10	9	11	12
	7	8	5	6	3	4	1	2	16	15	13	14	11	12	10	9
	8	7	6	5	4	3	2	1	15	16	14	13	12	11	9	10
	9	10	13	14	11	12	16	15	1	2	5	6	3	4	8	7
	10	9	14	13	12	11	15	16	2	1	6	5	4	3	7	8
	11	12	16	15	9	10	13	14	3	4	7	8	1	2	6	5
	12	11	15	16	10	9	14	13	4	3	8	7	2	1	5	6
	13	14	9	10	16	15	11	12	5	6	1	2	7	8	4	3
	14	13	10	9	15	16	12	11	6	5	2	1	8	7	3	4
	15	16	12	11	14	13	10	9	8	7	4	3	6	5	1	2
	16	15	11	12	13	14	9	10	7	8	3	4	5	6	2	1
Type	*ss*	*ss*	*sm*	*sm*	*ms*	*ms*	*mm*	*mm*	*aa*	*aa*	*ad*	*ad*	*da*	*da*	*dd*	*dd*

terchanges of outer switches are summarized in Fig. 14 and they correspond to the type symbols as follows

switch interchange	1	r	l	h	v	x	a	b
type symbol	1	*sm*	*ms*	*aa*	*mm*	*dd*	*da*	*ad*

The switch interchange symbols of course also form a group $\cong D_4$ under the multiplication transferred from the type symbols; indeed $\{1,h,v,x\}$, $\{1,r,l,v\}$, and $\{1,a,b,v\}$ each forms a sub-"Vierergruppe":

XV. CALCULATION OF G_n FOR CROSSBAR NETWORKS

In connecting networks built out of stages of rectangular crossbar switches, a permutation of the terminals cannot preserve adjacency in the network graph unless it maps the terminals of each switch onto those of some switch, either itself or another one, so as to either preserve or

Table VII — "Multiplication" table for type symbols. Top row: generators identifying with D_4. Left column: identification with outer switch interchanges.

		A^3B *aa*	AB *dd*	A *da*	A^3 *ad*	A^2 *mm*	B *ms*	A^2B *sm*	1 *ss*
h	*aa*	*ss*	*mm*	*sm*	*ms*	*dd*	*ad*	*da*	*aa*
x	*dd*	*mm*	*ss*	*ms*	*sm*	*aa*	*da*	*ad*	*dd*
a	*da*	*ms*	*sm*	*mm*	*ss*	*ad*	*dd*	*aa*	*da*
b	*ad*	*sm*	*ms*	*ss*	*mm*	*da*	*aa*	*dd*	*ad*
v	*mm*	*dd*	*aa*	*ad*	*da*	*ss*	*sm*	*ms*	*mm*
l	*ms*	*da*	*ad*	*aa*	*dd*	*sm*	*ss*	*mm*	*ms*
r	*sm*	*ad*	*da*	*dd*	*aa*	*sm*	*mm*	*ss*	*sm*
1	*ss*	*aa*	*dd*	*da*	*ad*	*mm*	*ms*	*sm*	*ss*

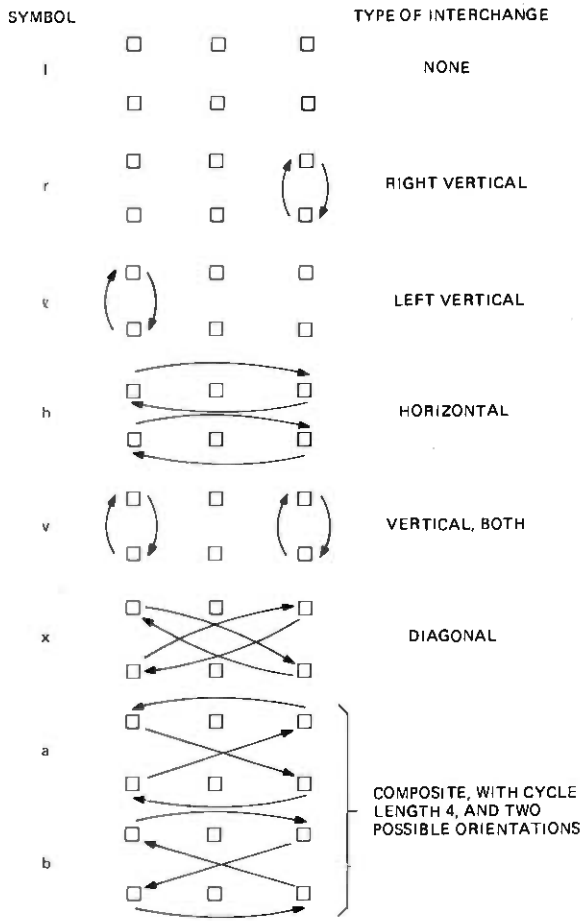


Fig. 14—Possible interchanges of outer switches.

interchange inlets and outlets. It follows that any such permutation determines a permutation of the switches, and that the set G_s of such induced "switch permutations" is again a group, which we call the switch permutation group. The task of calculating G_n for a crossbar network is substantially simplified by first finding G_s.

In the language of group theory, the symmetry group G_n for a crossbar network must be an *imprimitive* group, because it consists of permutations that either preserve switches or map them onto each other. Indeed it can be seen that the map ϕ which assigns to each $g \epsilon G_n$ the switch permutation that g induces is a homomorphism of G_n onto G_s. Once a permitted switch permutation $p \epsilon G_s$ is chosen, each element in $\phi^{-1}\{p\}$ is determined by choosing, for each permuted outer switch, a map which

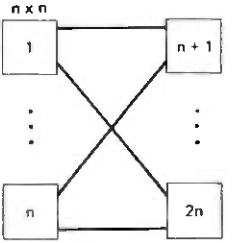


Fig. 15—Frame with "canonical" complete bipartite cross-connect field.

assigns the outer terminals of that switch to the outer terminals of its image under p. These maps are independent, so we have proved

Theorem 4: If ν is a crossbar network, then

$$G_n \approx G_s \times \prod_{0} S_{|0|}$$

where the product is over outer switches o, $|o|$ is the number of inlets (outlets) on o and S_k is the symmetric group on k objects.

For many crossbar networks made of stages it is rather straightforward to calculate the switch permutation group G_s, because G_s turns out to be isomorphic to a semidirect product of groups that are determined by the way the stages are joined together.

Example: *Frame*. It is clear that for a frame made of two identical stages, interconnected by the "canonical" (complete bipartite) cross-connect field as shown in Fig. 15, G_s will consist of all maps that take the inlet switches onto the outlet switches and vice versa, together with all maps that permute the inlet switches among themselves, and also the outlet switches among themselves. If each stage has n switches, then G_s is isomorphic to the largest imprimitive permutation group on $2n$ objects with two equinumerous sets of imprimitivity. If $1, \ldots, n$ are the inlet and $n+1, \ldots, 2n$ are the outlet switches respectively, then

$$G_s = F_n \cup \sigma F_n$$

where F_n is the isomorph of $(S_n)^2$ in S_{2n} which permutes $1, \ldots, n$ among themselves and also $n+1, \ldots, 2n$ among themselves and $\sigma = (1\ n+1)(2\ n+2)\cdots(n\ 2n)$. Each switch is $n \times n$, so

$$G_n \approx (F_n \cup \sigma F_n) \times (S_n)^{2n}$$

It can be verified that the union $F_n \cup \sigma F_n$ is in fact a group and a semidirect product. The order of G_s for the frame is $2(n!)^2$. For $n = 2$, $G_s \approx D_4$, the dihedral group of order 8.

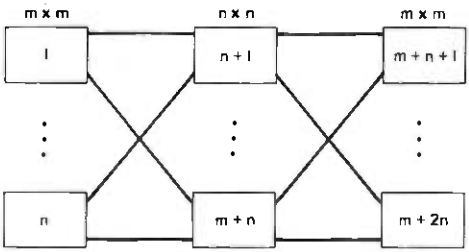


Fig. 16—Three-stage Clos network.

Example: *3-Stage Clos Network.* From Fig. 16 it is evident that G_s is the imprimitive group which permutes the middle switches, *and* either permutes the inlet switches and the outlet switches independently, or else maps inlet switches onto outlet and vice versa. If now K_n is the isomorph of $(S_n)^2$ in S_{m+2n} which permutes $1, \ldots, n$ and $m + n + 1, \ldots, m + 2n$ independently, and L_m is the isomorph of S_m in S_{m+2n} which permutes only $n + 1, \ldots, m + n$, then for the Clos network of Fig. 16,

$$G_s = (K_n \cup \tau K_n) \times L_m$$

where $\tau = (1\ 2n + 1) \cdots (n\ 3n)$, and $G_n \approx G_s \times (S_m)^{2n}$. For the case $m = n = 2$, treated in detail in Section XIV, we have $G_s \approx S_2 \times D_4$ and $G_n \approx (S_2)^5 \times D_4$. It is now easy to see that the $S_2 \times D_4$ structure of G_s arises from viewing the outer stages as a frame that yields D_4 as in the previous example, while the S_2 arises from permuting the middle switches.

Example: *Cascade of s stages with "complete bipartite" cross-connect.* The structure of this network is shown in Fig. 17. The parity of s plays a role here, as follows: If s is odd, then under a switch permutation the switches of the middle stage can only go into each other, and as we saw in the previous example, they contribute to G_s a group factor isomorphic

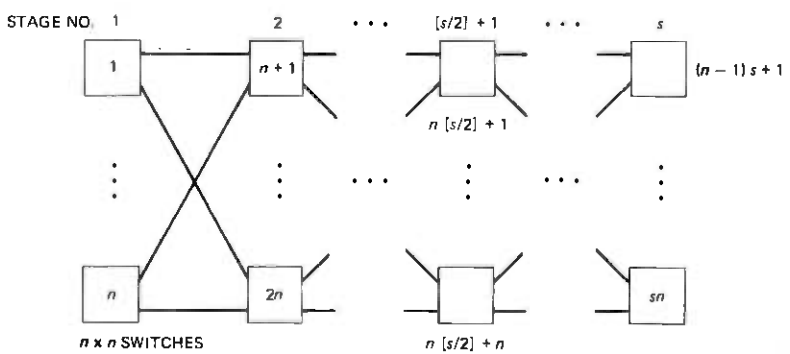


Fig. 17—Cascade of s stages with "complete bipartite" cross connect.

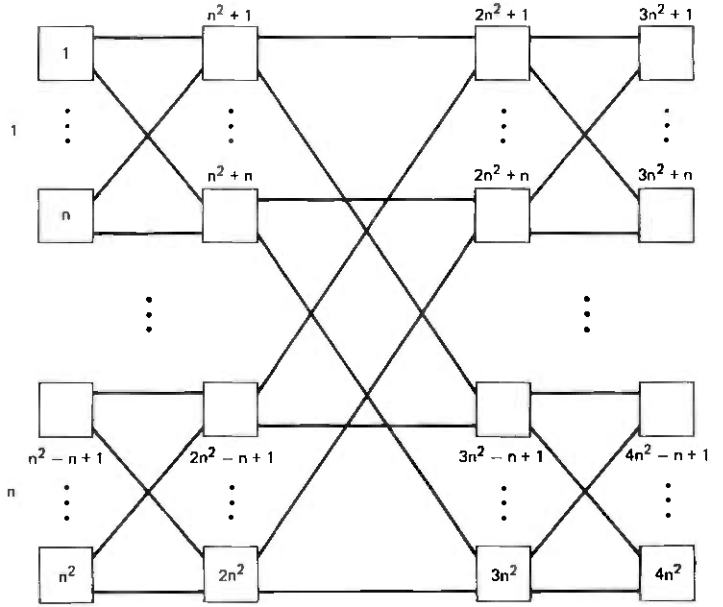


Fig. 18—No. 5 crossbar type network.

to S_n. Whatever the parity of s the noncentral switches contribute a factor isomorphic to an imprimitive group (*not* the largest) with $2[s/2]$ sets of imprimitivity; these are partitioned into pairs and the group elements either map each set onto itself, or each set onto its paired set, and never some of each. If the switches are numbered as in Fig. 17, this group is essentially the semidirect product

$$M_n \cup \theta M_n$$

where M_n is the largest subgroup of S_{sn} which fixes all the central switches and permutes vertically within every other stage, and

$$\theta = \prod_{k=0}^{[s/2]} (kn + 1\ (s - 1 - k)n + 1)\cdots(kn + n\ sn - kn)$$

Finally

$$G_s \approx (M_n \cup \theta M_n)(S_n)^{s-2[s/2]}$$

Example: *No. 5 Crossbar.* With the switches numbered as in Fig. 18, let P_n be the isomorph of $(S_n)^3$ in S_{4n^2} which for $k = 0,1,2,3$, permutes $kn^2 + 1, \ldots, (k + 1)n^2$ among themselves in such a way as to map frames onto frames, outer stages onto outer of the image frame, and inner stages onto inner stages of the image frame, and inner stages onto inner stages

of the image frame. Let

$$\sigma = (1\ 3n^2 + 1)(2\ 3n^2 + 2)\cdots(n^2\ 4n^2) \times (n^2 + 1\ 2n^2 + 1)(n^2 + 2\ 2n^2 + 2)\cdots(2n^2\ 3n^2)$$

Then

$$G_\sigma = P_n \cup \sigma P_n$$

$$G_n \cong (P_n \cup \sigma P_n) \times (S_n)^{2n^2}$$

The remainder of this paper derives the reduced state equations for the stochastic model of Ref. 2.

XVI. ROUTING OF CALLS

We shall use a routing matrix $R = (r_{xy})$ as a convenient formal description of how routes are chosen for calls. The class of routing matrices R can be described thus: for each $x \epsilon S$ let Π_x be the partition of A_x induced by the relation $\sim$ of "having the same calls up," or satisfying the same assignment of inlets to outlets; it can be seen that Π_x consists of exactly the sets A_{cx} for c free and not blocked in x; for $Y \epsilon \Pi_x$, r_{xy} for $y \epsilon Y$ is to be a probability distribution over Y, that is $r_{xy} \geq 0$ and $\sum_{y \epsilon Y} r_{xy} = 1$; r_{xy} is to be 0 in all other cases.

The interpretation of the routing matrix as a method of choice is to be this: any $Y \epsilon \Pi_x$ represents all the ways in which a particular call c (free and not blocked in x) *could* be completed when the network is in state x; for $y \epsilon Y$, r_{xy} is the chance (or fraction of times) that if call c arises in state x it will be completed by being routed in the network so as to take the system to state y. The distribution $\{r_{xy}, y \epsilon Y\}$ indicates how the calling-rate due to c is to be spread over the possible ways of putting up this call. Evidently, such a description of routing could be made time-dependent, and extended to cover refusal of unblocked calls as an option; we do not consider these possibilities here. The problem of choosing an optimal routing matrix R has been worked on at some length, in Refs. 9 and 10; its relations to state reduction are described in forthcoming sequels[5,6] to the present paper.

XVII. STOCHASTIC MODEL

We now recall[2] a stochastic model for the traffic offered to a network. A Markov stochastic process x_t taking values on S can be based on these simple probabilistic and operational assumptions:

(*i*) Holding times of calls are mutually independent variates, each with the negative exponential distribution of unit mean.

(*ii*) If u is an inlet idle in state $x \epsilon S$, and $v \neq u$ is any outlet, there is a conditional probability $\lambda h + o(h)$, $\lambda > 0$, as $h \to 0$, that u attempt a call to v in $(t, t + h)$ if $x_t = x$.

(*iii*) A routing matrix $R = (r_{xy})$ is used to choose routes, as follows: If $c = \{(u,v)\}$ is a call free and not blocked in x, then the fraction of times that the system pass from x to $y \epsilon A_{cx}$ if c arises when $x_t = x$ is just r_{xy}.

(*iv*) Blocked calls are declined, with no change of state.

It is convenient to collect these assumptions into a transition rate matrix $Q = (q_{xy})$, the generator of x_t; this matrix is given by

$$c_{xy} = \begin{cases} 1 \quad \text{if } y \epsilon B_x \\ \lambda r_{xy} \text{ if } y \epsilon A_x \\ -|x| - \lambda s(x) \text{ if } y = x, \text{ with } s(x) = |F_x| \\ 0 \quad \text{otherwise} \end{cases}$$

and the associated statistical equilibrium (or state) equations take the simple form

$$[|x| + \lambda s(x)]p_x = \sum_{y \epsilon A_x} p_y + \lambda \sum_{y \epsilon B_x} p_y r_{yx} \quad x \epsilon S$$

where $\{p_x, x \epsilon S\}$ is the asymptotic distribution of x_t.

The remainder of this paper takes up the problem of correctly writing some *reduced* state equations for the stochastic model we have described.

XVIII. MULTIPLICITY

In an (unreduced) state x belonging to a structural equivalence class α there may be several calls in progress whose termination would carry the system into an equivalence class β. For example, in Fig. 13, for a state x in $\alpha = 7$ there are two calls whose ending would yield a state from 3, and there is one call leading to 2. Since we have assumed a unit hangup rate per call in progress, the downward transition rates between adjacent unreduced states is always unity; but as soon as we lump states into equivalence classes $\alpha, \beta, \ldots$ these rates in effect add up. A similar situation obtains for new calls: there may be several calls free and not blocked in a state $x \epsilon \alpha$ whose completion could (under a suitable choice of route) lead to a state in equivalence class β.

We call this phenomenon *multiplicity*, and we need an account of it in order to define transition rates between reduced states, and to correctly use[5] and understand the reduced state equations.

The "hangup" matrix $H = (h_{xy})$ associated with a network ν is a matrix of zeros and ones that tells which states can be reached from which by a hangup:

$$h_{xy} = \chi_{y \epsilon B_x} = \chi_{x \epsilon A_y}$$

It can be seen that H admits the automorphism subgroup G_η, i.e.,

that

$$h_{xy} = h_{(gx)(gy)} \qquad g \epsilon G_\eta$$

H is so-called because if we make the standard traffic assumption that all calls in progress terminate independently of any past history at unit rate, then H is precisely the part of the transition rate matrix due to "hangups."

The transition rate, due to hangups, from a state x to an equivalence class β is

$$\sum_{y \epsilon \beta} h_{xy}$$

We claim that this number is the same for all x in an equivalence class α, and that it is the hangup rate from α to β. There are two ways of proving this result, which will be put in the following form:

Lemma 1: The numbers $|B_x \cap \beta|$ and

$$\sum_{y \epsilon \beta} h_{xy}$$

are equal; they assume the same value for all x in an equivalence class α, and represent the hangup rate $h_{\alpha\beta}$ from α to β.

Proof: Equality is obvious. The numbers are zero unless α covers β in the partial ordering on E. Let $x \epsilon \alpha$, $z \epsilon \alpha$, $z = gx$ where g is an automorphism of $(S, \leq)$. If now $y \epsilon B_x \cap \beta$, then $gy \epsilon B_{gx} = B_z$, and $gy \epsilon \beta$ because β is an equivalence class. Conversely if $y \epsilon B_z \cap \beta$, then $g^{-1} y \epsilon B_x \cap \beta$, by the same argument, so $B_z \cap \beta = g(B_x \cap \beta)$. The result follows because g is a bijection.

Alternative proof: With $x \epsilon \alpha$, $z \epsilon \alpha$, $z = gx$ as above, consider that where w is any element of β

$$|B_{gx} \cap \beta| = \sum_{y \epsilon \beta} \chi_{y \epsilon B_{gx}}$$

$$= \frac{|\beta|}{|G_\eta|} \sum_{f \epsilon G_\eta} \chi_{gfw \epsilon B_{gx}} = \frac{|\beta|}{|G_\eta|} \sum_{f \epsilon G_\eta} h_{(gx)(gfw)}$$

$$= \frac{|\beta|}{|G_\eta|} \sum_{f \epsilon G_\eta} h_{x(fw)} = \sum_{y \epsilon \beta} h_{xy} = |B_x \cap \beta|$$

The multiplicity question is more complicated for new calls than it is for hangups; this is because only one state can arise by removing a call c in progress, whereas possibly one of several could arise by completing a call c which is new and is not blocked; as a result, the calculation of transition rates between reduced states depends on routing decisions as well as on multiplicity. For by itself multiplicity will only provide the

calling rates which might take the state into a reduced state β: whether a particular transition occurs depends, however, on what state x the system is in and what routing policy is being followed. We shall see, though, that if the routing is *consistent* in the sense that it routes analogous calls analogously in equivalent states, then these calling rates will depend only on πx and β, i.e., on the reduced states in question. Mathematically this idea of consistency is expressed by the condition that the routing matrix $R = (r_{xy})$ *admit* the group G_η: if c is a call free and not blocked in x, and $\tau \epsilon G_\eta$, then in the terminology above, c and τc are analogous calls, x and τx are equivalent states, and $\tau(A_{cx}) = A_{(\tau c)(\tau x)}$, so consistency amounts to asking that we have $r_{xy} = r_{(\tau x)(\tau y)}$ for $y \epsilon A_{cx}$, i.e., that R admit G_η.

The matrix $N = (n_{xy})$ associated with a network ν is a matrix of zeros and ones that tells which states can be reached from which by adding a new call; it is obviously the transpose of the hangup matrix H:

$$n_{xy} = h_{yx} = \chi_{x \epsilon B_y} = \chi_{y \epsilon A_x}$$

Thus N admits the automorphism subgroup G_η, and we have this analog of Lemma 1, with a similar proof.

Lemma 2: The numbers $|A_x \cap \beta|$ and

$$\sum_{y \epsilon \beta} n_{xy} (= n_{\alpha\beta})$$

are equal; they assume the same value $n_{\alpha\beta}$ for all x in an equivalence class α, and they represent the number $n_{\alpha\beta}$ of calls free and not blocked in x which could be put up so as to lead to a state of β.

To calculate the actual transition rate from a state x into an equivalence class β, we must use the routing matrix R. Notice that R is just N with enough of the ones reduced (but still ≥ 0) so that r_{xy} summed over $y \epsilon A_{cx}$ is unity for each x and $c \epsilon F_x$.

For the traffic model assumed in Section XVII run according the routing matrix R the transition rate from a state x into an equivalence class β that intersects A_x is

$$\sum_{\substack{c \text{ free in } x \\ c \text{ not blocked}}} \sum_{y \epsilon \beta \cap A_{cx}} r_{xy} \left(= \sum_{y \epsilon \beta \cap A_x} r_{xy} \right)$$

Notice that replacement of each r_{xy} by 1 in this expression increases its value to precisely

$$n_{\alpha\beta} = |A_x \cap \beta| = |\{c : \beta \epsilon \pi(A_{cx})\}|$$

Lemma 3: If R admits G_η, then the numbers

$$\sum_{y \epsilon \beta \cap A_x} r_{xy} \ (= r_{\alpha\beta})$$

are the same for $x \epsilon \alpha$.

Proof: Like that of Lemmas 1 and 2.

We put

$A_\alpha = \{\beta \epsilon E$: β covers α in the partial order of E induced by $\leq\}$

$B_\alpha = \{\beta \epsilon E$: α covers β in the partial order of E induced by $\leq\}$

These are the reduced analogs of the sets A_x and B_x, useful for writing the state equations. It can be seen that

$$A_\alpha = \bigcup_{x \epsilon \alpha} \pi(A_x)$$

$$B_\alpha = \bigcup_{x \epsilon \alpha} \pi(B_x)$$

Let us set

$$q_{\alpha\beta} = \begin{cases} h_{\alpha\beta} & \beta \epsilon B_\alpha \\ \lambda r_{\alpha\beta} & \beta \epsilon A_\alpha \\ -[\lambda s(x) + |x|] & x \epsilon \alpha, \quad \alpha = \beta \\ 0 & \text{otherwise} \end{cases}$$

We can now informally describe the reduced equations as follows: when R admits G_η, they correspond to equilibrium equations for a Markov process πx_t on E whose transition rates up and down are $\lambda r_{\alpha\beta}$ and $h_{\alpha\beta}$ respectively, with the obvious necessary convention on the "diagonal." In terms of the notation introduced above, these reduced equations are

$$-p_\alpha q_{\alpha\alpha} = \sum_{\beta \epsilon A_\alpha} p_\beta h_{\beta\alpha} + \lambda \sum_{\beta \epsilon B_\alpha} p_\beta r_{\beta\alpha} \tag{2}$$

with $q_{\alpha\alpha} = -[\lambda s(x) + |x|]$ for any $x \epsilon \alpha$.

