Foreword

This issue marks the 40th anniversary of the Bell System Technical Journal, which has been published continuously since July, 1922. Through these 40 years the B.S.T.J. has had one basic purpose: to be a journal of definitive technical papers that help to record the progress of Bell System communications. The B.S.T.J. is thus an important factor in carrying out the Bell System policy of prompt publication of new research, development and systems engineering knowledge.

Although the pace of change has increased in recent decades, B.S.T.J. papers from the beginning have dealt with many themes that are still of fundamental importance to communications research and technology. Among the important early contributions, for example, were articles by Harvey Fletcher on the nature of speech, by R. V. L. Hartley and T. C. Fry on binaural hearing, by George A. Campbell on wave filters, and by E. C. Molina on traffic theory. Early articles of significance to telephone transmission included those on carrier transmission by Hartley, long cable circuits by A. B. Clark, and radio communications by Lloyd Espenschied. K. K. Darrow published the first of a notable series of articles reporting advances in physics, and among the first articles in the area of materials research, G. W. Elmen and H. D. Arnold reported on Permalloy.

These early contributions helped to establish standards of technical excellence that have been a challenge and inspiration to subsequent authors and to those who have served as editors and advisors. As new knowledge about communications advances even farther in scope and depth, the B.S.T.J. will continue to serve both the Bell System and the entire scientific and engineering community.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLI JULY 1962 NUMBER 4

Automatic Trouble Diagnosis of Complex Logic Circuits

by S. H. TSIANG and W. ULRICH

(Manuscript received November 2, 1961)

This paper deals with the problem of maintaining the most complex portion of an experimental electronic telephone switching system, the central control. New and more effective automatic trouble detection and diagnostic techniques were used. In order to utilize these techniques effectively, a maintenance dictionary, i.e., a table relating trouble indications with corresponding faulty plug-in package, had to be produced. The system itself was utilized to create this dictionary. Over 50,000 known faults were purposely introduced into the central control to be diagnosed by its diagnostic program. The corresponding test results were then recorded via a high speed output. Finally, these test data were sorted and printed in dictionary form by a computer.

I. INTRODUCTION

In November, 1960 Bell Laboratories started its field trial of an experimental electronic telephone switching system in the town of Morris, Illinois. This system (extensively described elsewhere)[1] was one of the first attempts to introduce electronics on a large scale into telephone switching and as such, brought us face-to-face with a new class of problems, especially in the field of maintenance of centralized telephone equipment.

The problems of maintaining an electronic telephone switching system are formidable, but as will be seen presently, the tools naturally available for this maintenance are powerful. This paper will deal with the problem of maintaining the most complex portion of the experimental electronic telephone switching system, the central control.

II. RELIABILITY AND MAINTENANCE OBJECTIVES

The character of a commercial telephone system as a whole imposes unusual maintenance objectives for its component parts. The vital role of a central office demands that it have an extremely low downtime. At the same time, since the telephone system is so widespread and cannot be concentrated in a few key locations, another objective is that it be maintainable by telephone system craftsmen.

These are extremely difficult requirements. In order to maintain a sufficiently low downtime with devices currently available, it is necessary to provide some redundancy in the equipment so that single troubles do not cause the entire system to fail. The simplest form of redundancy, and in the present state of the technology probably the least expensive, is the simple duplication of all common equipment in the system. Thus, where only one memory store is required for running the system, two are provided; where only one control unit is required for running the system, two are provided, etc.

III. DESCRIPTION OF THE MORRIS ELECTRONIC SWITCHING SYSTEM

Fig. 1 is a block diagram of the system. The customer's line comes into the office and has an appearance on the network and on the scan-

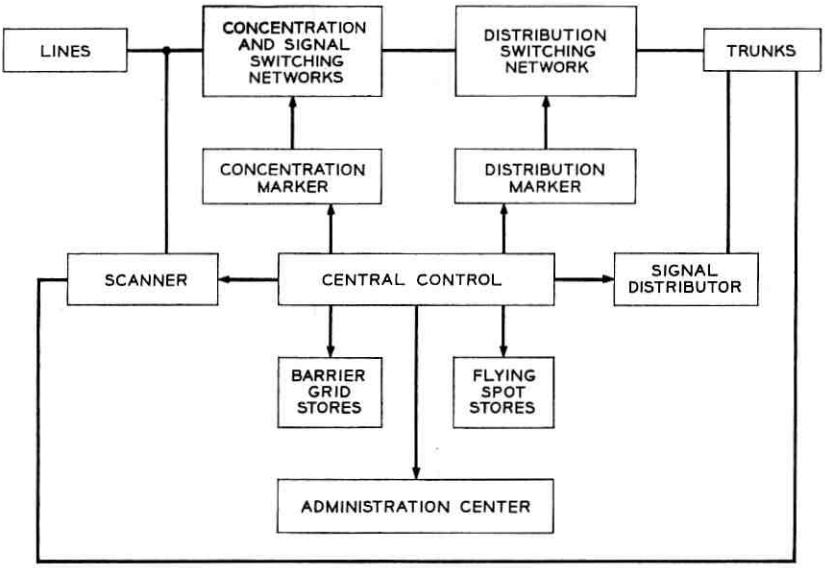


FIG. 1—General system block diagram.

ner. The appearance on the network permits him to be connected to another customer, to dial tone, or to a ringing signal. The appearance on the scanner permits the system to detect the status of his line. When a customer lifts his telephone handset off the cradle, a loop is closed to the central office, changing the voltage at a point that is detectable by the line scanner. Similarly, when dialing, this loop is opened for brief intervals; for example, if a 9 is dialed the circuit is opened nine times.

The flying spot store is used to store the program of the system and the individual translation information associated with each customer. The barrier grid store is used to store information as to the status of the lines in the office (busy, idle, or dialing) and to assemble dialing information during the time that a customer is actually dialing. The signal distributor is used to operate the test relays in the office and to signal to distant offices. It is also used to switch between working and standby units. The central control must coordinate the actions of all these units, i.e., take the output of the line scanner and the barrier grid store, act upon these outputs according to the instruction given by the flying spot store, and use these results either to set up a connection in the network or write further data in the barrier grid store; it must also request the next instruction from the flying spot store.

These facilities can be used for testing the system in a relatively sophisticated manner. Available for the purposes of setting up a call is a rather complex system for processing data. Testing is also actually a data-processing action. Thus, when a faulty unit is being tested, we get the instructions for testing from the flying spot store and we get the test results either from match circuits between units performing supposedly identical functions or from the line scanner which has test probes into various circuits. Test results are then assembled in the barrier grid store and are eventually typed out using the teletypewriter (which is also under the control of the signal distributor).

As mentioned earlier, all important and common equipment is duplicated. This leads to a two-fold advantage for testing. First, a rather sophisticated data-processing machine is always available which can automatically apply tests, and interpret and report test results. Second, an identical unit, presumably in good working order, is always available; the output of this unit may be compared with that of the circuit being tested. A match of outputs indicates a successful performance of a test, while a mismatch indicates a test failure.

The program for controlling tests is stored in the flying spot store. Storage space in this store is both expensive and limited. Therefore, it

is not possible to store a program sufficiently flexible so that it can print out the identity of a defective package. Instead, the program prints out test results and these test results are then compared with test results that are anticipated by the designer and are placed in a form referred to as the "maintenance dictionary."

IV. CENTRAL CONTROL MAINTENANCE

The central control is the basic control unit of a real-time, special-purpose, program-controlled data-processing machine. It controls the flow of information to the stores, markers, scanner and signal distributor. It is duplicated. There are over 8000 circuit packages (6500 transistors, 45,500 diodes) in both central controls. The two central controls normally operate in parallel, executing the same instruction and performing the same operations, even though only one of the controls, the "active" central control, is used to control such system output circuits as the markers and signal distributor. Certain key outputs of the two central controls are matched; trouble in either central control leads to a mismatch. Any mismatch initiates a special fault-check program, which is used to decide which, if any, of the central controls has the trouble. The program causes central control to make a number of decisions; if the active central control makes any incorrect decisions, it switches itself out of service. If a decision made by the active central control is correct but a mismatch occurs, it indicates a trouble in the standby central control. In case the active central control is defective to such a degree that it cannot execute the fault-check program, a time-out circuit will automatically switch this control out of service.

After the fault-checking program has been completed, a diagnostic program is started. Whenever the system finds a spare time period of one millisecond when no telephone operations are required, tests are performed on the suspected central control. The results of these tests are then typed out using the system teletypewriter. These results must then be interpreted with the aid of a maintenance dictionary.

V. THE CENTRAL CONTROL MAINTENANCE DICTIONARY PROBLEM

For most of the units of the system, a maintenance dictionary can be specified with a reasonable amount of effort by the designer of the test programs because the units are functionally comparatively simple. For example, the characteristics of different types of faults which may occur in the scanner lead to a relatively small number of simple types of patterns which can be readily examined. However, central control is much more complex both in its functions and in its circuitry. The type of

symmetry which is characteristic of scanner, signal distributor, network, and even the stores is totally absent in the case of central control. Even the number of the tests which has been selected (about 900) attests to this difference of complexity. The work of preparing a dictionary by hand would have been formidable and the resulting dictionary would have been very incomplete. As a result, the automatic means described in this paper were used to produce a dictionary that was both as complete as the diagnostic tests would permit and which required only a reasonable amount of development effort.

The scheme was to introduce about 50,000 faults into central control and get the test results associated with each fault. These test results could then be sorted and finally printed as the desired dictionary.

This scheme had a number of advantages over other techniques. First of all, while considerable effort would be required to design the basic mechanism and circuits for carrying out this scheme, once this effort had been expended, the number of troubles which could be analyzed could be made very large without enormous additional expenditure of effort. Secondly, no errors of analysis could creep into the system. Thirdly, because computer analysis would be required for the final production of the dictionary anyway, it would be possible with little additional effort to create a dictionary printout format that would closely resemble the format of the test results which are normally obtained by the system. Finally, the actual process of creating the dictionary could be deferred to a relatively late date so that the dictionary would be based on the most up-to-date version of central control and the most up-to-date version of the diagnostic program.

The basic scheme for deriving the dictionary data is discussed below. First, the system must be switched to a special dictionary mode of operation, dropping all telephone work. (Since this is done in the laboratory, not in a working office, this is not at all serious.) Then, information is fed to the system concerning the identity of the package whose possible faults are to be simulated. Next, the faults are simulated. After each fault, the system diagnostic program is started; a punched paper tape output is used to record the identity of the package, the number of the fault, and a complete report of the tests that failed when that fault was simulated. Such an output record is generated for each fault that is simulated. These records are then sorted and printed in suitable form by a computer.

VI. GENERAL DESCRIPTION OF DICTIONARY PREPARATION

A number of special pieces of equipment were designed for the preparation of the dictionary. Two fault-simulation units were used to simu-

late essentially all the packages and their faults. A test control unit was used to sequence the faults automatically in a given package and coordinate all special recording equipment with the system.

Information identifying the package whose faults were to be simulated was stored on a regular 5-channel teletype tape and was automatically fed into the system at the proper time by a conventional 100 words per minute reader when the corresponding circuit card was to be tested.

A high-speed 1000 words per minute (100 characters per second) TELETYPE tape punch was used to record the test results. The output data for each fault consists of the package identity and its location, the fault number, and the test results.

Fig. 2 shows the functional block diagram and peripheral equipment involved in collecting the test data.

The test control unit controls the over-all operation of the fault simulation and data gathering. The "normal-test" switch on the test control unit is first operated to the test position. This requests the system to go into the dictionary mode. When the system has reached a convenient point in its program, it stops the central control clock and turns over the control to the test control unit. The system cannot start again until a signal is received from the test control unit.

When the "system-off" lamp on the test control panel is turned on, the package on which faults are to be simulated can be pulled out and replaced by the fault-simulation unit. Then the "automatic test" push-button on the test control unit is operated. This tells the system to start an automatic test.

As soon as the automatic test signal is received, the system requests the package information tape reader to read in one package identity. Since the sequence of packages to be tested is the same as that listed on the package information tape, the identity will be the package currently under test. This information is stored by the system and later affixed to the corresponding diagnostic test results.

When the package identity has been stored, the system commands the test control unit to switch in the first fault and immediately initiates a complete diagnosis. Upon completion of the diagnostic tests, the system delivers the final test results to the high-speed tape punch. The system then requests the next fault to be switched in and another round of diagnostic tests and data punching begins.

The same cycle of operation is repeated until the fault-simulation unit informs the test control unit that the last fault on this package has been tested, in which case the test control unit puts the fault-simulation unit in a no-fault condition and asks the system to carry out the same

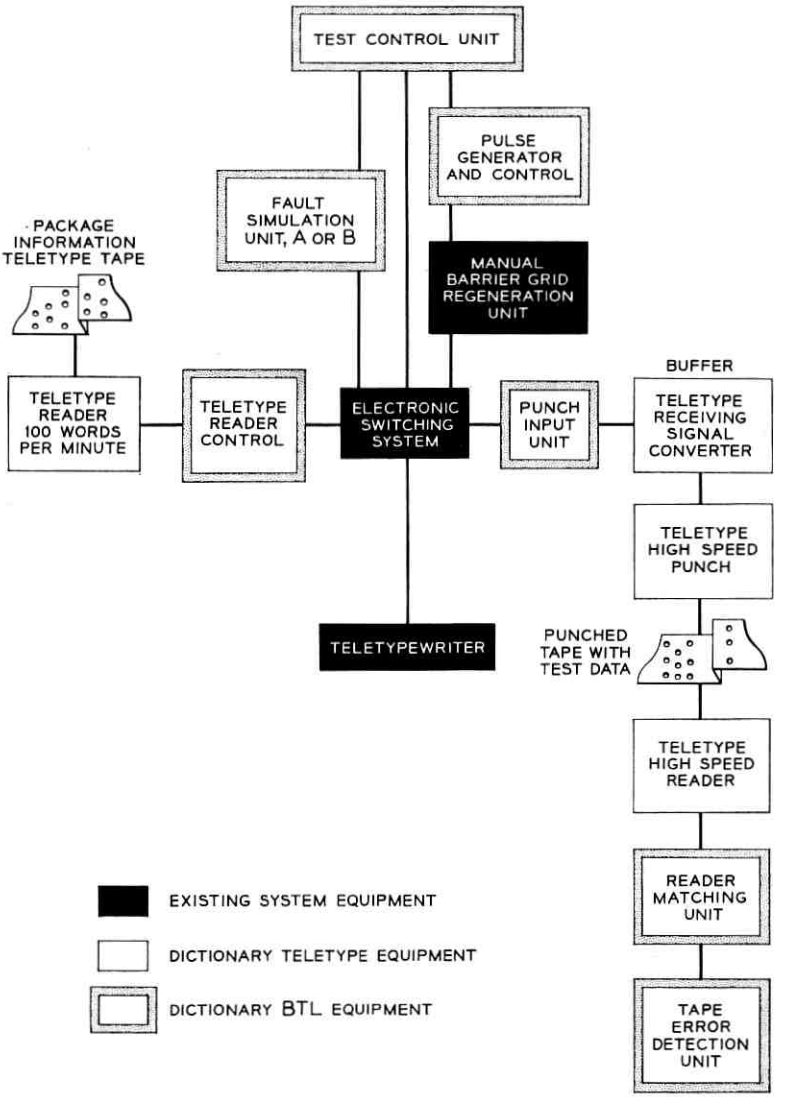


FIG. 2—Dictionary data acquisition: functional block diagram.

diagnostic routine. The test result output channel is now switched from the high-speed tape punch to the existing system teletypewriter, so that the test results will be easily readable. Normally, it is expected that there will be no test failures at the no-fault condition. Therefore, the teletypewriter will print out only the ALL TESTS PASSED message.

The main reason for making such a test is to check the correct functioning of the fault-simulation unit, i.e., that the substitution of the fault-simulation unit does not itself introduce a fault. The test also provides an auxiliary means to monitor the proper operation of the system.

When diagnosis for the no-fault condition is completed, the system performs the routine tests on all major system units; this enables the system to detect any trouble. After the system has succeeded in going through all the routine tests, it again turns itself off and lights the "end of automatic test" lamp on the test control unit panel. This completes the fault simulation for one package. The entire operation from the first fault to the last fault, then to the no-fault state, is done automatically following a single operation of the "automatic test" pushbutton.

During the time that the dictionary data was being gathered, a modified system program was used. The system did not spend a large amount of time performing unnecessary telephone work while it was actually creating the central control dictionary.

A high-speed 1000 words per minute TELETYPE reader was employed to read the tape as soon as it was punched. On the paper tape seven channels were used, six for the test data and one for the lateral parity bit generated by the system. The output of the reader was fed to a tape error detection unit which checks the parity of each character and the block length of the test records. The number of characters in the test record for each fault should be the same.

Before the data could be sorted, they were first converted from the punched paper tape to magnetic tape. A computer was then used to sort the test data. The actual dictionary was printed directly by a tape controlled printer.

In the dictionary project, work was divided into several major parts. A brief description of each part is given in the following sections.

VII. FAULT SIMULATION

The two central controls in the system are identical. Normally, one serves as the active unit and the other one as the standby or vice versa. The active unit always diagnoses the one in trouble. Therefore it is necessary to insert faults in only one of the central controls. The total number of circuit packages in which faults have to be simulated is about 4000.

About 49 different types of packages are used in central control.

All the faults simulated are of the catastrophic type. Faulty diodes

are simulated as being either shorted or open, resistors as open and transistors are simulated as either permanently on or off.

Only single troubles are simulated. It would be totally impractical to simulate multiple faults. Furthermore, routine tests (at 100 millisecond and 1 second intervals) and matching between two central controls are performed often enough so that it is reasonable to assume that a single trouble will be detected and diagnosed before another fault develops.

Obviously it would be equally impractical to simulate all marginal conditions. It is hoped that a majority of the marginal conditions, if they result in trouble at all, will give the same characteristic result as a corresponding catastrophic fault. For example, if the reverse impedance of a diode is too low, a gate may behave in the same manner as if the diode were shorted.

The number of faults per package varies from 2 to 30. Over 50,000 faults were simulated in order to create the dictionary.

In the packages containing only diodes, troubles are simulated for each gate; in the other type of packages, mainly those containing transistors, only faulty output conditions are simulated. The results of the diagnostic tests are based on the output of a package rather than the individual component contained therein.

An example of fault simulation for a 2-input OR gate is given in Fig. 3. Each diode is shorted and opened by the operation of a different relay. Only an open resistor is simulated. If a shorted resistor were simulated it might cause damage to other components in the system. Furthermore, shorted resistors are relatively rare trouble conditions. The fault-simulation unit for the AND gate is similar, except that no resistor is involved.

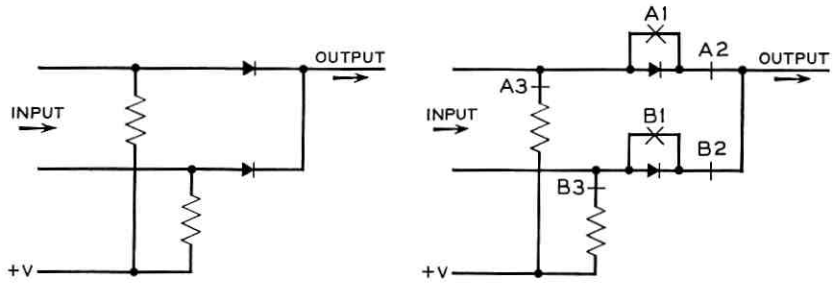
The transistor packages are simulated differently. For example, four possible faulty output conditions are simulated in a transistor flip-flop circuit.

1. Set permanently high, reset permanently low.
2. Reset permanently high, set permanently low.
3. Both high.
4. Both low.

All faults are controlled by the operation of different relays. Fig. 4 shows how this is done. The same simulation technique is used for all other types of transistor and miscellaneous circuit packages.

VIII. MODE OF OPERATION

When test data was being collected in the preparation of the dictionary, the system operated in two basic modes: normal and testing.



(a) A NORMAL TWO-INPUT OR GATE

(b) TWO-INPUT OR GATE FAULT SIMULATION CIRCUITS

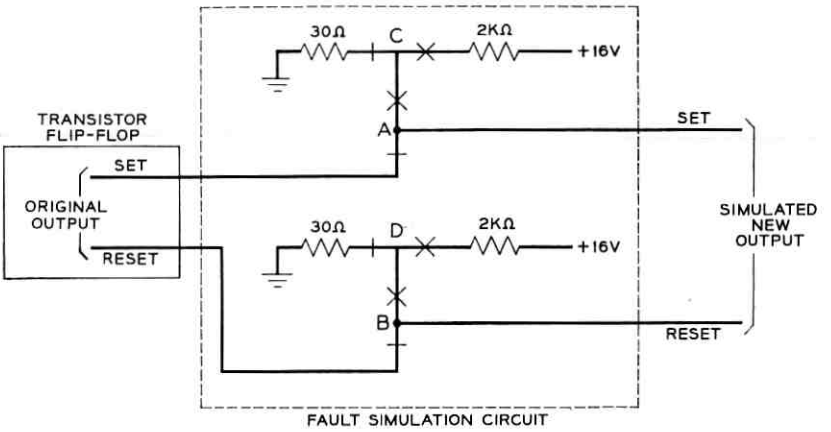
FAULT NUMBER		RELAY CONDITION	
		NONOPERATED	OPERATED
1. SHORTED DIODE	FIRST INPUT	ALL OTHERS	A1
2. OPENED DIODE	FIRST INPUT	ALL OTHERS	A2
3. OPENED RESISTOR	FIRST INPUT	ALL OTHERS	A3
4. SHORTED DIODE	SECOND INPUT	ALL OTHERS	B1
5. OPENED DIODE	SECOND INPUT	ALL OTHERS	B2
6. OPENED RESISTOR	SECOND INPUT	ALL OTHERS	B3
NO FAULT CONDITION		ALL	–

FIG. 3—Fault simulation on OR gate.

In the normal mode the dictionary test equipment is effectively disconnected from the system and the system is allowed to operate in its normal condition. If a trouble other than the one intentionally introduced is detected in the system during the course of a test, the system is placed in its normal mode so that any faults may be diagnosed and corrected.

In the test mode, the following different tests can be initiated by operation of appropriate push buttons:

1. Automatic test. This has already been described under the general description.
2. Manual test. The system can be asked to conduct diagnosis on any desired fault including the no-fault case. The results of a manual test are recorded on the system teletypewriter. Fault simulation is accomplished by means of the "fault advance" pushbutton on the test control unit. The manual test, made from time to time, is used to check the special dictionary programs and equipment. The manual test data are also used to spot-check the output data on the paper tape.



FAULTS	RELAY CONDITION	
	NONOPERATED	OPERATED
1. SET HIGH, RESET LOW	D	A,B,C
2. SET LOW, RESET HIGH	C	A,B,D
3. BOTH HIGH	–	A,B,C,D
4. BOTH LOW	C,D	A,B
NO FAULT CONDITION	A,B (C,D DON'T CARE)	

FIG. 4—Fault simulation on transistor flip-flop.

3. Skip package identity. This is a special feature provided so that a package identification block stored on the package information tape can be bypassed. When the "skip-package" button is operated, the system controls the tape reader and advances the package tape to the next block. This provision is made so that errors made during the preparation of the information tape can be corrected.

IX. SYSTEM CIRCUIT REQUIREMENTS

All communications between external equipment and the system are accomplished using the scanner and signal distributor. The scanner provides central control with an information access to lines, trunks, and test points. Signal-distributor outputs provide access from the central control to one of a large number of outputs. Very minor modifications in the system were needed to prepare the system to create the dictionary; the modifications consisted of adding a few temporary cables and connectors. The only wiring required was the jumper wiring from

the external equipment to those scanner and signal-distributor points used for the dictionary project. Altogether, 12 scanner points and 23 signal distributor points were used.

X. PACKAGE INFORMATION TAPE

The package information tape records the type and location of each package to be tested. This information is read into the system via a conventional 100-speed (10 characters per second) tape reader; the identity is read once for each package to be tested. The system controls the tape reader under the command of the test control unit.

Only the pertinent codes associated with the package identity are stored. The system rejects all functional codes such as space, carriage return, etc. This package information is later attached with its associated diagnostic test results in the final output for each fault.

Eleven characters are assigned for the identification of each package. The first five specify the location, the next six indicate the type of package.

The package information was first tabulated at random manually from an apparatus designation chart. IBM cards, one for each package, were prepared from the tabulated list, and were manually checked. After the cards were sorted according to the predetermined order in which the packages were to be tested, a printed list was prepared. This list was used as the master package test schedule during data gathering. The punched cards were then converted into a fully perforated, 5-channel paper tape in TELETYPE code. From this tape, a final printed, chadless tape was prepared. A printed tape was used so that the operator would be able to read it. Thus, a check could be made before a package identity was read into the system. A printed tape so obtained would go through a final check.

XI. SYSTEM PROGRAM MODIFICATIONS

Modification of existing programs, mainly the central control diagnostic programs, and the flow charting and coding of a number of new program segments were required. About 1000 new program words were added as a result of the changes and additions made. Modifications were necessary in order to have the system operate in different modes and perform different tasks.

During the period of dictionary preparation, the system had two additional teletypewriter communication channels: one input channel from the conventional 100-speed tape reader, for reading the package information into the system, and one output channel to the high-speed

tape punch for test data recording. In addition, the existing teletypewriter output was retained for conveying auxiliary information to the test team.

Most of the new program segments were for data handling, coordination of different test modes, control and reading of the package information tape, recording of test data on the high-speed punch, etc.

XII. PUNCHED PAPER TAPE TO MAGNETIC TAPE CONVERSION

The sorting of the test data was done on an IBM-704 computer. The test data recorded on paper tape in binary code had to be converted into magnetic tape, compatible with the 704. Conversion was made on an IBM-9200 machine. Information was processed at the rate of 500 characters per second.

The magnetic tape recording was an identical image of the paper tape. The code conversion was somewhat different from conventional conversions. In the 7-channel paper tape, the test data utilizes all 64 combinations of the first 6 channels, and the 7th channel is used for an odd parity bit for each character.

XIII. DATA RECORDING AND PROCESSING

13.1 *Data Recording*

Over 900 central control diagnostic tests are grouped into eight phases. Each phase diagnoses faults of certain parts of the central control.

During the normal diagnostic operations, the test results are recorded at the end of each phase. The printout consists of two parts. The first part is the system component identification, i.e., whether it is central control 0 or central control 1, together with the phase number (A, B, $\cdots$ or H). The second part is the test results. Only the numbers of those tests that fail are printed out. Each has a 3-digit octal number. The test results, therefore, are variable in length, depending on how many tests have failed.

For dictionary preparation, a binary coding system was adopted for recording test data. Every test in the diagnostic program was represented by one bit on the paper tape. This was also true for the magnetic tape, whose recording was identical to the paper tape except with higher longitudinal density: 200 versus 10 characters per inch. A "1" or "hole" means that the corresponding diagnostic test has failed and a "0" or "no hole" indicates that the test passed. Each character consists of 7 bits, 6 for registering six different test results, and 1 for parity. The

basic advantage of using this binary coding system is that it makes the sorting process easier. The number of characters in the test record for each fault is the same regardless of the number of test failures.

13.2 *Data Processing*

The data processing for the dictionary was done by an IBM-704 computer and the actual printing of the dictionary by a high-speed tape-controlled printer.

The 704 program is quite complex and involved. It can be divided into three major parts: phase sorting, test sorting, and data printing. The entire program is about 2500 words.