XIX. REDUCTION OF STATE EQUATIONS

We can now give precise explicit conditions under which the original "microscopic" equations of state can be exactly replaced by the less numerous "macroscopic" or reduced equations for the probabilities of equivalence classes. It will be shown that if the routing matrix admits G_η then the original and the reduced equations imply each other, and that the state probabilities $\{p_x, x \epsilon S\}$ are constant over equivalence classes α. Only the equilibrium case is considered.

Theorem 5: Let the routing matrix R admit G_η. Then

(*i*) The "microscopic" transition rate matrix

$$Q = H + \lambda R - \text{diag}[\lambda s(x) + |x|]_{x \epsilon S}$$

also admits G_η.

(*ii*) The projection map $\pi: x \to \pi x$ defines a "macroscopic" Markov process πx_t on the $\equiv -$ equivalence classes with transition rate matrix

$$q_{\alpha\beta} = \begin{cases} h_{\alpha\beta} + \lambda r_{\alpha\beta} & \alpha \neq \beta \\ q_{xx} & \alpha = \beta, \quad x \epsilon \alpha \end{cases}$$

(*iii*) If $p_\alpha = \sum_{x \epsilon \alpha} p_x$, then the equilibrium equations $\sum_x p_x q_{xy}$ and $\sum_\alpha p_\alpha q_{\alpha\beta}$ imply each other.

(*iv*) If $\{p_\alpha, \alpha \epsilon \pi S\}$ solves the reduced equations, then $\{p_x, x \epsilon S\}$ defined by

$$p_x = \frac{p_\alpha}{|\alpha|}, \quad x \epsilon \alpha$$

solves the original equilibrium equations.

Proof: (*i*) is clear from the hypothesis that R admits G_π, from Lemmas 1 and 3, and from the fact that $s(x) = s(\tau x)$, $|x| = |\tau x|$ for $\tau \epsilon G_\pi$. To prove (*ii*) it is enough, by (Lloyd's) Theorem 1, to prove that

$$\sum_{y \epsilon \beta} q_{xy}$$

are constant for $x \epsilon \alpha$. For $\alpha = \beta$ this follows from invariance of $s(\cdot)$ and $|\cdot|$ under $\tau \epsilon G_\pi$. If β covers α in the induced partial order on πS, then

$$\sum_{y \epsilon \beta} q_{xy} = \lambda \sum_{y \epsilon \beta} r_{xy} = \lambda r_{\alpha\beta}$$

Here we have used Lemma 3. Similarly if α covers β, then for $x \epsilon \alpha$

$$\sum_{y \epsilon \beta} q_{xy} = \sum_{y \epsilon \beta} h_{xy} = |B_x \cap \beta| = h_{\alpha\beta}$$

by Lemma 1; thus (*ii*) is proved.

To prove the reduced equations from the original ones, we sum the latter over $x \epsilon \alpha$, to get

$$\sum_{x \epsilon \alpha} p_x [\lambda s(x) + |x|] = \sum_{x \epsilon \alpha} \sum_{y \epsilon A_x} p_y + \lambda \sum_{x \epsilon \alpha} \sum_{y \epsilon B_x} p_y r_{yx}$$

Since $s(\cdot)$ and $|\cdot|$ are constant over α, the left-hand side is just $-p_\alpha q_{\alpha\alpha}$. For the first term on the right, argue thus: since $h_{yx} = 0$ unless $y \epsilon A_x$ and

$$\sum_{\beta \epsilon A_\alpha} \sum_{y \epsilon \beta}$$

sums over (at least) all $y \epsilon A_x$ if $x \epsilon \alpha$, we have

$$\begin{aligned} \sum_{x \epsilon \alpha} \sum_{y \epsilon A_x} p_y &= \sum_{\beta \epsilon A_\alpha} \sum_{y \epsilon \beta} p_y \sum_{x \epsilon \alpha} h_{yx} \\ &= \sum_{\beta \epsilon A_\alpha} \sum_{y \epsilon \beta} p_y |B_y \cap \alpha| \\ &= \sum_{\beta \epsilon A_\alpha} p_\beta h_{\beta\alpha} \end{aligned}$$

using Lemma 1. Similarly, since $r_{yx} = 0$ unless $h_{xy} = 1$, we find

$$\sum_{x\epsilon\alpha}\sum_{y\epsilon B_x} p_y r_{yx} = \sum_{x\epsilon\alpha}\sum_{\beta\epsilon B_\alpha}\sum_{y\epsilon\beta} p_y r_{yx} h_{xy}$$

$$= \sum_{\beta\epsilon B_\alpha}\sum_{y\epsilon\beta} p_y \sum_{x\epsilon\alpha} r_{xy}$$

$$= \sum_{\beta\epsilon B_\alpha} p_\beta r_{\beta\alpha}$$

This proves the reduced equations from the original; the original ones can be proved from the reduced by reversing the identities used above. Neither argument depends on the fact (*iv*), to be proved next, that p_x is constant over $x\epsilon\alpha$.

To prove (*iv*) we shall set $p_x = p_{\pi x}/|\pi x|$ in the unreduced equations and obtain an identity. Substitution in eq. (2) and multiplication by $|\alpha|$ gives, for $x\epsilon\alpha$,

$$p_\alpha[\lambda s(x) + |x|] = |\alpha| \sum_{y\epsilon A_x} \frac{p_{\pi y}}{|\pi y|} + \lambda|\alpha| \sum_{y\epsilon B_x} \frac{p_{\pi y}}{|\pi y|} r_{yx}$$

$$= |\alpha| \sum_{\beta\epsilon A_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} h_{yx}$$

$$+ \lambda|\alpha| \sum_{\beta\epsilon B_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} r_{yx}$$

$$= \sum_{\beta\epsilon A_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} \frac{|\alpha|}{|G_\eta|} \sum_{g\epsilon G_\eta} h_{(gy)x}$$

$$+ \lambda \sum_{\beta\epsilon B_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} \frac{|\alpha|}{|G_\eta|} \sum_{g\epsilon G_\eta} r_{x(gy)}$$

$$= \sum_{\beta\epsilon A_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} \frac{|\alpha|}{|G_\eta|} \sum_{g\epsilon G_\eta} h_{y(gx)}$$

$$+ \lambda \sum_{\beta\epsilon B_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta} \frac{|\alpha|}{|G_\eta|} \sum_{g\epsilon G_\eta} r_{(gx)y}$$

$$= \sum_{\beta\epsilon A_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta}\sum_{x\epsilon\alpha} h_{yx}$$

$$+ \lambda \sum_{\beta\epsilon B_\alpha} p_\beta \frac{1}{|\beta|} \sum_{y\epsilon\beta}\sum_{x\epsilon\alpha} r_{yx}$$

$$= \sum_{\beta\epsilon A_\alpha} p_\beta h_{\beta\alpha} + \lambda \sum_{\beta\epsilon B_\alpha} p_\beta r_{\beta\alpha}$$

Since the left side is $-p_\alpha q_{\alpha\alpha}$, the substitution has resulted in the reduced

equations. Since the solution p of the original equations is unique, (iv) is established.

XX. ACKNOWLEDGMENTS

The author is indebted to S. P. Lloyd for discussion and access to personal notes, to R. L. Graham for helping to formulate ideas and produce counterexamples, and to N. J. A. Sloane for helping to identify some of the groups that arose.

REFERENCES

1. S. P. Lloyd, "Markov Chains Admitting Symmetry Operations," unpublished Bell Laboratories memorandum, September 19, 1955.
2. V. E. Beneš, "Markov Processes Representing Traffic in Connecting Networks," B.S.T.J., *42*, No. 6 (November 1963), pp. 2795–2837.
3. A. F. Bulfer, "Blocking and Routing in Two-Stage Concentrators," Conference Record, National Telecommunications Conference, New Orleans, Dec. 1–3, 1975./4w 4. A. F. Bulfer, U.S. Patent No. 3,935,394, January 27, 1976.
5. V. E. Beneš, "Optimal Routing in Networks Whose States are Reduced Under Symmetries," to appear.
6. V. E. Beneš, "Optimal Routing in Some Two-Stage Concentrators," to appear.
7. C. J. Burke and M. Rosenblatt, "A Markovian Function of a Markov Chain," Ann. Math. Stat., *29* (1958), pp. 1112–1122.
8. S. P. Lloyd, private communication.
9. V. E. Beneš, "Programming and Control Problems Arising from Optimal Routing in Telephone Networks," B.S.T.J., *45*, No. 9 (November 1966), pp. 1373–1438.
10. V. E. Beneš, Optimal routing in connecting networks over finite time intervals, B.S.T.J., *46*, No. 10 (December 1967), pp. 2341–2352.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Elastic State of Stress in a Stalpeth Cable Jacket Subjected to Pure Bending

By G. S. BROCKWAY and G. M. YANIZESKI

(Manuscript received June 21, 1977)

The stresses in the plastic jacket of a slightly bent telephone cable are analyzed within the linear theory of elasticity. The jacket is considered to be bonded to the underlying corrugated steel by a flooding compound. The constraining effect of the steel results in a three-dimensional state of stress that differs substantially from the predictions of elementary beam theory. For the thin jackets typically used on telephone cables it is found that the stress state is essentially biaxial, the axial and circumferential normal stresses being at least an order of magnitude larger than the others. On the tensile side, the stresses are closely approximated (at any given point) by those in the well-known biaxial strip experiment, in which the principal stresses are in proportion by the Poisson's ratio of the plastic. The compressive side is likewise in biaxial compression, and there the flooding compound is subjected to tensile stresses even before the onset of any jacket buckling. The results confirm the validity of previous approaches to the effects introduced by imperfections and indicate further that the probability of spontaneous cracking is increased by the adherence of the jacket to the soldered steel layer.

I. INTRODUCTION

The selection and development of plastic jacketing compounds for multipair cables depend to a large extent on cable behavior during bending. Cable jackets are expected to be relatively flexible for ease of handling and installation while at the same time surviving large strains (up to 15 percent) without cracking, splitting, or severe wrinkling. Temperature extremes encountered in the field render these criteria even more stringent.

Recently completed analyses have led to easily performed laboratory tests for the screening of candidate compounds with regard to some of

these requirements. Now, for example, the relative influence of various sheath-grade plastics on the bending stiffness of cables can be evaluated by conducting ordinary tensile tests.[1] In addition, the relative sensitivity of compounds to low temperature and high strain-rate cracking can be determined through impact tests on notched specimens.[1] Still, there remains the observation of slow crack growth at high temperatures during bending and the occurrence of wrinkles in cable jackets during duct installation at low temperatures. None of these phenomena nor how they are affected by cable parameters such as jacket moduli and thickness, flooding compound tackiness, and the depth of the corrugations in the underlying steel, is presently understood.

This paper is devoted to a study of the state of stress in the jacket of a Stalpeth cable subjected to classical pure bending. The problem is treated within the framework of the linear theory of elasticity, which supposes small strains and rotations and an elastic material. Although the bending strains in telephone cable are frequently large, our analysis is intended to provide insight into the circumstances at incipient cracking, buckling, or yielding of the cable jacket. Standard techniques from linear viscoelasticity theory[2,3] can be applied to the elastic results given here to account for the time dependence inherent to plastics.

The primary emphasis in the present investigation is directed toward estimating the effect on the stress field in the jacket resulting from the constraint imposed on its inner surface by the underlying cable structure (see Fig. 1 for a detail of the Stalpeth construction). For slight bends, the influence of the soldered steel shield dominates that of the wire core, and we shall, therefore, consider the pure bending of a plastic jacket bonded to a corrugated metal shell. The analysis is further simplified by the realization that the corrugation wave length used in Stalpeth cable is one to two times larger than the jacket thickness, while the valleys imprinted on the jacket's inner surface are reasonably shallow because of the presence of the flooding compound. We are thus afforded the privilege of averaging field variables over a corrugation wave length to arrive at a boundary value problem involving a uniform cylindrical geometry. The amplification of jacket stresses created by the corrugation imprints can be deduced from the results obtained here together with published concentration factors in the usual way.

We begin in the next section by formulating the relevant three-dimensional field equations and show that they can be reduced through a change in dependent variables to plane strain equations. The approach taken is reminiscent of the scheme used in elementary elasticity for the St. Venant-bending of cylinders with irregularly shaped cross sections, but differs in that here the strain field, rather than the stress field, is supposed to conform to the elementary Bernoulli-Euler theory.

Next, the boundary conditions at the jacket-steel interface are con-

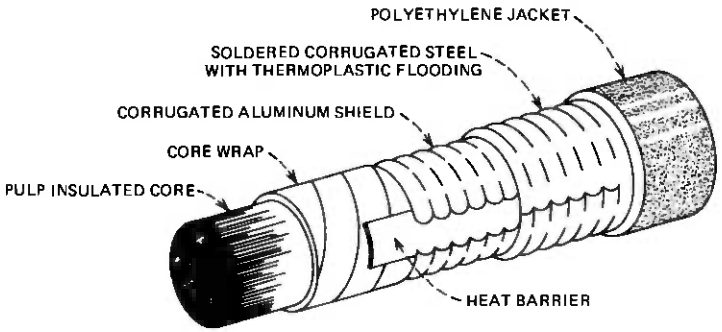


Fig. 1—Stalpeth cable.

sidered in detail. The steel is assumed to have negligible stiffness in the corrugation direction and to otherwise obey the usual hypotheses on the deformation of thin shells. Integration over a corrugation wave length then permits the stresses and displacements of the jacket to be related to those of the steel shield.

The appropriate (plane strain) boundary value problem for the stress state in the sheath having been set up, we find that it admits an elementary solution in closed form through the introduction of an Airy stress function. The full three-dimensional stress and displacement fields are then calculated.

When these results are applied to Stalpeth cables as presently manufactured, it is found that, on the tensile side of the cable, the jacket is essentially in a state of biaxial tension. That is, the ratios of the axial and circumferential normal stresses to the shear and transverse (thickness-direction) stresses are of the order of the diameter-to-thickness ratio. Furthermore, if E denotes the Young's modulus of the plastic jacket, D the cable diameter and ρ the bend radius, then the longitudinal and hoop stresses are shown to be in approximate proportion to the bending stresses $ED/2\rho$ from elementary beam theory. Of particular interest is the conclusion that these constants of proportionality vary appreciably only with the Poisson's ratio of the jacket, at least for the ranges of cable size, corrugation geometry and jacket moduli encountered in Stalpeth applications. With Poisson's ratio chosen in the typical range for plastics (0.3–0.5), the longitudinal stress varies from about 110 to 130 percent of the beam theory stress, while the hoop stress factor ranges from 30 to 45 percent. The stress state on the tensile side of the cable is thus closely approximated by the biaxial strip experiment discussed at the end of the paper. The importance of biaxiality in bending has been recognized previously.[4]

All of the results outlined above for the tensile side of the cable apply

as well on the compressive side, except that the stress state is, of course, one of biaxial compression. It is also observed that the flooding compound is subjected to a tension on the compressive side that exceeds that which currently used compounds are likely to support, even for very small bend radii.

The implications of these results on the cracking to which we alluded earlier are discussed qualitatively in the final section of the paper. For the time being, we remark only that the biaxiality tends to increase the likelihood of spontaneous cracking. Finally, it should be mentioned that the jacket stresses generated in the bending of cables other than Stalpeth can differ drastically from those obtained here. Indeed, the biaxial state of stress produced by the constraining effect of the soldered steel layer would not be present in those designs that allow relative motion between the jacket and underlying metallic layers.

II. REDUCTION OF THE PROBLEM TO ONE OF PLANE STRAIN

Consider a cylindrical shell of inner radius r_i and outer radius r_o. Referring to Fig. 2, choose cylindrical coordinates (r,θ,z) in the obvious way, and let the shell be subjected to pure bending in the y-z plane. The center line of the bent shell is then a circle of prescribed radius ρ.

Recall the field equations of linear elasticity in cylindrical coordinates:[5]

Strain-displacement relations

$$\epsilon_r = \frac{\partial u_r}{\partial r}, \quad \epsilon_\theta = \frac{1}{r}\left(\frac{\partial u_\theta}{\partial \theta} + u_r\right), \quad \epsilon_z = \frac{\partial u_z}{\partial z} \tag{1}$$

$$2\epsilon_{r\theta} = \frac{1}{r}\left(\frac{\partial u_r}{\partial \theta} - u_\theta\right) + \frac{\partial u_\theta}{\partial r}, \quad 2\epsilon_{rz} = \frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}, \quad 2\epsilon_{\theta z} = \frac{\partial u_\theta}{\partial z} + \frac{1}{r}\frac{\partial u_z}{\partial \theta}$$

Equations of equilibrium

$$\frac{\partial \sigma_r}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{r\theta}}{\partial \theta} + \frac{\sigma_r - \sigma_\theta}{r} = 0 \tag{2a}$$

$$\frac{\partial \sigma_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_\theta}{\partial \theta} + \frac{2\sigma_{r\theta}}{r} = 0 \tag{2b}$$

$$\frac{\partial \sigma_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_{\theta z}}{\partial \theta} + \frac{\partial \sigma_z}{\partial z} = 0 \tag{2c}$$

Hooke's law

$$2\mu(1+\nu)\epsilon_r = \sigma_r - \nu(\sigma_\theta + \sigma_z) \tag{3a}$$

$$2\mu(1+\nu)\epsilon_\theta = \sigma_\theta - \nu(\sigma_r + \sigma_z) \tag{3b}$$

$$2\mu(1+\nu)\epsilon_z = \sigma_z - \nu(\sigma_r + \sigma_\theta) \tag{3c}$$

$$2\mu\epsilon_{r\theta} = \sigma_{r\theta}, \quad 2\mu\epsilon_{rz} = \sigma_{rz}, \quad 2\mu\epsilon_{\theta z} = \sigma_{\theta z} \tag{3d}$$

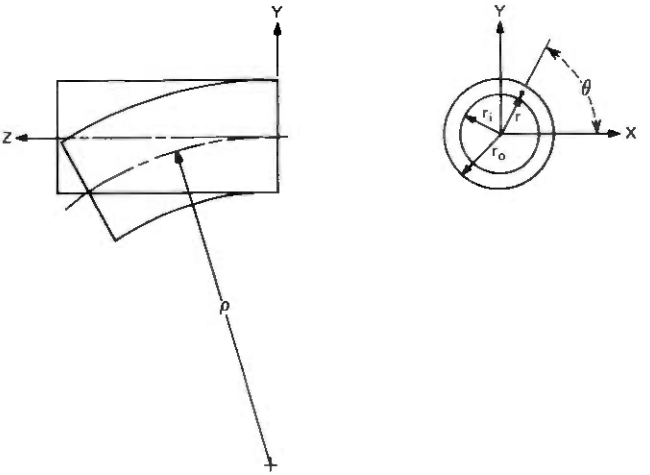


Fig. 2—Choice of cylindrical and Cartesian coordinates.

Here, the symbols u, ϵ, and σ, suitably subscripted, stand for components of displacement, strain, and stress, respectively. The constants μ and ν represent the shear modulus and Poisson's ratio for the cable jacket, while the coefficient

$$E = 2\mu(1 + \nu) \tag{4}$$

of ϵ_r, ϵ_θ, and ϵ_z in the first three equations of (3) is Young's modulus.

Next, make the assumption that the annular cross sections remain plane and normal to the shell's axis during bending.* It then follows from a familiar geometric argument[7] that

$$\epsilon_{rz} = \epsilon_{\theta z} = 0, \ \epsilon_z = \frac{r \sin \theta}{\rho} \tag{5}$$

With no consequent loss of generality, we assume further that the cross section at $z = 0$ remains in the x-y plane and undergoes no rotation about the z-axis nor overall rigid translation in the x-y plane, so that

$$u_z(r,\theta,0) = 0 \tag{6a}$$

$$u_\theta(r_o,0,0) = u_\theta\left(r_o, \frac{\pi}{2}, 0\right) = u_\theta(r_o,\pi,0) = 0 \tag{6b}$$

We show next that the formulae (5) and (6a) require the in-plane stresses σ_r, σ_θ, and $\sigma_{r\theta}$ to be associated with a state of plane strain. That is, we establish the existence of plane displacements $\hat{u}_r$ and $\hat{u}_\theta$, independent of z, that generate strains $\hat{\epsilon}_r$, $\hat{\epsilon}_\theta$, and $\hat{\epsilon}_{r\theta}$ obeying the plane strain

* The viewpoint here is similar to that of the traditional St. Venant "semi-inverse" method.[6] The solution thus obtained is justified by exhibiting stress distributions on the ends of the cylinder that support the assumed deformation.

form of Hooke's law:[5]

$$2\mu\hat{\epsilon}_r = (1-\nu)\sigma_r - \nu\sigma_\theta \tag{7a}$$

$$2\mu\hat{\epsilon}_\theta = (1-\nu)\sigma_\theta - \nu\sigma_r \tag{7b}$$

$$2\mu\hat{\epsilon}_{r\theta} = \sigma_{r\theta} \tag{7c}$$

We first confirm (7) and then prove the existence of the requisite displacements.

Since ϵ_z is known from (5), we have from (3c) that

$$\sigma_z = 2\mu(1+\nu)\epsilon_z + \nu(\sigma_r + \sigma_\theta) \tag{8}$$

Thus, σ_z may be eliminated from eqs. (3a), (3b) for ϵ_r, ϵ_θ to obtain

$$\epsilon_r = \hat{\epsilon}_r - \nu\epsilon_z, \ \epsilon_\theta = \hat{\epsilon}_\theta - \nu\epsilon_z \tag{9}$$

where $\hat{\epsilon}_r, \hat{\epsilon}_\theta$ are given by (7a), (7b). The last of (7) is, of course, satisfied by taking

$$\hat{\epsilon}_{r\theta} \equiv \epsilon_{r\theta} \tag{10}$$

To see that the strain field $\hat{\epsilon}_r$, $\hat{\epsilon}_\theta$, and $\hat{\epsilon}_{r\theta}$ is indeed generated from (7) by in-plane displacements $\hat{u}_r$ and $\hat{u}_\theta$, independent of z, observe that the displacement field*

$$\bar{u}_r = -\frac{(\nu r^2 + z^2)\sin\theta}{2\rho}, \ \bar{u}_\theta = \frac{(\nu r^2 - z^2)\cos\theta}{2\rho}, \ \bar{u}_z = \frac{rz\sin\theta}{\rho} \tag{11}$$

has an associated strain field

$$\bar{\epsilon}_r = \bar{\epsilon}_\theta = -\nu\epsilon_z, \ \bar{\epsilon}_z = \epsilon_z, \ \bar{\epsilon}_{r\theta} = \bar{\epsilon}_{rz} = \bar{\epsilon}_{\theta z} = 0 \tag{12}$$

Thus, the defining equations

$$u_r = \bar{u}_r + \hat{u}_r, \ u_\theta = \bar{u}_\theta + \hat{u}_\theta, \ u_z = \bar{u}_z + \hat{u}_z \tag{13}$$

for $\hat{u}_r$, $\hat{u}_\theta$, and $\hat{u}_z$, together with (1), (9), and (12), reveal that the displacements $\hat{u}_r, \hat{u}_\theta$ satisfy the strain-displacement relations for $\hat{\epsilon}_r, \hat{\epsilon}_\theta, \hat{\epsilon}_{r\theta}$. Moreover, when (13) is combined with (7c), (5), and (12), there results

$$\frac{\partial\hat{u}_r}{\partial z} + \frac{\partial\hat{u}_z}{\partial r} = 0, \ \frac{\partial\hat{u}_\theta}{\partial z} + \frac{1}{r}\frac{\partial\hat{u}_z}{\partial\theta} = 0, \ \frac{\partial\hat{u}_z}{\partial z} = 0 \tag{14}$$

Integration of (14) subject to the constraint (6a) yields

$$\hat{u}_z = \frac{\partial\hat{u}_r}{\partial z} = \frac{\partial\hat{u}_\theta}{\partial z} = 0$$

and the desired result is established.

* This displacement field (11) is that which the shell would exhibit were it hollow (see Sokolnikoff,[6] Article 32, for example).

The three-dimensional problem is now reduced to determining the plane-strain elastic state with displacements $(\hat{u}_r, \hat{u}_\theta)$, strains $(\hat{\epsilon}_r, \hat{\epsilon}_\theta, \hat{\epsilon}_{r\theta})$, and stresses $(\sigma_r, \sigma_\theta, \sigma_{r\theta})$, subject to boundary conditions on the cylindrical surfaces $r = r_i$ and $r = r_o$. The remaining stresses $\sigma_z, \sigma_{rz}, \sigma_{\theta z}$ are then provided by (8) and

$$\sigma_{rz} = \sigma_{\theta z} = 0 \tag{15}$$

which follows from (5) and (3c).

III. THE BOUNDARY CONDITIONS

As previously mentioned, the intent of this paper is to account for the influence of the soldered steel shield on the plastic jacket. Since the longitudinal (z direction in Fig. 3) stiffness of the corrugated steel is small for small longitudinal extensions,[8] the constraint imposed on the inner surface of a Stalpeth jacket is essentially confined to the x-y plane, *provided the jacket field variables are interpreted as averages over a corrugation wave-length.*

Moreover, since the corrugation depth H (see Fig. 3) is small compared to the radius r_i of the cable, the shield will deform in approximate accordance with Kirchoff's hypothesis of classical shell theory.[9] In particular, denoting by u and v the circumferential and radial displacements of the midsurface of the shield (see Fig. 3), one has the relations[9]

$$u_r^s(\xi,\theta) = v \tag{16a}$$

$$u_\theta^s(\xi,\theta) = u + \frac{2\xi}{2r_i - H}\left(u - \frac{dv}{d\theta}\right) \tag{16b}$$

for the radial and circumferential displacements of a particle at a distance ξ from the midsurface (again, see Fig. 3). Requiring the displacements of the steel to conform to those of the jacket at $\xi = H/2$ results in

$$\begin{aligned} v(\theta) &= u_r(r_i,\theta) \\ \left(\frac{2r_i}{2r_i - H}\right) u(\theta) &= u_\theta(r_i,\theta) + \frac{H}{2r_i - H}\frac{\partial u_r}{\partial\theta}(r_i,\theta) \end{aligned} \tag{17}$$

Next, take the shield to have an (in-plane) modulus E_s and use (16), (17) to compute the tension T and moment M in the shield averaged over a corrugation wave length ℓ (refer again to Fig. 3). There results

$$\begin{aligned} T(\theta) &= \frac{E_s A}{\ell}\epsilon_\theta(r_i,\theta) - \frac{AH}{I}M(\theta) \\ M(\theta) &= \frac{E_s I}{r_i^2 \ell}\left[u_r(r_i,\theta) + \frac{\partial^2 u_r}{\partial\theta^2}(r_i,\theta)\right] \end{aligned} \tag{18}$$

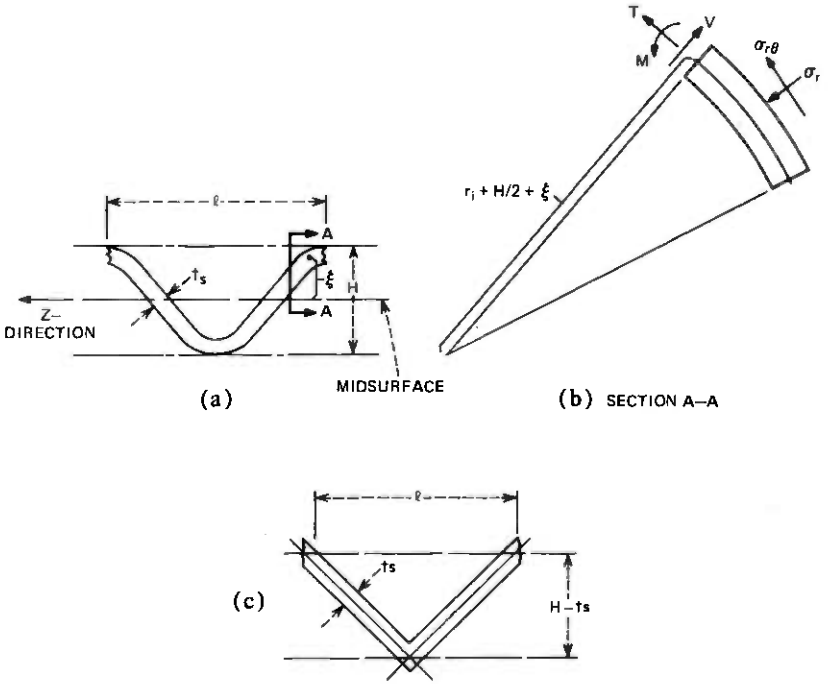


Fig. 3—Corrugation details.

where A and I are respectively the area and area moment of inertia about the midsurface of a corrugation wave length (see Fig. 3a), and ϵ_θ is the circumferential strain in the jacket.

Finally, recall that for the shield to be in equilibrium under the interfacial loads $\sigma_r, \sigma_{r\theta}$ exerted by the jacket, T and M must satisfy[10]

$$\begin{aligned} \left(r_i - \frac{H}{2}\right)\sigma_r &= T - \frac{1}{r_i - H/2}\frac{d^2M}{d\theta^2} \qquad \text{at } r = r_i \\ \left(r_i - \frac{H}{2}\right)\sigma_{r\theta} &= -\frac{dT}{d\theta} - \frac{1}{r_i - H/2}\frac{dM}{d\theta} \qquad \text{at } r = r_i \end{aligned} \tag{19}$$

The sought-after boundary conditions at the inner surface of the jacket can now be extracted from (18) and (19) with the aid of (5), (9), (11), and (13). Neglecting terms in $H/2r_i$ compared to unity, one has

$$-r_i\sigma_r + \frac{E_sA}{\ell}\hat{\epsilon}_\theta - \left(\frac{AH}{I}M + \frac{1}{r_i}\frac{d^2M}{d\theta^2}\right) = \frac{E_sA}{\ell}\frac{\nu r_i}{\rho}\sin\theta \tag{20a}$$

$$r_i\sigma_{r\theta} + \frac{E_sA}{\ell}\frac{\partial\hat{\epsilon}_\theta}{\partial\theta} - \left(\frac{AH}{I} - \frac{1}{r_i}\right)\frac{dM}{d\theta} = \frac{E_sA}{\ell}\frac{\nu r_i}{\rho}\cos\theta \tag{20b}$$

where

$$M = -\frac{E_s I}{r_i{}^2 \ell}\left(\hat{u}_r + \frac{\partial^2 \hat{u}_r}{\partial \theta^2}\right) \tag{20c}$$

The remaining boundary conditions

$$\sigma_r = \sigma_{r\theta} = 0 \qquad \text{at } r = r_o \tag{21}$$

insure the absence of load on the exterior of the cable.

IV. SOLUTION OF THE PROBLEM

According to the plane theory of elasticity, there exists an Airy stress function ϕ generating the stresses σ_r, σ_θ, and $\sigma_{r\theta}$ through the relations[5]

$$\begin{aligned} \sigma_r &= \frac{1}{r}\frac{\partial \phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \phi}{\partial \theta^2} \\ \sigma_\theta &= \frac{\partial^2 \phi}{\partial r^2} \\ \sigma_{r\theta} &= -\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial \phi}{\partial \theta}\right) \end{aligned} \tag{22}$$

In addition, ϕ must satisfy the biharmonic equation

$$\nabla^4 \phi = 0 \tag{23}$$

Michell[11] has given a general form for Airy's function for annular domains subject to the requirement that the stresses and displacements be single-valued.* Examining Michell's solution (see also Fung,[5] p. 246) and the boundary conditions (2), it seems worth trying ϕ in the form

$$\phi = (ar^3 + b/r)\sin\theta \tag{24}$$

To determine the constants a and b in (24), substitute first into (22) and impose the boundary conditions (21) to conclude that

$$b = ar_o{}^4$$

and hence

$$\begin{aligned} &\sigma_r = -2af(r)\sin\theta,\ \sigma_\theta = 2ag(r)\sin\theta,\ \sigma_{r\theta} = 2af(r)\cos\theta \\ &f(r) = r(r_o{}^4/r^4 - 1),\ g(r) = r(r_o{}^4/r^4 + 3) \end{aligned} \tag{25}$$

Before enforcing (20), we compute the displacements and strains gen-

* Although Michell's argument for the generality of his solution is sketchy, his result is nevertheless correct. This can be established directly with the aid of Muskhelishvili's[12] complex valued stress functions.

erated by the stress state in (25). Toward this end, substitute (25) into the plane strain form of Hooke's law to arrive at

$$\hat{\epsilon}_r = -\frac{a}{\mu}\hat{\mathcal{E}}_r(r)\sin\theta, \hat{\epsilon}_\theta = \frac{a}{\mu}\hat{\mathcal{E}}_\theta(r)\sin\theta, \hat{\epsilon}_{r\theta} = -\frac{a}{\mu}f(r)\cos\theta$$
$$\hat{\mathcal{E}}_r = r\left(\frac{{r_o}^4}{r^4} + 4\nu - 1\right), \hat{\mathcal{E}}_\theta = r\left(\frac{{r_o}^4}{r^4} + 3 - 4\nu\right) \tag{26}$$

If eqs. (26) are now incorporated into the strain displacement relations (1) and an elementary integration is performed, it is found that

$$\hat{u}_r = \frac{a}{2\mu}\left[\frac{{r_o}^4}{r^2} - (4\nu - 1)r^2 + \hat{A}\right]\sin\theta + \hat{B}\cos\theta$$
$$\hat{u}_\theta = -\frac{a}{2\mu}\left[\frac{{r_o}^4}{r^2} + (5 - 4\nu)r^2 + \hat{A}\right]\cos\theta - \hat{B}\sin\theta + \hat{C}r \tag{27}$$

where $\hat{A}$, $\hat{B}$, and $\hat{C}$ are constants fixed by the condition (6b). In fact, (27), (13), (11), and (6b) give

$$\hat{B} = \hat{C} = 0, \hat{A} = \left(\frac{\mu\nu}{\rho a} + 2(2\nu - 3)\right){r_o}^2 \tag{28}$$

We are now in a position to determine the unknown constant a. Observe first that (27) and (20) require that the moment M in the steel vanish identically. Thus, the second of (20) follows from the first provided

$$\frac{\partial\sigma_{r\theta}}{\partial\theta} = \sigma_r \qquad \text{at } r = r_i$$

a condition that is met for all r by the stress field in (25). Equations (25), (26), and the first of (20), therefore, reveal that the boundary relations (20) are all satisfied provided

$$a = \frac{\mu\nu r_i/\rho}{\hat{\mathcal{E}}_\theta(r_i) + Sf(r_i)}, S = \frac{2\mu\ell r_i}{E_s A} \tag{29}$$

The result in (29) may be rewritten in the form

$$a = \frac{\mu\nu}{\lambda\rho},$$
$$\lambda = 3 - 4\nu + \frac{{r_o}^4}{{r_i}^4} + S\left[\frac{{r_o}^4}{{r_i}^4} - 1\right] \tag{30}$$

with the aid of (25) and (26). The stress field is now completely determined since by (8) and (25),

$$\sigma_z = \left(1 + \nu + \frac{4a\rho\nu}{\mu}\right)\frac{2\mu r}{\rho}\sin\theta$$

or, using (30) and (4),

$$\sigma_z = \left[1 + \frac{4\nu^2}{\lambda(1+\nu)}\right]\frac{Er}{\rho}\sin\theta \tag{31}$$

By the same token, the displacements are given by (11), (13), (27), (28), and (30) in the form

$$\begin{aligned} u_r &= \frac{ar_o^2}{2\mu}\left[\frac{r_o^2}{r^2} - (\lambda + 4(\nu-1))\frac{r^2}{r_o^2} - \frac{\lambda}{\nu}\frac{z^2}{r_o^2} + \lambda + 2(2\nu-3)\right]\sin\theta \\ u_\theta &= -\frac{ar_o^2}{2\mu}\left[\frac{r_o^2}{r^2} + (5 - 4\nu - \lambda)\frac{r^2}{r_o^2}\right. \\ &\qquad \left. + \frac{\lambda}{\nu}\frac{z^2}{r_o^2} + \lambda + 2(2\nu-3)\right]\cos\theta \end{aligned} \tag{32}$$

The solution to the original three-dimensional problem has thus been found. Although there is no available uniqueness theorem encompassing the present problem (the boundary conditions (20) are nonstandard), it can be shown from Michell's[11] representation for the Airy function that the solution obtained here is unique except for certain peculiar choices of the elastic and geometric parameters of the jacket and shield. The physical significance of these instances of non-uniqueness is related to the important sheath-buckling phenomenon that has been observed in telephone cables. Since the present analysis is insufficient to adequately describe the crucial z-dependence of the ripples, we shall explore this issue no further in this paper.

V. DISCUSSION OF THE STRESSES IN THE JACKET

Observe first [eq. (31)] that the axial stress σ_z varies linearly with the distance $y = r\sin\theta$ from the neutral plane ($y = 0$), just as in the elementary Bernoulli-Euler theory. In addition, (25) reveals that σ_r and σ_θ are likewise antisymmetric about the neutral plane, σ_r being compressive and σ_θ tensile when σ_z is tensile and vice versa. The shear stress $\sigma_{r\theta}$, as expected, is symmetric about the neutral plane.

The variation in σ_r (and hence $\sigma_{r\theta}$) and σ_θ across the thickness of the jacket is illustrated in Fig. 4, where the dimensionless functions

$$\text{FBAR} = -f/4r_o,\ \text{GBAR} = g/4r_o \tag{33a}$$

[Eq. (25)] are plotted against the dimensionless radius

$$\text{RBAR} \equiv \bar{r} \equiv r/r_o \tag{33b}$$

for $0.5 \le \bar{r} \le 1$. It is seen that as long as the jacket thickness is less than 10 percent of the cable radius ($0.9 < \bar{r} < 1$), the transverse stress σ_r is less than 10 percent of the hoop stress σ_θ, which is itself essentially constant across the jacket.

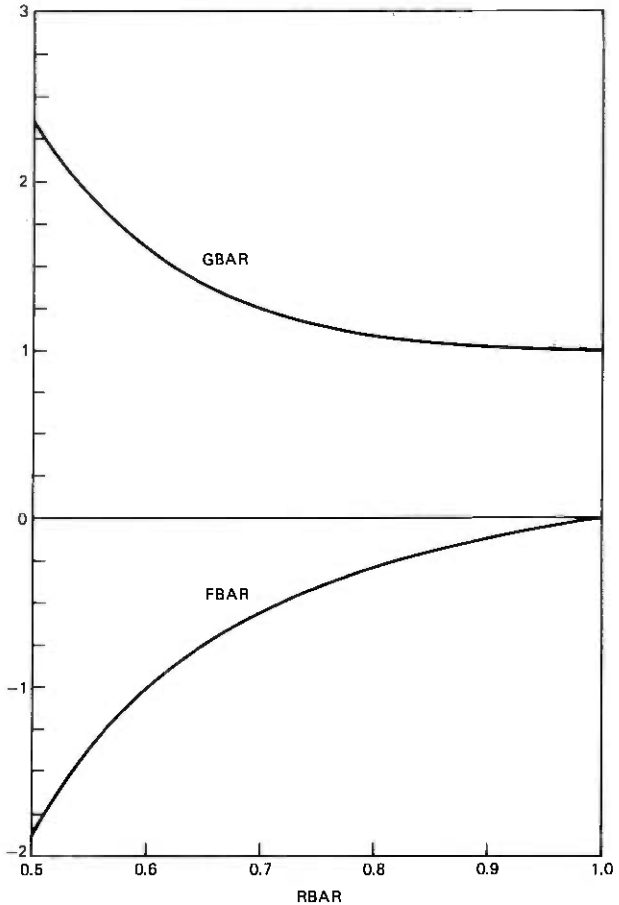


Fig. 4—Radial dependence of σ_r (FBAR) and σ_θ (GBAR).

With a view toward examining the influence of the properties of the jacket on the stress state, let $\sigma_z{}^m$ and $\sigma_\theta{}^m$ stand for the maximum values of the axial and hoop stresses, respectively. Then (25) and (31) together with (30) and (4) give

$$\begin{aligned} \sigma_\theta{}^m &\equiv \sigma_\theta(r_i, \pi/2, z) = \frac{E\nu}{(1+\nu)\lambda\rho} g(r_i) \\ \sigma_z{}^m &\equiv \sigma_z(r_o, \pi/2, z) = \left[1 + \frac{4\nu^2}{(1+\nu)\lambda}\right] \frac{Er_o}{\rho} \end{aligned} \tag{34}$$

The ratio of the maximum bending stress in the cable to the prediction of elementary beam theory is thus

$$\text{SIGMAX} \equiv \frac{\sigma_z^m}{Er_o/\rho} = 1 + \frac{4\nu^2}{(1+\nu)\lambda} \tag{35}$$

Further, the degree of biaxiality created in the jacket by its constrained inner surface is described by the parameter

$$\text{ALPHA} \equiv \frac{\sigma_\theta{}^m}{\sigma_z{}^m} = \frac{\nu g(r_i)/r_o}{(1+\nu)\lambda + 4\nu^2} \tag{36}$$

In Fig. 5, the dimensionless stresses SIGMAX and ALPHA are plotted versus Poisson's ratio ν for a 3-inch diameter Stalpeth cable conforming to the current Bell System design. In computing the parameter S that enters the formula (30) for λ, the area A of a half-corrugation wave length has been approximated by

$$A = \sqrt{m^2+1}\,\ell t_S,\ m = \frac{H - t_S}{\ell} \tag{37}$$

which is the area of the parallelogram shown in Fig. 3c. Also, the modulus E_S was assumed to be that of steel and the jacket modulus E was supposed to be 45,000 psi, representative of low-density polyethylene at room temperature.[13]

The curves shown are quite insensitive to variations in jacket modulus, SIGMAX changing by less than ½ percent and ALPHA by less than 2 percent when E is increased to 90,000 psi. The results indicate that for a Poisson's ratio $\nu = 0.35$, the bending stress in the jacket is about 12 percent higher than the prediction of the elementary theory while the hoop stress is roughly ⅓ of the axial stress.

The formulae (35) and (36) can also be used to investigate the effect of cable size on the state of stress in the jacket. Once the cable diameter is chosen, all other geometric parameters (e.g., jacket thickness, corrugation height) are fixed. This information was incorporated into (35) and (36) resulting in equations relating SIGMAX and ALPHA to cable diameter and the elastic constants of the jacket. For the range of Stalpeth cable currently manufactured, SIGMAX and ALPHA are independent of cable diameter.

The conclusions already reached as well as others of interest may be inferred more easily from the approximations of the previous equations arising from assuming the jacket to be thin. Toward this end, let

$$t = r_o - r_i \tag{38}$$

so that, for any n

$$\left(\frac{r_i}{r_o}\right)^n = \left(1 - \frac{t}{r_o}\right)^n = 1 - n\frac{t}{r_o} + 0\left[\left(\frac{t}{r_o}\right)^2\right] \quad \text{as } t/r_o \to 0 \tag{39}$$

If the approximation in (39) is applied to the formula (30) for λ, one gets

$$\lambda = 4(1-\nu) + 4t/r_o + 4St/r_o + 0[(t/r_o)^2]$$

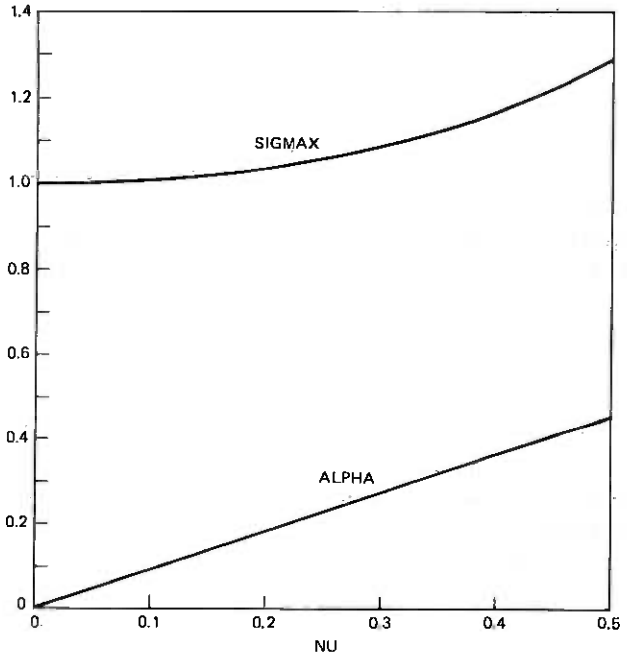


Fig. 5—Dependence of bending and hoop stresses on Poisson's ratio (cable diameter = 3 inches, E = 45,000 psi).

But from (30), (37), and (39)

$$\frac{St}{r_o} = \frac{2\mu}{E_S\sqrt{m^2+1}}\frac{t}{t_S}\frac{r_i}{r_o} = \frac{2\mu t}{E_S t_S\sqrt{m^2+1}}\left(1-\frac{t}{r_o}\right) + 0\left[\left(\frac{t}{r_o}\right)^2\right]$$

whence

$$\begin{aligned} \lambda &= 4\left(1-\nu+\frac{S_o}{1+\nu}\right) + 0(t/r_o) \\ S_o &= \frac{2\mu(1+\nu)t}{\sqrt{m^2+1}\,E_S t_S} = \frac{Et}{\sqrt{m^2+1}\,E_S t_S} \end{aligned} \tag{40}$$

Equations (40) and (35) give

$$\text{SIGMAX} = 1 + \frac{\nu^2}{1-\nu^2+S_o} + 0(t/r_o) \tag{41}$$

Similarly,

$$\frac{g(r_i)}{r_o} = 4 + 0[(t/r_o)^2]$$

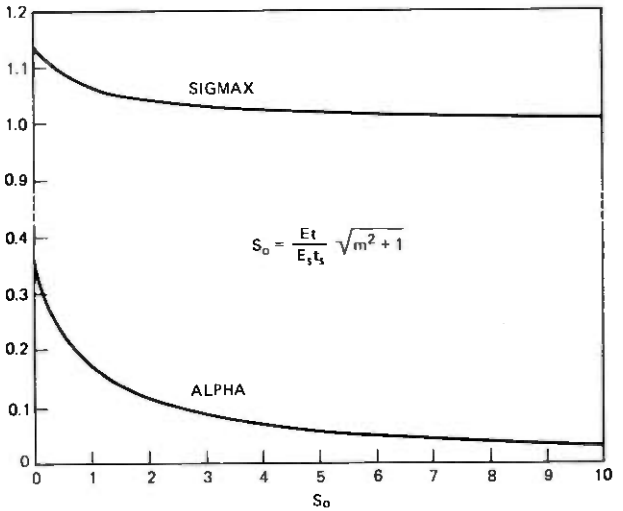


Fig. 6—Dependence of bending and hoop stresses on the modulus parameter S_o ($\nu = 0.35$).

so that, using (40) as well, ALPHA may be approximated from (36) by

$$\text{ALPHA} = \frac{\nu}{1 + S_o} + 0(t/r_o) \tag{42}$$

The dependence of the nondimensionalized stresses on Poisson's ratio is thus seen from (41) and (42) to have the form indicated in Fig. 5. Moreover, (41) and (42) exhibit no explicit dependence on cable radius, though there is a slight implicit dependence entering through S_o, since t and m are governed by cable diameter.