In the central control dictionary, all diagnostic test results were arranged in an orderly manner so that, given a certain sequence of test failures, the maintenance man could easily look for the same sequence in the dictionary. Associated with each sequence, one or several package identities and locations are listed. When several appear, failure of any one of them could result in such a test failure sequence.

13.2.1 *Phase Sorting*

All test records were sorted first according to their phase information. The test record for each fault contains three parts:

1. The package identification and fault number.
2. The individual test results.
3. Phase information, which is represented by 8 bits, one for each phase. If all tests have passed in a phase, the corresponding bit will be 0; if one or more tests have failed, the bit will be 1.

The purpose of this phase sorting is to arrange all records according to the alphabetical order given by the phase information. For example, consider a record α with phase information 00101100 and a record β 11000100. The 8 bits represent phases A, B, C, D, E, F, G, H. Omitting all zeros (that is, those phases which do not have any test failures), phase information for record α becomes CEF and record β ABF. After the phase sorting, record β will be placed in front of record α, that is, ABF in front of CEF, as are the words in a regular dictionary. Sorting by phase involves only part 3 of the test record.

13.2.2 *Test Sorting*

After the test records were sorted by phase information, they were subsorted in accordance with test failures. This was necessary because, with 255 possible phase combinations and over 50,000 different records, many records have the same phase combinations. The test sort takes all the records with identical phase information and further arranges

them in a numerical order according to the binary test results which are given in part 2 of each test record. The actual test results consist of 916 bits without counting the added dummy bits. Dummy bits were added to make up a complete 704 machine word (36 bits) required by the computer. Each bit corresponds to a particular test. A "1" or "hole" means that the test which the bit represents has failed.

As an example, assume three records, α, β, and γ, having identical phase information and, for simplicity, only 6 diagnostic tests. The test failure information on these records is the following: $\alpha = 0\ 0\ 1\ 0\ 1\ 1$, $\beta = 1\ 1\ 0\ 0\ 1\ 0$, and $\gamma = 0\ 1\ 0\ 0\ 0\ 0$. The leftmost bit corresponds to test No. 1. Therefore, tests No. 3, 5, and 6 have failed in record α; 1, 2, and 5 in β; and 2 in γ. After test sorting, the three records will be arranged in this order:

Record	Binary Test Data	Translated Test Result	Analogous Alphabetic Sequence
β	1 1 0 0 1 0	1, 2, 5	ABE
γ	0 1 0 0 0 0	2	B
α	0 0 1 0 1 1	3, 5, 6	CEF

It can be seen that the binary test data is so arranged that the analogous alphabetic sequence is in true alphabetic order.

13.2.3 *Dictionary Printing*

The printing of the actual dictionary was done by an IBM high-speed tape-controlled printer. When all the records were properly sorted by phase and test information, they were converted and printed in a dictionary form.

The test data collected for the dictionary was in binary form. This was converted to the form used by the system in typing out test results. Whenever a bit is "1", the location of that bit in the test recording indicates the phase and test number of this test failure. This is converted into a 3-digit octal number. The test data format in a dictionary, therefore, looks exactly like the system teletypewriter printout in a normal diagnosis. Identical test data produced by different faults is entered only once in the dictionary, but all the faults are listed. Fig. 5 shows a sample sheet of the dictionary.

XIV. EQUIPMENT EMPLOYED

14.1 *TELETYPE Equipment*

14.1.1 *Highspeed Tape Punch and Receiving Signal Converter*

A 1000 words per minute (100 characters per second), tape punch was used to record test data from the system. It was equipped for punch-

PHASE D PARTIAL
010 044 046
PHASE E
036 037 040 041 042
043 044 070 077

CC-010 J-F52615-12

PHASE D PARTIAL
010 055
PHASE E
077

CC-023 J-F52616-14

PHASE D PARTIAL
010 055
PHASE E
077 115

CC-024 L-F52601-3
CC-025 Y-F52618-29

PHASE D PARTIAL
010 055 056
PHASE E
070 077 115

CC-025 L-F52600-4

PHASE D PARTIAL
010 056
PHASE E
070

CC-023 J-F52616-10

PHASE D PARTIAL
010 056
PHASE E
(CONTINUED)

070 115

CC-024 L-F52601-5

PHASE D PARTIAL
010 056
PHASE E
115

CC-024AA-F52615-8

PHASE D PARTIAL
010 057
PHASE E
026

CC-012 L-F52600-3
CC-011 L-F52601-3

PHASE D PARTIAL
010 057 060
PHASE E
026

CC-012 L-F52600-4

PHASE D PARTIAL
010 073
PHASE E
010 032 034 070 077

CC-313 E-F52616-15
CC-309AD-F52630-5
7

PHASE D PARTIAL
011 012 013 014 015
016 017 043 044 045
046 073
PHASE E
(CONTINUED)

010 030 032 033 034
035 036 037 040 041
042 043 044 077

CC-010 N-F52617-2

PHASE D PARTIAL
011 012 013 015 016
017 043 044 045 046
073
PHASE E
010 012 013 014 016
027 030 031 032 033
034 035 036 037 040
041 042 043 044 070
071 072 073 074 075
076 077 100 101 102
103 104 105

Test Results

CC-210 P-F52628-20

PHASE D PARTIAL
011 012 016 017 043
044 045 046 073
PHASE E
012 013 017 027 030
031 032 033 034 035
036 037 040 041 042
043 044 070 071 072
073 074 075 076 077
100 101 102 103 104
105

Package Identity
central control
location of package
type of package
type of fault

CC-210 P-F52628-18

PHASE D PARTIAL
011 013 015 017 043
045 073
PHASE E
011 013 015 017 027
031 032 033 034 035
036 037 040 041 042
(CONTINUED)

ing eight channels, although only seven channels were used. The high-speed tape punch is a rotating device which can be operated only at a particular point in the cycle. The receiver signal converter acts as a buffer between the system and the punch. This equipment has coded input circuits, consisting of two wires each, so arranged that information appearing on these leads will be punched on the tape at the proper point in the punch operating cycle. Since the punch can accept information only at a certain point in a cycle, a buffer store is included in the receiving signal converter, where the coded signal is stored until the punch is ready to accept it.

14.1.2 *High-Speed Tape Reader*

The reader operates at a speed of 1071.4 words per minute (107.14 characters per second). It is a parallel output device, equipped for eight-channel operation. The reader was used in conjunction with the tape error-detection unit to check the output paper tape.

14.1.3 *Transmitter Distributor*

A transmitter distributor was used to read the package information tape. This unit operates at 100 words per minute (10 characters per second) and is equipped with a set of five parallel code-reading contacts. These contacts were used to feed the code signals to the system.

14.2 *Equipment Built*

14.2.1 *Test Control Unit (Fig. 6)*

The test control unit was the master control for the dictionary tests. This unit coordinated the operation of the system and all peripheral dictionary equipment. Since the operating time of any type of logic employed in this unit was insignificant compared to the total time consumed in changing the packages, recording the test data, loading, and unloading paper tapes for the punch, etc., relay logic was used. The circuits were so designed that any improper operation of pushbuttons and control levers did not cause erroneous data to be recorded.

The test control unit performs the following major functions:

1. Controls the system clock: when the system clock stops, it will not start again until a command is received from this unit.
2. Governs the reading of the package information tape.
3. Communicates with the system and provides instructions for per-

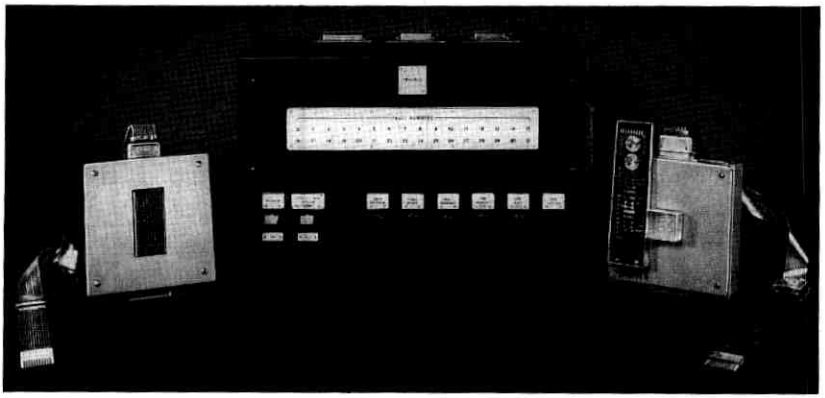


FIG. 6—Left: fault-simulation unit for gate packages. Center: test control unit. Right: fault-simulation unit for transistor packages, with flip-flop package plugged in.

forming an automatic or manual test or a package identification skipping operation.

4. Instructs the fault-simulation unit as to when a particular fault by itself or as a part of a series of faults should be simulated. The simulation and sequencing of faults during an automatic test are accomplished automatically.
5. Allows manual insertion of any desired faults with the aid of the fault reset button, fault advance button, and fault number indicator.
6. Provides visual indications for the following: (a) mode of operation: (test or normal), (b) which fault is being tested, (c) occurrence of an unsimulated trouble in the system, (d) completion of either an automatic or manual test or the skipping operation, (e) the off state of the system.

14.2.2 *Fault-Simulation Unit*

A total of 49 different types of packages are employed in the central control units. A study of the type and number of faults to be introduced in all the circuit packages used revealed a logical division between gate packages and other types of packages, such as flip-flops, amplifiers, etc. Two fault-simulation units were designed which simulated essentially all the circuit cards in the central control.

The fault-simulation units were wired to create the required circuit packages with the necessary fault-simulation features by means of a

plugboard arrangement. The plugboard consisted of a 150-pin connector. The socket portion of this connector was wired to various components provided in the fault-simulation unit. For each circuit package, a separate plug was used which contained the necessary jumper wires to simulate the corresponding package.

To minimize the stray capacitance, the fault-simulation units were laid out to be as compact as possible. Each unit was about 7″ x 7″ x 4″, as shown in Fig. 6. Microminiature relays were used inside the units to switch the various faults. During testing, they were placed in close proximity to the package socket. The AND and OR gate fault-simulation unit consisted of 30 relays plus the components necessary to simulate any one of the 16 types of gate package.

The second fault-simulation unit was used for transistor and miscellaneous packages. Most of the faulty conditions for these packages were simulated on the output terminals; therefore, only a few types of circuit cards needed to be wired up from components provided in the unit itself. In the majority of cases, a good package was plugged into a socket supplied on the fault-simulation unit, and the conditions on the output terminals were controlled by relays inside the unit. As in the first unit, plugboard connectors were provided for each simulated type of circuit packages.

The sequencing of faults was controlled by the test control unit. A signal from the fault-simulation unit was sent to the test control unit when all the faults were completed for a package.

14.2.3 *Tape Error Detection Unit*

The tape error detection unit was designed for checking the punched paper tape. The code signals were fed into this unit in parallel by the high-speed tape reader. Two types of tape errors were detected by this unit: parity of each character, and predetermined block length of test data. The output tape was checked as soon as it was punched.

XV. DETAILED PLANNING OF TEST PROCEDURES

In order to take up a minimum amount of system time to create the dictionary, careful plans were made. The plans included the detailed test procedures, manpower requirements, job descriptions for the operators, and the equipment operation instructions involved in the acquisition of test data. Suggested procedures were also given in case of trouble during the dictionary test. All conceivable major and minor troubles were analyzed. The effort spent in planning was worthwhile.

XVI. EXPERIENCE WITH MORRIS CENTRAL CONTROL MAINTENANCE

The following points summarize our experiences:

1. Certain troubles were not detected by the diagnostic program. These can be attributed to one of the following:
   (a) The particular component in which the trouble was simulated was not actually used by the system.
   (b) There was some redundancy in the circuits, creating the possibility of troubles that could not be detected.
   (c) The diagnostic program was not exhaustive. Circuits in certain areas have not been covered. Additional tests could have been added to the present diagnosis, but it was felt that such an effort was not justified for the Morris System.
   (d) Due to the design of the Morris central control, there were limitations which made it impossible to detect certain troubles. For example, a shorted clock diode on an AND gate usually resulted in no test failures. The low impedance of the clock supply prevented the clock pulse from being clamped. However, the margins were decreased and occasionally a test failed when this type of trouble was introduced. An open clock diode on an AND gate at the input to a flip-flop which had no feedback around it usually resulted also in no test failures. The flip-flop merely changed state, without waiting for the clock pulse which normally initiated the change. However, some flip-flops could be operated falsely by the noise on the gate leads which could occur before the clock pulse arrived.
2. Some troubles resulted in inconsistent test results. Most of these were due to circuit design. For example, certain troubles introduced caused complementary functions to be performed simultaneously, such as writing a "0" and a "1" in a memory spot. These troubles produced race conditions in the central control logic, resulting in inconsistent test results.
3. Some troubles introduced in the standby central control affected the operation of the active system. This indicated that better isolation was required between the two central controls.
4. Many test results were extremely difficult to analyze. Since there are some 900 tests made on each fault, even sketchy analysis is time consuming. It is very difficult to explain why certain tests fail with respect to trouble introduced and to predict all the tests that should fail. For these reasons, the dictionary is extremely useful. Even now, if a set of test failure results is not found in the dictionary, analysis presents grave problems.

5. The fault-simulation units introduced sufficient capacitance in the circuit to cause marginal operation in some of the circuits. By testing most transistor packages in a fixed state (that is, a flip-flop permanently set, etc.), and by reducing the value of diode gate resistors, it was possible to obtain meaningful data on all but seven packages in central control. These seven persisted, in spite of all efforts, in failing tests when the simulated package was plugged in and no actual faults introduced.
6. It appeared evident that a central control dictionary cannot be written from an analysis of the circuit with respect to the diagnostic tests. The complexity of the central control made it difficult to predict how the central control will perform with any given fault.
7. Even if a dictionary were not produced, extensive trouble insertion or trouble simulation was necessary to debug and measure the effectiveness of the diagnostic program.

XVII. SOME INTERESTING STATISTICS

The central control maintenance program has about 7200 program words, 6000 of which are for diagnosis. Over 50,000 faults were simulated. About 250 hours of actual system machine time were used explicitly for data gathering. The total data (about 60 million bits) occupied 93 reels of 1,000 foot, 7-channel paper tape. These in turn were converted by an IBM-9200 machine into four reels of magnetic tape. The conversion time was about 12 hours. Out of 51,671 records, about 35 were destroyed and 25 rendered doubtful because of machine errors.

The sorting program for the IBM-704 computer is about 2,500 words long, and was written by one man in about eight months. The total sorting time on the IBM-704 computer was 34 hours.

The time spent in designing the diagnostic program and producing the dictionary was about 12 man-years, of which two man-years was for designing the diagnostic program and two man-years was for debugging and modifying it.

The dictionary consists of 1,290 pages (11″ x $14\frac{7}{8}$″), bound into four volumes.

XVIII. DICTIONARY RESULTS

There are 10,315 different test patterns, 73 per cent of which list only one possible package failure, 13 per cent two possible package failures. Therefore, for a large number of catastrophic troubles (86 per cent), the dictionary will be able to pinpoint the fault to within two packages.

Our experience indicates that a man with a few minutes of training will be able to consult the dictionary and determine the faulty package, usually within two to three minutes, sometimes as long as five to ten minutes, and this appears adequate. Our preliminary evaluation of the dictionary indicates that it will be able to locate about 75 per cent of the troubles.

XIX. CONCLUSIONS

The feasibility of producing a central control dictionary by the system itself has been proven, and a dictionary has been produced. Considerable experience has been gained with the maintenance of a large electronic logic circuit. As a result, a number of improvements can be made. A great deal has also been learned regarding the limitations of diagnosis due to the central control circuit design. These limitations cause inconsistent test results or no test failures. With the improvements which can be made in the diagnosis and the circuit design, it appears feasible to have a central control dictionary which will be able to locate 90 per cent or more of all the probable troubles.

Using the dictionary techniques, the average repair time may be kept very low, and the maintenance was made much easier. Success in this area of work has contributed greatly to meeting the initial maintenance and reliability objectives.[2]

Considering that this was an initial attempt to solve a very complex and difficult problem, the results have been gratifying. Considerable headway has been made in automatic diagnostic techniques. However, we must develop these techniques further if we are to cope successfully with the problems of maintaining the even more complex electronic telephone switching system now being developed.

XX. ACKNOWLEDGMENTS

The authors would like to acknowledge the help of their many colleagues who contributed so much to the success of this project. We are also grateful to the Teletype Corporation for supplying a development model of their high speed tape punch and reader.

APPENDIX

Description of the Central Control Diagnostic Program

A complete understanding of the central control diagnostic program demands an intimate knowledge of the central control. This appendix

will therefore only illustrate the types of tests used and the methods for observing test results.

For clarity, it is desirable to define a simplified repertoire of instructions. A, B, C, D, AA, AB refer to symbolic designations for flip-flops, flip-flop groups, or storage locations.

SM	Sample match circuits for a mismatch condition during the process of executing the next instruction.
G A, B	Gate the contents of flip-flop group A to flip-flop group B, via bus.
ST1 AA	Set up transfer register 1 (T1) to quantity AA.
R0, AB R1, AB	Transfer to the address stored in transfer register 1 if the reading at AB is 0 or 1 respectively.
RFF0, C RFF1, C	Transfer to the address stored in transfer register 1 if FFC is 0 or 1 respectively.
WFF C, D	Write the contents of flip-flop C into memory at address D.
W0, AB W1, AB	Write 0 or 1 respectively into memory of address AB.

Match circuits are provided to compare the outputs of the instructions to the stores (transfer or advance to next instruction for the flying spot store, read or write 0 or write 1 to the barrier grid store) and to match the busses of the two central controls.

In the following examples, each program step is listed, followed by a symbolic modifier and by comments. In studying these programs it is important to remember that the two central controls are working in synchronism, and that the working central control is addressing and writing into the stores.

First, to check the ability to make decisions, the following program was applied:

```
W0  AB    (AB is any convenient address.)
SM
R1  AB
```

If a mismatch is detected on the R1 instruction, the standby central control has falsely transferred on a reading. The program

```
W1  AB    (AB is any convenient address.)
SM
R1  AB
```

will check the ability of the standby central control to transfer on a reading.

In order to check the ability to write correctly, register R was first set up, to all 1's. Register R has the property that its individual flip-flops may be written into memory using the WFF instruction, or may be read using the RFF0 and RFF1 instructions.
The program:

```
ST1     All ones
G       T1, R
SM
WFF     C, D      Any flip-flop C, of register R, to any convenient
                  address D.
```

In order to check the flip-flop groups of central control, the following program was used:

```
ST1     All zeroes
G       T1,  A    (A is the flip-flop group being tested.)
SM
G       A, R.
```

As previously mentioned, individual flip-flops of register R may be examined. If a mismatch has occurred, the proper flip-flop may be isolated by repeated use of the two instructions SM, RFFO, followed by a check to see if the flying spot store order-match circuit indicated a mismatch on the RFFO instruction (the instruction which followed SM was sampled for a mismatch); by checking all the flip-flops in R, the flip-flops of A which were not capable of being set were detected. R and T1 were previously checked to make certain that all their flip-flops could be set and reset.

The above programs are typical. Common subroutines were used to record the results of tests in the barrier grid store and to control the typing out of these results with the teletypewriter.

REFERENCES

1. Joel, A. E., B.S.T.J., **37**, 1958, pp. 1091–1124.
2. Haugk, G., and Yokelson, B. J., "Experience with the Morris Electronic Switching System," to be published in A. I. E. E. Trans.

Heuristic Remarks and Mathematical Problems Regarding the Theory of Connecting Systems

By V. E. BENEŠ

(Manuscript received February 1, 1962)

A connecting system consists of a set of terminals, a control unit for processing call information, and a connecting network. Together, these three elements provide communication, e.g., supply telephone service, among the various terminals. In this paper we present a comprehensive view of the theory of connecting systems, an appraisal of its current status, and some suggestions for further progress.

The existing probabilistic theory is reviewed and criticized. The basic features of connecting systems, such as structure, random behavior, complexity, and performance, are discussed in a nontechnical way, and the chief difficulties that beset the construction of a theory of traffic in large systems are described. It is then pointed out that despite their great complexity, connecting systems have a definite structure which can be very useful in analyzing their performance. A natural division of the subject into combinatory, probabilistic, and variational problems is drawn, and is illustrated by discussing a simple problem of each type in detail.

I. INTRODUCTION

Mass communication long ago spread beyond the manual central office and assumed a nationwide character; it is presently becoming world-wide in extent. Many of the world's telephones already form the terminals of one enormous switching system. The scale, cost, and importance of the system make imperative a comprehensive theoretical understanding of such global systems.

Nevertheless, a lack of knowledge about the combinatory and probabilistic properties of large switching systems is still a major lacuna in the art of mass communication. It is a fact of experience that each time a new switching system is planned, its designers ask once again some of

the perennial unanswered questions about connecting network design and system operation: How does one compute the probabilities of loss and of delay? What method of routing is best? What features make some networks more efficient than others? Etc.

The present paper is an informal discussion of problems in the theory of traffic flow and congestion in connecting systems (called traffic theory, or congestion theory, for short). The comments to be made are prefatory, tutorial, and illustrative. They are intended as background for several papers of a more technical nature; one of these papers[1] appears in this issue, and the remaining three[2,3,4] are to appear later. In these papers, topics touched on in the present work are considered in greater depth and detail. Together, the papers are an attempt to describe a comprehensive point of view towards the subject of connecting systems. I believe that this point of view will be useful in constructing a general theory of connecting networks and switching systems. What follows is then in part a prospectus for research to be reported on in the future.

My concern in this paper is with some of the physical bases and principal problems, with the fundamentals and difficulties, of the subject. I wish to emphasize some important properties and distinctions on which a systematic approach may be based. I am making a plea for a much more general, abstract, and systematic approach to large-scale congestion problems than has been envisaged heretofore.

Naturally, it is impossible to explore all the consequences of such a comprehensive approach in one paper; I do not pretend to have solved even some of the basic problems of the theory. I am only saying "Look, perhaps these observations will help provide a general approach."

Examples and simple problems appear in the text as illustrations of the principal points made. For tutorial purposes, I have chosen particularly simple and clear illustrations, which may seem trivial to cognoscenti of traffic theory. Nevertheless, it has been my experience in talking with engineers that the comprehensive view here presented is sufficiently new to warrant clear, simple examples. More complex problems do not belong in an introductory work; they are to appear in later papers.

II. SUMMARY

In Section III we give a historical sketch of traffic theory, which is followed by a critique of existing theories in Section IV. The general properties of switching systems are discussed in Section V. The performance of switching systems and desiderata for a theory of congestion are considered in Section VI and Section VII, respectively. Sections V to VII are heuristic and nonmathematical in character. Mathematical

models are considered in a general way in Section VIII, while Section IX concerns itself with some of the basic difficulties and questions that arise in constructing a theory of traffic in a large-scale system.

In Sections X and XI we show that, despite their great complexity, connecting systems actually have a definite structure which can be very useful in analyzing their performance. This usefulness is exemplified by four specific instances in Section XII. In Section XIII we make a general division of the subject into combinatory, probabilistic, and variational problems. The remaining sections, Sections XIV to XVI, are devoted to illustrating this division by working out a simple problem of each type in full detail.

III. HISTORICAL SKETCH

We shall not attempt to canvass systematically the literature of congestion theory. For the interested reader, the best single theoretical reference on the theory of probability in connecting systems is undoubtedly the treatise of R. Syski;[5] the historical development of the subject has been described in papers by L. Kosten[6] and R. I. Wilkinson.[7] Nevertheless, we include a brief account of previous work in order to substantiate our critique (Section IV) of present theories of traffic in connecting systems.

The first contributions to traffic theory appeared almost simultaneously in Europe and in the United States, during the early years of the 20th century. In America, G. T. Blood of the American Telephone and Telegraph Company had observed as early as 1898 a close agreement between the terms of a binomial expansion and the results of observations on the distribution of busy calls.* In 1903, M. C. Rorty used the normal approximation to the binomial distribution in a theoretical attack on trunking problems, and in 1908 E. C. Molina improved Rorty's work by his[8] (or Poisson's) approximation to the binomial distribution.

In Europe, the Danish mathematician A. K. Erlang, from 1909 to 1918, laid the foundations of the first dynamic theory of telephone traffic, which is in general use today. Perhaps influenced by statistical mechanics, Erlang introduced the notion of statistical equilibrium, and used it as a theoretical basis for deriving his now well-known loss and delay formulae. An account of Erlang's work is given by Jensen.[9]

From 1918 to 1939 traffic theory developed in many directions that are (on retrospect) closely allied to specific problems that arose in the design of the automatic telephone systems that were coming into use, and in

* Blood's unrecorded work was reported by E. C. Molina and described by R. I. Wilkinson.[7]

related queueing systems. We mention only a few topics: T. Engset[10] introduced the notion of a finite number of sources of traffic, G. F. O'Dell[11] published a classical paper on gradings, C. D. Crommelin[12] studied constant holding-time delay systems with many servers, E. C. Molina[13] made contributions to trunking theory. F. Pollaczek[14] and A. I. Khinchin[15] studied the queue with one server, and derived the delay distribution that bears their linked names. Pollaczek has also solved single-handedly many other difficult loss and delay problems. All these important contributions are concerned with congestion in specific parts of connecting systems. During this period, T. C. Fry wrote the first systematic and comprehensive book[16] on applied probability; this book devoted a chapter to telephone traffic, and appeared in 1928.

Between 1939 and 1948 there developed an increasing awareness (among workers in traffic theory) that the mathematical bases of traffic theory were closely related to the modern theory of stochastic processes initiated by A. N. Kolmogorov[17] in 1933. In particular, Erlang's idea of statistical equilibrium was identified with the stationary measure of a Markov process (or more generally with a semigroup of transition probability operators). Also, C. Palm[18] stressed the importance of recurrent processes, and W. Feller[19] that of birth-and-death processes, to traffic theory. However, particular problems continued to form the bulk of the new literature. Palm[18] made a penetrating theoretical analysis of traffic fluctuations, and L. Kosten studied such topics as retrials for lost calls,[20] and error in measurements of loss probability.[21]

The introduction of crossbar switching and common control of connecting networks in 1938 (see Ref. 22) was accompanied by a new kind of problem: calculating the loss due to *mismatching of available links* (rather than to unavailability of trunks). The first comprehensive treatment of loss in such systems was given by C. Jacobaeus[23]; his theory is adequate for practical purposes, but is based on assumed *a priori* distributions for the state of the system. R. Fortet[24] has also made contributions to this topic in the spirit of Jacobaeus' approach. A less satisfactory method for the same problems based only on the possible paths for a call has been developed (independently) by C. Y. Lee[25] and P. Le Gall.[26]

The statistical equilibrium approach to congestion in crossbar systems is rendered extremely arduous by the large number of possible states. The difficulties in this method have been faced with some success by K. Lundkvist[27] and A. Elldin[28]. However, no practically feasible approach exists at present that simultaneously includes both the concept of statistical equilibrium and the structure of the connecting network. *A fortiori*, no approach exists that also includes the effect of the common control equipment that places calls in the network.

IV. CRITIQUE

In comparison with the highly sophisticated communications systems that are being built, the models and assumptions on which theoretical studies are based are often crude and fragmentary, almost more indicative of our ignorance than of the properties of systems. It may be argued that such a harsh appraisal of the condition of traffic theory is unjustified, and is disproved by the practical successes of current engineering methods. However, it is not the efficacy of these methods, but their theoretical basis and scope, that we are questioning. Who knows to what extent present systems are "overdesigned"?