The approximate relations (41) and (42) have been used to generate the curves in Fig. 6. The parameter S_o for present cable designs is very small (10^{-2} to 10^{-3}) which accounts for the insensitivity of the curves in Fig. 5 to variations in jacket modulus in contrast to those in Fig. 6. In fact, the ratio $Et/E_s t_s$ is so small for all conceivable cable applications that S_o may as well be neglected. If this is done, then (41), (42) can be used in conjunction with (25), (31), (35), and (36) to provide

$$\sigma_z \approx \frac{E}{1 - \nu^2}\epsilon, \quad \sigma_\theta \approx \nu\sigma_z$$
$$\epsilon = \frac{r_o \sin\theta}{\rho} \tag{43}$$

All other stresses zero

as a good approximation to the jacket stresses.

VI. IMPLICATIONS ON CABLE PERFORMANCE

One important result that emerges from the present investigation is that (43) (for fixed θ, $0 < \theta < \pi$) is the stress state that arises in the so-called "biaxial strip" experiment (Fig. 7). When a strain $\epsilon > 0$ is imposed in the y direction, the strip is in plane stress ($\sigma_z \approx 0$) throughout and plane strain ($\epsilon_y \approx 0$) in a region near the center. The circumferential and longitudinal directions in the cable thus correspond, respectively, to the x and y directions in Fig. 7.

The approximate plane strain condition in the bent Stalpeth cable is created by the adherence of the jacket to the steel. Should this constraint remain intact during continued bending of the cable, the analogy with the strip experiment also continues. In that event, the elongation at break in the biaxial strip experiment would be an important material parameter. Another biaxial experiment[4] (equal principal stresses) indicates that some low-density polyethylenes exhibit an ultimate elongation biaxially that is substantially reduced from the uniaxial value, say to 20 percent or perhaps even less at lower temperatures or higher rates. Since jacket strains typically reach 15 percent during duct installation, failure due simply to biaxiality is indeed a concern. The strip experiment should, therefore, be instituted as a materials screening test.

Even if the material exhibits a high elongation in the biaxial strip experiment, however, the cable jacket may fail because of the localization of deformation in the neighborhood of imperfections, such as corrugation imprints or surface scratches. In these cases, failure occurs at bend radii for which the present analysis is applicable to points in the jacket away from the imperfection. Thus, if the geometry and orientation of the imperfection are known, concentration factors may be used to determine under what conditions the transition to highly localized deformation occurs. The critical parameters at which this transition occurs are currently found from impact tests on relatively narrow notched bars.[1] The

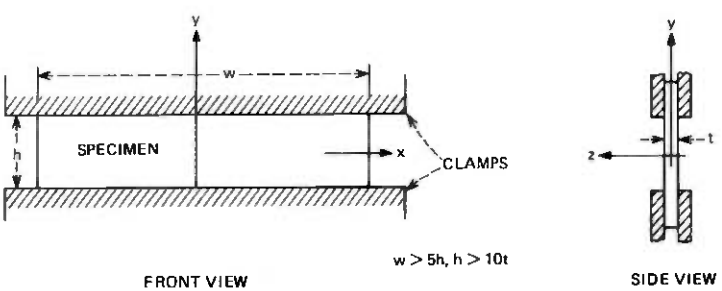


Fig. 7—Biaxial strip experiment (clamps pulled in y direction).

biaxiality evidenced here should have little effect on the conclusions drawn from such a procedure, since the notches create a complex state of stress which is largely independent of width.[14]

In addition, the growth of sharp internal or surface flaws under the influence of the far-field stresses calculated here can be analyzed with the aid of a recently developed viscoelastic theory[15] coupled with stress intensity factors available from the literature.[16] Theoretical failure times would then be compared against the duration of loading to give a probability of cracking failure for a given flaw population. Here, we remark only that the probability of premature fracture is obviously enhanced by the constraining effect of the steel, since the severity of loading for those flaws not oriented in the θ direction is increased by the biaxiality.

In addition to the effect on fracture performance, biaxiality can also have a substantial effect on the yielding behavior of plastics.[4,17] But in view of the fact that cable jackets, including Stalpeth jacket, are imprinted with corrugation valleys, *localized* yielding behavior is more relevant.

The failure mechanism on which the present analysis sheds the most light is that of sheath buckling. Firstly, on the compressive side of the cable, (25) indicates that the jacket exerts tensile loads on the flooding compound. Estimates from the preceeding formulae reveal the magnitude of this stress to be as high as 200 psi during installation. It would therefore seem that only an exceptional flooding compound would successfully restrain the jacket at low temperatures and suggests a need for the mechanical characterization of these compounds. The situation is further aggravated by the biaxiality and the fact that the stresses vary in proportion to the modulus. This accounts for the occurrence of buckling at low installation temperatures. The temperature at which buckling will occur for a given duration of loading can be estimated from this analysis and elementary viscoelasticity theory. This buckling analysis is presently being pursued.

VII. CONCLUSIONS

We have shown that:

(*i*) The state of stress in the jacket of a bent telephone cable is essentially biaxial and constant across the thickness.

(*ii*) At any fixed point in the jacket of a Stalpeth cable, the stress state is essentially that in the biaxial strip:

$$\sigma_z \approx \frac{E}{1-\nu^2}\epsilon, \quad \sigma_\theta \approx \nu\sigma_z, \quad \epsilon \equiv \frac{r_0 \sin\theta}{\rho}$$

with all other stresses and strains negligible. Thus, the maximum

bending stress is 110–130 percent of the elementary beam theory prediction while the maximum hoop stress is 30–45 percent of the bending stress.

(*iii*) The flooding compound on the compressive side of the cable is subjected to tensile and shear stresses as high as 200 psi.

(*iv*) Practical variations in cable design parameters such as corrugation depth and frequency, steel thickness and jacket thickness have no significant effect on the global stresses due to bending.

Having the results of this as well as previous investigations in mind, we have concluded that:

(*i*) The currently used techniques[1] for evaluating the notch sensitivity of plastics are applicable to the bending conditions encountered in cable installation.

(*ii*) The failure strength of flooding compounds can play an important role in preventing jacket buckling. For this reason, any anticipated change in flooding material should be thoroughly examined with respect to tensile strength and adhesive strength.

(*iii*) The probability of spontaneous cracking is increased by the adherence of the jacket to the soldered steel layer.

ACKNOWLEDGMENT

The authors appreciate the computational assistance provided by C. Jones.

REFERENCES

1. G. M. Yanizeski, E. D. Nelson, and C. J. Aloisio, "Predicting Fracture, Creep and Stiffness Characteristics of Cable Jackets from Material Properties," Proc. of the Twenty-Fifth Int. Wire and Cable Symp. (sponsored by U.S. Army Electronics Command), November 1976, pp. 272–280.
2. M. L. Williams, "Structural Analysis of Viscoelastic Materials," AIAA Journal, *2*, No. 5 (May 1964), pp. 785–808.
3. M. E. Gurtin and Eli Sternberg, "On the Linear Theory of Viscoelasticity," Archive for Rational Mechanics and Analysis, *11*, No. 4 (1962), pp. 291–356.
4. I. L. Hopkins, W. O. Baker, and J. B. Howard, "Complex Stressing of Polyethylene," J. Appl. Phys., *21*, No. 3 (March 1950), pp. 206–213.
5. Y. C. Fung, *Foundations of Solid Mechanics*, Englewood Cliffs, N.J.: Prentice-Hall, 1965.
6. I. S. Sokolnikoff, *Mathematical Theory of Elasticity*, New York: McGraw-Hill, 1956.
7. S. Timoshenko and D. H. Young, *Elements of Strength of Materials*, fourth edition, Princeton: D. Van Nostrand, 1962.
8. D. J. Meskell, tensile test data provided in a private communication, July 22, 1976.
9. V. V. Novozhilov, *Thin Shell Theory*, English translation, Gronigen: Noordhoff, 1964.
10. A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, fourth edition, New York: reprinted by Dover, 1944.
11. J. H. Michell, "On The Direct Determination of Stress in an Elastic Solid, with Application to the Theory of Plates," Proc. of the London Mathematical Soc., *31*, April 1899, pp. 100–146.
12. N. I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*, fourth edition, English translation, Gronigen: Noordhoff, 1963.

13. R. P. DeFabritis, private communication, April 27, 1976.
14. A. S. Tetelman and A. J. McEvily, Jr., *Fracture of Structural Materials*, New York: Wiley, 1967.
15. R. A. Schapery, "A Theory of Crack Initiation and Growth in Viscoelastic Media I. Theoretical Development," Int. J. Fract., *11*, No. 1 (January 1975), pp. 141–159.
16. P. C. Paris and G. C. Sih, "Stress Analysis of Cracks," Fracture Toughness Testing and Its Application, ASTM-STP-381, Philadelphia (1965), pp. 30–83.
17. R. S. Raghava, "Macroscopic Yielding Behavior of Polymeric Materials," Ph.D. dissertation, Mechanical Engineering Department, University of Michigan, 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

A Generalization of Takagi's Theorem on Optimal Channel Graphs

By F. R. K. CHUNG and F. K. HWANG

(Manuscript received June 6, 1977)

A channel graph, also called a linear graph,[5] is a multistage graph with the properties that (i) each of the first and the last stages consists of a single vertex (denoted by I and O respectively); (ii) for any vertex $v \neq I$ or O, v is adjacent to at least one vertex from the preceding stage and at least one vertex from the following stage. In a switching network, the union of all paths connecting a fixed input terminal to a fixed output terminal can usually be studied as a channel graph by taking each switch as a vertex. In comparing the blocking probabilities of two channel graphs with the same number of stages, we say one is superior *to another if its blocking probability is less than or equal to that of the other under any link occupancies. Takagi proved a basic theorem in showing one type of channel graph is superior to another. In this note we present a more powerful result which includes Takagi's theorem as a special case.*

I. INTRODUCTION

A graph is called a *multistage graph* if its vertex set can be partitioned into subsets $V_1, \ldots, V_s$, for some number s, and its edge set into subsets $E_1, \ldots, E_{s-1}$ such that E_i connects V_i with V_{i+1}. A *channel graph,* also called a linear graph,[5] is a multistage graph with the properties that (i) each of the first and the last stages consists of a single vertex (denoted by I and O respectively); (ii) for any vertex $\neq I$ or O, v is adjacent to at least one vertex from the preceding stage and at least one vertex from the following stage. In a switching network, the union of all paths connecting a fixed input switch to a fixed output switch can usually be studied as a channel graph by taking each switch as a vertex and each link as an edge.

In a switching network, a link can be in either one of two states, *busy* or *idle,* depending on whether it is part of a connection carrying a call. A path from a fixed input switch to a fixed output switch is *blocked* if

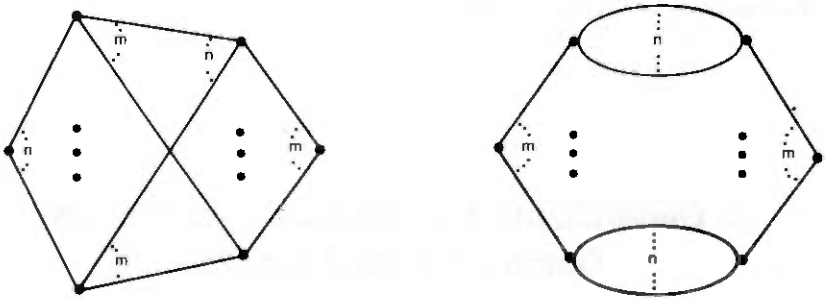


Fig. 1—Channel graphs in Takagi's theorem.

it contains a busy link. A pair of switches is blocked if every path between them is blocked. The same notion of "blocking" applies to the study of channel graphs and therefore we can talk about the blocking probability of a channel graph.

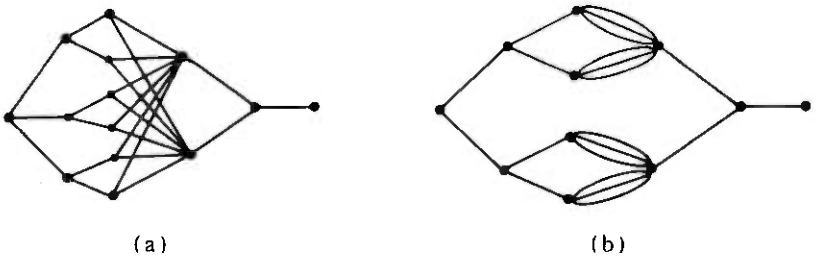
A *series-parallel* channel graph is a channel graph which is either a series combination or a parallel combination of two smaller series-parallel channel graphs with an edge being the smallest such graph. Channel graphs which are not series-parallel are often called spider-web channel graphs. Recent studies have shown, either by analysis or by simulation (see Refs. 3 and 9, for example), that spider-web channel graphs can sometimes significantly reduce blocking probabilities over series-parallel channel graphs for given switching network hardware. In particular, Takagi[8,9] gives a useful theorem which compares the blocking probability of the spider-web channel graph in Fig. 1a and the series-parallel channel graph in Fig. 1b.

In Fig. 1a, the connection between the two middle stages can be viewed as a complete bipartite graph on m and n vertices. In Fig. 1b, the connection between the two middle stages can be viewed as a matching of m pairs, each pair joined by n multiple edges.

The above theorem is the basis of Takagi's work[8,9] on optimal channel graphs which has been widely quoted in the literature (see Refs. 1, 2, 4, 6, 7, 10, 11, and 12, for example). In this note we present a more powerful result which deals with a much larger class of channel graphs and includes Takagi's theorem as a special case.

II. TAKAGI'S THEOREM

Takagi's comparison of two 4-stage channel graphs as shown in Fig. 1 actually has broader applications than it appears. The extension is made possible by interpreting each edge in the 4-stage channel graphs as a reduction of a multistage graph. The only requirement is that the



(a) (b)

Fig. 2—Two 6-stage channel graphs.

multistage graphs represented by the edges in the same set E_i are isomorphic. For instance, the two 6-stage channel graphs shown in Fig. 2 can be reduced to the two 4-stage graphs shown in Fig. 3.

Note that a vertex in the 4-stage graphs can represent a group of vertices (from the same stage) in the 6-stage channel graphs. Furthermore, two disjoint edges in the 4-stage channel graphs can come from two nondisjoint subgraphs of the 6-stage channel graphs. Finally, an edge in the 4-stage channel graphs can have more than one state where a state is basically a distinct subset of nonblocking paths in the corresponding multistage graph.

By associating a probability distribution to the joint states of the edges, the blocking probability of a channel graph can be computed. Let D be a collection of probability distributions on the joint state of the edges. Then an s-stage channel graph G is said to be *superior* to another s-stage channel graph G' with respect to D if given any member of D, the blocking probability of G never exceeds that of G'. Let $P(X_i)$, $i = 1$,

(a) (b)

Fig. 3—Two 4-stage channel graphs.

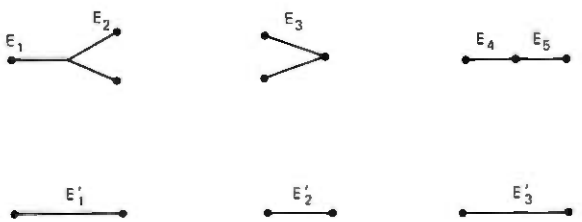


Fig. 4—The mapping from Fig. 2 to Fig. 3.

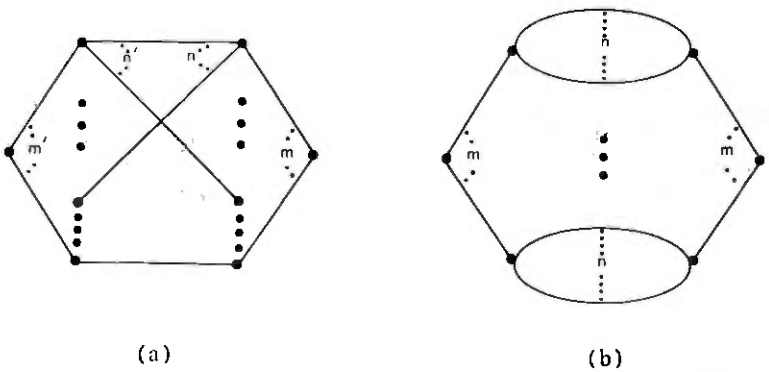


Fig. 5—Channel graphs in main theorem.

$\ldots,x$, denote the probability that an edge of E_1 is in state X_i, let $P(Z_j)$, $j = 1, \ldots, y$, denote the probability that an edge of E_3 is in state Z_j, and let $Y(i,j)$ denote the blocking probability of a path from I to O which contains an edge of E_1 in state X_i and an edge of E_3 in state Z_j. Finally, let S be a joint state of the edges in E_3. Then Takagi proves:

Takagi's theorem. The channel graph of Fig. 1a is superior to the channel graph of Fig. 1b for arbitrarily given S, $P(X_i)$, and $Y(i,j)$ under the following assumptions:

(*i*) The states of the edges are independent.
(*ii*) $n \geq m$.

In the next section, we give a generalization of Takagi's theorem.

III. THE MAIN THEOREM

Consider the two channel graphs in Fig. 5.

In Fig. 5a, each vertex in stage 2 has n' edges connected to n' distinct vertices in stage 3, and each vertex in stage 3 has n edges connected to n distinct vertices in stage 2. Furthermore $mn = m'n'$, $n' \leq m$ and $n \leq m'$. Figure 5b is the same as Fig. 1b.

Using the same notation as in Section II, we now prove:

Main theorem. The channel graph in Fig. 5a is superior to the channel graph in Fig. 5b for arbitrarily given S, $P(X_i)$ and $Y(i,j)$ under the following assumptions:

(*i*) The states of the edges are independent.
(*ii*) $m' \geq m$.

Before we prove the theorem we state a lemma proved by Takagi[8] which is a generalized version of Hölder's inequality.

Lemma. If $a_{ij} \geq 0$, and

$$\sum_{j=1}^{m} \frac{1}{b_j} = 1$$

for $b_j > 1$, and $\lambda_i \geq 0$, then the following inequality holds:

$$\sum_{i=1}^{n} \left(\lambda_i \prod_{j=1}^{m} a_{ij} \right) \leq \prod_{j=1}^{m} \left(\sum_{i=1}^{n} \lambda_i a_{ij}^{b_j} \right)^{1/b_j} \tag{1}$$

Proof of the theorem. Let S be an arbitrary state on the set of edges between stage 3 and stage 4 in Fig. 5. Suppose under S, z_j edges between stage 3 and stage 4 are in the state Z_j, $j = 1, \ldots, y$. In Fig. 5a, let w_{kj} be the number of edges joining the kth vertex in stage 2 to a vertex v in stage 3 such that the edge between v and O is in the state Z_j. Then it can be easily checked:

$$\sum_{k=1}^{m'} w_{kj} = nz_j \tag{2}$$

$$\sum_{j=1}^{y} w_{kj} = n' \tag{3}$$

and

$$\sum_{j=1}^{y} z_j = m \tag{4}$$

Let $Y(i,j)$ denote the blocking probability of a path from I to O which contains an edge in the state X_i between stage 1 and stage 2 and an edge in the state Z_j between stage 3 and stage 4. Furthermore, let $P(X_i)$ denote the probability of X_i and let x be the number of possible states X_i. Then the blocking probabilities of the channel graphs in Fig. 5a and 5b for the given state S, denoted by B_a and B_b, are

$$B_a = \prod_{k=1}^{m'} \left[\sum_{i=1}^{x} P(X_i) \prod_{j=1}^{y} Y(i,j)^{w_{kj}} \right]$$

and

$$B_b = \prod_{j=1}^{y} \left[\sum_{i=1}^{x} P(X_i) Y(i,j)^n \right]^{z_j}$$

Using (2), (3), and (4), we have

$$B_b = \prod_{j=1}^{y} \left[\sum_{i=1}^{x} P(X_i) Y(i,j)^n \right]^{\sum_{k=1}^{m'} w_{kj}/n}$$

$$= \prod_{j=1}^{y} \left\{ \prod_{k=1}^{m'} \left[\sum_{i=1}^{x} P(X_i) Y(i,j)^n \right]^{w_{kj}/n} \right\} = \prod_{k=1}^{m'} \left\{ \prod_{j=1}^{y} \left[\sum_{i=1}^{x} P(X_i) Y(i,j)^n \right]^{w_{kj}/n} \right\}$$

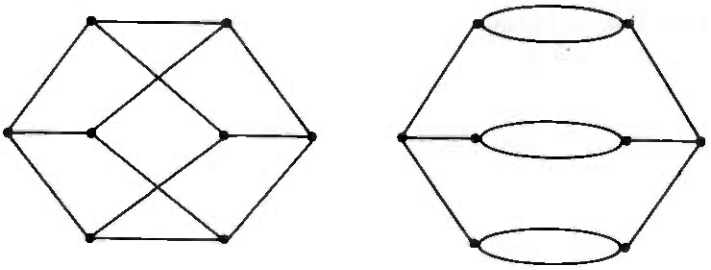


Fig. 6—Example 1.

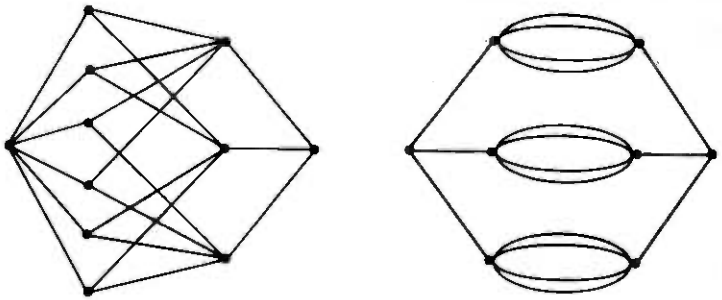


Fig. 7—Example 2.

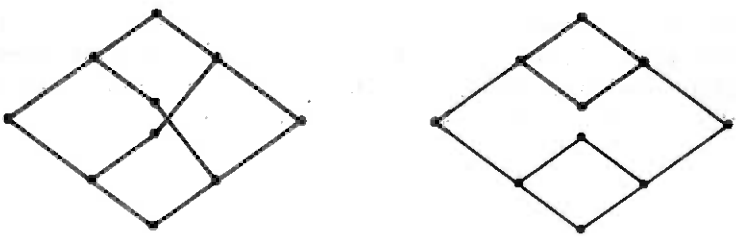


Fig. 8—A counterexample.

Fig. 9—A reduction of Fig. 8.