To be sure, measures of performance, loss and delay formulas, and routing methods are in daily use. Still, only in very special cases have they been investigated, let alone analyzed and understood in the full context of the system to which they are applied. Although the published literature on telephone traffic alone is vast, and many models and problems have been considered, the existing theories tend to be incomplete and oversimplified, applicable to at most a small portion of a system. Useful comprehensive models are needed; to date, only individual pieces of systems have been treated with theoretical justice. As R. Syski remarks on p. 611 of Ref. 5: "At the present stage of development . . . the theoretical analysis of the [telephone] exchange as a whole has not been attempted." The general theory of switching systems now consists of some apparently unrelated theorems, hundreds of models and formulas for relatively simple parts of systems, and much practical lore associated with specific systems. It will stay in this condition until sufficient theoretical underpinning is provided to unify the subject. We believe that this sad "state of the theory" is due largely to these three factors:

(*i*) The large scale, and consequent inherent difficulty of the problems.

(*ii*) The absence of a widely accepted framework of concepts in which problems could be couched and solved.

(*iii*) The lack of emphasis on and success with the combinatorial aspects of the problems.

More generally, many of the basic mathematical properties of connecting networks and switching systems have either never been studied, or, if studied, have not been digested, advertised, and disseminated for engineering use. As a result, the design and complexity of systems has consistently run ahead of the analysis of their performance.

V. GENERAL PROPERTIES OF CONNECTING SYSTEMS

We start by discussing some universal properties of connecting systems from the point of view of congestion, without reference to definite mathe-

matical models for their operation. Specifically, we describe, in a nontechnical way, (*i*) the general nature and outstanding features of connecting systems, (*ii*) the principal kinds of congestion that interest engineers, and (*iii*) some of the difficulties and desiderata in both the theory and practice of large-scale switching. No mathematical abstractions are used at first. Some observations made may seem obvious or trivial; nevertheless, they are necessary for the general understanding that we desire. On these observations, we shall base a systematic division of the theory into three kinds of problems, *combinatory*, *probabilistic*, and *variational*.

By a *connecting system* we shall mean a physical communication system consisting of (*i*) a set of terminals, (*ii*) control units which process requests for connection (usually between pairs of terminals), and (*iii*) a connecting network through which the connections are effected. The system is to be conceived as operating in the following manner: (1) calls (or requests for connection) between pairs of idle terminals arise; (2) requests are processed by a control unit, and desired connections are completed, if possible, in the connecting network; (3) calls exist in the network until communication ends; (4) terminals return to an idle condition when a call terminates. (Naturally, the arising requests may "defect" at any point during the process of connection.)

The gross structure of a connecting system is depicted in Fig. 1. Most modern connecting systems follow this basic pattern. Particularly important examples are telephone central offices, toll centers, telegraph networks, teletypewriter systems, and the many military communications systems.

All the examples cited share three important properties. These are (*i*) great combinatorial *complexity*, (*ii*) definite geometrical or other *structure*, and (*iii*) *randomness* of many of the events in the operating system.

It is obvious that many connecting systems are highly complicated.

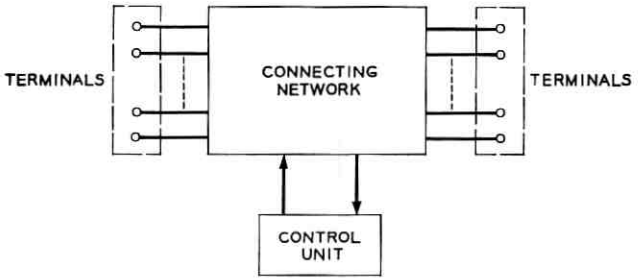


Fig. 1 — Connecting system.

Both the control unit and the connecting network contain thousands of parts which may (together) assume millions of combinations. That is, the system can be in any one of millions of possible "states." These numbers are increased when several switching centers are considered together as a unit, as in toll switching. Our purpose in calling attention to this complexity is to suggest that it calls for theoretical methods that, like those of statistical mechanics, are especially designed to distill important facts from masses of detail.

It is less often realized, however, that this complexity is accompanied by definite mathematical structure, and is frequently alleviated by many symmetries. The control unit and the connecting network always have a specific combinatory, geometric, and topological character, on which the performance of the system closely depends.

By imputing randomness to the systems of interest we do not imply that their operation is unpredictable; we mean only that the best way of describing this operation is by use of probability theory. It is not practical, even though it might be possible in principle, to predict the operation of a switching system by means of differential equations in the way that the flight of a rocket is predicted. However, differential equations have been used for many years to describe, not the motion of an actual system, but the changes in the *likelihoods* or *probabilities* of its possible states. Such equations govern the flow or change of probabilities and averages associated with the system, not the detailed time behavior of the system itself. It is in this weaker sense of assigning likelihood to various events that we can predict the behavior of switching systems, a fact first emphasized by A. K. Erlang's pioneering work on telephone traffic.[9] For instance, certain features (such as average loads offered and carried) of telephone traffic that are predictable in this weaker sense form the basis on which toll trunking routes are engineered.

We now turn to examples of the structure of connecting networks and of control units. The basic features of the connecting network for the No. 5 crossbar system are shown in a simplified form in Fig. 2. The network has two sides, one for subscribers' lines and the other for trunks. Small squares represent rectangular *crossbar switches*, capable of connecting any inlet terminal to any outlet terminal. These switches are arranged in groups called *frames*, either line link frames for subscribers' lines, or (on the other side) trunk line frames for trunks. Frames are indicated in Fig. 2 by large dashed squares enclosing four small squares; dots indicate repetition. The pattern of links which interconnect the switches is shown by solid lines between small squares. At most one link connects any pair of switches.

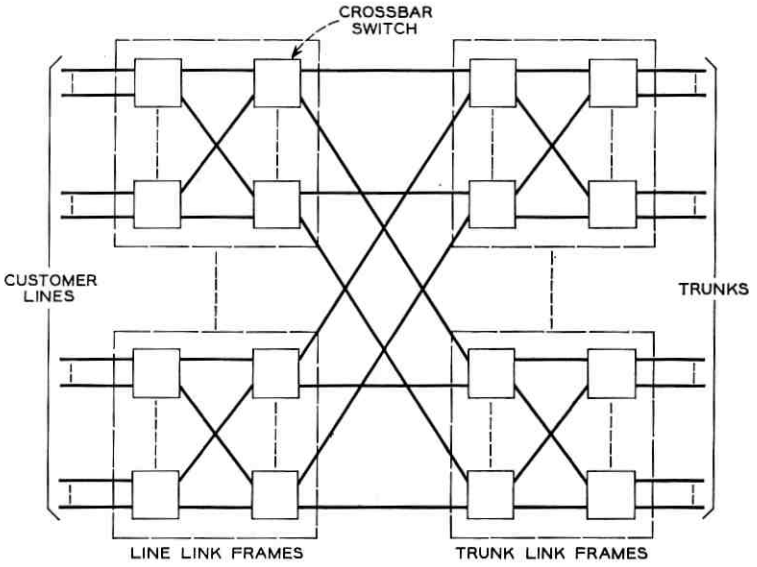


Fig. 2 — Basic No. 5 crossbar network.

As a second example of a connecting network, consider the three-stage Clos network (see Ref. 29) depicted in Fig. 3. The interpretation of this figure is the same as that of Fig. 2: small squares stand for crossbar switches, and lines between them represent links. Each call can be put into the network in m ways, one for each of the m switches in the middle column. This network has the property that if $m \geqq 2n - 1$, it is non-blocking.

A control unit consists of parts that are arranged in a manner reflecting their function, and are determined by the operations necessary to establish a connection, and by the philosophy of design and the tech-

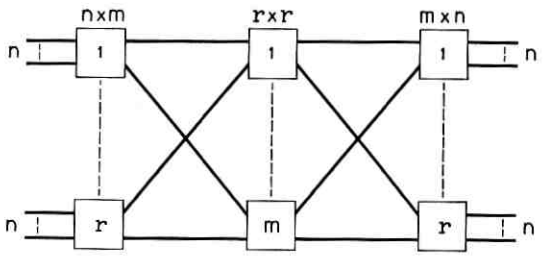


Fig. 3 — Clos three-stage network.

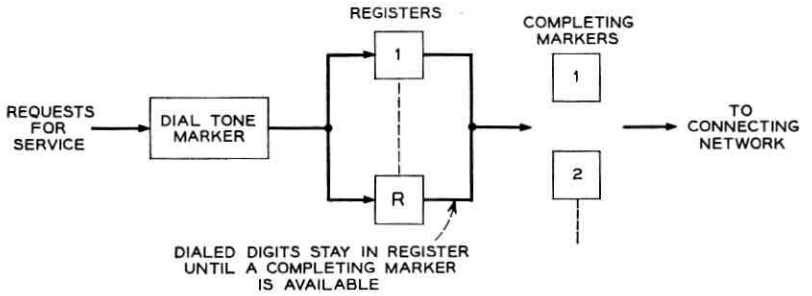


Fig. 4 — Simple control unit.

nology that are basic to the system. To establish a connection, the control unit must do some or all of the following: (*i*) identify the calling party or terminal, (*ii*) find out who the called party is, and (*iii*) complete the connection. Three examples will be considered, in order of increasing complexity and modernity.

A simple example of the structure of a control unit is given in Fig. 4. The unit consists of a dial-tone marker which assigns and connects available idle registers to subscribers for dialing. The dialed digits remain in the register until a completing marker (one of possibly several) removes them and uses them to complete the call. The calls, or requests for connection, may be thought of as arriving from the left, and proceeding through the diagram from left to right. There may be a delay in obtaining dial tone, a delay in securing the services of a completing marker, or a circuit-busy delay (or rejection) in the network. It should be observed that the switching equipment necessary for connecting subscribers to registers, or registers to completing markers, is left out of account in this model.

A second example is obtained from the first by inserting a buffer memory between the registers and the markers as shown in Fig. 5. (One

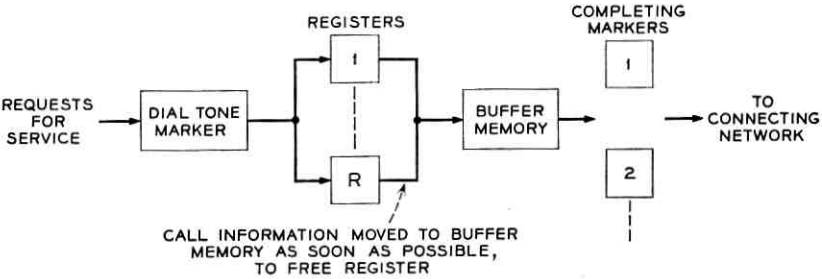


Fig. 5 — Control unit with buffer memory.

can argue that registers are expensive special-purpose units and should not be used for storing call information when cheap memory is available.) When dialing is finished, the call information is forthwith transferred to the buffer memory, there to wait for a completing marker without preempting a register. The markers and registers are now effectively isolated, so that delays in completing calls do not cause delays in obtaining dial tone. Again, traffic is viewed as moving from left to right.

The high speeds possible with electronic circuits have led to new configurations and problems (for control units and networks) which have not yet received much attention in congestion theory. Although it performs the same functions, the control unit of a modern electronic central office usually has an organization differing from that of the examples of Figs. 4 and 5, which are characteristic of electromechanical systems. Four principal reasons for this contrast are:

(*i*) The electronic office relies heavily on a large digital memory to aid in processing calls and (in time division systems) to keep track of calls in progress; electromechanical systems, on the other hand, are based largely on "wired-in" memory.

(*ii*) In the electronic office, processing a given call usually requires several consultations of the digital memory; thus, the flow of traffic in the control unit is re-entrant and not unidirectional as in Figs. 4 and 5.

(*iii*) The speed of electronic components often makes it possible to perform only one operation at a time; thus, a single unit may be (alternately) part of a dial-tone marker, part of a register, part of a completing marker, etc., depending on the details of organization of the control unit.

(*iv*) The replacement of "wired-in" memory, whose stored information is immediately available, by an electronic memory which has to be consulted, creates problems analogous to the problem of connecting completing markers to registers in the No. 5 crossbar system: special access units are needed. Subunits of the control unit, such as dial-tone markers, completing markers, senders, etc., must take turns in using the access circuit to the digital memory.

Fig. 6 depicts a (hypothetical) control unit for an electronic switching system built entirely around a memory which stores all information on the current status of calls. The control unit consists of various special-purpose units such as a sender, a receiver, a completing marker, a dial-tone marker, and registers. Each of the listed units can operate independently of and simultaneously with the others; however, they compete for (take turns at, possibly with priorities) the access circuit to the memory. Each unit depends on the memory to give it a new assignment, to file the results of the last one, or both. Every operation of a special-

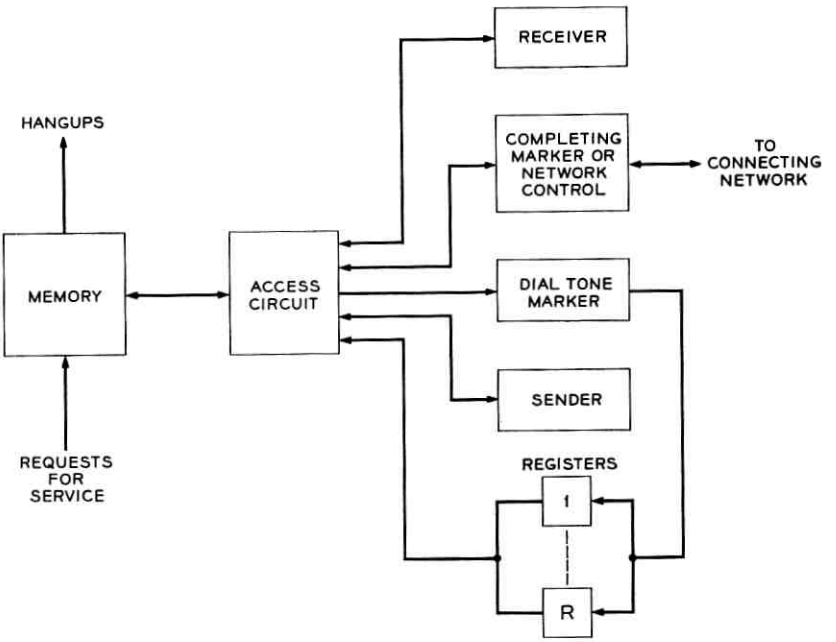


Fig. 6 — Block diagram of electronic control unit.

purpose unit requires access to the memory, either to obtain data from it, or to file data in it, or both. The memory contains several classes of calls: those waiting for dial tone, those waiting for a completing marker, those actually in progress in the connecting network, etc.

VI. PERFORMANCE OF SWITCHING SYSTEMS

In general, the gross or average features of switching systems are both more accurately predictable and more economically important than the specific details. The average load carried by a trunk group is usually more easily predicted than the condition of a particular trunk; and the "all trunks busy" condition of the group is of greater concern to the telephone administration than the busy condition of a single trunk.

From the point of view of economics and traffic engineering, only certain average features of the behavior of a system (used as measures of performance) are important. These few quantities of interest depend on the multitude of details of "fine structure" in the control unit and the connecting network. Although the intricate details give rise to the important averages, the details themselves are of relatively little interest.

In the rest of this paper, we shall repeatedly contrast the few average quantities that are of engineering interest with the many millions of detailed features and properties (of connecting systems) on which the averages are based. The central problem in the theory of connecting systems is to understand how the interesting quantities arise from the details, and to calculate them.

We shall start our discussion of the contrasting roles of averaged features and details by considering some of the different kinds of congestion that interest engineers, and in addition some associated measures for the performance of systems.

Congestion is said to occur in a connecting system when a requested connection cannot be completed immediately. By "immediately" we mean, of course, not "instantaneously", but "as fast as control equipment, assumed available, can do its work". The time it takes to complete a call contributes to congestion only if it keeps other calls from being completed at the normal rate. That a call cannot be completed immediately (in this sense) may be due to facts of three kinds: (*i*) certain necessary units of switching equipment (like trunks, or markers) are all busy; (*ii*) there are available units, but they occur in an unusable combination, or "fail to match"; (*iii*) congestion has occurred previously, and other requests are awaiting completion.

In telephone traffic theory, requests for connections which encounter congestion are traditionally termed *lost calls*. This terminology is used whether the request is refused (and never completed), or merely delayed (and completed later). Switching systems differ in the *disposition* of lost calls, i.e., in what is done with requests which encounter congestion. There are in theory two principal ways of disposing of lost calls. In the first way, termed "lost calls cleared", the request is denied and leaves the system; this way of dealing with lost calls naturally gives rise to the *proportion of requests denied*, or the probability of blocking or loss, as a measure of performance. The second way of disposing of lost calls is termed "lost calls delayed", and consists in delaying the request until equipment becomes available for completing the connection; associated with this is the *probability of delay* in excess of a specified time t, as a measure of performance.

On the simplified account of the last paragraph we must impose at least two qualifications. First, whether a request suffers blocking or delay (or both!) may depend on the condition of the system at times shortly after the request is made; second, the completion of a request usually involves a sequence of steps, any one of which may expose the request to delay or loss. For example, a request may encounter delay in obtaining

dial tone, delay in securing the services of a completing marker, and delay or blocking in the attempted completion of the desired connection through the connecting network.

We conclude this section by briefly considering what general features of connecting systems are particularly relevant to their performance as measured (for example) by probabilities of blocking or delay, or by average loads carried, offered, or both. Now, a connecting system has two principal parts, the control unit and the connecting network; the features of the system that are relevant to performance are conveniently distinguished according to whether they are features of the control or of the network. This distinction is fundamental because the performance of the control is largely determined by the speed and number of the various sub-units comprising it, while the performance of the network is largely dependent on what combinations of calls can be in progress simultaneously.

The control unit is basically a data processing system: it collects information about desired connections, digests it, makes routing decisions, and issues orders for completing requested calls in the connecting network. Its capacity is measured, e.g., by the number of customers who can be dialing simultaneously, or by the number of calls which are being completed in the network at the same time. Its performance is described by the probability distributions of delay before receiving dial tone, and of delay after completion of dialing until the desired connection is completed.

For a simple model of a control unit (such as depicted in Fig. 4), the features pertinent to performance are: (*i*) the calling rate, (*ii*) the number of registers for dialing, and (*iii*) the speed and number of completing markers. In the case of the prototype electronic control unit (depicted in Fig. 6) some additional features appear: (*iv*) the speed of the access circuit to the memory, (*v*) the order of priority of the functions being performed, the discipline of access to various services, and the competition for access among marker, dial tone marker, sender, etc., (*vi*) the presence of re-entrant traffic (every call must "use" the access circuit at least twice), and (*vii*) the number and arrangement of the various functions which are going on simultaneously.

The connecting network, in contrast to the control unit, determines what calls can be in progress, rather than how fast they can be put up. Its configuration determines what combinations of terminals can be connected simultaneously together. For example, if $m \geqq n$, the Clos network of Fig. 3 has the property of *rearrangeability*: any preassigned set of calls can be simultaneously connected. The No. 5 network of Fig. 2 does not

have this property: the number of calls between a line link frame and a trunk link frame is limited by the number of links between those two frames. Such combinatory properties of the structure of the connecting network play a determining role in estimating the cost and the performance (probability of blocking) of the network. If the structure is too simple, very few calls can be in progress at a given time and blocking is high; if it is extensive and complex, it may indeed provide for many large groups of simultaneous calls in progress, and so a low probability of blocking, but the network itself may be prohibitively expensive to build and to control.

VII. DESIDERATA

Our discussion of the three prominent features of switching systems — (*i*) great complexity, (*ii*) definite structure, and (*iii*) randomness — has exposed or suggested some of the problems and desiderata which a theory of congestion in large-scale systems must (respectively) encounter and supply. Specific statements of requirements and tasks are now given.

General desiderata can be obtained by examining the purpose served by a theory of congestion. The function of such a theory is twofold: it is (*i*) *to describe* the operation of switching systems, and (*ii*) *to predict* the performance of systems. More specifically, the descriptive function (*i*) is to provide a theoretical framework into which any system can be fitted, and which permits one to evaluate the performance of the system, e.g., to compute the chance of loss, to estimate a sampling error, or to prove a network nonblocking. The predictive function (*ii*) has logically the same structure as (*i*), but emphasizes the use of theory to make future capital out of past experience, to extrapolate behavior and thus to guide engineering practice.

More specific tasks than these appear when we list some of the activities comprised by the theory and practice of traffic engineering. A possible list is as follows:

i. Describing and analyzing mathematical models.
ii. Computing measures of performance for specific models.
iii. Studying the accuracy of traffic measurements, the effects of transients, and problems explicitly involving random behavior in time.
iv. Comparing networks, control systems, methods of routing, etc.
v. Using traffic data to verify empirically the assumptions of theories.
vi. Making predictions and estimates for engineering use.

On the basis of this list, and of our previous discussions of complexity,

randomness, gross features, and details, we can say that a satisfactory theory of congestion must meet the following requirements:

i. It must be sufficiently general to apply to any system.

ii. It must yield computational procedures for system evaluation and prediction of performance, based on masses of detail. These procedures must be at once feasible and sufficiently accurate, and if approximations are made, their effect must be analyzable.

iii. It must encompass all the three basic elements simultaneously, viz., the random traffic, the control unit, and the connecting network.

VIII. MATHEMATICAL MODELS

We shall now consider what mathematical structures are appropriate theoretical descriptions of operating connecting systems. The discussion will provide an intuitive picture of an operating system, and will help to motivate a natural division of our subject into *combinatory*, *probabilistic*, and *variational* problems.

By a *state* we shall mean a partial or complete description of the condition (of the system under study) in point of (*i*) busy or idle network links, crosspoints, and terminals, and (*ii*) idle or busy control units or parts thereof. Complete, highly detailed descriptions correspond to fine-grained states specified by the condition of every crosspoint, link, or other unit in the system, in absolute detail. Incomplete descriptions correspond to coarse-grained states, or to equivalence classes of fine-grained states.

During operation, the connecting system can pass through any permitted sequence of its states. Each time a new call arises, or some phase of the processing of a call by the control unit is finished, or a call ends, the system changes its fine-grained state. These changes do not usually occur at predetermined epochs of time, nor in any prescribed sequence; they take place more or less at random. At any particular time, it is likely that some terminals, links, and parts of the control unit are idle, that various requested calls are being processed, and that certain calls are in progress in the connecting network.

The last paragraph suggests the following intuitive account of an operating switching system: it is a kind of dynamical system that describes a random trajectory in a set of states. Such an intuitive notion can be made mathematically precise in many ways. Any one precise version is a *mathematical model* for the operation of the switching system. In constructing such a model, it is neither necessary nor desirable always to use the most detailed (the fine-grained, or microscopic) states; often a partial

description in terms of coarse-grained states suffices, and is less difficult to study. Indeed, in building a model it is to some extent possible to choose the set of states to suit special purposes. One can, for instance, control the amount of information included in the state so as to strike a balance between excessive detail and insufficient attention to relevant factors. It is possible to make the notion of state more or less complete so as to achieve certain (desired) mathematical properties (such as the Markov property, or a suitable combinatory structure) which simplify the analysis of the random trajectory. Finally, one can add supplementary variables analogous to counter readings or cumulative measurements, and obtain their statistical properties.

The abstract entity appropriate for describing the random behavior of a switching system is a *stochastic process*. For our present heuristic purposes, we can define a stochastic process as follows: by a *possible history* of the system we mean a function of time taking values in the chosen set of states; a stochastic process is then a collection Ω of possible histories of the system in time, with the property that many (presumably interesting) subsets A of Ω have numerical probabilities $\Pr\{A\}$ associated with them. The probability $\Pr\{A\}$ of the set A of possible histories is interpreted as the chance or likelihood that the actual history of the system be one of the histories from the set A. Models of this kind furnish information because desired quantities can be calculated from the basic probabilities $\Pr\{A\}$.

IX. FUNDAMENTAL DIFFICULTIES AND QUESTIONS

The systematic use of mathematical models (such as stochastic processes) in congestion theory and engineering has been largely limited to small pieces of systems like single-server queues, groups of trunks with full access, etc. More complex models of systems involving connecting networks have hardly been touched by theory. This limitation has been due almost entirely to the large number of states such models require, and to the complex structure of the transitions (changes of state) that can occur. In short, the essential characteristics (of large-scale connecting systems) themselves generate the basic difficulties of the theory.

In most congestion problems, it is easy enough to construct (say) a Markov process that is a probabilistic model of the system of interest. But it is difficult, because of the large number of states and the complexity of the structure, to obtain either analytic results or fast, reliable simulation procedures. This circumstance has been a major obstacle to progress in the congestion theory of large systems. One of its consequences has been that in some cases, models known to be poor repre-

sentations of systems have been used merely because they were mathematically amenable, and no other tractable models were available. Even overlooking such extremes, it is fair to state that, to date, problems of analysis and computation have limited the amount of detail embodied in the notion of state for models of switching systems. Every effort has been made to keep the number of states in models small, and their complexity low.

Having exposed some basic properties of and theoretical problems arising from congestion in connecting systems, let us acknowledge that an operating, large-scale connecting system cannot be done full theoretical justice except by a stochastic model with an astronomical number of states and a very complicated structure of possible transitions. At this point, let us try to take a synoptic view of the subject, and ask some general questions whose discussion might indicate new approaches and emphases. Let us, in the current idiom, lean back in our chairs, make a (n) (agonizing?) reappraisal, and draw ourselves the "big picture."*

The following three questions seem (to this writer) to be pertinent, and are taken up in the next sections:

i. What is the value of mathematical models that have a very detailed notion of state?

ii. Is it possible to make explicit theoretical use of the very properties of connecting systems that appear to be most troublesome? How can the two principal difficulties (large number of states, complex structure of changes) be turned into positive advantages?

iii. What features of connecting systems are especially relevant to the mathematical analysis of system operation?

We do not pretend to provide iron-clad answers to these questions. We try to give a helpful discussion of relevant matters, illustrated by examples.

X. THE MERITS OF MICROSCOPIC STATES

We have raised the question: To what extent can detailed probabilistic models of the minutiae of operating switching systems (i.e., models with "microscopic" states) improve our understanding of these systems, and so our ability to engineer them? Against the value of such detailed models it can be argued that for engineering purposes only certain performance data are of interest, and that the detailed model produces a vast amount of information with no apparent practical method for reducing this information to probabilities of delay or blocking.

* Supplying those clichés whose substitution leaves the content of this last sentence invariant is left as an exercise for the reader.

Since the usefulness of mathematical models depends entirely on the desired information they can be forced to yield, it is not reasonable to dismiss detailed models *a priori*. For in truth, few if any such models have been considered, and it has not been shown that they are useless in the sense that no practical method for extracting useful quantities from these models exists.

To be sure, the congestion engineer is not as concerned with the minutiae themselves as with their effect *en masse*. But he has to base his conclusions and recommendations in *some* way on the total effects of a large number of individually trivial events. Hence, at some point in his procedure, he must take account of the large number of states and the complex structure of possible transitions of his system.

Traffic engineering practice is based on (relatively few) probabilities and averages, such as average loads, deviations about them, and blocking or delay probabilities. Any reliable theoretical estimate of these averages must be based on the combinatory and probabilistic properties of a theoretical model (stochastic process) for system operation. At worst, an approach or model that provides detailed information might yield a much-needed check point for the methods that are in current engineering use, and so increase the engineer's understanding of and confidence in these methods.

However, there is a much more general, positive sense in which attention to the details of connecting systems can contribute to theoretical progress. This is taken up in the next section.