If $m' = m$, then $n' = n$. By substituting $\lambda_i = P(X_i)$, $a_{ij} = Y(i,j)^{w_{kj}}$ and $b_j = n/w_{kj}$ in (1), we obtain

$$B_b \geq \prod_{k=1}^{m'} \left\{ \sum_{i=1}^{x} P(X_i) \prod_{j=1}^{y} Y(i,j)^{w_{kj}} \right\} = B_a$$

If $m' > m$, then $n > n'$. Define

$$Y(i,y+1) = 1 \quad \text{and} \quad b_{y+1} = \frac{n}{n-n'}$$

Then $b_{y+1} > 1$ and

$$\sum_{j=1}^{y+1} \frac{1}{b_j} = \sum_{j=1}^{y} \frac{w_{kj}}{n} + 1 - \frac{n'}{n} = 1$$

Now

$$B_b = \prod_{k=1}^{m'} \left\{ \prod_{j=1}^{y+1} \left[\sum_{i=1}^{x} P(X_i) Y(i,j)^n \right]^{b_j} \right\}$$

$$\geq \prod_{k=1}^{m'} \left\{ \sum_{i=1}^{x} P(X_i) \prod_{j=1}^{y+1} Y(i,j)^{n/b_j} \right\}$$

$$= \prod_{k=1}^{m'} \left\{ \sum_{i=1}^{x} P(X_i) \prod_{j=1}^{y} Y(i,j)^{w_{kj}} \right\} = B_a$$

where the inequality is obtained by making a similar substitution as before. The proof is complete.

We note that by setting $m' = n$ (hence $n' = m$) in the given theorem, we obtain Takagi's theorem immediately.

IV. DISCUSSION

Two examples to which Takagi's theorem does not apply while our theorem does are shown in Figs. 6 and 7. In both figures, the channel graph (a) is superior to that of (b). The comparison in Fig. 6 is especially useful since the degrees of corresponding vertices are exactly the same.

Next we ask can we generalize our theorem in the direction that the states of the edges between stage 1 and stage 2 can also be dependent. We conjecture the theorem will still be true but no proof is known yet. Certainly we cannot hope to imitate the proof given here. A ready counterexample to this approach is illustrated in Fig. 8.

If the two states we assume are such that the first edge between the first two stages and the first edge between the last two stages are blocked, then the two channel graphs can be reduced to the channel graphs in Figs. 9a and b, respectively. Clearly, the channel graph in Fig. 9b is superior. But from Takagi's theorem or our theorem, we know that the channel graph in Fig. 9a is superior to that in Fig. 9b.

REFERENCES

1. S. Asano, T. Saito, and H. Inose, "An Expression for Structures of Connecting Networks," Electron. Comm. Japan, *57-A* (1974), pp. 68–76.
2. K. W. Cattermole, "Graph Theory and the Telecommunication Networks," Bull. Inst. Math. and Its Applications, *11*, No. 5 (1975), pp 74–106.
3. G. Harland, "Design of Mixed Analogue and Digital Switching Networks," 8th ITC, Melbourne (1976), pp. 546-1 to 546-6.
4. K. Kodaira and K. Takagi, "A General Purpose Blocking Probability Calculation Program for Multi-stage Link Systems," 8th ITC, Melbourne (1976), pp. 331-1 to 331-8.
5. C. Y. Lee, "Analysis of Switching Networks," B.S.T.J., *34*, No. 6 (December 1955), pp. 1287–1315.
6. A. Lotze, A. Röder, and G. Thierer, "PPL—A Reliable Method for the Calculation of Point-to-Point Loss in Link Systems," 8th ITC, Melbourne (1976), pp. 547-1 to 547-14.
7. V. I. Neiman, "Structural Properties of Connecting Networks," 8th ITC, Melbourne (1976), pp. 131-1 to 131-8.
8. K. Takagi, "Design of Multi-stage Link Systems by Means of Optimum Channel Graphs," Electron. Commun. Japan, *51-A* (1968), pp. 37–46.
9. K. Takagi, "Optimal Channel Graph of Link Systems," Electron. Commun. Japan, *54-A* (1971), pp. 1–10.
10. G. Timperi and D. Grillo, "Structural Properties of a Class of Link Systems," Alta Frequenza, *41* (1972), pp. 278–289.
11. J. G. van Bosse, "On an Inequality for the Congestion in Switching Networks," IEEE Trans. Commun., *22* (1974), pp. 1675–1677.
12. J. G. van Bosse, "A Generalization of Takagi's Results on Optimal Link Graphs," 8th ITC, Melbourne (1976), pp. 513-1 to 513-7.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

Computer-Aided Magnetic Circuit Design for a Bell Ringer

By R. M. HUNT and J. W. NIPPERT

(Manuscript received June 17, 1977)

A general computer-aided design method for use with electromagnetic devices such as ringers, relays, and solenoids is described. The method is demonstrated by applying it to the design of polarized bell ringers. A lumped-element model with electrical, magnetic, and mechanical portions is used in the analysis. First, interaction equations are derived using a Lagrangian formulation applied to a simple model. Second, the model is refined by subdividing the iron members and including more leakage paths. An electrical circuit analysis program assembles the equations for the electromagnetic portion of this more complete model and produces a subroutine that solves these equations. A computer program has been written to predict the effects of changing motor parameters. The versatility and usefulness of the design technique has been demonstrated by applying it to the Bell System TRIMLINE® *telephone set ringer to achieve major design improvements.*

I. INTRODUCTION

In 1952 the Bell System introduced the standard 500-type telephone set which uses a "universal" two-gong ringer (C-type) to meet a wide variety of service conditions.[1] A universal ringer must: (*i*) have two voltage sensitivity modes that ensure adequate operation under worst-case conditions and provide protection against cross ringing on party lines, (*ii*) be electrically polarized to protect against bell tapping due to dialing transients, (*iii*) have high impedance (≈8 kohms) at ringing voltages so that multiple ringers can be bridged across the line, (*iv*) have high impedance (≈120 kohms) at voice frequencies to prevent speech signal attenuation, and (*v*) have two coil windings for use with multiparty ringing circuits.[2]

Recently a project was undertaken to design a miniature single-gong ringer that would have the low cost, reliability, sensitivity, and sound

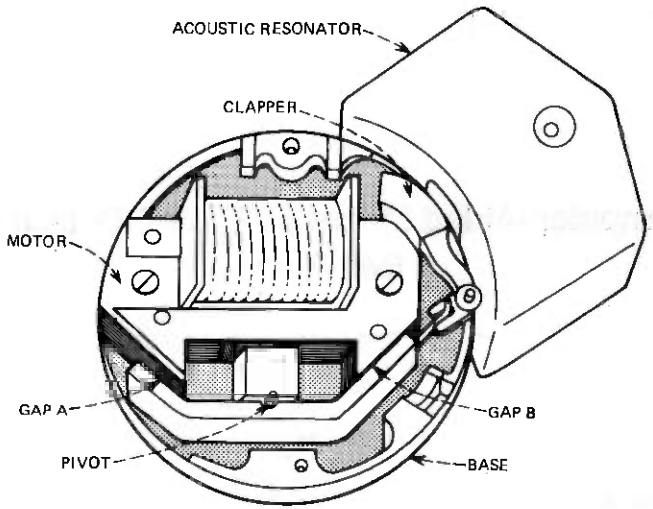


Fig. 1—Ringer with rocking-armature motor (gong and clamp plate removed).

output of the C-type ringer but would be significantly smaller. The small single-gong P-type ringer commonly used in *TRIMLINE*® telephones and other new telephone sets did not optimally meet all of these objectives and previous attempts to meet them were unsuccessful. To achieve these objectives, effort was channeled toward a rocking armature type motor[3,4] that drives a single clapper and fits under the gong. A mathematical model was essential for the design process in order to perform parametric analyses and design optimization. The mathematical model had to realistically account for flux saturation of materials and flux leakage paths in the design, both of which have significant influence on the performance of a compact ringer motor.

II. RINGER OPERATION

The structure of a ringer with a rocking armature motor is shown in Fig. 1 and an exploded view of the motor is shown in Fig. 2. In the absence of coil current a bias spring (not shown) returns the armature to the position shown in Fig. 1. The permanent magnet flux passes through both armature gaps and nearly saturates the shunt member. The ringer is electrically polarized since only coil current of one polarity will cause operation. Coil current in the operate direction increases the flux in gap A and decreases the flux in gap B causing clockwise armature rotation which drives the clapper away from the gong. As the current returns to zero the armature returns to its normal position and drives the clapper into the gong. For negative coil current the shunt member of the magnetic circuit has a relatively low reluctance so that most of the coil flux

passes through the shunt instead of the working gaps. The low reluctance shunt results in the ringer having a high electrical inductance which is necessary for resonance of the ringer circuit at the frequency of the ringing power supply, 20 Hz.

The ringer motor consists of a bobbin-wound coil, a pole piece assembly, a permanent magnet, a pivot pin, and an armature of low carbon steel. Figure 2 shows one version of the rocking armature motor which

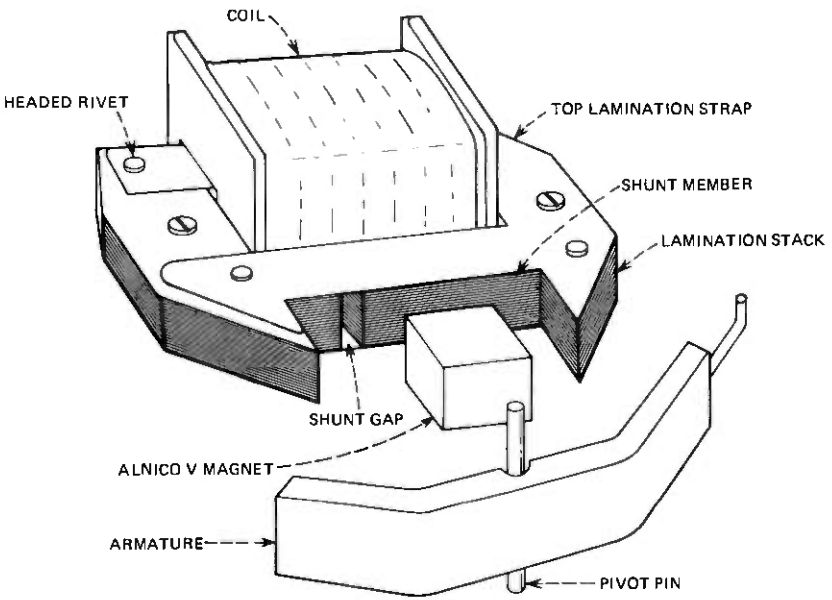


Fig. 2—Rocking-armature motor.

is described in this paper. In this version the pole piece assembly consists of two lamination stacks sandwiched between two silicon steel lamination straps and held together with three aluminum rivets.

Basic motor performance is usually studied using measured static torque curves that show the relationship between blocked armature torque, armature position, and dc coil current. Figure 3 shows two idealized torque curves, one with the armature blocked in the nonoperate position (gap B closed) and the other with the armature blocked in the operate position (gap A closed). A single cycle of very slowly varying coil current is shown below the torque curves to illustrate the quasistatic operation of the motor. A major objective of the mathematical model described here was the prediction of static torque curves from basic motor parameters such as dimensions, material magnetic characteristics, and armature displacement.

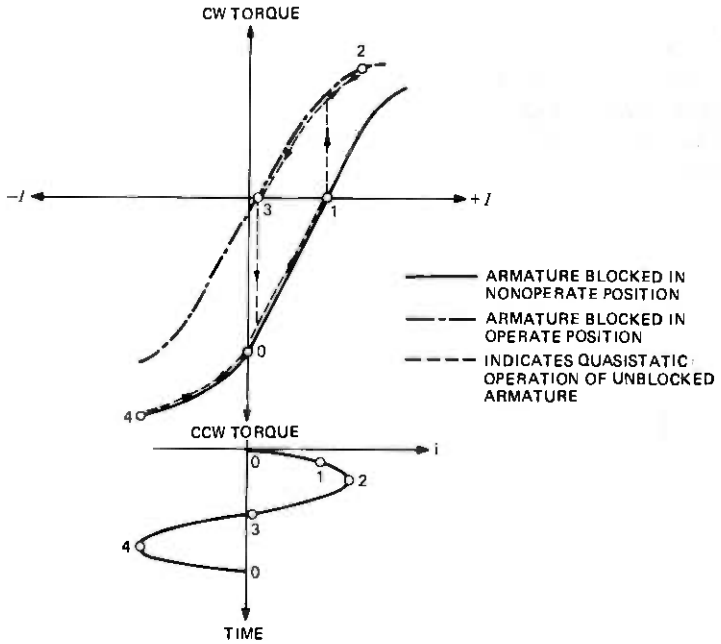


Fig. 3.—Ringer motor static torque curves and quasistatic operation.

III. MATHEMATICAL MODEL

Successful methods of modeling electromechanical devices such as relays and solenoids have appeared elsewhere.[5,6] The electromechanical ringer is a complex device and a complete analysis would involve the solution of Maxwell's three-dimensional field equations in space occupied in part by nonlinear material with memory. Fortunately, because ringers are low-frequency and low-velocity devices, the following simplications can be made for design analysis:

(*i*) The field problem can be approximated by a network where each volume of space occupied by a uniform material is represented as one or more lumped elements. Each element represents a physical effect in the uniform member, such as reluctance or loss.

(*ii*) All magnetic members except the permanent magnet are assumed not to have memory (no hysteresis).

(*iii*) Stray flux paths through air can be adequately represented by lumped leakage reluctances.

A lumped-element model for the magnetic portion of the motor of Fig. 2 appears in Fig. 4. This model uses the common flux-current analogy between magnetic circuits and electric circuits. Iron members are modeled by nonlinear reluctances and are characterized by normal *B-H*

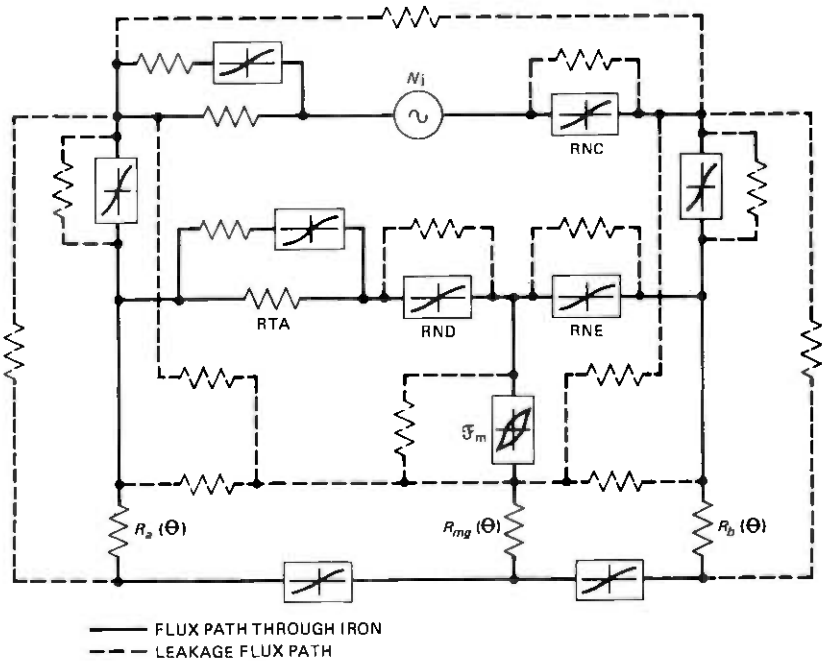


Fig. 4—Lumped element model (equivalent circuit) for magnetic portion of rocking-armature motor.

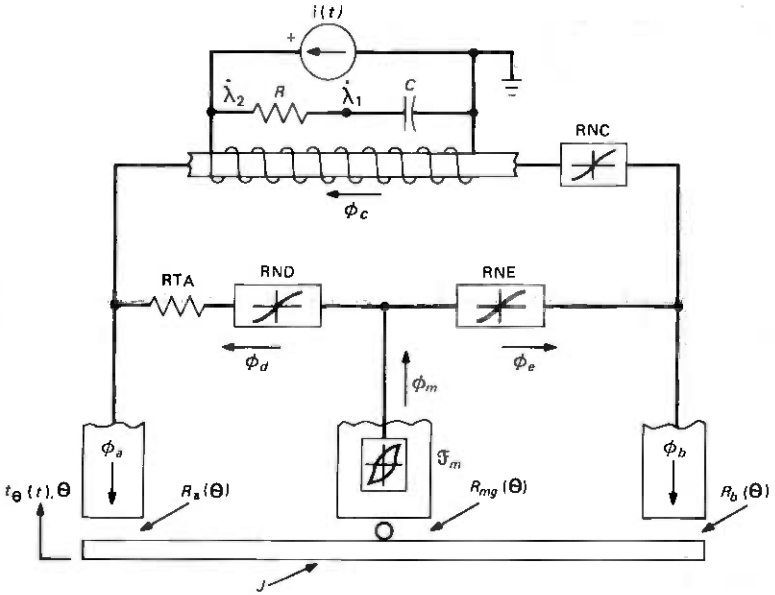
curves. The magnet is modeled using its demagnetization (de-mag) curve and recoil permeability. A function was fitted to reference data for the *B-H* characteristics. The fitting process and results are described in Appendix A.

Leakage flux paths are modeled by lumped reluctances. The number and magnitude of these reluctances were determined from available formulae and prototype measurements as described in Appendix B.

IV. EQUATION FORMULATION

A consistent way to develop the equations for a system that has mechanical, magnetic, and electrical energy is to use the Lagrangian formulation. The procedure is illustrated in Appendix C by applying it to the simplified model of Fig. 5. The equilibrium equations for the simplified model are a set of five differential equations, one electrical, one mechanical, and three magnetic, the simultaneous solution of which describes the dynamic performance of the motor.

The equilibrium equations are reduced to the static case since only quasistatic motor performance is considered here. In the resulting set of nonlinear equations the current source becomes a magnetomotive source NI in one of the magnetic equations and the mechanical equation



$\dot{\lambda}_1$ = VOLTAGE ACROSS CAPACITOR
$\dot{\lambda}_2 = N\dot{\phi}_c$ = VOLTAGE ACROSS COIL
J = ARMATURE MOMENT OF INERTIA
$t_\Theta(t)$ = EXTERNALLY APPLIED TORQUE
$\mathcal{F}_m$ = MAGNET MAGNETOMOTIVE FORCE
ϕ = MAGNETIC CIRCUIT FLUX

Fig. 5—Simplified lumped element model for rocking-armature motor.

gives the static torque on the armature. See Appendix C, eq. (21). Armature torque t_o is given by

$$t_o = -\frac{\phi_a^2}{2P_a^2}\frac{\partial P_a}{\partial \theta} - \frac{\phi_b^2}{2P_b^2}\frac{\partial P_b}{\partial \theta} - \frac{(\phi_a + \phi_b)^2}{2P_{mg}^2}\frac{\partial P_{mg}}{\partial \theta}$$

where the ϕ's and P's are the gap fluxes and permeances respectively. Magnetic permeance is the reciprocal of reluctance.

It is apparent from this equation that for accurate calculations of torque, appropriate functions for gap permeance must be determined. This is done in Appendix D. The gap fluxes are found by solving the three magnetic equations for specified values of coil current and armature displacement. The gap fluxes are then used in the torque equation above. To achieve accurate calculations of circuit fluxes it was found necessary to include up to 30 elements in the magnetic portion of the model.

Having used the Lagrangian formulation to determine the terms of mechanical-magnetic interaction for the simplified model, circuit

analysis techniques were then used to generate the equations for the electromagnetic portion of the more complex model shown in Fig. 4. This step permitted increasing model complexity to account for significant measured effects without tedious derivation of new equilibrium equations.

V. SOLUTION OF EQUATIONS

The next task is to find the solution to a set of simultaneous nonlinear equations. Since linear equations are easy to solve, at least in principle, the set of nonlinear equations may be solved most easily by finding a sequence of solutions to related linear equations that converges to the solution of the nonlinear equations. In general, closed-form (exact) solutions do not exist for nonlinear systems, so some such iterative method of solution must be used.

The procedure adopted is well known as Newton's method. The nonlinear system is expanded in a Taylor series about some trial solution. Retaining only the linear terms in this expansion, the resulting set of linearized equations is solved to yield a new approximation to the solution. Successive solutions converge to the solution of the original nonlinear system.

The magnetic portion of the more complex model was analyzed using the Circuit Analysis Program for Efficient Computer Optimized Design (CAPE COD).[7] From a topological input description of the magnetic circuit, CAPE COD produces a subroutine that solves the network equations. This routine is combined with the interaction terms derived from the Lagrangian formulation to provide the complete device model for the static case.

5.1 Sequence of analyses

The first calculation procedure determines a maximum operating point for the magnet. For this calculation, the magnet is described by its de-mag B-H curve, and the independent coil current source is set to zero. A tentative operating point is found by iteration, as described above. Once this maximum operating point is determined, the operating point of the magnet is lowered until the ringer will just begin to operate at a specified current. The results of this analysis are then printed.

A graph for the de-mag B-H curve for the magnet and the "load line" of the rest of the circuit as seen by the magnet is produced. For this calculation of the load line, a value for H is selected, and the slope of the B-H curve for the magnet is set to zero. The solution of this system gives the corresponding value of B on the load line for the selected H. The independent source associated with the coil is again set to zero for these calculations.

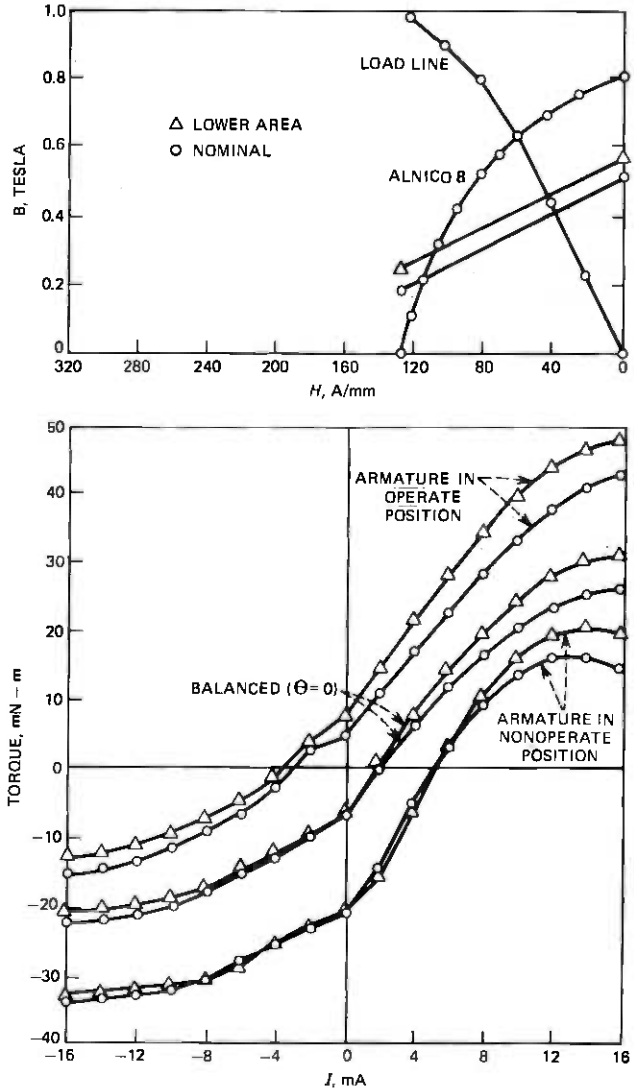


Fig. 6—Typical program output showing the predicted effect of decreasing the cross-sectional area of a magnetic shunt member.

To calculate torques, the magnet characteristic used is that of the recoil line through the final operating point. The independent source associated with the coil is set to the appropriate ampere-turns, and the calculations proceed as before. The torque on the armature is calculated from the flux through the gaps at the armature. The displacement of the armature and the ampere-turns of the source are varied to produce a family of curves. Figure 6 is a sample of the plotted program output

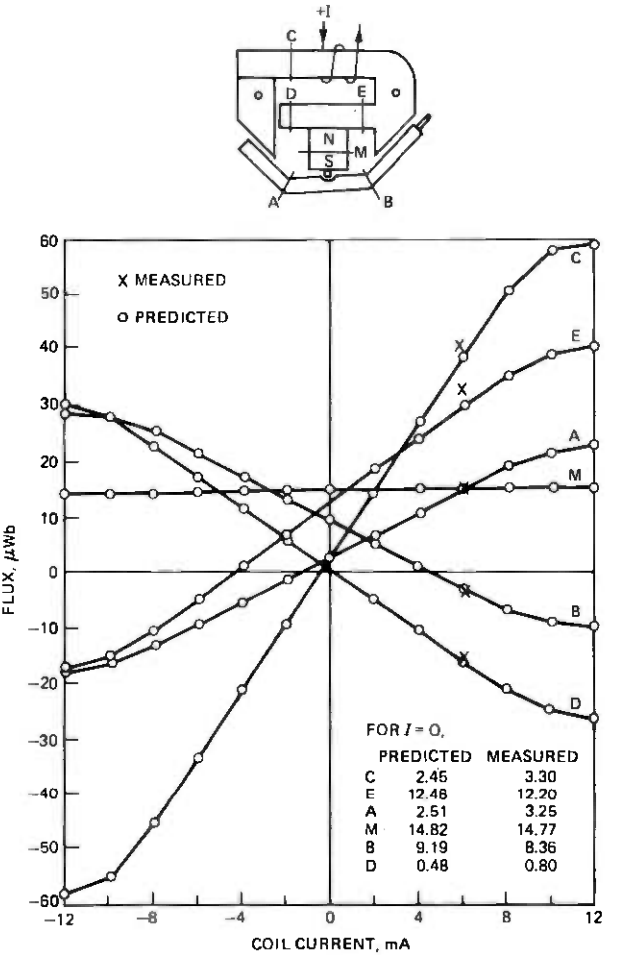


Fig. 7—Flux versus coil current, predicted and measured.

showing both the permanent magnet and torque characteristics of a motor and the predicted effect of changing a single parameter, the cross-sectional area of the shunt.

VI. VERIFICATION OF MATHEMATICAL MODEL

Measurements were made on an early prototype of the rocking armature motor in order to verify the mathematical model. The solid curves of Fig. 7 show the predicted flux versus coil current in key magnetic members. Measured fluxes, some of which are indicated, showed good agreement with predicted values. Agreement was particularly good for fluxes ϕ_a and ϕ_b, which determine armature torque.

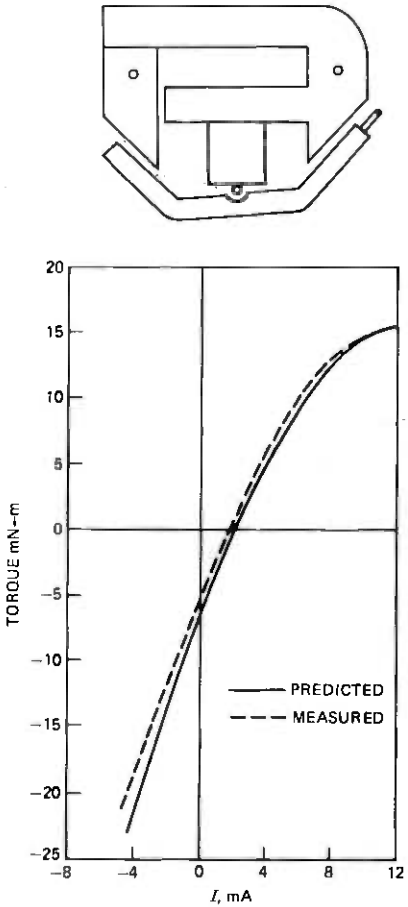


Fig. 8—Predicted and measured torque curves.

The comparison of predicted to measured armature torque is shown in Fig. 8. Agreement between these two curves is very satisfactory. A comparison of the predicted to measured effect of changing two important variables, armature displacement and magnet strength, is shown in Fig. 9. The measured change is considered to show adequate agreement, especially in the important midrange currents where ringer bias adjustments are made.

VII. OPTIMIZATION

The mathematical model of the ringer motor aided first in identifying key parameters and second in finding a set of parameter values which optimize motor performance. The General Purpose Optimization Package (GPOP)[7] was combined with the ringer analysis program to

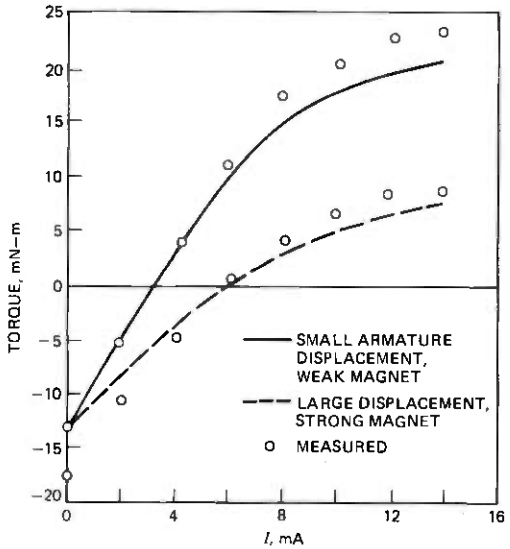


Fig. 9—Predicted and measured effects of parameter changes.

perform that optimization. First a criterion function based on the desired motor characteristics was developed. Second, the program determined a set of values for the parameters that improves the performance of the ringer relative to a set of initial values provided to the program.

7.1 Desired motor output

Basic motor performance is evaluated through use of static torque curves. The two idealized torque curves of Fig. 3 are shown in Fig. 10, with important points indicated. The two "stick" torques are defined to occur at zero current with the armature on either pole. They are indicated on the figure by the labels T_a and T_b. Torque T_a must be either negative (counterclockwise) or, if positive (clockwise), must be small enough to be overcome by a practical bias spring. This ensures that the armature always returns to its nonoperate position in the absence of coil current.

Torque T_b must be large enough to drive the clapper against the gong with sufficient velocity. At the same time T_b must not be so large that excessive armature impact noise and wear are produced by the armature striking pole B.

Of primary importance is the slope of the nonoperated torque curve immediately above turn-on current I_1. This is called the motor torque factor and should be as high as possible to facilitate ringer sensitivity adjustments and to maximize ringer reliability. The slope between I_2 and I_3 is usually lower than that between I_1 and I_2 but must be high

enough to prevent reliability problems in high ringing-voltage situations with the ringer bias spring in its high-tension position.

Another important criterion is the inductance of the ringer. The change in flux through the coil between peak currents was taken as an approximation to the inductance.

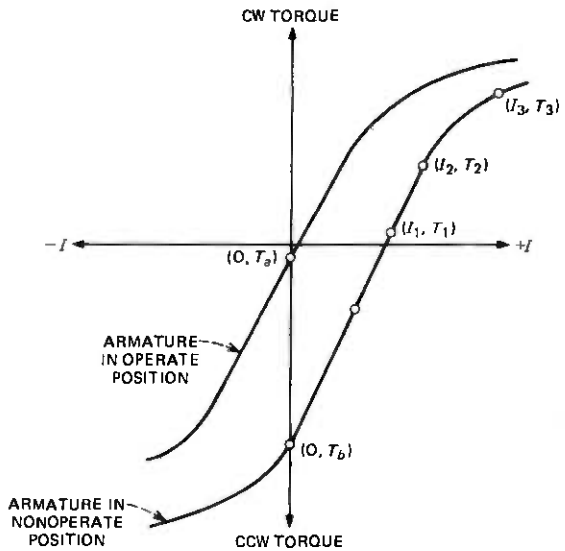


Fig. 10—Idealized torque curves.

7.2 Criterion function

From these diverse requirements a function is formed that expresses in a single number a measure of performance for a ringer built to match any given set of parameters. The form of the function selected involves the use of penalty functions. The idea is fairly simple. To a function describing how good the performance is, functions are added that are zero if a given constraint is satisfied and nonzero if the constraint is violated. Appendix E shows the function used.

VIII. RESULTS

8.1 Rocking armature motor

The modeling technique was used to evaluate variations of the basic rocking armature motor so as to best meet the diverse requirements. Based on the theoretical results obtained, laboratory models of ringer

designs were built and evaluated in terms of the Bell System ringer requirements. Figure 11 shows a motor design which meets requirements and appears simple to manufacture. Iron straps are used to sandwich the lamination stacks and also to reduce flux densities in the assembly. A short Alnico 8 magnet is used which allows the use of a straight armature. Figure 12 shows the predicted and measured results of an optimization run. As shown, a significant reduction in pole A stick torque

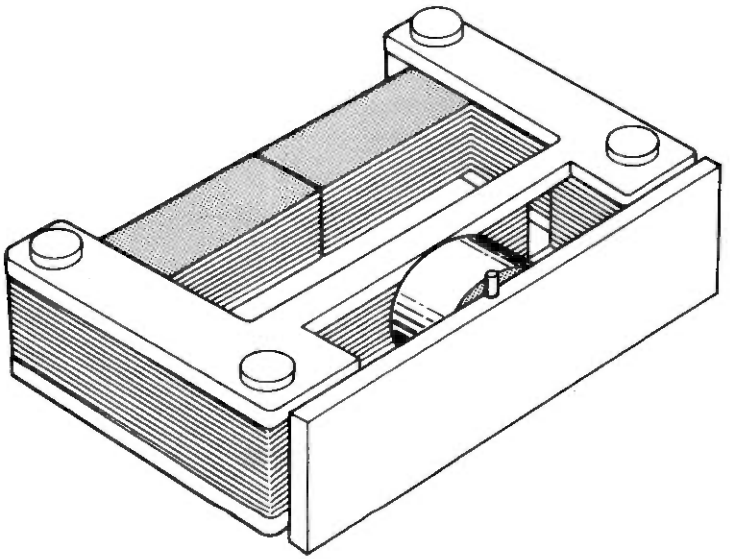


Fig. 11—Ringer motor with straight armature.

is achieved, but with a sacrifice in torque factor. Further evaluation and refinement of this rocking armature design was not pursued since effort was turned toward improving the P-type ringer as described below.

8.2 Bell System P-type ringer

The analysis techniques developed for the rocking armature motor were applied to the Bell System small P-type ringer. The objective was to investigate potential improvements to the torque curve with simple design changes. Figure 13 shows the motor structure which is a single-ended rocking-armature type with air gaps on opposite sides of the armature. Figure 14 shows the equivalent circuit used to analyze it. It was predicted and then verified with measurements that a significant improvement in torque factor is achieved with two basic design changes. First, the magnetic bias force is reduced by moving the pivot pin closer to the magnet center. This results in higher magnet flux after the mag-

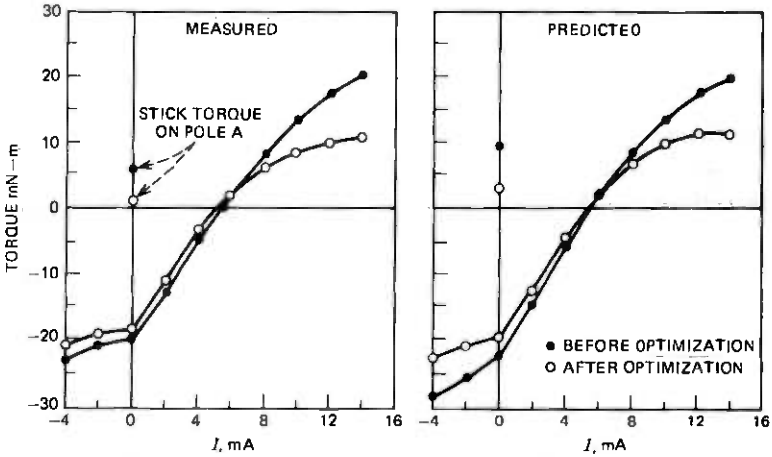


Fig. 12—Predicted and measured effect of optimized parameters.

netization adjustment. Secondly, the flux density at a critical point of saturation is reduced by increasing the cross-sectional area of the pole piece. The measured results are shown in Figure 15.

IV. CONCLUSIONS

A general method of computer-aided design for electromagnetic devices has been presented. The method was applied to the design of two polarized bell ringer motors. A mathematical model and computer simulation of the rocking armature motor made it possible to accurately and efficiently study motor performance as a function of the many parameters involved. The simulation methods were also applied to the P-type ringer commonly used in the *TRIMLINE* telephone and resulted

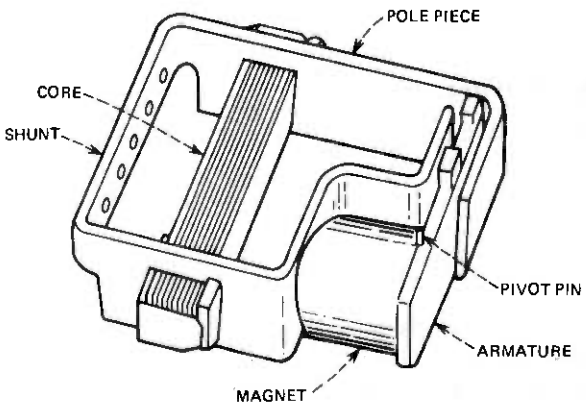


Fig. 13—P-type ringer motor.

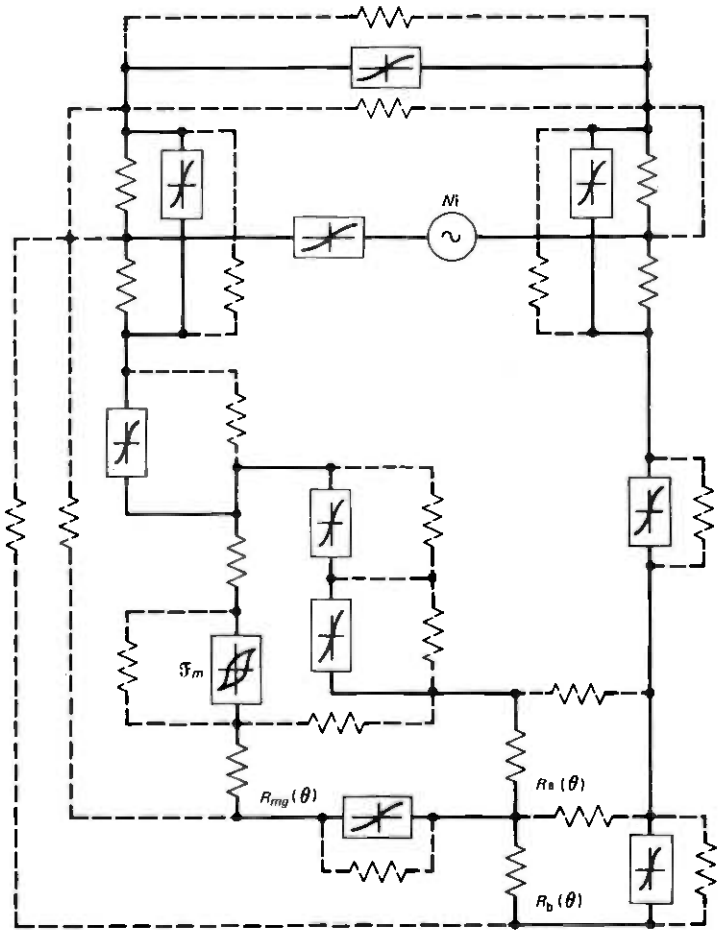


Fig. 14—Lumped element model of P-ringer motor.

in a significantly higher torque factor. This improvement in addition to others not covered here have led to a modified P ringer virtually meeting the original goals of the rocking armature ringer development.

X. ACKNOWLEDGMENTS

The authors are indebted to P. O. Schuh for suggesting the Lagrangian formulation. D. P. Borenstein, G. M. C. Fisher, D. J. Leonard, E. A. Mills, M. L. Warnock, R. S. Weiner, and K. B. Woodard supported and encouraged the effort. L. A. Marcus, N. B. Karu, and W. F. Wernet aided in the physical design and dynamic measurements of later models.

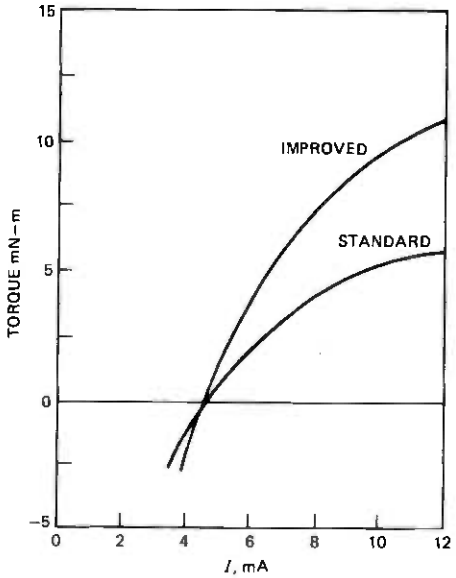


Fig. 15—P-type ringer motor measured torque curves.

APPENDIX A

Iron member characterization

The characteristics of an iron member are modeled using the element's length, width, and the normal B-H curve for the material. Several methods have been used for representing this curve: a table of (B,H) pairs may be given, simple hyperbolas may be fit to a range of the curve, a series of exponentials may be fit to the curve, a function of a polynomial divided by another polynomial (rational function) may be fit to the data. For this analysis H is given as the ratio of two polynomials in B.

A family of functions with numerator and denominator polynomials of varying orders was fitted to the data. The function selected from this family provided a good combination of low order, close approximation to the data being fitted, and smooth behavior of the derivative. Figure 16 shows the selected curve fit and the data points used in the fitting process.

APPENDIX B

Leakage elements

Leakage elements are reluctances which represent stray flux paths through air from two points in the magnetic circuit. Leakage elements placed across nonlinear magnetic elements reduce the sharpness with which the elements go into saturation and are essential to obtaining agreement between predicted and measured fluxes.

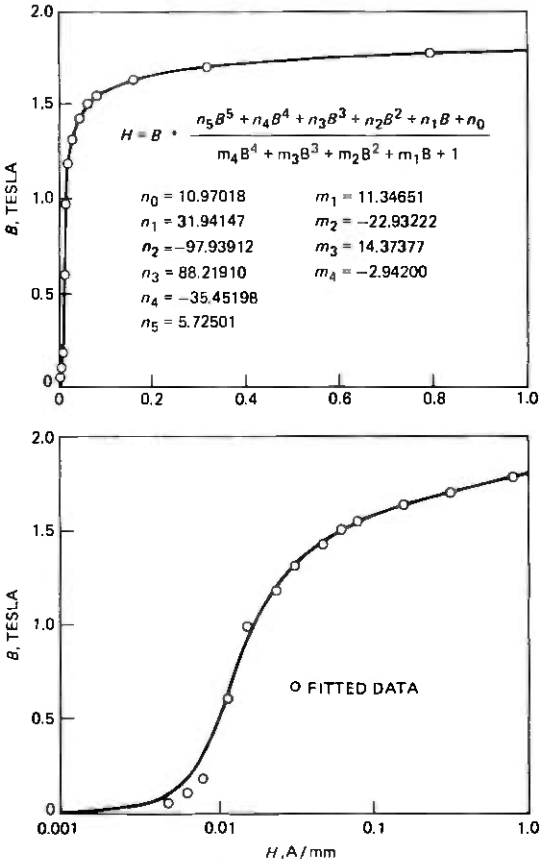


Fig. 16—*BH* curve fit for silicon steel.

The number of leakage elements and their placement in the model can be observed by sprinkling iron fillings on a prototype as shown in Fig. 17. Precise numerical values of these leakage elements are difficult to obtain. However, the method of "estimating the permeances of probable flux paths" from Ref. 8 has proven useful in obtaining first-order estimates. Also, low-frequency inductance measurements of partial motor assemblies can be used to estimate leakage around the coil. Flux measurements on the physical model are then used to refine the initial estimates. Once determined, the leakage reluctances remain constant unless major changes in geometry are made.

APPENDIX C

Lagrangian formulation

The first step for the Lagrangian formulation is to select an appropriate set of generalized coordinates which define the state of the system.

Typically these are flux linkage, node voltage, capacitor charge, inductor currents, and physical displacements. For this ringer analysis, coordinates representing linkage, node voltage, and angular displacement of the armature where selected.

The second step of the procedure requires formulation of the Lagrangian function and a Rayleigh dissipation function in terms of the generalized coordinates. Since most power dissipation occurs in winding resistance and since the technique becomes burdensome with individual hysteresis and eddy current loss elements, a single electrical loss element was assumed to account for the total system dissipation.

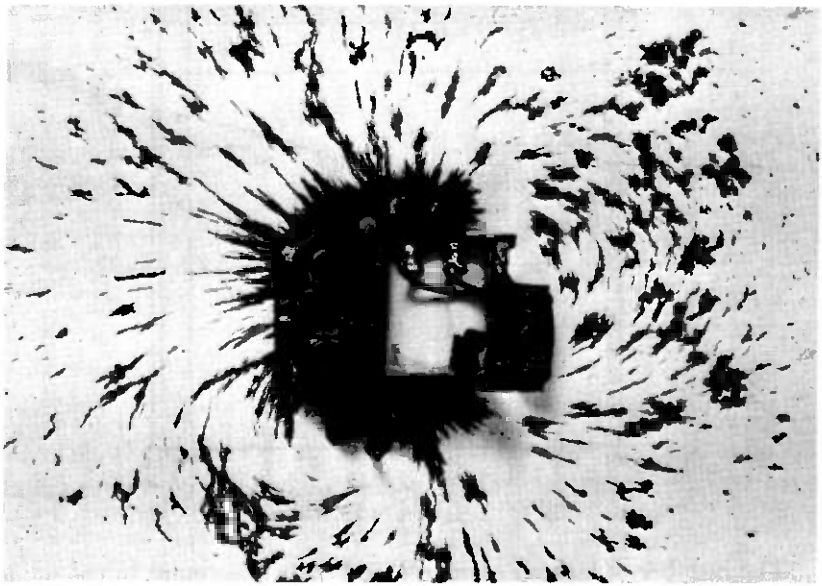


Fig. 17—Prototype ringer with iron filings.

The third step applies Lagrange's formula to the energy functions to produce a set of simultaneous differential equations. Primarily this involves taking partial derivatives and collecting terms.

The advantage of this technique is that it systematically accounts for the interaction between the magnetic and mechanical portions of the system. The disadvantages are as follows: first, the choice of generalized coordinates is not necessarily obvious; second, hysteresis and other losses are difficult to fit into the formulation.

The Lagrangian $\mathcal{L}$ is defined as the difference between the total system coenergy function $\mathcal{T}'$ and the total system energy function $\mathcal{V}$ as defined below. Choosing a nodal formulation for the electrical portion

of the system and using the functional dependencies prescribed in Ref. 9 we have:

$$\mathcal{L}(\dot{\theta},\theta,\dot{\lambda},\lambda,t) = \mathcal{T}'(\dot{\theta},\theta,\dot{\lambda},t) - \mathcal{V}(\theta,\lambda,t) \tag{1}$$

where θ is the angular displacement of the armature, λ is the flux linkage, t is time and dots above quantities denote differentiation with respect to time.