XI. FROM DETAILS TO STRUCTURE

The prospect of solving (say) statistical equilibrium equations for models with a very detailed notion of state is discouraging indeed, although it has been faced, notably by Elldin[28] in Sweden. Nevertheless, a sanguine and useful approach (along this line) to connecting systems can be obtained by a shift of emphasis from "details" to "structure." We have emphasized that describing an operating connecting system means keeping track of numerous details, none of which is interesting in itself. We have said that the operation of such a system could be pictured as a trajectory in a very complicated set of states. We now claim that the inclusion of enough details (in the notion of state for a model) gives the set of possible states a *definite structure* that is useful because it makes possible or simplifies the analysis of the probabilistic model.

Whatever may be the value of detailed probabilistic knowledge for the immediate problems of engineering, such knowledge is useful if not essential in theoretical studies. By using a highly detailed, "microscopic"

description for the state of the system, it is possible to exploit the extensive mathematical structure (properties) that such a set of states naturally has. Indeed, the combinatory properties and geometrical structure of the set of states are two of the very few weapons available for attacking large-scale problems of traffic theory. I believe that in the past these properties and this structure have not been sufficiently exploited. They can only be put to use by a systematic application of "microscopic" states.

The three basic properties of switching systems discussed in Section V were (*i*) extreme combinatory complexity, (*ii*) definite geometrical structure, and (*iii*) randomness. The preceding paragraphs of this section can be related systematically to these properties, and elaborated into a sort of program: Instead of throwing up our hands at (*i*) in trying to do justice to (*iii*), we should realize that a detailed notion of state allows us to turn (*ii*) to our advantage in studying (*iii*). Let us then disregard the fact that there are many states, and analyze the structure of possible changes of state, to see how to capitalize on it.

For, indeed, the possible microscopic states of a particular connecting system are not arbitrary. They are rigidly determined by the combinatory and topological properties of the connecting network, and by the organization of the control unit. Such a set of possible states has a mathematical structure of its own, and this structure is relevant to the performance of the system, and to any stochastic process that represents its operation.

It can be seen quite generally that when a switching system changes its microscopic state, it can only go to a new state chosen from among a few "neighbors" of the state it is leaving. These neighbors comprise the states which can be reached from the given state by starting a new call, ending an existing call, or completing some operation in the control unit. In a large system, a state may have many such neighbors, but they will be few in comparison with the total number of microscopic states.

A striking and useful example of how details give rise to structure can be obtained by considering the possible states of a connecting network. These states can be arranged in a pattern as follows: At the bottom of the pattern we put the zero or ground state in which no calls are in progress; above this state, in a horizontal row, we place all the states which consist of exactly one call; continuing in this way, we stack up level after level of states, the kth level L_k consisting of all the states with k calls in progress.

We now construct a graph by drawing lines between states that differ from each other by exactly one call. (Such states, needless to say, are

always in successive levels of our diagram.) This graph we call the *state-diagram*. It is a natural (and standard) representation of the partial ordering $\leqq$ of the states: where x and y are states,

$$x \leqq y$$

means that y can be obtained from x by adding zero or more calls to x, or alternately, that x can be got from y by removing zero or more calls. The importance of this state-diagram lies in two facts:

i. The state diagram gives a geometrical representation of the possible states of the system. The myriad choking "details" of the connecting network have been converted into a vast geometrical structure with special properties. The operating system describes a trajectory through the state diagram, moving between levels as calls begin and end.

ii. Any stochastic process describing the operation of the connecting network is a point moving randomly on the state diagram. The motion is only between adjacent levels. New calls put into the network correspond to jumps to the next higher level; hangups correspond to jumps to the next lower level.

As a simple example, we consider the possible states of a single 2 by 2 switch. These consist of (*i*) the zero state, (*ii*) the four ways of having one call up, and (*iii*) the two ways of having two calls up. These states are depicted in Fig. 7. Fig. 8 shows the states of a 2 by 3 switch.

XII. THE RELEVANCE OF COMBINATORY AND STRUCTURAL PROPERTIES: EXAMPLES

In this section we elaborate, by discussing examples, our theme that the combinatory and structural properties of connecting systems are of

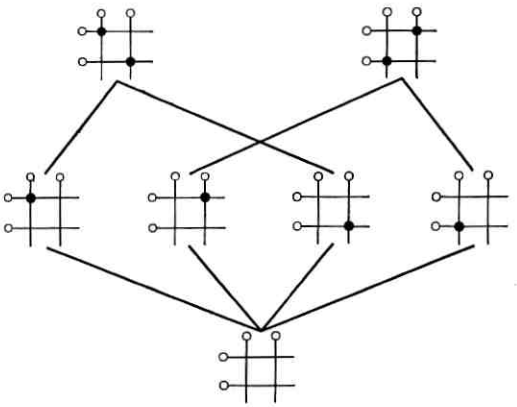


Fig. 7 — States of a 2 by 2 switch.

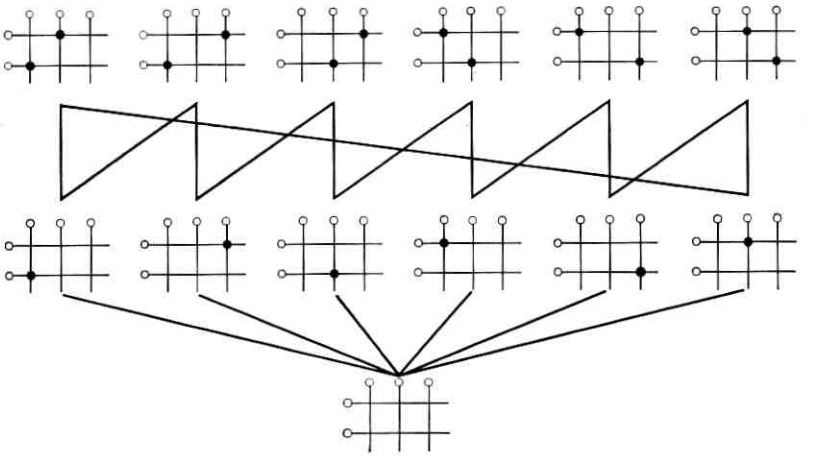


Fig. 8 — States of a 2 by 3 switch.

the greatest import (*i*) to their performance, and (*ii*) to the analysis of mathematical models of their operation. The organization of the control unit and the configuration of the connecting network largely determine the possible microscopic states of the system. Let us see what effects these features can have on problems of system analysis.

Example 1: Any connecting system has a "zero" or ground state in which all terminals and links are idle, no calls are being processed by the control unit, and the connecting network is empty. The existence of this zero state is a structural property common to all switching systems. This zero state seems most uninteresting. Nevertheless, many probabilistic models (for switching system operation) have the property that if the equilibrium probability of the zero state is known, then that of any other state can be determined in a simple way. Several specific examples of this phenomenon are worked out later in this paper, so none will be given here. (See Sections XV and XVI.)

Example 2: The relevance of combinatory properties of the connecting network to the calculation of probabilities can be vividly illustrated by reference to Clos' work on nonblocking networks (see Ref. 29). The blocking probability of a connecting network is the fraction of attempted calls that cannot be completed because no path for the call exists in the current state of the network. Until Clos' article appeared it was not generally known that, *no matter what probabilistic model was used*, an exact calculation of blocking probability for a Clos network with $m \geqq 2n - 1$ (see Fig. 3) would yield the value zero!*

* Zero, not zero factorial, which equals unity!

Example 3: Consider the class of connecting networks which have the property that in any state of the network, two idle terminals (forming an inlet-outlet pair) can be connected in at most one way. For each member of this class of networks we construct a Markov stochastic process to represent its operation under random traffic, as follows: in any state, if an inlet-outlet pair is idle, the conditional probability is $\lambda h + o(h)$ that it request connection in the next interval h, as $h \to 0$; also, an existing call terminates in the next interval h with a probability $h + o(h)$, as $h \to 0$; requests that encounter blocking are denied, and do not change the state of the system (lost calls cleared).

If X is a finite set, let $| X |$ be its cardinality, i.e., the number of elements of X, and let S be the set of all states of the network under discussion. For x in S, define

A_x = set of states accessible from x by adding a call

B_x = set of states accessible from x by removing a call

$| x |$ = number of calls in progress in state x

L_k = set of states with k calls in progress.

Note that $| B_x | = | x |$.

Let p_x be the stationary or equilibrium probability that the system is in state x. By reference to Fig. 9, it can be seen that the statistical equilibrium equations for our probabilistic model are

$$(\lambda | A_x | + | x |)p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y , \qquad x \in S.$$

Since in any state an idle pair can be connected in at most one way, no routing decisions need to be made, and the solution of this equation

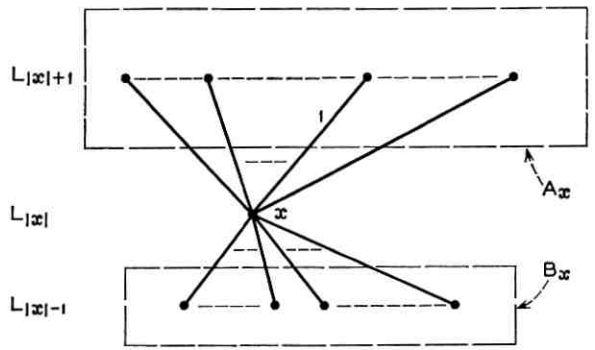


Fig. 9 — A state x, and the sets A_x, B_x in the state diagram.

(regardless of the network configuration!) is given by

$$p_x = p_0 \lambda^{|x|} \qquad x \neq 0$$

$$p_0^{-1} = 1 + \sum_{\substack{y \in S \\ y > 0}} \lambda^{|y|}$$

$$= \sum_{k \geqq 0} \lambda^k \mid L_k \mid$$

where 0 is the zero state. We have therefore shown that the simple combinatory property, that a call can be put up in at most one way, implies that the stationary probabilities of the Markov process we defined are of a simple geometric type. Note the important role played by the zero state, as discussed in Example 1.

Example 4: The Markov stochastic processes of the previous example can be used to illustrate another important point. There are many switching system models for which quantities of interest (such as the probability of blocking) can be given rigorously, without approximations, by a formula in which the distinction between system combinatorics and random customer behavior appears explicitly. In Example 3, the state probabilities $\{p_x, x \in S\}$ are completely determined by the quantities

$$\mid L_k \mid , \qquad k \geqq 0$$

i.e., by the number of states with k calls in progress, for $k \geqq 0$. For these models we can express the blocking probability as a function of the traffic parameter λ and of $\mid L_k \mid$, $k \geqq 0$. The numbers $\mid L_k \mid$ represent purely combinatory properties of the network.

The blocking probability b can be calculated as follows: b is the fraction of attempted calls that are unsuccessful, so that

$$1 - b = \frac{\text{total rate of successful attempts}}{\text{total rate of attempts}} .$$

In equilibrium, the total rate of successful attempts must equal the total rate of hang ups. The total rate of hang ups is

$$\sum_{x \in S} p_x \mid x \mid = \text{mean number of calls in progress}$$

(because the mean holding time is used as the unit of time). Let N be the number of terminals offering traffic. Since an idle inlet-outlet pair calls at a rate λ, the attempt rate in a state x is

$$\lambda \cdot (\text{number of idle pairs in a state } x) = \lambda \binom{N - 2 \mid x \mid}{2} .$$

The total rate of attempts is then

$$\lambda \sum_{x \epsilon S} p_x \binom{N - 2 \mid x \mid}{2}.$$

Hence,

$$b = 1 - \frac{\sum_{x \epsilon S} p_x \mid x \mid}{\lambda \sum_{x \epsilon S} p_x \binom{N - 2 \mid x \mid}{2}}$$

$$= 1 - \frac{\sum_{k>0}^{[N/2]} \lambda^k k \mid L_k \mid}{\lambda \sum_{k \geqq 0}^{[N/2]} \lambda^k \mid L_k \mid \binom{N - 2k}{2}},$$

where $[N/2]$ is the greatest integer less than or equal to $N/2$. This formula exhibits the blocking probability as a rational function of the calling rate λ per idle pair and as a bilinear function of the combinatory constants $\{\mid L_k \mid, k \geqq 0\}$. The degree of the denominator in λ is one more than that of the numerator, so $b \to 1$ as $\lambda \to \infty$; also, note that

$$\lim_{\lambda \to 0} b = 1 - \frac{\mid L_1 \mid}{\binom{N}{2}}.$$

This limit is greater than zero if there are calls which cannot be put up in *any* way. Finally, we observe that if the network is non-blocking, then

$$k \mid L_k \mid = \sum_{x \epsilon L_{k-1}} \binom{N - 2 \mid x \mid}{2}$$

$$= \mid L_{k-1} \mid \binom{N - 2k + 2}{2}$$

and so $b = 0$, as it should, if we interpret

$$\binom{N - 2[N/2]}{2}$$

as zero.

XIII. COMBINATORY, PROBABILISTIC, AND VARIATIONAL PROBLEMS

The preceding discussions have established that the ingredients going into a mathematical model of a connecting system are of two kinds.

On one hand are the combinatory and structural properties, and on the other, the probabilistic features of traffic. We emphasize the distinction between these aspects, and claim that by carefully drawing it, we can extend the general understanding of connecting systems, unify or modify existing theoretical methods, and obtain new engineering results.

Our discussion also suggests that to study stochastic processes that represent operating connecting systems, it is essential to have an extensive theory of the combinatory and topological nature of the microscopic states of such systems.

In any specific model of a connecting system, one can distinguish the combinatory from the stochastic features. However, it is also of interest to compare models of systems in an effort to determine optimal systems. These facts suggest a useful though imprecise division of the entire subject (of connecting system models) into three broad classes of problems. In order of priority, these are

i. Combinatory problems.
ii. Probabilistic problems.
iii. Variational problems.

This order of priority arises in a natural way: one needs to study combinatory problems in order to calculate probabilities; one needs both combinatory and stochastic information in order to design optimal systems.

The tripartite division just made provides a rational basis for organizing research effort. Since so many of our pronouncements have been generalities, we devote the remainder of the paper to illustrating carefully each of the three divisions (combinatory, probabilistic, variational) by working out and discussing in detail a very simple (yes, a trivial) problem from each division. These problems have been chosen for their tutorial value rather than their realism or usefulness. In discussing them, we place emphasis on furthering insight rather than solving practical problems, on exposing principles rather than providing engineering data.

XIV. A PACKING PROBLEM

It has long been suspected (and in some cases, verified experimentally) that routing calls through a connecting network "in the right way" can yield considerable improvements in performance. This procedure of routing the calls through the network is called "packing" (the calls), and the method used to choose routes is called a "packing rule." The use of the word "packing" in this context was surely suggested by an analogy with packing objects in a container. However,

the existence and description of packing rules that demonstrably improve performance (e.g., by minimizing the chance of blocking) are topics about which very little is known.

What, then, is the "right way" to route calls? It has been argued heuristically that it is better to route a call through the most heavily loaded part of the network that will still take the call. Appealing and simple as this rule is, nothing is known about it. We know of no published proof of either its optimality or its preferability over some other rule. The rule will be proven optimal for an example in Section XVI.

The question naturally arises, though, whether for a given network in which blocking can occur there exists a packing rule so cunning that by following it all blocking is avoided. Then, use of the rule makes the network nonblocking. Such a network may be termed *nonblocking in the wide sense*, while a network none of whose states has any blocked calls may be termed *nonblocking in the strict sense*.

The existence of such a rule is a purely combinatory property of the network, and so serves as an example of the first type of problem described in Section XIII. Unfortunately, *practically useful* connecting networks that are nonblocking in the wide sense are yet to be found. Since we are primarily interested in exemplifying principles, we shall be content with discussing an impractical network that is nonblocking in the wide sense. The example to be given was suggested by E. F. Moore.*

Let us first consider the three-stage connecting network depicted in Fig. 10. All switches in the middle column are 2 by 2, and there are $2n - 1$ of them, so, by a result of C. Clos,[29] the network is nonblocking. Suppose that we use the rule that an empty middle switch is not to be used unless there is no partially filled middle switch that will take the call. In other words, do not use a fresh middle switch unless you have to! In general, this rule is not quite the same as the one exhorting use of the heavily loaded switches wherever possible, because it only tells us what to avoid, but it is in the same spirit. In the case to be considered, however, a middle switch is either empty, half-full, or full; hence the two rules coincide.

We shall show that if this rule is used, then no more than $[3n/2]$ middle switches are *ever* used, where $[x]$ is the greatest integer less than or equal to x. Thus the rest, about one quarter of the middle switches, could be removed and no blocking would result if the rule were used. It can be verified by examples that if there are only $[3n/2]$ middle switches and the rule is violated, then calls can be blocked. Thus, the network of

* Private communication.

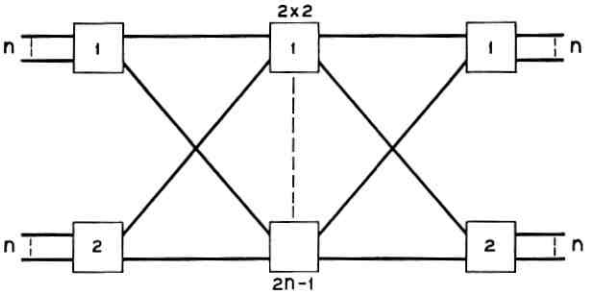


Fig. 10 — Three-stage nonblocking connecting network (Clos type).

Fig. 11 is not nonblocking in the strict sense, but is nonblocking in the wide sense.

A state x of a connecting network is called reachable (under a rule ρ) if using the rule ρ to make routing decisions does not prevent the system from reaching x from the zero state. We set

$$S(x) = \text{number of middle switches in use in state } x.$$

Let us use the diagram of Fig. 12 as a canonical representation for a 2 by 2 middle switch. The numbers at the left [top] indicate to which outer switch on the left [right] the numbered link connects. The seven possible states of a middle switch are depicted in Fig. 13, and are indexed therein by letters $a,b, \cdots, g$. A state x may then be represented (to within renaming switches and terminals) by giving seven integers $a(x), b(x), \cdots, g(x)$ where

$$a(x) = \text{number of middle switches of type } a \text{ when network is in state } x$$

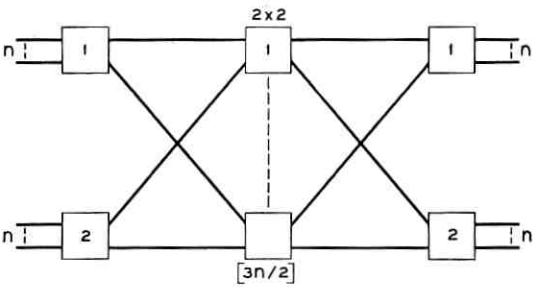


Fig. 11 — Three-stage network which is nonblocking if proper routing is used.

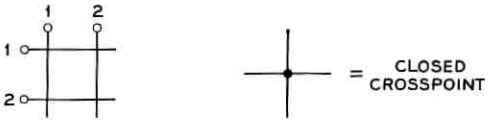


Fig. 12 — Representation of a 2 by 2 middle switch.

$\vdots$

$g(x)$ = number of middle switches of type g when network is in state x.

It is clear that for any state x

$$a(x) + b(x) + \cdots + g(x) = 2n - 1$$

$$b(x) + c(x) + \cdots + g(x) = S(x).$$

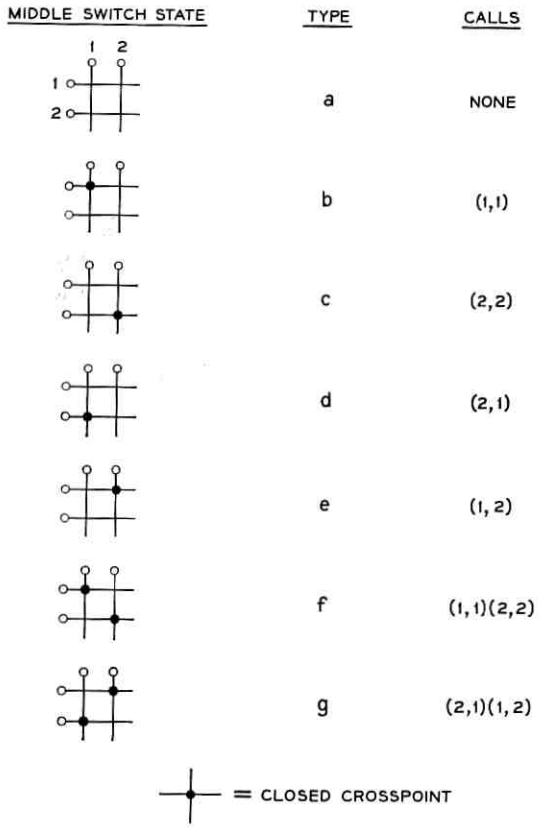


Fig. 13 — Seven possible states of a middle switch.

Theorem 1: *Let* ρ *denote the rule: Do not use an empty middle switch unless necessary. Let* x *be a state of the network of Fig. 10. Let* x *be reachable under* ρ. *Then for* $n \geqq 2$

$$S(x) \leqq [3n/2] \tag{1}$$

$$\left.\begin{aligned} b(x) + c(x) + f(x) &\leqq n \\ d(x) + e(x) + g(x) &\leqq n \end{aligned}\right\} \tag{2}$$

Proof: Each reachable state is reachable in a certain minimum number of steps. The theorem is true if x consists of one call and is reachable from the zero state in one step. As an hypothesis of induction, assume that the theorem is true for all states reachable in k steps or fewer. All changes in the state are either hangups, or new calls of the following kinds:

Type 1:

$$\begin{array}{lllll} & a(y) \to a(y) - 1 & & & \\ (1,1) & b(y) \to b(y) + 1 & \text{with} & c(y) = 0 & \\ (2,2) & c(y) \to c(y) + 1 & \text{with} & b(y) = 0 & \\ (2,1) & d(y) \to d(y) + 1 & \text{with} & e(y) = 0 & \\ (1,2) & e(y) \to e(y) + 1 & \text{with} & d(y) = 0. & \end{array}$$

Type 2: (preferred by ρ)

$a(y)$ remains fixed and

$$\begin{array}{llllll} (1,1) & f(y) \to f(y) + 1, & c(y) \to c(y) - 1 & \text{with} & c(y) > 0 \\ (2,2) & f(y) \to f(y) + 1, & b(y) \to b(y) - 1 & \text{with} & b(y) > 0 \\ (2,1) & g(y) \to g(y) + 1, & e(y) \to e(y) - 1 & \text{with} & e(y) > 0 \\ (1,2) & g(y) \to g(y) + 1, & d(y) \to d(y) - 1 & \text{with} & d(y) > 0. \end{array}$$

All states, reachable or not, satisfy the inequalities

$$\begin{aligned} b(y) + e(y) + f(y) + g(y) &\leqq n \\ c(y) + d(y) + f(y) + g(y) &\leqq n \\ b(y) + d(y) + f(y) + g(y) &\leqq n \\ c(y) + e(y) + f(y) + g(y) &\leqq n. \end{aligned}$$

The alternative preferred by ρ changes neither the value of $S(\cdot)$ nor the truth of (2) of the theorem. Consider a state x first reachable in

$k + 1$ steps. If x is first reachable by a hangup or by putting up a call of Type 2, then (1) and (2) are true of x. Suppose then that x is first reachable in $k + 1$ steps only by putting up a call of Type 1. Without loss of generality we can consider only the case where the new call is a (1, 1) call; the other three cases are symmetric. Let y be a state from which x is thus first reachable. Since the avoided alternative is used, we have

$$c(y) = 0.$$

Since a (1, 1) call is possible in state y, we must have

$$b(y) + d(y) + f(y) + g(y) \leqq n - 1$$
$$b(y) + e(y) + f(y) + g(y) \leqq n - 1$$

and from the induction hypothesis

$$d(y) + e(y) + g(y) \leqq n.$$

Hence,

$$2\{b(y) + d(y) + e(y) + f(y) + g(y)\} \leqq 3n - 2$$

or, since $c(y) = 0$

$$S(y) \leqq \frac{3n}{2} - 1.$$

However, $S(x) = S(y) + 1$, so $S(x) \leqq [3n/2]$. To show that (2) also holds of x consider that

$$b(y) + e(y) + f(y) + g(y) \leqq n - 1$$
$$c(y) = 0.$$

It follows that

$$b(y) + c(y) + f(y) \leqq n - 1.$$

However, since x is obtained from y by putting up a (1, 1) call of Type 1, we have

$$b(x) = b(y) + 1, \qquad e(x) = e(y)$$
$$c(x) = c(y) = 0, \qquad f(x) = f(y)$$
$$d(x) = d(y), \qquad g(x) = g(y).$$

Hence, (2) of Theorem 1 is true of x. This proves the result.

XV. A PROBLEM OF TRAFFIC CIRCULATION IN A TELEPHONE EXCHANGE

We shall describe and analyze a simple stochastic model for the operation of the control unit of a switching system. The connecting network is assumed to be nonblocking and is left out of account.

To set up a telephone call in a modern electromechanical automatic exchange usually involves a sequence of steps which are (traditionally and functionally) divided into two groups. The first group consists in collecting in a *register* the dialed digits of the called terminal. The second group, performed by a machine called a *marker*, consists in actually finding a path through the connecting network for the desired call, or otherwise disposing of the request for service. For even if a path to the called terminal be found, this terminal may already be busy.

In the exchange, enough registers and markers must be provided to give customers a prescribed grade of service. For engineering purposes, then, it is desirable to know the probability that r registers and m markers are busy. Let us assume that the exchange serves N customers, and that there are R registers and M markers. All calls are assumed to go to terminals *outside* the exchange.

We may think of each customer's line as being in one of a number of conditions, and moving from one condition to another. It makes no difference whether we ascribe these "conditions" to the line itself, or to a fictitious single customer if several people use the line. A given line may be *idle* (i.e., not in use); at some point in time it may request a connection, i.e., the customer picks up the receiver and starts *waiting for dial tone*; after obtaining a register he spends a certain amount of time *dialing*; he then *waits for a marker* to complete his call (freeing the register meanwhile); upon obtaining a marker, he must wait until the marker *completes* the connection; at this point he begins his *conversation*; at the end of his conversation his line becomes *idle* again.

One may now ask, what is the distribution of the N customers among these various conditions? Clearly, if not enough markers are provided there will be a tendency for the customers to collect in the "waiting for a marker" condition; a lack of registers will make the customers collect in the "waiting for dial tone" condition.

To obtain a simple probabilistic model for the "circulation" of customers, we assume that the probability that an idle customer starts a call in the next interval of time of length h is $\lambda h + o(h)$, the chance that a dialing customer completes his dialing in the next interval h is $\delta h + o(h)$, the chance that a busy marker finishes the call it is working on is $\mu h + o(h)$, and the probability that a conversation ends is $h + o(h)$,

all as $h \to 0$. The probability of more than one such event in h is $o(h)$ as $h \to 0$.

These assumptions are in turn consequences of assuming that the time a customer stays idle, the time a customer takes to dial, the time a marker takes to complete a call, and the holding time (conversation length) are all mutually independent random variables, each with a negative exponential distribution, and the respective means λ^{-1}, δ^{-1}, μ^{-1} and unity. The number λ is the calling rate per idle customer, δ and μ are the average rates of dialing and call completion by a marker (respectively), and time is measured in units of mean holding time, so that the hangup rate per call in progress is unity. The assumption that the marker operation times are exponentially distributed is not realistic, but we make it here in the interest of obtaining a global model whose statistical equilibrium equations can be solved in a simple way. This restrictive assumption could be avoided at the cost of complicating the mathematics. The important features of our model are depicted in Fig. 14; the labeled arrows indicate the rates of motion for various transitions.

The state of the system is adequately described by stating the number i of idle customers, the number r of customers that are dialing or waiting for dial tone, the number m that are being serviced by a marker or are waiting for a marker, and the number c of calls in progress. Actually, any three of these numbers suffice, since for physically meaningful states

$$i + r + m + c = N.$$

Let p_{irmc} be the equilibrium (or stationary) probability of the state

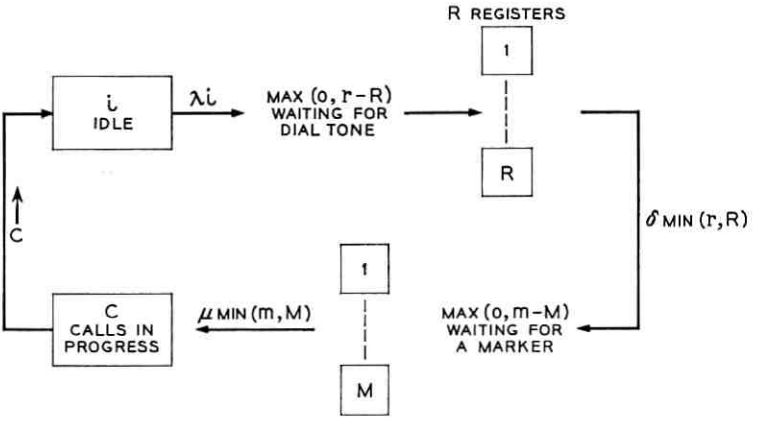


Fig. 14 — Diagram of a telephone system.

(i,r,m,c). The "statistical equilibrium" equations are, with suitable conventions at the boundaries.

$$(\lambda i + \delta \min(r,R) + \mu \min(m,M) + c)p_{irmc}$$
$$= (c+1)p_{(i-1)rm(c+1)} + \lambda(i+1)p_{(i+1)(r-1)mc}$$
$$+ \delta \min(r+1,R)p_{i(r+1)(m-1)c} + \mu \min(m+1,M)p_{ir(m+1)(c-1)}.$$

These equations state that the average rate at which a state is left equals the average rate at which it is reached from other states. We observe that the flow of calls in the exchange is in a sense *cyclic*; in making a call, each customer passes through four stages: idle, dialing, marker, conversation, then back to idle, in that order. This fact yields a way of solving the equations. Each side of the equilibrium equations has four terms, one for each of the four stages of a call. We shall find a way of assigning to each term on the left a corresponding *equal* term on the right which will cancel it.

The solution of the equations for $(i,r,m,c) \neq (N,0,0,0)$ is proportional to

$$f_{i,r,m,c} = \frac{N!}{i!r!m!c!} \cdot \frac{\prod_{j=0}^{r} \max(1,j/R) \prod_{j=0}^{m} \max(1,j/M)}{\lambda^i \delta^r \mu^m}.$$

The constant of proportionality is the probability of the "zero" state

$$p_{N000} = \left(1 + \sum_{\substack{i+r+m+c=N \\ i,r,m,c \geqq 0 \\ i<N}} f_{i,r,m,c}\right)^{-1}$$

obtained from the normalization condition for probabilities. The algebraic character of the solution is closely analogous to the actual pattern of circulating traffic in Fig. 14, for the easiest way of showing that f_{irmc} is actually a solution of the statistical equilibrium equations is to make the following correspondence between terms on opposite sides of the equations:

$$\lambda i p_{irmc} \sim (c+1)p_{(i-1)rm(c+1)}$$
$$\delta \min(r,R)p_{irmc} \sim \lambda(i+1)p_{(i+1)(r-1)mc}$$
$$\mu \min(m,M)p_{irmc} \sim \delta \min(r+1,R)p_{i(r+1)(m-1)c}$$
$$cp_{irmc} \sim \mu \min(m+1,M)p_{ir(m+1)(c-1)}.$$

It can be seen that each term on the left cancels the corresponding

one on the right when f_{irmc} is substituted. Each term represents the (total) rate of occurrence of one of the four kinds of possible event: request for service, completion of dialing, completion of a call, and hangup. In the life history of a given call, these events occur in the natural cyclic order given. Events associated with corresponding (i.e., canceling) terms are next to each other in this cyclic order.

XVI. AN OPTIMAL ROUTING PROBLEM

Our final example is a variational problem involving both combinatoric and probability. We shall exhibit some particular answers to the following question: If requested connections can be put up in a connecting network by several different routes, leading to different states, which routes should be chosen so as to minimize the probability of blocking? This question poses a variational problem in which many possible methods of operating a connecting network of given structure are compared, rather than one in which different network structures are compared.

We shall consider this question for a connecting network that is of little practical significance because it is obviously wasteful of crosspoints. Its virtues, however, are that it is perhaps the simplest network for which our question can be asked, and that it clearly exhibits the principles and arguments involved, so that these can be understood. The network is shown in Fig. 15, the squares standing for square 2 by 2 switches.

The possible states of this network are determined by all the ways

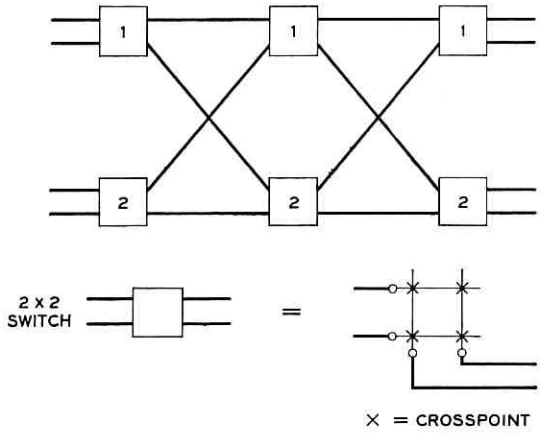


Fig. 15 — A simple network in which optimal routing is studied.

in which four or fewer inlets on the left can be connected pairwise to as many outlets on the right, no inlet being connected to more than one outlet, and vice-versa. These possible states are depicted in a natural arrangement in Fig. 16; states which differ only by permutations of customers or switches have been identified in order to simplify the diagram. That is, there is essentially only one way to put up a single call, there are four ways of having two calls up, two ways each of having three and four calls up. These "ways" have been arranged in rows according to the number of calls in progress, and lines have been drawn between states that differ from each other by only the removal or addition of exactly one call.

For ease of reference, let us number the states in the (partly arbitrary) way indicated in Fig. 16; insofar as possible, we have used small num-

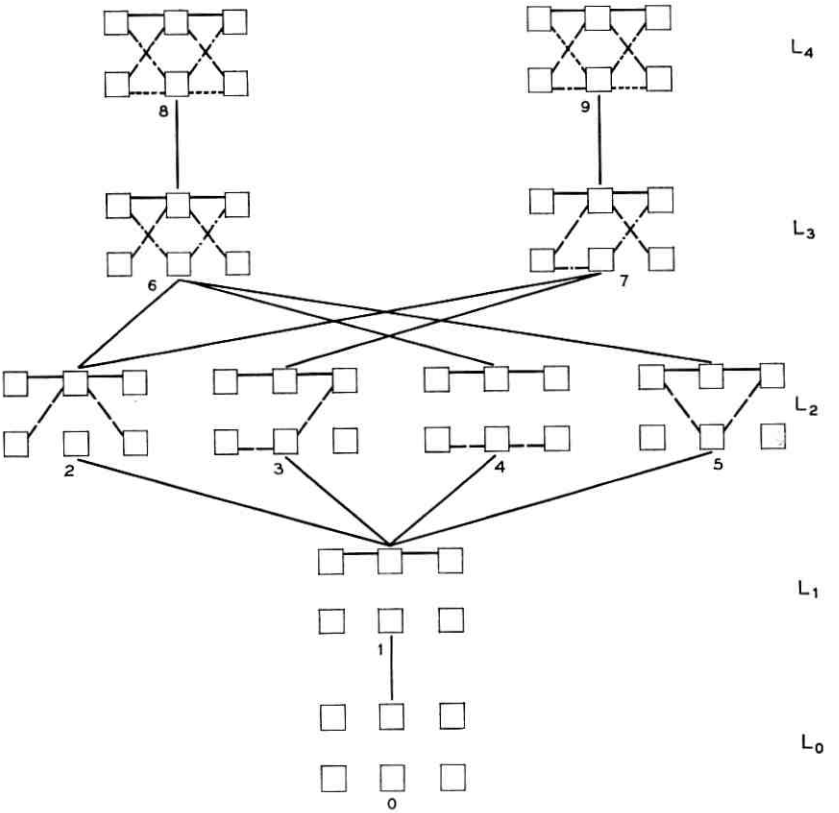


Fig. 16 — (Reduced) state diagram for the network shown in Fig. 15.

bers for states with small numbers of calls. The set of possible states of our example then consists of (essentially) ten different configurations of calls in the basic network of Fig. 15. The state diagram, with each state identified now only by its number within a small circle, is schematized in Fig. 17. Also indicated in this schema are two important sets of quantities associated with the states. To the left of each state is the number of idle inlet-outlet pairs, and to the right of each state is the number of idle inlet-outlet pairs that can actually be connected, i.e., that are not blocked.

Only in the state numbered 4 are there any blocked calls. It is to be noticed that state 4 realizes essentially the same assignment of inlets to outlets as state 2, which has no blocked calls. The difference between the two is that in state 2 all the traffic passes through one middle switch, leaving the other entirely free for any call that may arise. Clearly, then, this difference illustrates the "packing rule" that one should always put through a call using the most heavily loaded part of the network that will still accept the call.

The question naturally arises, therefore, whether this packing rule is

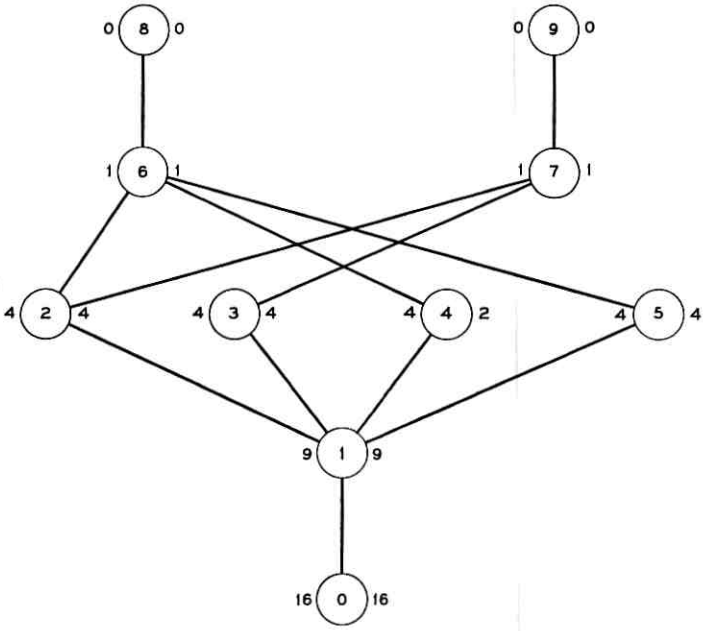


Fig. 17 — Schema of state diagram.

in any sense optimal for our particular example. We shall prove that it is, in two senses. It is clear from an inspection of the state diagram that only in state 1 is there ever a choice of route, and that this choice is always between states 2 and 4. From the fact that state 4 is the only state with any blocked calls, it is intuitively reasonable to expect that the probability of blocking is the least if the "bad" state 4 is avoided as much as possible, i.e., if from state 1 we always pass to either 2, 3, or 5, and visit 4 only when we have to, via a hangup from state 6.

The next task is to choose a probabilistic model for the operating network; this will be done in the simplest possible way. We postulate that in any state of the system, the probability that a given idle inlet-outlet pair request connection in the next interval of time h is $\lambda h + o(h)$, the chance that an existing connection cease is $h + o(h)$, and the chance that more than one event (new call or hangup) occur in h is $o(h)$, as $h \to 0$. The number λ is the calling rate per idle pair, and time is measured in units of mean holding time, so the "hangup" rate is unity. New calls that are not blocked are instantly connected, with some specific choice of route, while blocked calls are lost and do not affect the state of the system, their terminals remaining in the idle condition.

To complete the probabilistic description of the behavior of the system, it remains to specify how routes are chosen. In our example, this amounts to specifying whether, for certain calls arising in state 1, the route leading to state 2 or that leading to 4 is chosen. At first we shall only consider methods of choice that are independent of time, i.e., the choice is made in the same way each time.

The methods of choice over which we shall take an optimum may be parametrized as follows: each time a choice is to be made between going to state 4 and state 2, a coin is tossed with a probability α of coming up heads. If a head comes up we choose state 4; if a tail, we choose state 2; the toss of the coin is independent of previous tosses and of the history of the system. The parameter α may take on any value in the interval $0 \leqq \alpha \leqq 1$; the value $\alpha = 0$ corresponds to choosing state 2 every time; the value $\alpha = 1$ corresponds to choosing state 4 every time; a value of α intermediate between 0 and 1 means that 4 is chosen over 2 a fraction α of the time.

Introducing a natural terminology (from the theory of games), we may say that a choice of α represents a *policy* or *strategy* for making routing decisions; a value 0 or 1 of α represents a *pure strategy*, in which the route is specified by a rigid rule, and there is no randomization; an intermediate value of α represents a *mixed strategy*.

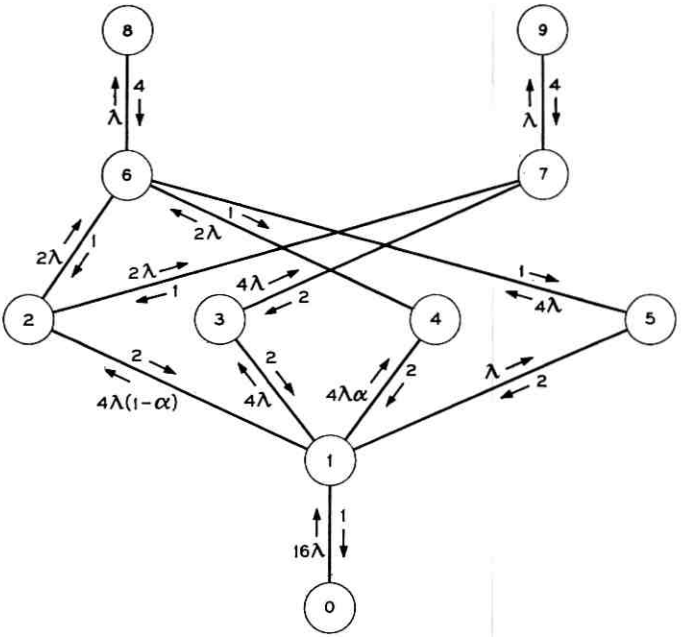


Fig. 18 — Schema of state diagram showing transition rates.

A choice of α determines a matrix $Q = Q(\alpha,\lambda)$ of transition rates (Fig. 18) among states of the system, and so a Markov stochastic process taking values on those states. As a measure of performance we shall use the fraction b *of requests for connection that encounter blocking*, defined as follows: let $b(t)$ be the number of blocked calls occurring in the interval $(0,t]$; and let $r(t)$ be the number of requests for service occurring in $(0,t]$; then

$$b = \lim_{t\to\infty} \frac{b(t)}{r(t)}.$$

It can be shown that this limit exists and is constant with probability one, so b is well defined.

The number $b = b(\alpha,\lambda)$ can be calculated from the matrix Q as follows: if $(i,i = 0, \cdots, 9)$ is a state, let $\beta(i)$ be the number of blocked idle pairs in state i, and let $\gamma(i)$ be the number of calls in existence in state i. The stationary state probabilities $\{p_i, i = 0, \cdots, 9\}$ exist and are the unique solution of the matrix-vector equation $Qp = 0$. Then b is given by

$$b = \frac{\sum_{i=0}^{9} p_i \beta(i)}{\sum_{i=0}^{9} p_i [4 - \gamma(i)]^2}$$

$$= \frac{2p_4}{\sum_{i=0}^{9} p_i \delta(i)}, \quad \text{with} \quad \delta(i) = [4 - \gamma(i)]^2$$

$$= \frac{(p,\beta)}{(p,\delta)}$$

where the inner product (p,x) is $\sum_{i=0}^{9} p_i x_i$.

We may therefore formally state our variational problem for this example as follows: to find that α in the interval $0 \leqq \alpha \leqq 1$ for which the ratio

$$b = \frac{(p,\beta)}{(p,\delta)} = \text{minimum}$$

subject to the conditions $Qp = 0$, $\sum_{i=0}^{9} p_i = 1$.

It is natural to expect that in choosing an optimum routing method in the example above there is no point in randomizing, i.e., using a mixed strategy with α unequal to either 0 or 1. That this is so is not obvious from our mathematical statement of the problem, and requires proof. We shall demonstrate a more general result:

Theorem 2: *Let x and y be vectors of* 10 *dimensions, with y nonnegative and not identically zero.*

$$\begin{matrix}\min \\ \text{or} \\ \max\end{matrix} \left\{ \frac{(p,x)}{(p,y)} \;\middle|\; Qp = 0, \quad \sum_{i=0}^{9} p_i = 1, \quad 0 \leqq \alpha \leqq 1 \right\}$$

is always achieved for $\alpha = 0$ or $\alpha = 1$.

Proof: The equation $Qp = 0$ may be written out in the detailed form

$$(i) \qquad 16\lambda p_0 = p_1$$

$$(ii) \qquad (9\lambda + 1)p_1 = 16\lambda p_0 + 2\sum_{i=2}^{5} p_i$$

$$(iii) \qquad (4\lambda + 2)p_2 = 4\lambda(1 - \alpha)p_1 + p_6 + p_7$$

$$(iv) \qquad (4\lambda + 2)p_3 = 4\lambda p_1 + p_7$$

$$(v) \qquad (2\lambda + 2)p_4 = 4\lambda\alpha p_1 + p_6$$

$$(vi) \qquad (4\lambda + 2)p_5 = \lambda p_1 + p_6$$

$$(vii) \qquad (\lambda + 3)p_6 = 2\lambda p_2 + 2\lambda p_4 + 4\lambda p_5 + 4p_9$$

$$(viii) \qquad (\lambda + 3)p_7 = 2\lambda p_2 + 4\lambda p_3 + 4p_8$$

$$(ix) \qquad 4p_8 = \lambda p_6$$

$$(x) \qquad 4p_9 = \lambda p_7 .$$

These are the standard "statistical equilibrium" equations for the probabilistic model we have assumed. They can be solved by successively eliminating every p_i except p_0 and obtaining a solution of the form

$$p_i = f_i p_0 , \qquad i \neq 0.$$

The value of p_0 is then determined by the normalization condition $\sum_{i=0}^{9} p_i = 1$ as

$$p_0 = \frac{1}{1 + \sum_{i=1}^{9} f_i} .$$

The f_i are of course functions of λ and α. We shall prove that they are *linear* functions of the parameter α.

We first eliminate p_1 and note that $f_1 = 16\lambda$. Since the relations (iii)–(iv) contain the variables $\{p_i , i = 2,3,4,5\}$ only on the left, these variables may be eliminated entirely from (ii), and from (vii)–(x). But substitution for these variables in (vii) and $(viii)$ in terms of (iii)–(vi) introduces α and p_0 only in inhomogeneous terms. Hence, f_6 and f_7 are linear in α, and so all $\{f_i , i = 1, \cdots , 9\}$ are linear in α.

Clearly, we have

$$\frac{(p,x)}{(p,y)} = \frac{(f,x)}{(f,y)}$$

because the normalization terms $1 + \sum_{i=1}^{9} f_i$ cancel out, and so it follows that $(p,x)/(p,y)$ is a *bilinear* function of α, i.e., it has the form

$$g(\alpha) = \frac{A_1 + B_1\alpha}{A_2 + B_2\alpha}$$

where A_1, A_2, B_1, and B_2 are constants. Now

$$\frac{d}{d\alpha} g(\alpha) = \frac{B_1(A_2 + B_2\alpha) - B_2(A_1 + B_1\alpha)}{(A_2 + B_2\alpha)^2}$$

$$= \frac{B_1A_2 - B_2A_1}{(A_2 + B_2\alpha)^2}$$

which is of the same sign as its numerator. Thus $g'(\alpha)$ is either always nonpositive or nonnegative, and so any extremum of $g(\alpha)$ in $0 \leqq \alpha \leqq 1$ is assumed at the boundary, either for $\alpha = 0$ or $\alpha = 1$. Since the solution p of $Qp = 0$ is known to have all strictly positive components for all α in the unit interval, we have $A_2 + B_2\alpha = (p,y) > 0$.

It follows in particular that the minimum of blocking probability b is achieved for $\alpha = 0$ or $\alpha = 1$. It is unthinkable that visiting a blocking state (state 4) more frequently should decrease b, so we conjecture (and shall shortly prove that) α should be zero rather than one.

Before doing this though, let us observe that there is only one blocking state (viz., 4), and that the blocking probability b can be written as

$$b = \frac{2p_4}{16p_0 + 9p_1 + 4\sum_{i=2}^{5} p_i + p_6 + p_7}.$$

These facts and our intuition suggest that b should be a monotone increasing function of

$$f_4 = \frac{p_4}{p_0}.$$

This conjecture is correct, and provides an easy way of showing that $\alpha = 0$ gives the least blocking probability. Let us prove it.

From (*i*) and (*ii*) we find that

$$\sum_{i=2}^{5} p_i = 8\lambda(9\lambda + 1)p_0 - 2\lambda p_0 = 72\lambda^2 p_0$$

whence

$$b = \frac{2f_4}{16 + 144\lambda + 288\lambda^2 + f_6 + f_7}.$$

From (*vii*)–(*x*) we find that

$$p_6 + p_7 = \frac{1}{\lambda + 3}\left(\lambda(p_6 + p_7) + 4\lambda\sum_{i=2}^{5} p_i - 2\lambda p_4\right)$$

$$= \tfrac{4}{3}\lambda\sum_{i=2}^{5} p_i - \tfrac{2}{3}\lambda p_4$$

$$= 96\lambda^3 p_o - \tfrac{2}{3}\lambda p_4.$$

Therefore

$$b = \frac{2f_4}{16 + 144\lambda + 556\lambda^2 + 192\lambda^3 - \frac{2}{3}\lambda f_4}.$$

This is of the form

$$\frac{2x}{a - cx}$$

where a and c are strictly positive constants. Now

$$\frac{d}{dx}\frac{2x}{a - cx} = \frac{2}{a - cx} + \frac{2cx}{(a - cx)^2}$$

$$= \frac{2a}{(a - cx)^2} \geqq 0.$$

Hence, b is a monotone increasing function of f_4. It follows that b is a minimum if f_4 is a minimum.

To prove that the blocking probability b is a minimum for $\alpha = 0$, it remains to calculate p_4 from the equilibrium equations. By eliminating all the equilibrium probabilities except p_6 and p_7, we find

$$p_6 = \frac{1}{\lambda + 3}\left(\frac{8\lambda^2(1 - \alpha)16\lambda p_0}{4\lambda + 2} + \frac{8\lambda^2\alpha 16\lambda p_0 + 2\lambda p_6}{2\lambda + 2} + \frac{4\lambda^2 16\lambda p_0 + 4\lambda p_6}{4\lambda + 2} + \lambda p_7\right)$$

$$p_7 = \frac{1}{\lambda + 3}\left(\frac{8\lambda^2(1 - \alpha)16\lambda p_0}{4\lambda + 2} + \frac{(16\lambda)^2\lambda p_0 + 4\lambda p_7}{4\lambda + 2} + \lambda p_6 + \frac{2\lambda}{4\lambda + 2}(p_6 + P_7)\right).$$

We have purposely not simplified the terms so that their origin can be verified. From these two equations we find that

$$f_6 = \frac{p_6}{p_0}$$

$$= X^{-1}128\lambda^3\left(\frac{1 - \alpha}{4\lambda + 2} + \frac{\alpha}{2\lambda + 2} + \frac{1}{8\lambda + 4} + \frac{\lambda\left(\frac{1 - \alpha}{4\lambda + 2} + \frac{1}{2\lambda + 1}\right)}{\lambda + 3 - \frac{3\lambda}{2\lambda + 1}}\right)$$

where

$$X = \lambda + 3 - \frac{\lambda}{\lambda + 1} - \frac{2\lambda}{2\lambda + 1} - \frac{2\lambda^3 + 2\lambda^2}{2\lambda^3 + 4\lambda + 3}$$

$$= \frac{2\lambda^3 + 5\lambda^2 + 7\lambda + 3}{2\lambda^2 + 3\lambda + 1} - \frac{2\lambda^3 + 2\lambda^2}{2\lambda^3 + 4\lambda + 3}$$

$$> 0.$$

The coefficient of α in f_6 is

$$\frac{128\lambda^3}{2\lambda + 2}\left(1 + \frac{2\lambda + 2}{4\lambda + 2}\left(1 + \frac{\lambda}{\lambda + 3 - \dfrac{3\lambda}{2\lambda + 1}}\right)\right).$$

This is positive, because

$$1 - \frac{2\lambda + 2}{4\lambda + 2}\left(1 + \frac{\lambda}{\lambda + 3 - \dfrac{3\lambda}{2\lambda + 1}}\right) = 1 - \frac{\lambda + 1}{2\lambda + 1}\left(\frac{4\lambda^2 + 5\lambda + 3}{2\lambda^2 + 4\lambda + 3}\right)$$

$$= 1 - \frac{4\lambda^3 + 9\lambda^2 + 8\lambda + 3}{4\lambda^3 + 10\lambda^2 + 10\lambda + 3}.$$

However,
$$\frac{p_4}{p_0} = \frac{32\lambda^2\alpha}{\lambda + 1} + \frac{p_6/p_0}{2\lambda + 2}.$$

Hence,
$$\frac{df_4}{d\alpha} > 0.$$

We shall now consider the problem of optimal routing in our (trivial) network from a different point of view. Instead of minimizing the *ratio* of unsuccessful attempts to attempts, let us simply minimize the average number of unsuccessful attempts in any finite number of events, counting changes of state and unsuccessful attempts as events.

In our example, the only choice is between states 2 and 4, when a particular call requests connection in state 1. By a *policy*, let us mean a function $p(\cdot)$ on the nonnegative integers taking the values 0 and 1. Let x_n be the state of the network after n events, $n \geqq 0$. We say that the system is operated according to policy $p(\cdot)$ if, for each $n \geqq 0$, given that $x_n = 1$ and a choice occurs, the system moves to

state 2 if and only if $p(n) = 1$

state 4 if and only if $p(n) = 0$.

Now our intuitive feeling is that going to state 2 is preferable over

going to state 4 under all circumstances. At the cost of anticipating results to be proven, let us partially order all the possible policies by the definition: If $p(\cdot)$ and $q(\cdot)$ are policies, then

$$p \geqq q \quad \text{if and only if} \quad p(n) \geqq q(n) \quad \text{for all} \quad n \geqq 0.*$$

The shift transformation T of policies $p(\cdot)$ is defined by the condition

$$Tp(n) = p(n + 1) > n \geqq 0.$$

It is evident that $p \geqq q$ implies $Tp \geqq Tq$. Let $E_{0,p}(x) \equiv 0$, and define

$$E_{n,p}(x) = E\left\{\begin{array}{l}\text{number of unsuccessful attempts after } n \text{ events} \\ \text{starting from state } x \text{ if the system is operated ac-} \\ \text{cording to policy } p(\cdot)\end{array}\right\}$$

Let S be the set of states $\{0, 1, \cdots, 9\}$.

We shall prove

Theorem 3: *If* $p \geqq q$, *then for all* $n \geqq 1$ *and* $x \in S$

$$E_{n,p}(x) \leqq E_{n,q}(x).$$

As a preliminary result (not without its own interest) we shall need the

Lemma: *For* $n \geqq 1$ *and any policy* $p(\cdot)$

$$E_{n,p}(4) = \max_{x \in S} E_{n,p}(x).$$

This says that starting in the (sole blocking) state 4 is always the worst way to start, no matter how long we run the system.