The system coenergy function is the sum of the mechanical (kinetic) and electrical coenergy functions:

$$\mathcal{T}'(\dot{\theta},\theta,\dot{\lambda},t) = T'(\dot{\theta},\theta,t) + W_e'(\dot{\lambda},\theta) \tag{2}$$

The system energy function is the sum of mechanical (potential) and magnetic energy functions:

$$\mathcal{V}(\theta,\lambda,t) = V(\theta,t) + W_m(\lambda,\theta) \tag{3}$$

For the system of Fig. 5 the mechanical coenergy function is

$$T'(\dot{\theta},t) = \frac{1}{2}J\dot{\theta}^2 \tag{4}$$

where J is the moment of inertia of the armature about the pivot axis. The electrical coenergy function is:

$$W_e'(\dot{\lambda}_1) = \frac{1}{2}C\dot{\lambda}_1^2 \tag{5}$$

where C is the value of the ringing capacitor. The total system coenergy function now becomes:

$$\mathcal{T}'(\dot{\theta},\lambda_1,t) = \frac{1}{2}J\dot{\theta}^2 + \frac{1}{2}C\dot{\lambda}_1^2 \tag{6}$$

Since no change in mechanical potential energy is assumed, it remains to find the magnetic energy function, W_m. The energy W in a lumped magnetic element of length ℓ, area A, flux density B and magnetic field H can be approximated by the following integral:

$$W = \ell A \int_o^B H(B')\,dB' \tag{7}$$

Since $B = \phi/A$,

$$W = \int_o^{\phi} \ell H\left(\frac{\phi'}{A}\right) d\phi' = \int_o^{\phi} \mathcal{F}\left(\frac{\phi'}{A}\right) d\phi', \tag{8}$$

where ℓH is called the magnetomotive force (MMF) and is labeled $\mathcal{F}$.

For linear magnetic elements $\mathcal{F}$ is a linear function of flux, i.e., $\mathcal{F} = R\phi$, where R is the reluctance of the element and defined as $\ell/\mu_o A$.

Therefore the energy in a linear element is:

$$W = \int_o^{\phi} R\phi' \, d\phi' = \frac{1}{2} R\phi^2. \tag{9}$$

In a system model that includes more magnetic elements than those which directly link the coil, it is convenient to use flux (ϕ) for coordinates in the magnetic portion of the circuit rather than flux linkage (λ). Both have the same dimensions, differing only by the dimensionless quantity, number of turns.

The total magnetic energy function, W_m, can now be expressed by summing the energy of magnetic elements:

$$\begin{aligned} W_m(\phi_i,\theta) = {} & \frac{1}{2} R_a(\theta)\phi_a^2 + \frac{1}{2} R_b(\theta)\phi_b^2 + \frac{1}{2} R_{mg}(\theta)\phi_m^2 \\ & + \frac{1}{2} R_{ta}\phi_d^2 + \int_o^{\phi_m} \mathcal{F}_m(\phi_m') \, d\phi_m' + \int_o^{\phi_d} \mathcal{F}_d(\phi_d') \, d\phi_d' \\ & + \int_o^{\phi_e} \mathcal{F}_e(\phi_e') \, d\phi_e' + \int_o^{\phi_c} \mathcal{F}_c(\phi_e') \, d\phi_e' \end{aligned} \tag{10}$$

Note that the first three terms are gap energies and are functions of armature displacement. The last four terms are the energies in the nonlinear elements.

Continuity of flux in the magnetic circuit provides the following constraint equations:

$$\begin{aligned} \phi_m &= \phi_a + \phi_b \\ \phi_d &= \phi_a - \phi_c \\ \phi_e &= \phi_c + \phi_b \end{aligned} \tag{11}$$

After substituting the constraint equations into eq. (10) and then using eqs. (3), (6), and (10) in (1), the Lagrangian becomes:

$$\begin{aligned} \mathcal{L}(\dot\theta,\theta,\dot\lambda_1,\phi_i,t) = {} & \frac{1}{2} J\dot\theta^2 + \frac{1}{2} C\dot\lambda_1^2 - \frac{1}{2} R_a(\theta)\phi_a^2 \\ & - \frac{1}{2} R_b(\theta)\phi_b^2 - \frac{1}{2} R_{mg}(\theta)(\phi_a + \phi_b)^2 - \frac{1}{2} R_{ta}(\phi_a - \phi_c)^2 \\ & - \int_o^{\phi_a+\phi_b} \mathcal{F}_m(\phi_m') \, d\phi_m' - \int_o^{\phi_a-\phi_c} \mathcal{F}_d(\phi_d') \, d\phi_d' \\ & - \int_o^{\phi_c+\phi_b} \mathcal{F}_e(\phi_e') \, d\phi_e' - \int_o^{\phi_c} \mathcal{F}_c(\phi_c') \, d\phi_c' \end{aligned} \tag{12}$$

The Rayleigh dissipation function D is

$$D = \int_o^{\dot\lambda_r} \frac{\dot\lambda_r'}{R} \, d\dot\lambda_r' = \frac{\dot\lambda_r^2}{2R} = \frac{(\dot\lambda_2 - \dot\lambda_1)^2}{2R} = \frac{(N\dot\phi_c - \dot\lambda_1)^2}{2R} \tag{13}$$

where $\dot{\lambda}_r$ is the voltage across resistor R and the voltage across the coil $\dot{\lambda}_2$ is equal to $N\dot{\phi}_c$.

Lagrange's formula can now be applied:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) - \frac{\partial \mathcal{L}}{\partial q_i} + \frac{\partial D}{\partial \dot{q}_i} = Q_i \tag{14}$$

For the electrical coordinate λ_1, we have

$$C\ddot{\lambda}_1 - \frac{1}{R}(N\dot{\phi}_c - \lambda_1) = 0 \tag{15}$$

For the mechanical coordinate θ we have

$$J\ddot{\theta} + \frac{\phi_a^2}{2}\frac{\partial R_a}{\partial \theta} + \frac{\phi_b^2}{2}\frac{\partial R_b}{\partial \theta} + \frac{(\phi_a + \phi_b)^2}{2}\frac{\partial R_{mg}}{\partial \theta} = t_\theta(t) \tag{16}$$

For the magnetic coordinates ϕ_a, ϕ_b, and ϕ_c we have

$$\begin{aligned}
&(R_a + R_{mg} + R_{ta})\phi_a + R_{mg}\phi_b - R_{ta}\phi_c \\
&\qquad + \mathcal{F}_m(\phi_a + \phi_b) + \mathcal{F}_d(\phi_a - \phi_c) = 0 \\
&R_{mg}\phi_a + (R_b + R_{mg})\phi_b + \mathcal{F}_m(\phi_a + \phi_b) \\
&\qquad + \mathcal{F}_e(\phi_c + \phi_b) = 0 \\
&\mathcal{F}_c(\phi_c) + \mathcal{F}_e(\phi_c + \phi_b) + \mathcal{F}_d(\phi_a - \phi_c) - R_{ta}\phi_a + R_{ta}\phi_c \\
&\qquad + \frac{N}{R}(N\dot{\phi}_c - \dot{\lambda}_1) = Ni(t)
\end{aligned} \tag{17}$$

where R_a, R_{mg}, and R_b are understood to be functions of θ. After the permanent magnet has been charged and then stabilized, its second quadrant minor loop operation is modeled by a source of constant magnetomotive force, $-F_s$, in series with an internal magnet reluctance R_m. Thus,

$$\mathcal{F}_m(\phi_a + \phi_b) = -F_s + R_m(\phi_a + \phi_b) \tag{18}$$

The five equilibrium equations are then reduced to the static case by setting all time derivatives to zero and representing constant coil current and armature torque by I and t_o respectively:

$$\begin{aligned}
&(R_a + R_{mg} + R_{ta} + R_m)\phi_a + (R_{mg} + R_m)\phi_b \\
&\qquad - R_{ta}\phi_c + \mathcal{F}_d(\phi_a - \phi_c) = F_s \\
&(R_{mg} + R_m)\phi_a + (R_b + R_{mg} + R_m)\phi_b + \mathcal{F}_e(\phi_c + \phi_b) = F_s \\
&-R_{ta}\phi_a + R_{ta}\phi_c + \mathcal{F}_c(\phi_c) + \mathcal{F}_e(\phi_c + \phi_b) - \mathcal{F}_d(\phi_a - \phi_c) = NI
\end{aligned} \tag{19}$$

$$\frac{\phi_a^2}{2}\frac{\partial R_a}{\partial \theta} + \frac{\phi_b^2}{2}\frac{\partial R_b}{\partial \theta} + \frac{(\phi_a + \phi_b)^2}{2}\frac{\partial R_{mg}}{\partial \theta} = t_o \tag{20}$$

The result is three nonlinear algebraic magnetic equations (19) which must be solved for the three unknown fluxes. Fluxes ϕ_a and ϕ_b are then used in the mechanical eq. (20) for calculating torque developed on the armature. For convenience in handling gap functions, the torque equation is rewritten in terms of gap permeances (reciprocal of reluctance):

$$-\frac{\phi_a^2}{2P_a^2}\frac{\partial P_a}{\partial\theta}-\frac{\phi_b^2}{2P_b^2}\frac{\partial P_b}{\partial\theta}-\frac{(\phi_a+\phi_b)^2}{2P_{mg}^2}\frac{\partial P_{mg}}{\partial\theta}=t_o \tag{21}$$

APPENDIX D

Working gap functions

Since armature torque is determined from gap permeances and the partial derivatives of these permeances with respect to armature displacement, accurate mathematical expressions (gap functions) for these permeances must be derived. Gap functions which account for both fringing flux and main gap flux between nonparallel surfaces can be derived by use of the principles given by Roters.[8] By introducing the "magnetic pivot" concept,[10] these functions are modified to account for the fact that the mechanical pivot is not located in the plane of the gap magnetic surfaces.

The derivation of the permeance of gap B (P_b) will serve to illustrate the method. The region between the armature and pole B is divided into 11 simple geometric elements as shown in Fig. 18. Each element represents a region of flux paths. Total gap permeance is the sum of the individual permeances, each defined in terms of gap geometry as shown below:

$P_1(\theta)$ = permeance of main gap

$$P_1(\theta)=\begin{cases}\dfrac{\mu_o F_b}{\theta}\ln\left(\dfrac{\theta B_2+L_{mg}}{\theta B_1+L_{mg}}\right) & \text{for } \theta\neq 0\\[2ex] \mu_o F_b\left(\dfrac{B_2}{\theta B_2+L_{mg}}-\dfrac{B_1}{\theta B_1+L_{mg}}\right) & \text{for } |\theta|\ll\dfrac{L_{mg}}{B_1}\end{cases} \tag{22}$$

P_2 = permeance of one-half of a semicircular cylinder with length F_b and radius $\theta B_1 + L_{mg}$

$$P_2 = 0.52\,\mu_o F_b \tag{23}$$

$P_3(\theta)$ = permeance of a quarter annulus with axial length F_b and bounding radii of $\theta B_1 + L_{mg}$ and $\theta B_1 + L_{mg} + t$

$$P_3(\theta)=\frac{2\mu_o F_b}{\pi}\ln\left(1+\frac{t}{\theta B_1+L_{mg}}\right) \tag{24}$$

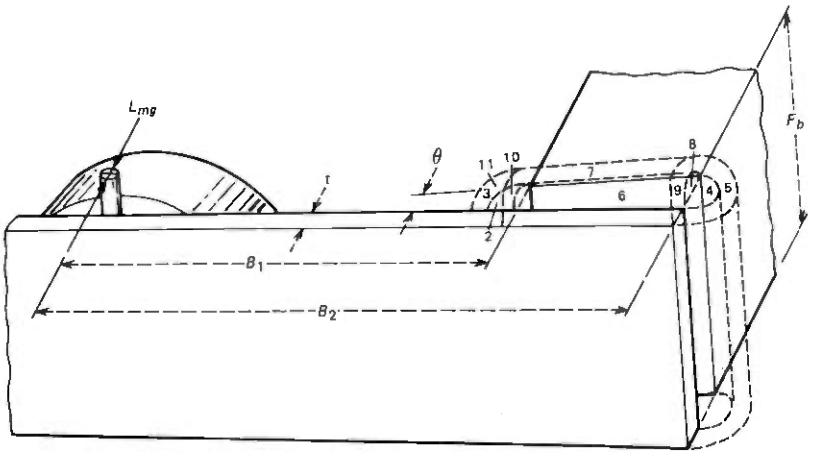


Fig. 18—Flux paths at gap B.

P_4 = permeance of a semicircular cylinder with length F_b and diameter $\theta B_2 + L_{mg}$

$$P_4 = 0.26\ \mu_o F_b \tag{25}$$

$P_5\ (\theta)$ = permeance of a half annulus with axial length F_b and bounding radii of $\frac{1}{2}(\theta B_2 + L_{mg})$ and $\frac{1}{2}(\theta B_2 + L_{mg}) + t$

$$P_5(\theta) = \frac{\mu_o F_b}{\pi} \ln\left(1 + \frac{2t}{\theta B_2 + L_{mg}}\right) \tag{26}$$

P_6 = permeance of a semicircular cylinder with length $B_2 - B_1$ and average diameter $\theta(B_2 + B_1)/2 + L_{mg}$

$$P_6 = 0.26\ \mu_o\ (B_2 - B_1) \tag{27}$$

$P_7\ (\theta)$ = permeance of a half annulus with length $B_2 - B_1$ and average bounding radii of

$$\frac{\theta}{2}\left(\frac{B_2 + B_1}{2}\right) + \frac{L_{mg}}{2}$$

and

$$\frac{\theta}{2}\left(\frac{B_2 + B_1}{2}\right) + \frac{L_{mg}}{2} + t$$

$$P_7(\theta) = \frac{\mu_o(B_2 - B_1)}{\pi} \ln\left(1 + \frac{2t}{\theta\left(\frac{B_2 + B_1}{2}\right) + L_{mg}}\right) \tag{28}$$

$P_8(\theta)$ = permeance of a spherical quadrant with diameter $\theta B_2 + L_{mg}$

$$P_8(\theta) = 0.077\ \mu_o(\theta B_2 + L_{mg}) \tag{29}$$

P_9 = permeance of a quadrant of a spherical shell with bounding radii of $\frac{1}{2}(\theta B_2 + L_{mg})$ and $\frac{1}{2}(\theta B_2 + L_{mg}) + t$

$$P_9 = \frac{\mu_o t}{4} \tag{30}$$

$P_{10}(\theta)$ = permeance of a spherical quadrant with diameter $\theta B_1 + L_{mg}$

$$P_{10}(\theta) = 0.077\ \mu_o(\theta B_1 + L_{mg}) \tag{31}$$

P_{11} = permeance of a quadrant of a spherical shell with bounding radii of $\frac{1}{2}(\theta B_1 + L_{mg})$ and $\frac{1}{2}(\theta B_1 + L_{mg}) + t$

$$P_{11} = \frac{\mu_o t}{4} \tag{32}$$

Now we can write $P_b(\theta)$, the total permeance of gap B:

$$P_b(\theta) = P_1(\theta) + P_2 + P_3(\theta) + P_4 + P_5(\theta) + 2P_6 + 2P_7(\theta) + 2P_8(\theta) + 2P_9 + 2P_{10}(\theta) + 2P_{11} \tag{33}$$

Since $P_2 = 2P_4$ and $P_9 = P_{11}$,

$$P_b(\theta) = P_1(\theta) + P_3(\theta) + 3P_4 + P_5(\theta) + 2P_6 + 2P_7(\theta) + 2P_8(\theta) + 2P_{10}(\theta) + 4P_{11} \tag{34}$$

The total permeance of gap B (34 and 22 to 32) has now been expressed in terms of ringer motor parameters. The permeance is then differentiated with respect to θ for use in the torque eq. (21). Gap A and the magnet-to-armature gap are handled in the same way.

APPENDIX E

Criterion function

The first term in the function to be minimized is the reciprocal of the torque factor squared. The rest of the terms are constraints imposed by means of penalty functions. Refer to Fig. 10 and the definitions in Section VII. Let

R_1 = 1/(torque factor)^2

$R_2 = T_a$

$R_3 = T_b$

$R_4 = T_1$

$R_5 = T_3 - T_2$

R_6 = estimate of inductance

R_7 = thickness of shim in gap A

$$R_8 = \text{roll over} = \frac{(\text{torque factor from } I_2 \text{ to } I_3)}{(\text{torque factor from } I_1 \text{ to } I_2)}$$

Then let

$$D_i = 0, R_{\min_i} \leq R_i \leq R_{\max_i}$$

$$D_i = R_i - R_{\min_i}, R_i < R_{\min_i}$$

$$D_i = R_{\max_i} - R_i, R_i > R_{\max_i}$$

The criterion function P is formed by taking the weighted sum of the ${D_i}^2$.

$$P = \sum_{i=1}^{8} A_i {D_i}^2$$

The designer controls the optimization process by judicially choosing the weighting factors A_i and the constraints $R_{\min_i}$ and $R_{\max_i}$.

REFERENCES

1. A. H. Inglis and W. L. Tuffnell, "An Improved Telephone Set," B.S.T.J., *30*, No. 2 (April 1951), pp. 239–270.
2. Henry W. Ott, "Ringing Problems on Long Subscriber Loops," Telephony, *186*, No. 25 (June 24, 1974), pp. 33–40.
3. J. R. Power, "Flux Modulated Ringer," U.S. Patent 2,716,232, August 23, 1955.
4. W. Kalin and J. R. Power, "Audible Signal For Field Telephone Sets," U.S. Patent 2,658,194, November 3, 1953.
5. R. W. Kulterman and L. F. Mattson, "Computerized Analysis of Magnetically Coupled Electromechanical Systems," IEEE Transactions on Magnetics, *MAG-5*, No. 3 (September 1969).
6. U. Rauterberg, "Calculation and Optimization of Magnet Systems by Computer," 22nd Annual National Relay Conf. at Oklahoma State University, Stillwater, Oklahoma, April 30–May 1, 1974, paper No. 7 in Conf. Proc.
7. R. M.-M. Chen, C. F. Hempstead, Y. L. Kuo, M. L. Liou, R. P. Snicer, and E. D. Walsh, "Role of Computing and Precision Measurement," B.S.T.J., *53*, No. 10 (December 1974), pp. 2249–2253.
8. H. C. Roters, *Electromagnetic Devices*, first ed., New York: Wiley, 1941, Chap. 5.
9. J. Meisel, *Principles of Electromechanical-Energy Conversion*, New York, McGraw-Hill, 1966, Chap. 3.
10. M. S. Stein and B. S. Bengtsson, "The Working Air Gap of the EMR," paper presented at the 21st Annual National Relay Conf. at Oklahoma State University, Stillwater, Oklahoma, May 1–2, 1973, paper no. 16 in Conf. Proc.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 1, January 1978
Printed in U.S.A.

The Preparation of Optical Waveguide Preforms by Plasma Deposition

By R. E. JAEGER, J. B. MACCHESNEY, and T. J. MILLER

(Manuscript submitted February 11, 1977)

Optical fiber preforms are prepared by a technique similar to the modified CVD process, except that an RF plasma is used. Optical losses as low as 6.5 dB/km at 1.06 μm were achieved. Higher reaction efficiencies resulted in deposition rates more than three times greater than the modified CVD process.

I. INTRODUCTION

The modified chemical vapor deposition (MCVD) process[1] has achieved acceptance as a means of preparing low-loss optical waveguides. A recent study of the preparation of silica-clad germania borosilicate optical fibers by this process[2] indicates that it can provide low-loss optical fibers having properties which surpass minimum requirements for wide-band communications applications. Typical fibers produced by MCVD are characterized by low loss (<4 dB/km at 1.06 μm), a reproducible graded index profile ($\alpha \sim 2.0$) and controlled core/cladding ratio, circularity and concentricity.

Although current deposition rates are competitive with other processes, higher rates are desirable. Furthermore, the efficiency, defined as the ratio of glass deposited/reactants in, is generally below 50 percent for most GeO_2 containing compositions. These considerations stimulated our investigation into using the ionized environment of a plasma to increase the rate and efficiency of the MCVD process. Initially both microwave and RF plasmas were considered. In the former, ionic species are produced which deposit as a vitreous deposit directly on the tube walls as the result of a heterogeneous reaction similar to conventional CVD.[3] This advantage, however, is purchased by the use of reduced pressures which dictates a low flux of reactants. As a result, even at 100 per cent efficiency, the deposition rates in the microwave method need be no higher than MCVD and frequently are lower.

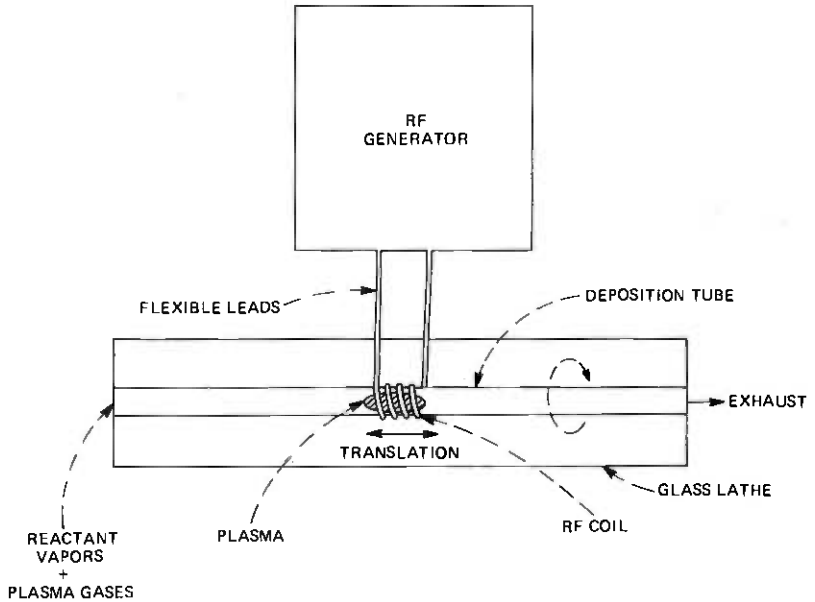


Fig. 1—Apparatus for preparing preforms.

In contrast, the use of an RF argon-oxygen plasma operating at a pressure of 1 atmosphere provides the potential for higher deposition rates as well as increased efficiency. In this type of plasma, temperatures in the vicinity of 10,000–20,000°K exist, producing an intense reaction region in which highly ionized species occur. Under these conditions, reactions between ionized species occur resulting in a deposit on the tube wall which is similar to that produced by MCVD. Significant improvements in the efficiency and deposition rate achieved by the RF plasma MCVD process are derived from the ability to produce gas phase reactions in a highly ionic environment.

II. EXPERIMENTAL

The apparatus consists of a glass working lathe positioned adjacent to an RF generator as shown in Fig. 1. An RF coil, mounted to a motorized carriage on the lathe was connected to the generator via water-cooled flexible leads. In a typical deposition, a 25 × 28 mm TO8 silica tube was positioned in the lathe and a flow of argon was started. All gas flows were left to right and were controlled with rotometers. The plasma was initiated at a frequency of 4.5 MHz by striking an arc with a graphite rod inserted into the tube in the vicinity of the coil. The coil was traversed at a constant rate over ~25 cm length and the tube was rotated as shown. Gaseous $SiCl_4$ and $GeCl_4$ were introduced to the reaction zone using an

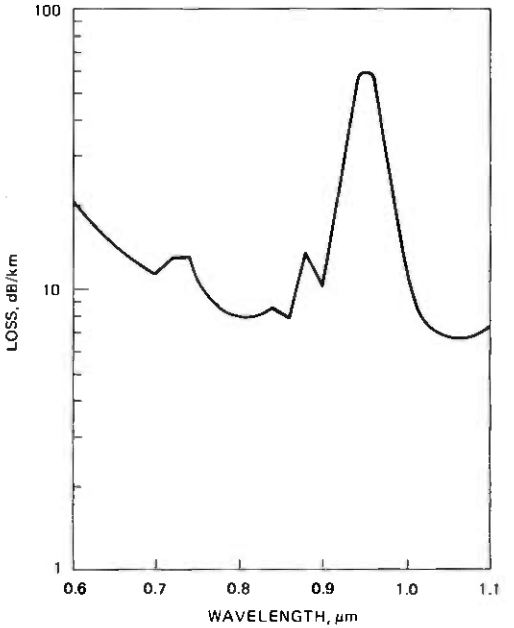


Fig. 2—Loss spectrum.

oxygen carrier which was bubbled through room temperature chloride liquids. BCl_3 gas was directly introduced from a cylinder of liquid BCl_3. $SiCl_4$ at 92 cc/min and BCl_3 at 10 cc/min were initially introduced into the gas stream to produce a borosilicate "barrier layer." Deposition was carried out for 5 minutes before $GeCl_4$ at 65 cc/min was added to produce a germanium borosilicate core. Core deposition continued for 15 minutes, at which time the RF power and all gas flows were turned off and the tube collapsed using an oxyhydrogen burner.