Proof: For $n = 1$ and $x \neq 4$, $E_{n,p}(x) = 0$ since no unsuccessful attempts can occur in any state except 4. However,

$$E_{1,p}(4) = \frac{2\lambda}{2 + 2\lambda}$$

so the lemma is true for $n \leqq 1$. Assume as an hypothesis of induction that it is true for $n \leqq k$. Now for $x \neq 4$, $E_{k+1,p}(x)$ is a convex combination of values of $E_{k,Tp}(\cdot)$, so clearly for $x \neq 4$

$$E_{k+1,p}(x) \leqq \max E_{k,p}(y) = E_{k,p}(4).$$

However, elementary probability arguments establish that

$$E_{k+1,p}(4) = E_{k,p}(4) + Pr\{x_k = 4 \mid x_0 = 4\} E_{1,T^k p}(4)$$

so the lemma is proven.

* Read "$p \geqq q$" as "p is better than q"!

Proof of Theorem 3: For any policy $s(\cdot)$

$$E_{1,s}(x) = 0 \quad \text{if} \quad x \neq 4$$

$$E_{1,s}(4) = \frac{2\lambda}{2 + 4\lambda}.$$

Hence,

$$E_{1,p}(x) = E_{1,q}(x) \quad \text{for all} \quad x \in S.$$

Assume as an hypothesis of induction that $p \geqq q$ implies

$$E_{n,p}(x) \leqq E_{n,q}(x)$$

for all x and all $n \leqq k$. Now for $x \neq 4$ or 1 and any policy $s(\cdot)$, $E_{k+1,s}(x)$ is a convex combination of values of

$$E_{k,Ts}(\cdot).$$

For $x = 4$, we have for any policy $s(\cdot)$

$$E_{k+1,s}(4) = \frac{2\lambda}{2 + 4\lambda} + \text{convex combination of } E_{k,Ts}(\cdot)$$

where the coefficients of the convex combination are transition probabilities independent of the policy $s(\cdot)$, and

$$\frac{2\lambda}{2 + 4\lambda} = \Pr\left\{\begin{matrix}\text{first event is a} \\ \text{blocked attempt}\end{matrix} \text{ start in state 4}\right\}.$$

Hence, $p \geqq q$ and $x \neq 1$ implies

$$E_{k+1,p}(x) \leqq E_{k+1,q}(x)$$

For $x = 1$ and any policy $s(\cdot)$ we have

$$E_{k+1,s}(1) = \frac{4\lambda}{1 + 9\lambda}\{s(1)E_{k,Ts}(2) + [1 - s(1)]E_{k,Ts}(4)\} + \frac{1 + 5\lambda}{1 + 9\lambda} \text{ convex combination of } E_{k,Ts}(\cdot)$$

where the coefficients of the convex combination are independent of $s(\cdot)$, and

$$\frac{4\lambda}{1 + 9\lambda} = \Pr\left\{\begin{matrix}\text{first event requires} \\ \text{routing decision}\end{matrix} \text{ start in state 1}\right\}.$$

Suppose now that $p \geqq q$. It is sufficient to show that

$$p(1)E_{k,\tau p}(2) + [1 - p(1)]E_{k,\tau p}(4)$$
$$\leqq q(1)E_{k,\tau q}(2) + [1 - q(1)]E_{k,\tau q}(4).$$

If $p(1) = q(1)$, this follows from the hypothesis of induction. The only other possibility is that $p(1) = 1$ and $q(1) = 0$. By the lemma and the hypothesis of induction we find

$$E_{k,\tau p}(2) \leqq E_{k,\tau p}(4)$$
$$\leqq E_{k,\tau q}(4).$$

This proves Theorem 3. The result at once shows that the policy $p \equiv 1$ is optimal in the sense that it minimizes

$$\limsup n^{-1}E_{n,p}(x).$$

XVII. ACKNOWLEDGMENT

The author takes pleasure in expressing his gratitude to J. Riordan, R. I. Wilkinson, E. Wolman, A. Descloux, and W. Helly for reading the preliminary draft. Their encouragement and suggestions have led to substantial improvements.

REFERENCES

1. Beneš, V. E., Algebraic and Topological Properties of Connecting Networks, this issue, pp. 1249–1274.
2. Beneš, V. E., On Rearrangeable Three-Stage Connecting Networks, to appear.
3. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, to appear.
4. Beneš, V. E., A Thermodynamic Theory of Traffic in Connecting Networks, to appear.
5. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1960.
6. Kosten, L., The Historical Development of the Theory of Probability in Telephone Traffic Engineering in Europe, Teleteknik, **1,** 1957, pp. 32–40.
7. Wilkinson, R. I., The Beginnings of Switching Theory in the United States, Teleteknik (English Edition), **1,** 1957, pp. 14–31.
8. Molina, E. C., Computation Formula for the Probability of an Event Happening at Least c Times in N Trials, Amer. Math. Monthly, **20,** 1913, pp. 190–193.
9. Jensen, A., An Elucidation of A. K. Erlang's Statistical Works Through the Theory of Stochastic Processes, in the Life and Works of A. K. Erlang, Trans. Danish Acad. Sciences, 1948, pp. 23–100.
10. Engset, T., Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wähleranzahl in Automatischen Fernsprechämtern, E.T.Z., **31,** 1918, pp. 304–306.
11. O'Dell, G. F., An Outline of the Trunking Aspect of Automatic Telephony, J. Inst. Elec. Engrs., **65,** 1927, pp. 185–222.
12. Crommelin, C. D., Delay Probability Formulae, P. O. Elec. Engrs. J., **26,** 1933–1934, pp. 266–274.
13. Molina, E. C., Application of the Theory of Probability to Telephone Trunking Problems, B.S.T.J., **6,** 1927, pp. 461–494.
14. Pollaczek, F., Über eine Aufgabe der Wahrscheinlichkeitstheorie, Math. Zeit., **32,** 1930, pp. 64–100 and 729–750.

15. Khinchin, A. I., Matematicheskaya Teoriya Statsionarnoi Ocheredi, Matematicheskii Sbornik, **39,** 1932, pp. 73–84.
16. Fry, T. C., *Probability and Its Engineering Uses*, D. Van Nostrand, New York, 1928.
17. Kolmogorov, A. N., *The Foundations of Probability*, Second Edition, Chelsea, New York, 1956.
18. Palm, C., Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, **44,** 1943, pp. 1–189.
19. Feller, W., On the Theory of Stochastic Processes, with Particular Reference to Applications, *Proc. [1st] Berkeley Symp. Math. Stat. and Prob.*, 1949, pp. 403–432.
20. Kosten, L., On the Influence of Repeated Calls in the Theory of Probabilities of Blocking, De Ingenieur, **59,** 1947, pp. 1–25.
21. Kosten, L., Manning, J. R., and Garwood, F., On the Accuracy of Measurements of Probabilities of Loss in Telephone Systems, J. Royal Statistical Soc., B, **11,** 1949, pp. 54–67.
22. Scudder, F. J., and Reynolds, J. N., Crossbar Dial Telephone Switching System, B.S.T.J., **18,** 1939, pp. 76–118.
23. Jacobaeus, C., A Study on Congestion in Link Systems, Ericsson Technics, **51,** 1950, pp. 1–68.
24. Fortet, R., and Canceill, B., Probabilité de Perte en Selection Conjuguée, Teleteknik, **1,** 1957, pp. 41–55.
25. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., **34,** 1955, pp. 1287–1315.
26. Le Gall, P., Methode de Calcul de L'encombrement dans les Systèmes Téléphoniques Automatiques à Marquage, Ann. des Telecom., **12,** 1957, pp. 374–386.
27. Lundkvist, K., Method of Computing the Grade of Service in a Selection Stage Composed of Primary and Secondary Switches, Ericsson Review, No. 1, 1948, pp. 11–17.
28. Elldin, A., Applications of Equations of State in the Theory of Telephone Traffic, Thesis, Stockholm, 1957.
29. Clos, C., A Study of Non-Blocking Switching Networks, B.S.T.J., **32,** 1953, pp. 406–424.

Algebraic and Topological Properties of Connecting Networks

By V. E. BENEŠ

(Manuscript received May 29, 1961)

A connecting network *is an arrangement of switches and transmission links allowing a certain set of terminals to be connected together in various combinations, usually by disjoint chains (paths): e.g., a central office, toll center, or military communications system. Some of the basic combinatory properties of connecting networks are studied in the present paper.*

Three of these properties are defined informally as follows: A network is rearrangeable *if, given any set of calls in progress and any pair of idle terminals, the calls can be reassigned new routes (if necessary) so as to make it possible to connect the idle pair. A state of a network is a blocking state if some pair of idle terminals cannot be connected. A network is* nonblocking in the wide sense *if by suitably choosing routes for new calls it is possible to avoid all the blocking states and still satisfy all demands for connection as they arise, without rearranging existing calls. Finally, a network is* nonblocking in the strict sense *if it has no blocking states.*

A distance between states can be defined as the number of calls one would have to add or remove to change one state into the other. This distance defines a topology on the set of states. Also, the states can be partially ordered by inclusion in a natural way. This partial ordering and its dual define two more topologies for the set of states. The three topologies so obtained are used to characterize (i.e., give necessary and sufficient conditions for the truth of) the three properties of rearrangeability, nonblocking in the wide sense, and nonblocking in the strict sense. Each of these three properties represents a degree of abundance of nonblocking states; the mathematical concept used to express these degrees is the topological notion of denseness.

TABLE OF CONTENTS

I. INTRODUCTION 1250
II. SUMMARY 1251
III. STRUCTURE AND CONDITION OF A CONNECTING NETWORK 1252
IV. GRAPHICAL DEPICTION OF STRUCTURES AND CONDITIONS 1253
V. NETWORK STATES 1254
VI. THE STATE DIAGRAM 1257
VII. SOME NUMERICAL FUNCTIONS 1259

VIII. Assignments 1262
IX. Three Topologies 1264
X. Some Definitions and Problems 1265
XI. Rearrangeable Networks 1268
XII. Networks Nonblocking in the Wide Sense 1270
XIII. Networks Nonblocking in the Strict Sense 1272

I. INTRODUCTION

Any large communication system contains a *connecting network*, an arrangement of switches and transmission links through which certain terminals can be connected together in many combinations, usually by many different possible routes through the network. Examples of connecting networks can be found in telephone central offices, toll centers, and military communications systems.

The connections in progress in a connecting network usually do not arise in a predetermined temporal sequence; instead, requests for connection (new calls) and terminations of connection (hangups) occur more or less "at random." For this reason it is customary to use the performance of a connecting network when subjected to random traffic as a figure of merit. One precise measure of this performance is the fraction of requested connections that cannot be completed in a given time interval, or the *probability of blocking*. In a telephone connecting network this probability measures to some extent the grade of service given to the customers.

The performance of a connecting network for a given traffic level is determined largely by its configuration or structure. This configuration may be described by stating what terminals or transmission links have a switch placed between them and can be connected together by closing the switch. The configuration of a connecting network determines what groups of terminals can be connected together simultaneously. Any one set of permissible connections may be called a *state* of the network. Quantities such as the number of combinations of terminals that can be connected, and the number of states in which a given combination is connected, clearly are indicative of both the performance and the cost of the system. If these numbers are small the performance may be poor and the cost low; if large, the performance may be unnecessarily good and the cost prohibitive. These numbers are among the purely combinatory and topological properties of the connecting network.

For example, in a telephone exchange, the network configuration determines what pairs of terminals can be simultaneously connected by disjoint paths, that is, what calls can be in progress. If this configuration is too simple, only a few pairs of terminals can have calls in progress between them at the same time. If the configuration is extensive and

complicated it may provide for many large groups of simultaneous calls in progress, but the network itself may be expensive to build and difficult to control.

To design connecting networks with confidence, then, it is desirable to have an adequate general understanding of their combinatory and topological properties. A discussion, in part heuristic and tutorial, of connecting systems and of some associated mathematical problems has been given in a paper[1] by the author; the reader is referred thereto for material suitable as background for the present paper. In that work a division of the topic into *combinatory*, *probabilistic*, and *variational* problems was drawn, and it was argued that the elements of this division had a natural order of priority: one must know the combinatory properties of a system in order to calculate its probabilistic properties, i.e., its performance in the face of random traffic; and one must know both the combinatory and the probabilistic properties of systems in order to compare them and to select optimal ones.

In this paper we shall be concerned exclusively with those combinatory and topological properties of a general connecting network that seem to be most relevant to its performance.

II. SUMMARY

Some of the basic combinatory properties of connecting networks are studied in the present work. Three of these properties, rearrangeability, nonblocking in the wide sense, and nonblocking in the strict sense, can be defined informally as follows: for brevity, define an *idle pair* to be a pair of idle terminals consisting of an inlet and an outlet. A network is *rearrangeable* if, given any set of calls in progress, and any idle pair, the existing calls can be assigned new routes (if necessary) so as to make it possible to connect the idle pair. A state of a network is a *blocking state* if some idle pair cannot be connected. A network is *nonblocking in the wide sense* if by suitably choosing routes for new calls it is possible to avoid all the blocking states and still satisfy all demands for connection as they arise, without having to rearrange existing calls. Finally, a network is *nonblocking in the strict sense* if it has no blocking states.

A distance between states of a connecting network can be defined as the number of pairs of terminals that are connected in one state and not in the other. This distance defines a topology on the set of states. Also, the states can be partially ordered by inclusion in a natural way. This partial ordering and its dual define two more topologies for the set of states. The three topologies so obtained are used to characterize (i.e., give necessary and sufficient conditions for the truth of) the three

properties of rearrangeability, nonblocking in the wide sense, and nonblocking in the strict sense. Each of these three properties represents a degree of abundance of states in which calls are not blocked; the mathematical concept used to express these degrees is the topological notion of *denseness*. A study of some particular connecting networks that are rearrangeable is given in another paper.[2]

III. THE STRUCTURE AND CONDITION OF A CONNECTING NETWORK

In discussing connecting networks, we shall abstract from the many possible technological realizations and actual designs of connecting networks, and shall consider only certain relevant features on which we can base a useful and sufficiently general mathematical theory.

Most real telephone switching networks consist of pairs of wires for talking paths and electromechanical switches for crosspoints; in certain experimental systems the talking paths are pulse code modulation channels, and the crosspoints are time-division gates made of transistors. However, any attempt to formulate some general properties of connecting networks must be independent of the network configuration chosen, and of the technology used to build the network, for a particular real system. A theory must apply equally well to Strowger switches, crossbar switches, gas-diode switches, and time-division switches. Unless it is independent of technology, a theory of connecting networks is limited in scope and may have missed the heart of the problem. We therefore use some of the terminology of switching engineers but understand it to refer to defined mathematical idealizations of switches, gates, crosspoints, transmission links, etc., rather than to the physical entities themselves.

We distinguish between switching networks used for communication and those used for control functions and logical transformations, like relay nets. Our concern is with networks of the former kind, and we call these *connecting networks*.

A communications switching network, or connecting network, consists of three kinds of entities: (*i*) *wires* or other transmission media along which communication may take place; (*ii*) *terminals* to which the wires are attached; and (*iii*) *crosspoints* or switches which can be used to connect the terminals, and hence the wires, together in various combinations. Each crosspoint can connect together exactly one pair of terminals, and it has two conditions: in the "on" or closed condition the two terminals are connected and communication can pass from one to the other; in the "off" or open condition the terminals are disconnected, and no information can pass through.

From the point of view of switching, two terminals connected together

by a wire are essentially one terminal, albeit a spatially extended one. We therefore regard terminals as identical if they are wired together; in mathematical terms, we identify terminals under the equivalence relation of "being wired together." Henceforth, then, our considerations will leave wires out of account, and will be based only on the notions of *terminal* and *crosspoint*.

By the configuration or *structure* of a connecting network, we mean a specification of the terminals between which individual crosspoints have been placed. By the *condition* of a connecting network, we mean a specification of the closed and open crosspoints. In most cases of interest the structure is invariant in time, while the condition changes in a random way. We shall assume that at most one crosspoint is placed between distinct terminals, and that no crosspoint is placed from a terminal to itself.

IV. GRAPHICAL DEPICTION OF NETWORK STRUCTURE AND CONDITION

A simple device can illustrate the four notions we have introduced so far. In Fig. 1(a) the nodes (points) represent *terminals*, and the branches (lines) labeled x_i, $i = 1, \cdots, 6$, represent *crosspoints* placed between the terminals. The resulting graph represents the *structure* of a network. If we interpret the labels x_i as binary variables specifying the condition of the (respective) crosspoints, with 0 meaning "open" and 1 meaning "closed," then an assignment of values 0 or 1 to $\{x_1, \cdots, x_6\}$ represents a possible *condition* of the network, illustrated in Fig. 1(b). We are purposely avoiding the term "network state" here in order to assign it a useful precise meaning in the next section.

We have illustrated the use of a *labeled graph* as a general representation for (simultaneously) the structure and condition of a connecting

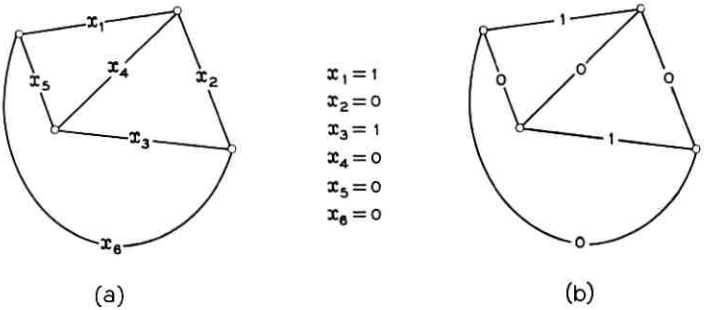


Fig. 1 — (a) Representation of structure; (b) simultaneous representation of structure and condition.

network. This representation is useful because it identifies the structure and condition of the network with a definite mathematical entity. It will become apparent that simple properties of this mathematical representation have great theoretical and practical relevance to congestion problems.

In general, the labeled graph g representing* the structure and condition of an arbitrary connecting network is constructed as follows:

i. nodes (points) of g correspond to terminals of the network;
ii. branches (lines, or edges) of g correspond to crosspoints of the network;
iii. open crosspoints are labeled 0;
iv. closed crosspoints are labeled 1.

Two terminals are *connected* in g if g contains a chain of closed crosspoints from one terminal to the other.

V. NETWORK STATES

Let G be a graph representing the structure of a switching network, and let V be the set of all labeled graphs g (labeled "versions" of G) obtained by assigning 0 or 1 to each line of G. There are several reasons why not every element g of V represents a physically meaningful state of the network.

In most switching systems there is an explicit functional distinction between terminals which are used only to connect other terminals together, and those between which desired connections arise, and which are never used to connect other terminals together. Terminals of the former kind we shall call *links*, because of their intermediary nature, and those of the latter kind, *inlets* and/or *outlets*. Desired connections always arise between two or more inlets or outlets. If more than two are involved, the connection is termed a "conference" call. Usually, though, the connections are disjoint chains of closed crosspoints, assuring private conversations between inlets and outlets by pairs only; we restrict attention to these. In terms of our graphical representation of the structure and condition of a switching network, the distinctions made above impose restrictions on the elements of V which represent realistic conditions of a network having the structure of G.

The restrictions on the assignment of the labels 0 or 1 listed above are (perhaps the most important) among many which are imposed by the functional and operational features of a real switching system. In general, a real connecting network specifies (or *uses*) only a subset of

* A glossary of mathematical notations appears at the end of this paper, Section IX.

the set V of all possible labeled versions of the graph G that represents the structure of the network being studied. We have therefore avoided calling elements of V "states of the network" because not all members of V can reasonably represent the condition of an actual network. We now attempt to characterize those subsets S of V which can represent real networks. Each such subset S will be called a *class of network states.*

The Boolean operations of *join* $\cup$ and *meet* $\cap$ (union and intersection, respectively) are definable for elements x, y of V in an obvious way:

$x \cup y$ = the V-element having a 1 wherever either x or y has a 1, and 0 elsewhere,

$x \cap y$ = the V-element having a 1 wherever both x and y have a 1, and 0 elsewhere.

The complement x' and the difference $x - y$ can be defined analogously. In view of this it is natural to inquire whether these Boolean operations can be used to characterize subsets S of V which are classes of network states.

If the elements x, y of V belong to such a subset (class of network states) S, it is not necessarily true that $x \cup y$, nor that $x \cap y$, belongs to S. In the case of $x \cup y$, there may be links and crosspoints used in both x and y, and so $x \cup y$ may violate the requirement of privacy. Even if $x \cap y = 0$ there may still be inlets used in both x and y, so that $x \cup y$ would lead to undesirable paths of extreme length. In the case of $x \cap y$, there may be so little in common to x and y that $x \cap y$ reduces to a single closed crosspoint between two links (i.e., *not* between an inlet and an outlet). Thus the Boolean operations do not yield a useful way of describing S.

The preceding remarks suggest that since any connection is a chain, none of whose terminals and crosspoints occurs in another connection, the labels 0 and 1 are really superfluous, although they served a tutorial purpose heretofore. That is, in describing the possible subsets S of network states, we can (and should) take advantage of inherent physical restrictions, and conveniently replace our representation* $x \in V$ of the structure and condition of a network by a corresponding set of disjoint chains, since each physically meaningful element x from V is equivalent to such a set. A formal development of this suggestion follows.

Let T be the set of terminals of a connecting network. The *graph* G representing the structure of the network is a subset G of the product

$$T \times T = \{(u, v) \mid u \in T, v \in T\}$$

* "$x \in V$" means that x is an element of the set V.

with the properties

$$(u, v) \in G \quad \text{if and only if} \quad (v, u) \in G$$

$$(u, u) \quad \text{is never in } G$$

and the interpretation

$(u, v) \in G$ if and only if (u, v) is an edge of graph G

if and only if nodes u and v are adjacent in the graph G

if and only if there is a crosspoint between terminals u and v.

A *chain* p of length n between terminals u and v is a sequence of elements $\{z_i \in T, 0 \leqq i \leqq n\}$ such that

$$z_0 = u, \qquad z_n = v,$$

$$z_i \neq z_j \quad \text{for} \quad i \neq j,$$

$$(z_i, z_{i+1}) \in G \quad \text{for} \quad i = 0, \cdots, n - 1.$$

Two chains p_1 and p_2 are called *disjoint* if they have no nodes (terminals $\in T$) in common; in this case we write symbolically $p_1 \cap p_2 = \phi$, with ϕ = null set.

We shall henceforth assume that the set T of terminals has been (functionally) decomposed into three sets:

$$T = I \cup \Omega \cup L,$$

where I is a set of *inlets*, Ω a set of *outlets*, and L is the set of *links*. It is possible that $I = \Omega$ or that $I \cap \Omega =$ empty set, or that some intermediate condition obtain. However, we shall insist that $(I \cup \Omega) \cap L$ be null, i.e., that no link be an inlet or an outlet.

The set C of *connections* consists of all chains $p = \{z_i \in T, i = 0, \cdots, n(p)\}$ such that

$$z_0 \in I, \qquad z_{n(p)} \in \Omega, \qquad z_0 \neq z_{n(p)}$$

$$z_i \in L, \quad \text{for} \quad i \neq 0 \quad \text{or} \quad n(p).$$

Each element p of C represents a possible connection from an inlet to an outlet through the network whose structure is represented by the graph G.

Elements of the set S of network states will be defined as subsets x of C, $x \subset C$, consisting entirely of disjoint chains, that is, such that

$$p_1, p_2 \in x \quad \text{implies} \quad p_1 \cap p_2 = \phi.$$

Two subsets x and y of C are called *compatible* if

$$p_1 \in x,\ p_2 \in y \quad \text{implies} \quad p_1 \cap p_2 = \phi.$$

The connections that comprise compatible states can all be put up simultaneously without interfering with each other or violating the requirement of privacy.

The functional and physical restrictions imposed by real networks determine (in any particular system) a subset E of C consisting of (what we shall call) the *elementary states*, or single connections that can actually be used. For example, chains in C that double back and are wastefully circuitous may be excluded from E.

Given such a subset E of elementary states, we can define a class of network states S, associated with E, in a natural way as follows: S is the smallest class of subsets of E containing all unit subsets of E, and closed under formation of arbitrary intersections (meets) and unions (joins) of *compatible* subsets of E. That is, S is the smallest class of E-subsets such that

$$p \in E \quad \text{implies} \quad \{p\} \in S,$$

$$x, y \in S \quad \text{implies} \quad x \cap y \in S,$$

$$\text{if} \quad x, y \in S \quad \text{and} \quad p_1 \in x,\ p_2 \in y \quad \text{implies} \quad p_1 \cap p_2 = \phi,$$

$$\text{then} \quad x \cup y \in S.$$

We henceforth use "S" as a generic notation for a class of network states defined as above. The word "network" will refer to a graph G representing structure, choices I and Ω of inlets and outlets respectively, and a choice E of elementary states. The choice of G, I, Ω, and E uniquely determines a class S of network states according to the definition given previously. The quadruple (G, I, Ω, E) will be called a network, N.

It is easily verified that the class S of network states is partially ordered by inclusion, $\leqq$. Moreover, any two elements x, y of S have a unique intersection (meet) consisting of just those connections common to both x and y, and S itself has a unique least element included in every other element, viz., the ground state in which no calls are in progress. However, since only infima exist, and since there may be many maximal elements in the partial ordering, S is not a lattice, in general.

VI. THE STATE DIAGRAM

The partial ordering $\leqq$ of S has a special nature that allows us to arrange the network states $x \in S$ in a particularly intuitive and useful

pattern. The following conventions and definitions will be helpful in discussing this pattern.

If K is any set, we use the notation $|K|$ to mean the number of elements of K. E.g., if x is a network state,

$$|x| = \text{the number of calls in progress in state } x.$$

The sets L_k are defined by the conditions

$$L_k = \{x \in S \mid |x| = k\}, \qquad k = 0, 1, \cdots,$$

that is, L_k is the set of all network states consisting of exactly k connections. L_0 is a unit set containing just the zero state. The sets L_k are a partition of S corresponding to the equivalence relation of "having the same number of calls in progress."

To obtain our pattern for arranging network states we start with the zero or ground state in which no calls are in progress: this is the empty set (of chains). Above this zero element, in a horizontal row, we place all the states consisting of a single connection, i.e., all the elements of E. Continuing in this way, we put the set L_{k+1} of states consisting of $(k + 1)$ disjoint chains (i.e., $k + 1$ calls) in a horizontal row above the set L_k of states with k disjoint chains (i.e., k calls in progress). We call L_k the kth *level*.

The diagram is completed by constructing the corresponding Hasse figure (Birkhoff,[3] p. 12); that is, we think of the states $x \in S$ (in their arrangement into levels L_k) as nodes, and we construct a graph by drawing lines between states x, y of respective adjacent levels L_k, L_{k+1} just in case

$$y - x \in E,$$

i.e., if and only if y results from x by putting up one more call. The resulting graph can be termed the *state diagram* D of the network N described by the quadruple (G, I, Ω, E). The state diagram D is a natural and standard representation of the partial ordering of S. The history of the connecting network when operating can be thought of as a trajectory on D.

We shall use the network depicted in Fig. 2 to illustrate the state diagram D. For most practical purposes this network is wasteful of crosspoints, but it makes a suitably simple example of the partial ordering of the states. The network has four inlets and four outlets, and no inlet is an outlet. The squares in Fig. 2 represent 2 by 2 switches, as indicated.

The possible states of this network are determined by all the ways

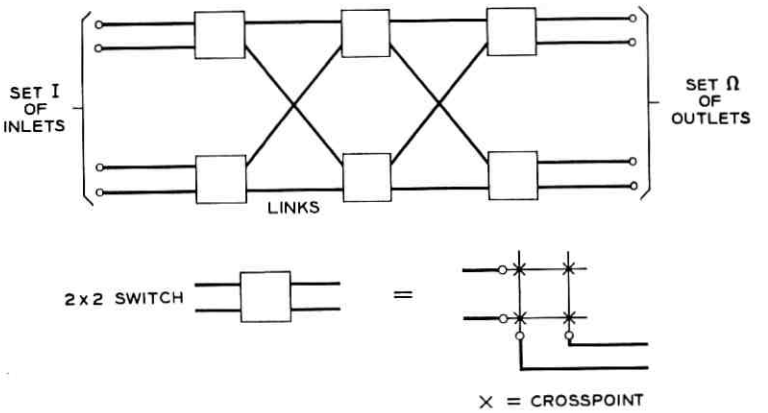


Fig. 2 — Illustrative three-stage connecting network.

in which four or fewer inlets can be connected pairwise to as many outlets on the right, no inlet being connected to more than one outlet, and vice versa. These possible states have been depicted in a natural arrangement in Fig. 3, which shows a reduced state diagram in which states which differ only by permutations of inlets, outlets, or switches have been identified. There is essentially only one way to put in a single call; there are four ways of putting in two calls; and there are two ways each of putting in three and four calls. The states have been arranged in levels according to the number of calls in progress. In each state only links actually in use are shown, and the different notations on the links indicate the routing.

VII. SOME NUMERICAL FUNCTIONS

The finite set S of network states is *partially ordered* by *inclusion*, which we shall denote by $\leqq$. A *chain* in S is a subset X of S which is *simply ordered* by (the restriction to X of) $\leqq$; that is, for any two elements $x, y \in X$, we have either $x \geqq y$ or $y \geqq x$. Such a chain is not to be confused with the "chains" on the graph G that are elements of states $x \in S$. The *dimension* or *height* $| x |$ of a state is the maximum "length" d of chains $0 < x_1 < \cdots < x_d = x$ that have x for greatest element. (This usage is consistent with the previous definition of $| \cdot |$.)

Remark 1: The dimension $| x |$ of a state x is the *number of busy pairs*, or the *number of calls in progress*, in the state x.

A state x is said to *cover* another state y if and only if $x > y$, and there are no $z \in S$ such that $x > z > y$. The state x is then "immediately above" y. It is apparent that x covers y if and only if $x > y$ and $| x | =$

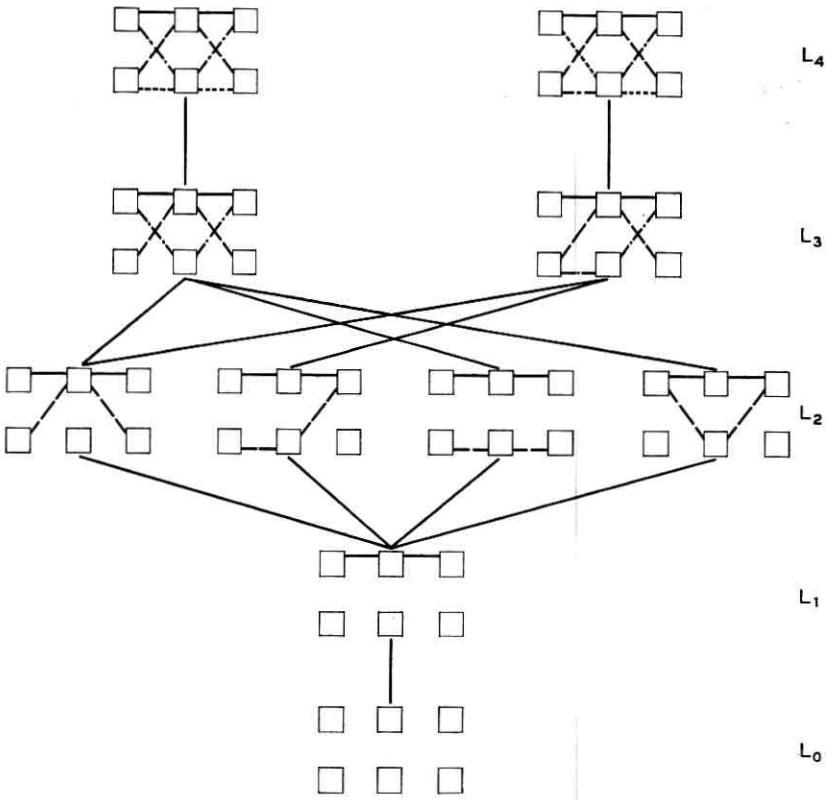


Fig. 3 — (Reduced) state diagram for the network shown in Fig. 2.

$|y| + 1$. In fact, the construction of the partial ordering of S arranges the states according to levels, each level being the (equivalence) class of all states having the same *dimension*. In determining dimension one need only consider chains that are "maximal" or "connected" in the sense that x_i covers x_{i-1} for all i. Also, it can be seen that the partial ordering $\leqq$ of S satisfies the *Jordan-Dedekind chain condition*: all connected chains between fixed end points have the same length.

The present section will be devoted to various relationships between numerical functions defined on S, counting or "enumeration" problems, etc., based largely on the dimension function and the chain condition.

The *Möbius function* $\mu(\cdot)$ of the partially ordered system $(S, \leqq)$ is defined recursively by

$$\mu(0) = 1, \qquad \mu(x) = -\sum_{y<x} \mu(y) \quad \text{if} \quad x > 0,$$

where 0 denotes the zero or ground state in which no calls are up. The Möbius function has the following two important properties:

i. Let $f(\cdot)$ be any function defined on S, and let

$$F(x) = \sum_{y<x} f(y).$$

Then $f(\cdot)$ and $F(\cdot)$ are related by the Möbius inversion formula

$$f(x) = \sum_{y<x} \mu(y)F(x - y).$$

Here $x - y$ denotes the state obtained from x by removing all the calls of state y; this makes sense, since $y < x$. (See Weisner.[4])

ii. Let $\lambda(x, n)$ be the number of chains of length n which can be interpolated between 0 and x. Then P. Hall[5] has shown that

$$-\mu(x) = \lambda(x, 1) - \lambda(x, 2) + \cdots .$$

By the Jordan-Dedekind chain condition, all the chains from 0 to x have the same length, viz., $|\, x \,|$. Hence for $x > 0$

$$\mu(x) = (-1)^{|\, x \,|}\, \lambda(x, |\, x \,|).$$

For simplicity of notation set

$$\begin{aligned} \lambda(x, |\, x \,|) &= \eta(x) \\ &= \text{number of ways of "climbing" from 0 to } x. \end{aligned}$$

Also, we introduce the following sets:

$$\begin{aligned} A_x &= \{y \mid y \text{ covers } x\} \\ B_x &= \{y \mid x \text{ covers } y\} \\ L_n &= \{x \mid |\, x \,| = n\}. \end{aligned}$$

These have the following respective intuitive meanings: A_x is the set of states *immediately above* x, i.e., obtainable from x by adding one more call; B_x is the set of states immediately below x, i.e., obtainable from x by removing one call; L_n is the nth level, the set of all states having n calls up. The cardinality of a finite set X is designated by $|\, X \,|$.

Remark 2: $|\, B_x \,| = |\, x \,|$ for each $x \in S$. Clearly, x covers exactly $|\, x \,|$ states, each obtainable from x by removing one call.

Remark 3: For each $x \in S$

$$\eta(x) = \sum_{y \in B_x} \eta(y).$$

Indeed, every state y covered by x gives rise to exactly $\eta(y)$ climbing paths from 0 that reach x via y.

Remark 4: For $x \in L_n$, $\eta(x)$ has the constant value $n!$ This is obvious intuitively, since there are $n!$ orders in which the n calls of $x \in L_n$ could be put up. More formally, the result is true for $x = 0$; assume it true for $y \in L_{n-1}$; then by the previous results,

$$\eta(x) = \sum_{y \in B_x} \eta(y) = |B_x| \cdot (n-1)!$$
$$= n!.$$

Remark 5: The Möbius function $\mu(\cdot)$ is given by

$$\mu(x) = (-1)^{|x|} (|x|)!$$
$$= (-1)^n n! \quad \text{for} \quad x \in L_n.$$

Theorem 1:

$$|L_n| = \frac{1}{n} \sum_{y \in L_{n-1}} |A_y|, \qquad n > 0.$$

Proof: The segments in the partial ordering passing from elements $y \in L_{n-1}$ to L_n are just those that pass from some $x \in L_n$ to L_{n-1} and by Remark 2, each $x \in L_n$ has exactly $|x|$ $(= n)$ such segments. Therefore,

$$n \cdot |L_n| = \sum_{y \in L_{n-1}} |A_y|$$

and the sum on the right is exactly divisible by n.

Definition: C_n is the total number of chains (of length n) from 0 into L_n, i.e., to some state in L_n.

Remark 6:

$$C(n) = \sum_{x \in L_n} \eta(x) = \sum_{y \in L_{n-1}} \eta(y) \cdot |A_y|.$$

It can be seen that $x \in L_n$ has $\eta(x)$ chains climbing to it from 0; for $x, y \in L_n$, $x \neq y$, these chains are distinct since their highest elements are unequal. This proves the first identity. Also each chain climbing to L_n from 0 must pass through some unique $y \in L_{n-1}$. Each $y \in L_{n-1}$ has $\eta(y)$ chains of length $n-1$ reaching it from 0, and each such chain can then be completed to reach L_n in $|A_y|$ ways. It follows also that

$$|L_n| = \frac{C_n}{n!} = \frac{1}{n} \sum_{y \in L_{n-1}} |A_y|.$$

VIII. ASSIGNMENTS

By an *assignment* we shall mean any one-to-one map $a(\cdot)$ of a subset of I into Ω. An assignment is to be interpreted as a specification of what

inlets are to be connected to what outlets, without regard to the possible routes that these connections might take through the network. If $I \cap \Omega$ is nonnull, we restrict assignments so as to satisfy $a(u) \neq u$.

Let x be a network state consisting of chains $p_1, p_2, \cdots, p_n$ with $n = n(x)$ and each p_i a chain between $u_i \in I$ and $v_i \in \Omega$. We say that x *realizes the assignment* $a(\cdot)$ if and only if

i. the domain of $a(\cdot)$ is $\{u_i, 0 \leqq i \leqq n(x)\}$

ii. the range of $a(\cdot)$ is $\{v_i, 0 \leqq i \leqq n(x)\}$

iii. $a(u_i) = v_i, 0 \leqq i \leqq n(x)$.

An assignment is *realizable* if some network state realizes it; a state realizes exactly one assignment; the zero state realizes the null assignment. A *maximal* assignment is one that has either domain I or range Ω. The set of all assignments is denoted by A, and that of all maximal assignments by $\hat{A}$.

Two terminals, $u \in I$ and $v \in \Omega$, are *connected* in state x if and only if some chain $p \in x$ is a chain between u and v, i.e., if and only if x realizes the (unit) assignment

$$\{(u, v)\}.$$

We define the function $\gamma(\cdot)$ from S into (the set of) subsets of $I \times \Omega$ by the condition

$$\gamma(x) = \{(u, v) \in I \times \Omega \mid u \text{ and } v \text{ are connected in } x\}.$$

Formally, then $\gamma(x)$ is the assignment realized by state x; heuristically, we may think of $\gamma(x)$ as the set of calls which are in progress in state x. The set of *unit assignments*, that is, of

$$c = \{(u, v)\} \quad \text{such that} \quad (u, v) \in I \times \Omega,$$

will be denoted by U, and a unit assignment $c \in U$ will be referred to informally as a *call*.

If $a = a(\cdot) \in A$ is an assignment, we use the notation

$$\gamma^{-1}(a)$$

for the inverse image of $a(\cdot)$ under $\gamma(\cdot)$, i.e., the set of (equivalent) states y such that $\gamma(y) = a$. In a similar vein, if X is a set of states, we define

$$\gamma(X) = \{a \in A \mid a = \gamma(x) \quad \text{for some} \quad x \in X\},$$

that is, $\gamma(X)$ is the set of assignments realized by members of X.

IX. THREE TOPOLOGIES

Two network states x and y are equivalent, written $x \sim y$, if and only if they realize the same assignment, i.e.,

$$\gamma(x) = \gamma(y).$$

Intuitively, equivalent but nonidentical states correspond to different ways of putting up the same set of calls.

A pseudo-metric (Kelley,[6] p. 118) on S can be defined by the formula

$$d(x, y) = |\gamma(x)\Delta\gamma(y)|, \qquad x, y \in S,$$

where Δ denotes the symmetric difference of sets, and $|\cdot|$ cardinality, as before. In plain words, the distance $d(x, y)$ between x and y is the number of pairs $(u, v) \in I \times \Omega$ that are either connected in x and not connected in y, or connected in y and not connected in x. Clearly

$$d(x, 0) = |x|, \qquad 0 = \text{zero state},$$

and also

$$d(x, y) = 0 \quad \text{if and only if} \quad x \sim y.$$

Thus $d(\cdot, \cdot)$ only identifies states up to equivalence. The function $d(\cdot\ , \cdot)$ is obviously symmetric, and the triangle inequality is a consequence of the set inclusion

$$(X\Delta Y) \subseteq (X\Delta Z) \cup (Y\Delta Z).$$

The pseudo-metric $d(\cdot\ , \cdot)$ can be used to define a topology for S in a standard way (see Kelley,[6] p. 118 et seq.) The *closure* of a set X in the d-topology consists of all states equivalent to members of X, and is denoted by X^d.

For each subset X of S, we define its $\leqq$-*closure* $\underline{X}$ by the condition

$$\underline{X} = \{y \in S \mid y \leqq x \quad \text{for some} \quad x \in X\}.$$

The operation on sets so defined satisfies the Kuratowski closure axioms (cf. Kelley,[6] p. 43):

$$\underline{\phi} = \phi$$

$$X \subseteq \underline{X}$$

$$\underline{X} = \underline{X}$$

$$\underline{X} \cup \underline{Y} = \underline{X} \cup \underline{Y}$$

and so defines a closure topology for S. The set $\underline{X}$ consists of all states

that are "below" some member of X in the state-diagram D, i.e., can be reached from a member of X by removing calls.

In a similar way, we define the $\geqq$-*closure* $\bar{X}$ of a set $X \subseteq S$ as

$$\bar{X} = \{y \in S \mid y \geqq x \quad \text{for some} \quad x \in X\}.$$

The converse of a partial ordering relation is also a partial ordering, called its *dual*. Hence the mapping $X \to \bar{X}$ is also a closure operation, defining a third topology on S.

X. SOME DEFINITIONS AND PROBLEMS

An inlet or outlet is *idle* in a network state x if it belongs to neither the range nor the domain of the assignment $\gamma(x)$ realized by x. An *idle pair* of the state x is an element (u, v) of $I \times \Omega$ such that both u and v are idle in x. A call $c = \{(u, v)\}$ is *new* in x if (u, v) is an idle pair.

We shall now define what is meant by a blocked call. Let $x \in S$ realize the assignment $\gamma(x)$ and let c be a new call in x, i.e., let

$$c = \{(u, v)\} \in U$$

be a unit assignment such that (u, v) is an idle pair of x. The new call c is *blocked in* x if there is no state $y > x$ such that

$$\gamma(y) = \gamma(x) \cup c.$$

A state x is a *blocking state* if some call is blocked in x. The state x is called *nonblocking* if and only if for every idle pair (u, v) of x, the call

$$c = \{(u, v)\}$$

is not blocked in x, i.e., there is a $y \in S$ above x which realizes the larger assignment $\gamma(x) \cup c$, so that

$$\gamma(y) = \gamma(x) \cup \{(u, v)\},$$
$$y > x.$$

The set of nonblocking states is designated by the symbol B'. A state that realizes a maximal assignment has no idle pairs, and is (trivially) nonblocking. In plain terms, a nonblocking state x is one in which any idle inlet u can be connected to any idle outlet v *without disturbing the calls that are already present*; in this case there is a path r, disjoint from all paths $p \in x$, between u and v, and

$$x \cup \{r\} \in S,$$

i.e., use of this path results in a network state.

A network $N = (G, I, \Omega, E)$ will be called *nonblocking in the strict sense* if and only if every state is nonblocking, i.e., $B' = S$. Such networks have been discovered and studied extensively by C. Clos. (See Clos[7] and Kharkevich.[8]) A network that is nonblocking in this strong sense has the property that no matter in what state it is, any idle pair can be connected (in a way that results in a legitimate network state).

In most switching networks there may be several or many ways of connecting an idle pair, i.e., putting up a new call, in a given state, all of which lead to legitimate network states. Thus, even if the set S of network states contains blocking states, it is conceivable that by making the right choices of paths for connections one might avoid all the blocking states, and still satisfy all demands for connection as they arise, without disturbing calls already present. That is, there may exist a *rule* for choosing paths which, if followed, confines the trajectory of the system to nonblocking states (without refusing any demands for connection by idle pairs).

We next discuss what is meant by a rule. If a call $c = \{(u, v)\}$ is blocked in a state x it cannot be put up without disturbing existing calls of x, and there is no question of using a rule. Also, if x is a maximal state, no new calls can be put up, and a rule is unnecessary. But if a call c can be put up in one or more ways in the state x, then there is at least one $y > x$ such that $\gamma(y) = \gamma(x) \cup c$. In such a case some method of specifying permitted or prohibited new states could be used in order to improve performance.

A *rule* $\rho(\cdot\ ,\ \cdot)$ for a network N is a mapping of the Cartesian product

$$[S - \gamma^{-1}(\hat{A})] \times U$$

into subsets of S, with the properties: if $x \in S$ and $c = \{(u, v)\} \in U$ with (u, v) an idle pair of x (so that c is a new call in x), then

$$0 \subseteq \rho(x, c) \subseteq \gamma^{-1}(\gamma(x) \cup c);$$

if x is maximal, or if (u, v) is not idle, $\rho(x, c)$ is defined (arbitrarily) as the null set. If for some call c not up in x we have

$$y \in \rho(x, c),$$

we say that the transition (between states) $x \rightarrow y$ is *permitted* by $\rho(\cdot\ ,\ \cdot)$.

We say informally that a state x is *reachable under a rule* $\rho(\cdot\ ,\ \cdot)$ if there is some sequence of changes of state, consisting of either hangups or transitions permitted by $\rho(\cdot\ ,\ \cdot)$, and leading from the zero state to x. More precisely, we define the notion

x is reachable under $\rho(\cdot\ ,\ \cdot)$ in n steps

recursively, as follows:

i. The zero state is reachable under $\rho(\cdot\ ,\ \cdot)$ in zero steps.

ii. If x is reachable under $\rho(\cdot\ ,\ \cdot)$ in n steps, and for some call $c \in U$, $\gamma(x) = \gamma(y) \cup c$, then y is reachable under $\rho(\cdot\ ,\ \cdot)$ in $(n + 1)$ steps.

iii. If x is reachable under $\rho(\cdot\ ,\ \cdot)$ in n steps, and for some call $c \in U$, c is new in x and $y \in \rho(x, c)$, then y is reachable under $\rho(\cdot\ ,\ \cdot)$ in $(n + 1)$ steps.

A state is *reachable under* $\rho(\cdot\ ,\ \cdot)$ if it is reachable under $\rho(\cdot\ ,\ \cdot)$ in n steps, for some $n \geqq 0$. The set of states that are reachable under $\rho(\cdot\ ,\ \cdot)$ will be denoted by R_ρ.

A network $N = (G, I, \Omega, E)$ will be called *nonblocking in the wide sense* if and only if there is a rule $\rho(\cdot\ ,\ \cdot)$ for N under which no blocking state is reachable, i.e.,

$$R_\rho \subseteq B'.$$

In words, we may say that a network is nonblocking in the wide sense if there is a rule, depending on the states, and on the connections that are requested, such that if the rule if used (starting from the zero state) no blocking state is ever reached, and hence no request for connection by an idle pair (of a state that can be reached) need ever be refused. In making this definition, we think of the system as starting (empty) at the zero state; in any state x that it reaches, any idle pair of x may demand connection; it must always be possible to make this connection without disturbing existing calls, and reach a (nonblocking) state y one level higher, $y \in L_{|x|+1}$; at any instant an existing call may terminate, and the system move to a state of $L_{|x|-1}$. An example of such a network was given in Ref. 1.

Finally, we consider a still weaker property of networks than the first two defined, namely, the possibility of satisfying a demand for connection by *rearranging* (if necessary) the existing calls in such a way that the desired call can then be accommodated. Let x be a network state realizing the assignment $\gamma(x)$. We call x *rearrangeable* if and only if for every idle pair (u, v) of x there is a $y \in S$, possibly depending on (u, v) and x, which realizes the larger assignment $\gamma(x) \cup \{(u, v)\}$, i.e.,

$$\gamma(y) = \gamma(x) \cup \{(u, v)\}.$$

Alternately x is rearrangeable if for every call c new in x there is a y such that

$$\gamma(y) = \gamma(x) \cup c.$$

This definition is the same as that of a nonblocking state except that the condition $x < y$ is omitted. That is, to realize the larger assignment $\gamma(x) \cup c$ it may be necessary to reroute existing calls to give a new state $z \sim x$ which is not comparable to x, and which has a path r, disjoint from $p \in z$, between u and v. The state y may then be taken to be $z \cup \{r\}$. A network N is called *rearrangeable* if its states $x \in S$ are rearrangeable.

With these definitions laid down, we can formulate several problems of the combinatory theory of connecting networks:

i. Can *general* characterizations of the properties of being rearrangeable, and of being nonblocking (strict or wide sense) be given?

ii. What relationships exist among the concepts we have defined?

iii. What *specific* networks are rearrangeable, or non-blocking (strict or wide sense)?

To attack problem (*i*) we make the following observations: the three properties of interest represent different degrees of abundance of states in which calls are not blocked. The relative abundance or density of such states throughout S determines which (if any) of the three properties N has. The heuristic concept of abundance suggests the topological one of *denseness*, and the possibility of characterizing the three properties in terms of denseness. This idea is developed in the remaining sections; it leads to answers to problems (*i*) and (*ii*) above.

XI. REARRANGEABLE NETWORKS

Let X be a subclass of the class S of network states. We say that X is *sufficient* if $\gamma(X) = A$, i.e., if every assignment is realized by some state of X. We make two comments:

Remark 7: If $\hat{A} \subseteq \gamma(x)$, then $\underline{X}$ is sufficient. This can be seen as follows: every assignment is a subset of some maximal assignment, and so belongs to the $\leqq$-closure $\underline{X}$ of X. For the same reason we have

Remark 8: The following properties of a network N are equivalent:

i. N is rearrangeable.

ii. Some sufficient class exists.

iii. The range of $\gamma(\cdot)$ includes $\hat{A}$.

It is convenient to approach the study of rearrangeable networks by taking the point of view of a particular pair of customers, i.e., of a particular inlet-outlet pair $(u, v) \in I \times \Omega$. Such a pair corresponds to a unit assignment or *call*

$$c = \{(u, v)\} \in U,$$

any realization of which is among the states of E, the set of elementary states. For each call $c \in U$ we define

$$I_c = \{x \in S \mid c \text{ is new in } x, \text{ i.e., } (u, v) \text{ is idle in } x\},$$

$$B_c = \{x \in S \mid c \text{ is blocked in } x\}.$$

It can be verified that

$$B_c \subset I_c, \qquad \text{for} \quad c \in U,$$

$$B' = \bigcap_{c \in U} (B_c)',$$

$$S - \gamma^{-1}(\hat{A}) = \bigcup_{c \in U} I_c .$$

We call a network N rearrangeable for the unit assignment or call c if and only if for every $x \in I_c$ there is a $y \in S - I_c$ which realizes the larger assignment $\gamma(x) \cup c = \gamma(y)$. In words, this condition states that for any state in which the pair (u, v) is idle there is a (possibly rearranged) state in which all the same calls are up, and in addition u is connected to v. It is easy to see that N is rearrangeable if and only if it is rearrangeable for all calls $c \in U$.

Let X, Y be arbitrary subsets of S. In accord with a standard definition (Kelley,[6] p. 49), X is said to be dense in Y in the d-metric if Y is included in the d-closure of X, i.e.,

$$Y \subseteq X^d.$$

Now in a metric space the closure of a set X is the set of all points that are at distance zero from X, when the distance of a point y from a set X is defined as

$$\inf_{x \in X} d(x, y) .$$

Hence the closure of X is the set of all y such that for some $x \in X$, $d(x, y) = 0$, or equivalently, $x \sim y$. That is, the d-closure of X is the set of all states that are equivalent to a member of X:

$$X^d = \{y \in S \mid y \sim x \quad \text{for some} \quad x \in X\}.$$

These observations lead to the following result:

Theorem 2: N is rearrangeable if and only if

$$(B_c)' \text{ is } d\text{-dense in } I_c, \qquad \text{for each } c \in U.$$

Proof: Let N be rearrangeable; let $c \in U$; and pick x in I_c. Then there exists $y \in S$ such that

$$\gamma(y) = \gamma(x) \cup c,$$

and so there exists a $z \in E \cap \gamma^{-1}(c)$ such that $z \leqq y$ and $x \sim y - z$. Obviously then

$$y - z \in (B_c)'$$

and since x is equivalent to $y - z$ we have

$$x \in ((B_c)')^d.$$

Since x was an arbitrary member of I_c, we have proved $I_c \subseteq ((B_c)')^d$. Conversely, assume that the condition in the theorem holds, and pick any $c \in U$, and $x \in I_c$. Then $x \sim y$ for some y in $(B_c)'$, so that c is not blocked in y. Thus N is rearrangeable for all $c \in U$, and so is rearrangeable.

A similar argument yields the weaker and simpler result:

Remark 9: If B' is d-dense in S, then N is rearrangeable. In this case, since $S \subseteq (B')^d$, given a state x there is always an equivalent nonblocking state y, with

$$y \sim x, \; y \in B'.$$

Hence rearrangements can be made uniformly in the calls new to x.

XII. NETWORKS NONBLOCKING IN THE WIDE SENSE

We now turn to the characterization of networks for which there is a rule for routing calls which allows the operating system to avoid blocking states entirely. The case in which the network is actually nonblocking in the strict sense, so that *any* rule will do, is excluded here as trivial. The point is to use a network with blocking states, but to manage to avoid them by clever routing. The following general criterion of a useful rule $\rho(\cdot\,,\,\cdot)$ suggests itself: $\rho(\cdot\,,\,\cdot)$ should make as many blocking states as possible unreachable, consistent with satisfying requests for connection by unblocked new calls.

To exhibit, in an intuitive way, all the relationships that obtain, it is convenient to introduce an additional concept: a class X of network states is *preservable* (*by new calls*) if and only if for any $x \in X$ and any call c that is new to x and unblocked in x, there is a state $y \in X$ such that

$$y > x \quad \text{and} \quad \gamma(y) = \gamma(x) \cup c.$$

That is, if an idle pair (u, v) of x corresponds to a call $c = \{(u, v)\}$ that is unblocked in x, then some state $y \in X$ realizes $\gamma(x) \cup \{(u, v)\}$,

and y is above x in the state-diagram, $y > x$. In words, X is preservable if any call that can be put up at all in a state of X can be put up *salva* staying in X, that is, in such a way that the system stays in X. A $\leqq$-closed class is always preservable (by new calls). We make

Remark 10: If X is preservable, $0 \in X$, and $X \subseteq B'$, then X is sufficient.