III. RESULTS AND DISCUSSION

It is well known that the stability of an RF plasma may be related to the diameter of the tube in which it is being sustained. We found that under our operating parameters, stable operation could be obtained at a frequency of 4.5 MHz by using 25 mm ID tubes. Fundamental considerations indicate that, in general, it is desirable that the tube radius be greater than the skin depth[4] of the plasma, which is proportional to the ½ power of the ratio of the plasma resistivity to the operating frequency. Hence, higher frequencies would be one variation that would simplify the use of smaller tubes.

Figure 2 gives the loss spectrum for a 1100 meter length of a plasma-deposited fiber with a 35 μm core and a 110 μm OD having a minimum loss of 6.5 dB/km at 1.06 μm. The peaks are due to overtones and com-

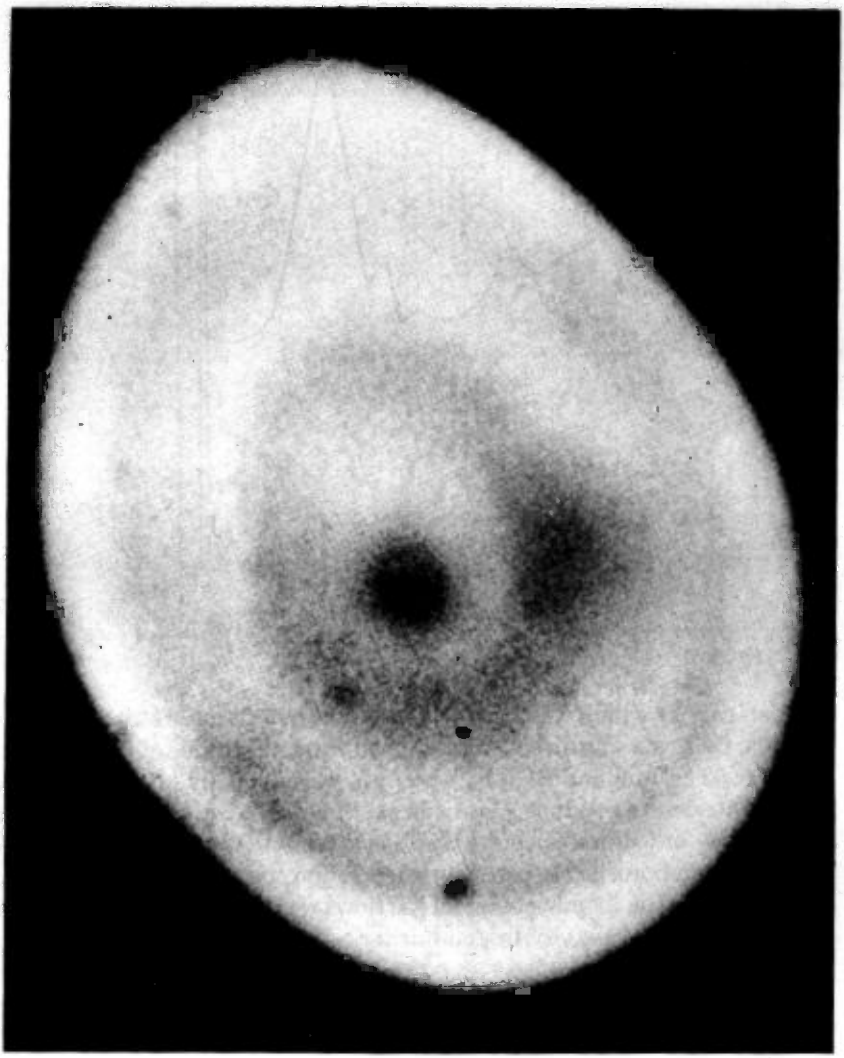


Fig. 3—Unetched microstructure.

binations of the fundamental OH absorption band and the loss at 0.95 μm indicates a water content of ~50 ppm.[5] Since typical MCVD fibers exhibit a minimum loss below 4 dB/km and contain only a few ppm water, an attempt was made to determine the extent to which the current plasma process parameters contributed to the excess water content and loss, by preparing a preform in the conventional manner using an oxyhydrogen torch as the heat source. This resulted in a fiber having the same OH content but with a minimum loss < 3 dB/km at 1.06 μm, suggesting that the excess loss at this wavelength is related to specific

process parameters while the OH concentration may be due to the general lack of humidity control in the surrounding environment.

The unetched microstructure of a core deposit for a 200 μm fiber viewed in an optical microscope, shown in Fig. 3, indicates some of the limitations encountered under the present operating conditions. The noncircularity of the core reflects the difficulty in collapsing a 25 mm tube while the rings result from compositional variations in the deposit. Examination of an etched core deposit in a scanning electron microscope revealed that the dark rings are GeO_2 deficient regions. Thus, the excess loss in this fiber may result from scattering caused by inhomogeneity of the deposit and/or compositional variations in the core. In either case, large excursions of the tube wall temperature during the deposition are the probable cause. However, other loss phenomena have not been ruled out.

In order to make very low loss preforms, it may be desirable to heat the deposit on the wall to a temperature where fining and homogenization of the glass take place with minimal vaporization of the more volatile constituents or heating the tube to temperatures where it will deform. Temperature control in the plasma deposition process was attained primarily by traversing the plasma ball along the tube at a rate sufficient to accomplish these objectives. Additional cooling of the outside of the tube to prevent deformation was provided by an air stream directed at the region surrounding the plasma. However, other important parameters that affect the wall temperature include the composition and velocity of the plasma gases and the RF power levels. For example, increasing the oxygen/argon ratio in the plasma requires increased RF power, both of which produce a high plasma temperature that is transferred to the walls. The evaluation and control of these parameters is currently underway.

The principal advantage of plasma deposition is the potential for increased deposition rates and improved overall reaction efficiency. For approximately equivalent reactant flows a rate of deposition of 0.19 g/min was obtained compared to a rate for MCVD of 0.05 g/min. This increase in the deposition rate supports the belief that the overall reaction efficiency is markedly increased for the RF plasma process. In addition, the NA of the plasma-deposited fibers were 0.28 compared to 0.24 for the conventional process, indicating increased efficiency for GeO_2 incorporation. Furthermore, reasonable extension of the results achieved to date suggest that deposition rates perhaps an order of magnitude greater than normally achieved by the MCVD process may be obtainable.

IV. CONCLUSIONS

The RF plasma CVD process is capable of producing low loss (< 10 dB/km) optical fibers at high deposition rates. A better understanding

of the effect of various operating parameters on the nature of the deposition process is necessary to exploit the full potential of this process.

ACKNOWLEDGMENTS

The authors wish to thank M. Grasso, J. Simpson, and P. D. Lazay for the optical loss measurements.

REFERENCES

1. J. B. MacChesney, P. D. O'Connor, F. V. DiMarcello, J. R. Simpson, and P. D. Lazay, Proc. 4th Int. Cong. on Glass, Kyoto, Japan (1974), pp. 6–40.
2. J. B. MacChesney, P. B. O'Connor, and H. M. Presby, Proc. IEEE, *62* (1974), p. 1280.
3. F. V. DiMarcello and J. C. Williams, Conf. Pub. No. 132, IEEE, First European Conf. on Optical Fiber Communications, (1975), p. 36.
4. J. Koenings, D. Kuppers, H. Lydtin, and H. Wilson, Proc. of 5th International Conf. of CVD (Sept. 21–26, 1975), Fulmer, England.
5. P. Geittner, D. Kuppers, and H. Lydtin, Applied Physics Letters, *28*, No. 11, p. 545 (1976).
6. D. Kuppers, J. Koenings, and H. Wilson, J. Electrochem. Soc., Solid-State Science and Technology, *123*, No. 7 (1976), p. 1079.
7. *High Temperature Technology*, Proc. of an Int. Symp. on High Temperature Technology, Asilomar, California 1963, Session I, Buttersworths, Wash. (1964), p. 525.
8. G. R. Newns and H. N. Daglish, Symp. on Electrotech. Glasses, London (1970).

Contributors to This Issue

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research on traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, and stochastic control. In 1959–60, he was visiting lecturer in mathematics at Dartmouth College. In 1971, he taught stochastic processes at SUNY Buffalo, and from 1971–72, he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of two books in his field. Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, Mind Association, IEEE.

G. Samuel Brockway, B.S., 1966, University of Miami; M.S., 1968, Georgia Institute of Technology; Ph.D., 1972, California Institute of Technology; assistant professor of mechanical engineering, Texas A&M University, 1972–75; research scientist, R & D Associates, Marina del Rey, California, 1975–76; Bell Laboratories, 1976—. Prior to joining Bell Laboratories, Mr. Brockway was engaged in theoretical studies into the propagation of shock waves in elastic materials, the nonlinear mechanical behavior of materials subject to microstructural damage and the growth of cracks in viscoelastic media. He is presently involved in the application of various areas of mechanics such as elasticity, viscoelasticity and fracture mechanics to the selection and development of plastics for use in fiber optic as well as traditional wire media.

Fan R. K. Chung, B.S., 1970, National Taiwan University; Ph.D., 1974, University of Pennsylvania; Bell Laboratories, 1974–. Mrs. Chung's current interests include combinatorics, graph theory, and the analysis of algorithms. She is presently investigating various problems in the theory of switching networks.

Leonard G. Cohen, B.E.E., 1962, City College of New York; Sc.M., 1964, and Ph.D. (Engineering), 1968, Brown University; Bell Laboratories, 1968—. At Brown University, Mr. Cohen was engaged in research

on plasma dynamics. At Bell Laboratories, he has concentrated on optical fiber transmission studies. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu; senior member, IEEE.

Francis P. Duffy, B.A., 1965 King's College; M.S., 1968, Stevens Institute of Technology; Bell Laboratories, 1965—. Mr. Duffy has been involved in conducting statistical surveys to determine telephone network performance and customer behavior characteristics. Currently he is involved in studying voice signal powers in the telephone network.

R. M. Hunt, A.E., 1962, Wentworth Institute; B.S. (physics), 1966, Kansas State University; M.S. (engineering science), 1968, Purdue University; Bell Laboratories; 1961–1963, 1966—. Mr. Hunt completed detectibility studies of acoustic ringing signals, worked on the design and development of piezoelectric and electromagnetic tone ringer transducers, and studied ringer sound power measuring techniques. Most recently, he has done computer-aided design of electromechanical ringers. Member, ASA.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., City University of New York; Ph.D. (Statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang spent the fall of 1970 visiting the Department of Mathematics of National Tsing-Hua University. He has been engaged in research in statistics, computing science, discrete mathematics, and switching networks.

Raymond E. Jaeger, Ph.D. in ceramics, Rutgers University. He is currently Director of Research and Development at Galileo Electro-Optics Corporation. Prior to this, he spent 17 years at Bell Laboratories, Murray Hill, New Jersey, as member of technical staff. While at Bell, he worked in the Metallurgical Laboratory on a variety of materials-related engineering development programs. Member, American Ceramic Society.

Ivan P. Kaminow, B.S.E.E., 1952, Union College, New York; M.S.E., 1954, University of California, Los Angeles; A.M., 1957, Ph.D., 1960, Harvard University. Hughes Aircraft, Co., Culver City, CA (1952–1954); Bell Laboratories, 1954—. Mr. Kaminow has done research on microwave antennas, ferrites, ferroelectrics, nonlinear optics, raman scattering, electro-optic devices and optical fibers. Fellow, IEEE, APS, OSA.

John B. MacChesney, B.A., 1951, Bowdoin College; Ph.D., 1959, Pennsylvania State University; Bell Laboratories, 1959—. Mr. MacChesney has worked on a variety of materials-related problems. Currently he is engaged in the development of processes for molding preforms for optical fibers.

Wanda L. Mammel, A. B. (mathematics), 1943, Winthrop College; M.Sc. (applied mathematics), 1945 Brown University; Bell Laboratories, 1956—. Ms. Mammel is engaged in finding mathematical methods for the numerical solution of a variety of problems. In particular, she has applied linear programming techniques to problems of crystal plasticity. At present she is working on problems in microwave propagation and optical waveguides.

L. M. Manhire, B.S., 1973, Lebanon Valley College; M.S. (applied math), 1976, Fairleigh Dickenson University; Bell Laboratories, 1973—. Ms. Manhire has been concerned with the 1973 Bell System customer loop survey. She is presently engaged in software development for nonurban studies including long route design and pair gain applications.

Robert A. Mercer, B.S. (physics), 1964, Carnegie-Mellon University; Ph.D. (Physics), 1969, Johns Hopkins University; Bell Laboratories, 1973—. Mr. Mercer has done research in experimental high-energy physics as a research associate at Johns Hopkins and as an assistant professor at Indiana University from 1970 to 1973. He joined Bell Laboratories as a member of the Network Performance Characterization Department, and currently is supervising a group involved in modeling and analysis of Bell System network performance. Member, American Physical Society, Sigma Xi.

Calvin M. Miller, B.S.E.E., 1963, North Carolina State University at Raleigh; M.S.E., 1966, Akron University; Goodyear Aerospace Corporation, 1963–1966; Martin Marietta Company, 1966–1967; Bell Laboratories, 1967—. Prior to joining Bell Laboratories, Mr. Miller designed electronic and optical components of side-looking radar processor equipment and control systems for reentry vehicles and aircraft flying simulators. At Bell Laboratories, Mr. Miller developed equipment and methods for transmission line characterization. His present interests are in the area of fiber optics as a practical transmission medium. He is

supervisor of an exploratory optical fiber splicing group. Member, OSA.

Thomas J. Miller, B.S., 1968, and M.S., 1975 (ceramic science), Rutgers University; Bell Laboratories, 1968—. Mr. Miller has worked on the non-conventional processing of ceramic powders. He is presently concerned with the fabrication of high-strength optical fibers and the development of preform preparation processes.

F. W. Mounts, E.E., 1953, M.S., 1956, University of Cincinnati; Bell Telephone Laboratories, 1956–. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, Eta Kappa Nu; Senior Member, IEEE.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969 and Ph.D. (E.E), 1970, Rice University; Optimal Data Corporation, Huntsville, Alabama, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on various aspects of signal processing. Member, Tau Beta Pi, Sigma Xi.

J. W. Nippert, B.S. (engineering science), 1970, Purdue University; M.S. (engineering science), 1971, Purdue University; Bell Laboratories, 1970—. Mr. Nippert authored circuit analysis programs, developed circuit models for telephone components, did computor-aided design of electromechanical ringers, and wrote the software for a microprocessor-based small PBX-like system. Currently, he is developing new features for the SPC network. Member, Tau Beta Pi, Sigma Gamma Tau, Omicron Delta Kappa. Professional Engineer, Indiana, 1977,

Birendra Prasada, B.S., 1953, M.S., 1955, Banaras University; Ph.D., 1960, University of London; Central Electronics Engineering Research Institute, Pilani, India, and Defense Science Laboratory, Delhi, India, 1961–1963; Massachusetts Institute of Technology, 1965–1966; Indian Institute of Technology, 1968–1972; Bell Laboratories, 1963–1965, 1973–1976; Bell Northern Research, 1976—. Mr. Prasada's main research and teaching interests are in the areas of visual communications, systems engineering, systems design, and human communication. He has worked

as an industrial consultant in India and the United States. Member, 1963, Senior Member 1976, IEEE.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests include macroeconomics, the theory of social groups, and the economic theory of large corporations. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, and Tau Beta Pi.

Charles M. Schroeder, A.S. (electromechanical), 1969; Western Electric Engineering Research Center, 1967—. Mr. Schroeder was initially involved with the design and retrofitting of optical packages for thin film projection and near contact printing systems. In fiber optics, his primary work was in fiber ribbon stripping and multiple fiber splicing, and is now in fiber coatings.

George M. Yanizeski, B.S., 1964, M.S., 1965, Ph.D., 1968 (civil engineering), Carnegie-Mellon University; instructor, Carnegie-Mellon University, 1967–68; member of technical staff, Bellcomm, Inc., 1968–72; Bell Laboratories, 1972—. At Bellcomm, Mr. Yanizeski participated in the analysis and engineering of thermal control systems on NASA's Skylab orbiting space station. At Bell Laboratories, he has been involved in a series of theoretical and experimental studies characterizing and quantifying the mechanical performance of telephone cable, and he has been involved in several projects to develop new sheath designs employing new polymeric materials.

ERRATA

In the Bell System Technical Journal, *56*, No. 8 (October 1977), Figs. 2 and 3 on pp. 1454 and 1455 should appear as follows.

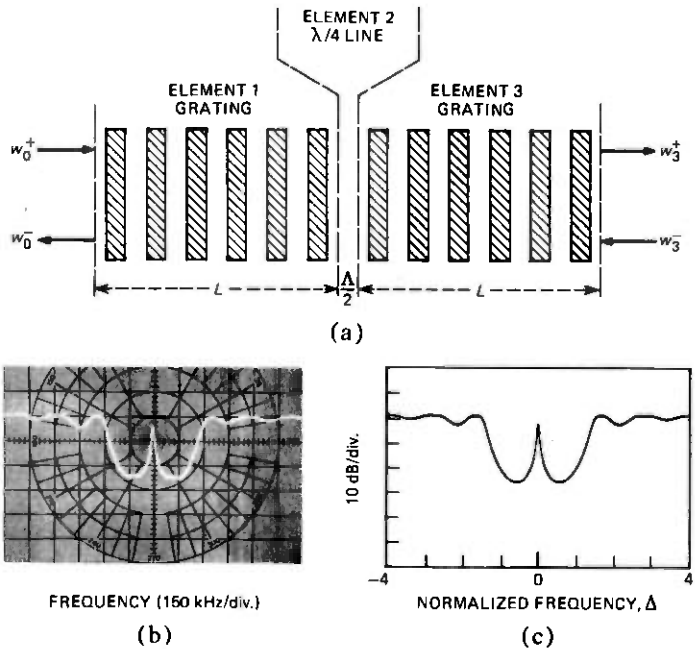


Fig. 2—(a) Diagram of a grating resonator in the external transmission configuration. (b) The transmission spectrum at ~145.5 MHz for a resonator on YZ-$LiNbO_3$ using Ti-diffused gratings with $\Lambda = 12\ \mu m$ and $L = 6.48$ mm. (c) The calculated transmission spectrum for the device in (b) using $\kappa = 3.74\ cm^{-1}$ and $\alpha = 0.036\ cm^{-1}$.

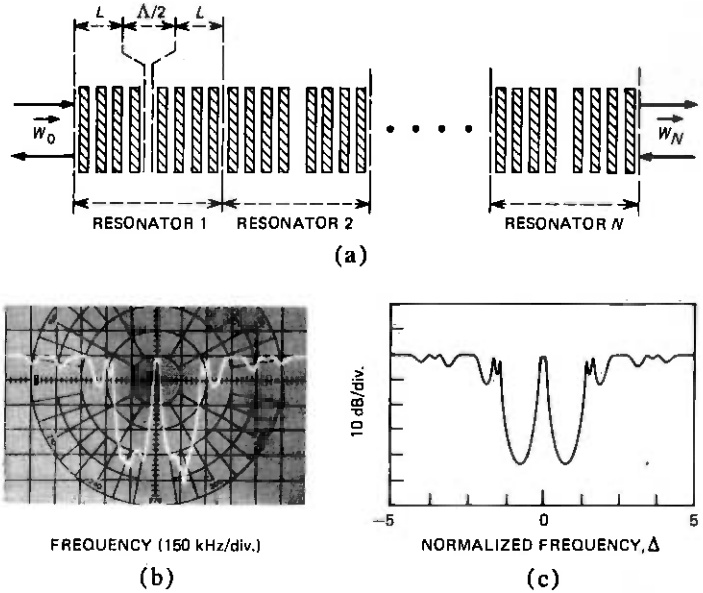


Fig.3—(a) Diagram of cascade of N identical resonators. (b) The transmission spectrum at 145.5 MHz of a cascade of three identical resonators with $\Lambda = 12\ \mu m$ and $L = 3.84$ mm. (c) The calculated transmission spectrum for the device in (b) using $\kappa = 3.55\ cm^{-1}$ and $\alpha = 0.027\ cm^{-1}$.

Equation (15b) on p. 1509 should be:

$$\sigma_{mn} \cong \tfrac{1}{2}\omega^2 C_{mn}^2 S_n - \frac{2}{M} E_f \sum_{m=1}^{M/2} \tfrac{1}{2}\omega^2 C_{mn}^2 \qquad (15b)$$

Papers by Bell Laboratories Authors

CHEMISTRY

Band Model for the Electronic Structure of Expanded Liquid Mercury. L. F. Mattheiss and W. W. Warren, Jr., Phys Rev, *16* (July 1977), pp. 624–638.

ELECTRICAL AND ELECTRONIC ENGINEERING

Rain-Scatter Interference on an Earth-Space Path. T. S. Chu, IEEE Trans. Ant. Propag., *25*, No. 2 (March 1977), pp. 287–288.

MATERIALS SCIENCE

Glass Transition Temperatures on Regularly Alternating Acrylonitrile-Vinyl Acetate Copolymers. A. E. Tonelli, Macromolecules, *10* (1977), pp. 716–717.

Composition and Stress State of Thin Films Deposited by Ion Beam Sputtering. R. N. Castellano, M. R. Notis and G. W. Simmons Vacuum, *27*, No. 3 (June 1977), pp. 109–117.

Concerning the Semimetallic Characters of TiS_2 and $TlSe_2$. J. A. Wilson, Solid State Commun., *22* (1977), pp. 551–553.

Computer Model of Metallic Spin Glasses. L. R. Walker and R. E. Walstedt, Phys. Rev. Lett., *38*, No. 9 (February 1977), pp. 514–518.

Diffusion Kinetics of Au Through Pt Films About 2000 and 6000 Å Thick Studied With Auger Spectroscopy. C. C. Chang and G. Quintana, Thin Solid Films, *31* (1976), pp. 265–273.

Si Depth Profile and Contaminants in Si-doped Al Film. C. C. Chang, T. T. Sheng, D. V. Speeney, and D. B. Fraser, J. Appl. Phys., *47* (May 1976), pp. 1790–1794.

PHYSICS

Dielectric Function Measurements on Ion-Implanted Metallic Systems. E. N. Kaufmann, D. E. Aspnes, J. W. Rodgers, and A. A Studna, J. Opt. Soc. Amer., *67* (August 1977), pp. 1099–1101.

A New Spectroscopic Technique for Imaging the Spatial Distribution of Nonradiative Defects in a Scanning Transmission Electron Microscope. P. M. Petroff and D. V. Lang, Appl. Phys. Lett., *31* (July 1977), pp. 60–62.