It is then possible to start at the zero state, and call by call realize any maximal assignment *salva* staying in X. We now state

Theorem 3: N is nonblocking in the wide sense if and only if there exists a nonempty subset X of states such that

i. X is preservable.

ii. $X \subseteq B'$.

iii. X is $\leqq$-closed, i.e., $X = \underline{X}$.

Proof: Let (*i*)–(*iii*) hold for some subset X, and define a rule $\rho(\cdot\ ,\ \cdot)$ by the condition that if $c \in U$ is new to x, then

$$\rho(x, c) = \gamma^{-1}(\gamma(x) \cup c) \cap X.$$

Use of $\rho(\cdot\ ,\ \cdot)$ is tantamount to requiring that any call must be put up so as to lead to a state of X. By (*i*) and (*ii*), this can always be done. Since X is $\leqq$-closed, hangups preserve membership in X; since X is nonempty it contains the zero state. Hence all states reachable under $\rho(\cdot\ ,\ \cdot)$ belong to $X \cap B'$ and

$$R_\rho \subseteq B',$$

so that N is nonblocking in the wide sense.

Conversely, if N is nonblocking in the wide sense, then some rule $\rho(\cdot\ ,\ \cdot)$ is such that no blocking state belongs to R_ρ. Set $X = R_\rho$. Then X is $\leqq$-closed, because any state below a reachable state is reachable by hangups. Also $X \subseteq B'$, because $\rho(\cdot\ ,\ \cdot)$ avoids all blocking states. Finally, X must be preservable since one can "preserve" X simply by using only state-transitions permitted by $\rho(\cdot\ ,\ \cdot)$, i.e., by putting up unblocked new calls so as to lead only to states vouchsafed by $\rho(\cdot\ ,\ \cdot)$.

We recall that for $x \in S$,

$$\begin{aligned} A_x &= \{y \mid y \text{ covers } x\}, \\ &= \{y \mid y = x \cup z \text{ for some } z \in E\}, \\ &= \{\text{set of states immediately above } x\}. \end{aligned}$$

The property of preservability (of a set X of states) will now be given a topological characterization in terms of *denseness*, in the following result:

Theorem 4: A nonempty subset X of S is preservable if and only if for every $x \in X$, $A_x \cap X$ is dense in A_x in the sense of the d-metric; i.e., $x \in X$ implies

$$A_x \subseteq (A_x \cap X)^d.$$

Proof: Take $x \in X$ and $y \in A_x$, so that y is "immediately above" x, or y covers x. Then there is a call c new in x such that

$$\gamma(y) = \gamma(x) \cup c,$$

and so if $A_x \cap X$ is dense in A_x, there is a $z \in A_x \cap X$ which is equivalent to y. Since z covers x, it follows that the call c new to x can be connected in state x so as to give rise to a state of X. That is, we have

$$z \in X$$

$$\gamma(z) = \gamma(x) \cup c.$$

Since c was an arbitrary new call of $x \in X$, the set X is preservable, if the condition of Theorem 4 is true. Conversely, let X be preservable, and take $x \in X$ and $y \in A_x$. Then there exists a call c not blocked in x with $\gamma(y) = \gamma(x) \cup c$. But since X is preservable, and c is not blocked in x, there is a z in $A_x \cap X$ such that $\gamma(z) = \gamma(x) \cup c$, that is $z \sim y$. Hence y is equivalent to an element of $A_x \cap X$. Since y was arbitrary, it follows that for $x \in X$,

$$A_x \subseteq (A_x \cap X)^d.$$

Remark 11: The sets $\{A_x, x \in X\}$ in the condition of Theorem 4 may be replaced by the "x-cones"

$$\{y \mid y > x\}.$$

XIII. NETWORKS NONBLOCKING IN THE STRICT SENSE

A network that is nonblocking in the strict sense has no blocking states whatever. A simple characterization of this property is given by

Theorem 5: N is nonblocking in the strict sense if and only if there is a subset X of B' such that

i. X is sufficient

ii. X is d-closed, i.e., $X = X^d$.

Proof: If N has the property, then $S = B'$, and we may take $X = S$. Conversely, if (i) and (ii) obtain, take any $x \in S$; since X is sufficient, there exists $y \in X$ for which $\gamma(x) = \gamma(y)$, i.e., $x \sim y$. But $X = X^d$, so $x \in X$, and hence $x \in B'$.

By Remark 10, the condition (i) that X be sufficient can be replaced by the condition that X be preservable and nonempty.

IX. GLOSSARY

G	an arbitrary graph
g	a copy of G with each edge labeled 0 or 1
V	the set of all labeled versions g of G
S	the set of network states (typical members x, y, z)
T	the set of terminals (nodes of G)
p	a typical path (chain) on G
ϕ	null set
I	the set of inlets
Ω	the set of outlets
L	the set of links
C	the set of all connections (paths from I to Ω)
E	the set of elementary states
N	an arbitrary network, specified by choosing G, I, Ω, and E
$\leqq$	partial ordering of V or S by inclusion
0	zero state
$\mid x \mid$	number of elements of the set X
L_k	set of states with exactly k calls up
D	state diagram (Hasse figure of $\leqq$ on S)
$\mu(\cdot)$	Möbius function of $\leqq$
$\lambda(x, n)$	number of chains of length n from 0 to x
$\eta(x)$	$\lambda(x, \mid x \mid)$
A_x	set of states directly above x
B_x	set of states directly below x
C_n	$\sum_{x \in L_n} \eta(x)$
$a(\cdot)$	an assignment (any 1–1 map of subset of I into Ω)
$\hat{A}$	set of maximal assignments
U	set of unit assignments or calls
c	a call, or typical member of U
$\gamma(x)$	the assignment realized by state x
$\sim$	equivalence of states
$d(x, y)$	$\mid \gamma(x) \Delta \gamma(y) \mid$
$\underline{X}$	$\leqq$-closure of X
$\bar{X}$	$\geqq$-closure of X
X^d	d-closure of X
B	set of states in which some call is blocked
$\rho(\cdot\,, \cdot)$	a rule for operating a network
R_ρ	the set of states reachable under $\rho(\cdot\,, \cdot)$.

REFERENCES

1. Beneš, V. E., Heuristic Remarks and Mathematical Problems Regarding the Theory of Switching Systems, this issue, pp. 1201–1247.
2. Beneš, V. E., On Rearrangeable Three-Stage Connecting Networks, to appear.
3. Birkhoff, G., Lattice Theory, Amer. Math. Soc. Colloq. Publ. XXV, rev. ed., 1948.
4. Weisner, L., Trans. Amer. Math. Soc., **38,** 1955, pp. 474–484.
5. Hall, P., Quarterly Journal of Mathematics, **7,** 1936, pp. 134–151.
6. Kelley, J. L., *General Topology*, D. Van Nostrand, New York, 1955.
7. Clos, C., A Study of Non-Blocking Switching Networks, B.S.T.J., **32,** 1953, pp. 406–424.
8. Kharkevich, A. D., Multi-Stage Construction of Switching Systems (in Russian), Doklady AN SSSR, **112,** 1957, pp. 1043–6.

Solution of Systems of Linear Ordinary Differential Equations with Periodic Coefficients

By H. E. MEADOWS

(Manuscript received February 7, 1962)

An analysis technique is presented to provide an essentially explicit solution for a system of n simultaneous first-order linear differential equations with periodic coefficients. This representation of a periodic variable-parameter linear system of arbitrary finite order is chosen for its theoretical and practical advantages over the classical nth order linear differential equation. Emphasis is placed on natural mode solutions of a homogeneous set of equations. The characteristic exponents for these solutions are determined from a polynomial equation the coefficients of which are linear combinations of n − 1 convergent infinite-order determinants. Approximate calculation of these determinants is feasible for problems of moderate order.

1. INTRODUCTION

Systems of linear ordinary differential equations with periodic coefficients are assuming an increasing importance in engineering problems. Two applications of present interest are periodically time-variable networks and multimode waveguide with periodic physical distortions. Such applications have usually been analyzed by methods appropriate to special cases such as the second-order case or by approximate techniques valid for almost constant-parameter systems. However, perturbation techniques for almost stationary systems are inadequate for careful analysis of large-signal behavior of time-variable networks. Similarly, a periodically distorted helix waveguide, for which more than two modes must be considered,[1] should be described by a differential system of order greater than two. These examples illustrate the importance of a technique for obtaining essentially explicit solutions of periodic variable-parameter linear systems. Solutions in terms of characteristic exponents are known to exist for systems of linear differential equations with periodic coefficients.[2] However, the methods usually

employed for solving such systems, such as power-series techniques, iterative processes, and incremental numerical solution methods, fail to provide a system response description valid for all values of the independent variable (time, distance, etc.).

The analysis method to be presented below provides an essentially explicit solution for periodic variable-parameter linear systems of arbitrary finite order. The solution describes the system behavior for all values of the independent variable. Emphasis will be placed on obtaining a set of basis functions for a homogeneous system, since the solution in the inhomogeneous case can be obtained from the basis functions. As shown by Darlington,[3] these functions may be regarded as analogues of partial fractions in fixed network theory.

II. FORMULATION OF DIFFERENTIAL SYSTEM

In this discussion the system of equations to be solved will be represented by the vector differential equation

$$F'(t) = B(t)F(t) \tag{1}$$

where $F(t)$ and $B(t)$ are nth-order column and square matrices, respectively, and the prime denotes differentiation with respect to the independent variable t. It is supposed that the elements of $B(t)$ are known functions of t with a common period of unity, i.e.,

$$B(t) = B(t + 1). \tag{2}$$

The formulation of this problem in (1) is chosen not only for its elegance, but also because of its practical advantages. As indicated by Kinariwala[4] these include the ability to write such an equation directly from a time-variable network, the fact that the eigenvalues of $B(t)$ are natural frequencies for stationary networks, and the convenience of (1) in obtaining the quadratic forms representing stored energy and dissipated power in stationary or nonstationary cases. These advantages have their translated versions in other physical problems, including multimode waveguide problems. Moreover, an equation such as (1) is easily obtained from an nth-order linear differential equation, but the transformation from (1) to such an equation can be quite difficult (or analytically inconvenient).[4] Thus, (1) represents a well-founded beginning for the analysis of variable-parameter problems of practical or theoretical interest.

III. FORM OF SOLUTIONS

The form of solutions of (1) is well known;[2] pertinent properties of such solutions will be reviewed here briefly. If $B(t)$ is piecewise con-

tinuous (1) has the unique solution

$$F(t) = X(t)F(0) \tag{3}$$

where $X(t)$ is the unique nonsingular square matrix satisfying

$$\begin{aligned} X' &= BX \\ X(0) &= I = \text{diag}\,\{1\}. \end{aligned} \tag{4}$$

When $B(t)$ satisfies (2), $X(t)$ may be written as

$$X(t) = J(t)\, e^{Kt} \tag{5}$$

where

$$J(t) = J(t+1) \tag{6}$$

and

$$e^{K} = X(1). \tag{7}$$

For convenience it will be assumed here that the eigenvalues of K are distinct, or at least that K can be diagonalized; thus, a constant nonsingular matrix P exists so that

$$K = PMP^{-1} \tag{8}$$

where

$$M = \text{diag}\,\{\mu_i\} \tag{9}$$

and the constants μ_i are the eigenvalues of K. The matrix exponential function in (5) may be similarly diagonalized, so that the solution (3) may be constructed in the form

$$F(t) = J(t)P[\text{diag}\,\{e^{\mu_i t}\}]P^{-1}F(0). \tag{10}$$

By establishing the special initial conditions

$$F_i(0) = P \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{\text{th}} \text{ row} \tag{11}$$

the corresponding unique solution

$$F_i(t) = e^{\mu_i t} J(t) F_i(0), \tag{12}$$

is obtained from (10). Thus, by proper choice of initial conditions a

set of n solutions of the form

$$F(t) = e^{\mu t}Q(t) \tag{13}$$

where μ is a scalar constant and $Q(t)$ is a column matrix with period unity, have been shown to exist.

The n solutions in the form (12) or (13) represent natural modes of the periodic system described by (1) and (2). If the n values of μ_i are distinct the corresponding n solutions are certainly independent and form a set of basis solutions of (1). Any other solution of (1) comprises a linear combination of solutions like (12) or (13). Moreover, as Darlington[3] has pointed out, these natural-mode solutions are essentially unique because of their simple form. Hence the natural modes given by (12) represent a complete and essentially unique description of the natural behavior of the periodic system. The eigenvalues μ_i, frequently referred to as characteristic exponents, play a role analogous to response poles or natural frequencies of stationary systems. The strength of each natural mode in the homogeneous case is determined by the initial conditions and the constant matrix P. Moreover, the natural-mode solutions allow a complete solution to be calculated in the inhomogeneous case. Thus, the determination of n corresponding solutions for μ and $Q(t)$ in (13) is central to the problems associated with (1) and (2).

The object of the present treatment is to indicate a technique for determining the characteristic exponents μ, as well as the corresponding matrices $Q(t)$ if desired. Primary attention is given in finding the characteristic exponents μ because of their practical importance and because the solution for $Q(t)$ is not greatly difficult in principle if the appropriate characteristic exponent is known. Solutions for $Q(t)$ are mentioned in Appendices A and B.

The method to be discussed resembles the technique used by Hill[5,6] in solving the second-order equation

$$x''(t) + A(t)x(t) = 0 \tag{14}$$

where $A(t)$ is periodic. It will be shown that the characteristic exponents may be determined from roots of either a transcendental or polynomial equation in which certain infinite-order determinants enter as parameters. A technique similar to Hill's was employed by H. von Koch in the last century to provide an explicit solution in terms of infinite-order determinants for a general nth-order linear differential equation with periodic coefficients. This technique is carefully discussed by Forsyth[7] and Riesz,[8] who also give references to von Koch's original papers. Thus, the method presented here, although developed independently, does not solve an unsolved mathematical problem when applied to a periodic

variable-parameter system described by an nth-order linear differential equation. It does, however, solve the stated problem in a way that appears to have several advantages, mostly associated with its formulation as a system of n simultaneous first-order linear equations. These advantages, already mentioned in Section II, seem likely to make the present solution technique more useful in the analysis and synthesis of periodic variable-parameter systems than one based entirely on the classical nth-order linear differential equation.

IV. INTEGRAL FORM FOR THE PERIODIC SYSTEM

The analysis of the periodic system begins by multiplying both members of (1) by e^{-at}, where a is an arbitrary constant, and adding and subtracting aFe^{-at} to yield, whenever F' exists,

$$(Fe^{-at})' + aFe^{-at} = BFe^{-at}. \tag{15}$$

Integration of (15) results in the integral equation

$$Fe^{-at} + a\int Fe^{-at}\,dt = \int BFe^{-at}\,dt + C \tag{16}$$

where C is a constant. Any solution of (15) is also a solution of (16); thus, let F be a solution given by (13) and let

$$a = \mu + j2\pi k \tag{17}$$

where k is an arbitrary integer. Equation (16) becomes

$$Qe^{-j2\pi kt} + (\mu + j2\pi k)\int Qe^{-j2\pi kt}\,dt = \int BQe^{-j2\pi kt}\,dt + C. \tag{18}$$

If (18) is evaluated at $t = 0$ and $t = 1$, and the results subtracted, the first term in (18) makes no contribution, being periodic. Hence, (18) implies

$$(\mu + j2\pi k)\int_0^1 Q\,e^{-j2\pi kt}\,dt = \int_0^1 BQe^{-j2\pi kt}\,dt \tag{19}$$

for all integers k. It will be seen below that this integral equation suffices to determine μ and $Q(t)$, which are essentially eigenvalues and eigenfunctions.

V. MATRIX DIFFERENCE EQUATION

To make use of (19) in finding solutions of (1) it will be assumed that the given matrix $B(t)$ and the solution matrix $Q(t)$ may be expanded in the Fourier series

$$B(t) = \sum_{p=-\infty}^{\infty} B_p e^{j2\pi pt} \tag{20}$$

and

$$Q(t) = \sum_{p=-\infty}^{\infty} Q_p e^{j2\pi pt} \tag{21}$$

where matrices B_p are square matrices and Q_p are column matrices. Requirements on the asymptotic behavior of the elements of matrices B_p and Q_p for large values of $| p |$ will be discussed in Appendix A in relation to convergence of certain infinite-order determinants. The Fourier series for the matrix product BQ may be written as

$$BQ = \sum_{p=-\infty}^{\infty} (BQ)_p e^{j2\pi pt} \tag{22}$$

in which the column matrices $(BQ)_p$ are given by the convolution

$$(BQ)_p = \sum_{r=-\infty}^{\infty} B_{p-r} Q_r . \tag{23}$$

Except for a factor of 2π the integrals in (19) express the Fourier coefficients of Q and BQ. Thus, if Q and BQ possess Fourier series (19) is equivalent to the infinite set of linear equations

$$(\mu + j2\pi k)Q_k = (BQ)_k \tag{24}$$

or

$$(\mu + j2\pi k)Q_k = \sum_{r=-\infty}^{\infty} B_{k-r} Q_r , \tag{25}$$

where k assumes all integral values. Equation (25) might be regarded as a matrix difference equation for Q_p ; however it is more convenient here to consider (25) as defining an eigenvalue problem for an infinite matrix. In terms of Kronecker's δ, (25) is

$$0 = \sum_{r=-\infty}^{\infty} [B_{k-r} - \delta_{kr}(\mu + j2\pi k)I]Q_r \tag{26}$$

where I is the nth-order unit matrix. The expanded form of (26) is shown in the following infinite-order matrix equation, in which the first matrix is partitioned into $n \times n$ size blocks and the second into $n \times 1$ size blocks. The "origins" fall at $(B_0 - \mu I)$ and Q_0 .

$$\begin{vmatrix} \cdot & & & & & & & \\ \cdot & B_0-(\mu-j4\pi)I & B_{-1} & B_{-2} & & & & \\ \cdot & B_1 & B_0-(\mu-j2\pi)I & B_{-1} & B_{-2} & & & \\ \cdot & B_2 & B_1 & B_0-\mu I & B_{-1} & B_{-2} & & \\ & & B_2 & B_1 & B_0-(\mu+j2\pi)I & B_{-1} & & \\ & & & B_2 & B_1 & B_0-(\mu+j4\pi)I & \cdots & \\ & & & & & & \ddots & \end{vmatrix} \begin{vmatrix} \vdots \\ Q_{-2} \\ Q_{-1} \\ Q_0 \\ Q_1 \\ Q_2 \\ \vdots \end{vmatrix} = \begin{vmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{vmatrix} \tag{27}$$

For convenience it will be assumed that B_0 is a triangular matrix so that its eigenvalues appear explicitly as main diagonal elements. To show that a constant linear transformation of the dependent variable F can always produce this property, let

$$B(t) = B_0 + A(t) \tag{28}$$

where $A(t)$ has a zero mean, and let

$$X = PF \tag{29}$$

where P is a nonsingular matrix of constants. Then (1) is transformed to

$$X' = (PB_0P^{-1} + PAP^{-1})X. \tag{30}$$

This equation has the same form as (1), but the constant term in its coefficient matrix is the matrix PB_0P^{-1} derived from B_0 by a similarity transformation. It is well known that a square matrix is reducible by a similarity transformation to the classical canonical form having eigenvalues on the main diagonal and possibly nonzero constants in some positions of the next higher diagonal.[9] (These constants cannot appear if B_0 has distinct eigenvalues; hence, B_0 can often be assumed to be diagonal.) The matrix B_0 can also be reduced to a triangular form by a similarity transformation in which P is a unitary matrix.[10]. This reduction, which is always possible, may sometimes have advantages in studying energy functions or related quadratic forms. Thus, by either technique B_0 can be reduced to triangular form. It will be assumed that such a transformation has been effected in obtaining (1).

VI. CONVERGENT LINEAR EQUATIONS AND INFINITE-ORDER DETERMINANT

To produce convergence of the determinant of coefficients of the infinite set of homogeneous equations defining Q, Equations (26) or

(27), divide each elementary row of the coefficient matrix in Equation (27) by its main-diagonal elements. When B_0 is diagonal this process is identical with multiplication of each equation in Equation (26) or each matrix row in the square matrix of Equation (27) by the diagonal matrix $[B_0 - (\mu + j2\pi k)I]^{-1}$. In general, let the matrix Λ be defined by

$$\Lambda = \text{diag}\,\{\lambda_p\} \tag{31}$$

where λ_p represent the n eigenvalues of B_0. Then the set of equations with convergent determinant may be written as

$$[M_{kr}][Q_r] = 0 \tag{32}$$

where submatrices M_{kr} are given by

$$M_{kk} = [\Lambda - (\mu + j2\pi k)I]^{-1}[B_0 - (\mu + j2\pi k)I] \tag{33}$$

and

$$M_{kr} = [\Lambda - (\mu + j2\pi k)I]^{-1}B_{k-r}\,, \qquad k \neq r. \tag{34}$$

When B_0 is diagonal, M_{kk} reduces to the unit matrix. Thus, a typical determinant $d[M_{kr}]$ of (32) may be illustrated for $n = 2$ by the following scheme:

$$d(\mu) = \begin{vmatrix} \ddots & & & & & & & \\ \cdots & 1 & 0 & \frac{a_{-1}}{\lambda_1+j2\pi-\mu} & \frac{b_{-1}}{\lambda_1+j2\pi-\mu} & \frac{a_{-2}}{\lambda_1+j2\pi-\mu} & \frac{b_{-2}}{\lambda_1+j2\pi-\mu} & \\ \cdots & 0 & 1 & \frac{c_{-1}}{\lambda_2+j2\pi-\mu} & \frac{d_{-1}}{\lambda_2+j2\pi-\mu} & \frac{c_{-2}}{\lambda_2+j2\pi-\mu} & \frac{d_{-2}}{\lambda_2+j2\pi-\mu} & \\ \cdots & \frac{a_1}{\lambda_1-\mu} & \frac{b_1}{\lambda_1-\mu} & 1 & 0 & \frac{a_{-1}}{\lambda_1-\mu} & \frac{b_{-1}}{\lambda_1-\mu} & \\ & \frac{c_1}{\lambda_2-\mu} & \frac{d_1}{\lambda_2-\mu} & 0 & 1 & \frac{c_{-1}}{\lambda_2-\mu} & \frac{d_{-1}}{\lambda_2-\mu} & \\ & \frac{a_2}{\lambda_1-j2\pi-\mu} & \frac{b_2}{\lambda_1-j2\pi-\mu} & \frac{a_1}{\lambda_1-j2\pi-\mu} & \frac{b_1}{\lambda_1-j2\pi-\mu} & 1 & 0 & \\ & \frac{c_2}{\lambda_2-j2\pi-\mu} & \frac{d_2}{\lambda_2-j2\pi-\mu} & \frac{c_1}{\lambda_2-j2\pi-\mu} & \frac{d_1}{\lambda_2-j2\pi-\mu} & 0 & 1 & \cdots \\ & & & & & & & \ddots \end{vmatrix} \tag{35}$$

The notation used in (35) is

$$B_p = \begin{bmatrix} a_p & b_p \\ b_p & d_p \end{bmatrix}, \qquad p \neq 0 \tag{36}$$

$$B_0 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}. \tag{37}$$

The determinant of the infinite-order matrix $[M_{kr}]$ of (32), illustrated by (35) for $n = 2$, will be denoted by $d(\mu)$ to show its functional dependence on the argument μ. The function $d(\mu)$ is actually a determinant of infinite order. If this determinant converges it represents a function of μ which must vanish in order to obtain nontrivial solutions for Q_r in (32). Requirements necessary for the convergence of $d(\mu)$ are discussed in Appendix A, where it will be shown that $d(\mu)$ converges for a large class of problems. Hence the basic equation

$$d(\mu) = 0 \tag{38}$$

defines the characteristic exponents of the differential system (1).

VII. FUNCTIONAL EXPRESSIONS FOR THE CHARACTERISTIC DETERMINANT

Equation (38) taken alone is rather unwieldy, involving as it does the equation to zero of an infinite-order determinant whose elements are functions of μ. However, it will now be shown that expressions for $d(\mu)$ in terms of elementary functions may be written to allow a simple solution of (38).

The determinant $d(\mu)$ is shown in Appendix A to converge for all values of μ except those for which the denominators of rows of $d(\mu)$ vanish. Multiplication of one row of an infinite-order determinant by any scalar is equivalent to multiplication of the determinant by the same scalar. Similarly multiplication of any row of $d(\mu)$ by its corresponding denominator $\lambda_p - (\mu + j2\pi k)$ produces a determinant convergent at $\lambda_p = \mu + j2\pi k$, so that each row of $d(\mu)$ introduces exactly one pole in $d(\mu)$. Moreover, $d(\mu)$ is periodic in μ with period $j2\pi$, since replacing μ by $\mu + j2\pi$ only shifts the origin of the infinite-order determinant. Evidently $d(\mu)$ has simple poles at

$$\mu = \lambda_p + j2\pi q, \qquad p = 1,2, \cdots, n \qquad q \text{ integral.} \tag{39}$$

It will be assumed for the moment that these poles are distinct; this restriction may be relaxed slightly, as shown in Appendix B. Finally, as μ approaches infinity along any radial line in the complex μ plane except a vertical line, the off-diagonal elements in $d(\mu)$ tend toward zero, or briefly

$$d(\infty) = 1. \tag{40}$$

The periodicity of $d(\mu)$ implies that the residue of $d(\mu)$ at any of the poles in (39) is independent of the particular integer q. Thus, a formal expansion of $d(\mu)$ in partial fractions is

$$d(\mu) = K_\infty + \sum_{p=1}^{n} \sum_{q=-\infty}^{\infty} \frac{K_p}{\mu - \lambda_p - j2\pi q}. \tag{41}$$

According to the Mittag-Leffler theorem[11] this expansion defines the function

$$d(\mu) = K_\infty + \frac{1}{2}\sum_{p=1}^{n} K_p \coth\left(\frac{\mu - \lambda_p}{2}\right). \tag{42}$$

A relation fixing K_∞ may be derived from (40) by noting that as μ approaches infinity along any nonvertical radial line

$$\lim_{\mu\to\infty} \coth\left(\frac{\mu - \lambda_p}{2}\right) = 1 \tag{43}$$

so that

$$K_\infty = 1 - \frac{1}{2}\sum_{p=1}^{n} K_p. \tag{44}$$

To compute the residues K_p the well-known rule

$$K_p = \lim_{\mu\to\lambda_p} (u - \lambda_p)d(\mu) = [(\mu - \lambda_p)d(\mu)]_{\mu=\lambda_p} \tag{45}$$

is employed. The procedure is simply to multiply every element in the row of $d(\mu)$ containing $\lambda_p - \mu$ (in the denominators) by the factor $(\mu - \lambda_p)$ and to evaluate the resulting determinant. For example, in the case of $n = 2$ used above for illustration, the row of $d(\mu)$ containing $\lambda_1 - \mu$ in the denominators is replaced by

$$\cdots\ -a_2\ -b_2\ -a_1\ -b_1\ 0\quad 0\ -a_{-1}\ -b_{-1}\ -a_{-2}\ -b_{-2}\ \cdots \tag{46}$$

and the resulting determinant evaluated at $\mu = \lambda_1$. Reasonably accurate and efficient methods for computing K_p from such a determinant can be programmed just as for Hill's determinant in the second-order case. Such a technique is discussed briefly in Appendix C.

It is well known that the solution of Hill's equation generally requires the evaluation of only one infinite-order determinant, while the solution of a second-order problem using (38) and (42) appears to require the evaluation of two determinants. Actually it will be shown that only $n - 1$ determinants need be calculated for an nth-order system of equations. In addition (42) may be simplified because of the relation among the residues K_p to be demonstrated below.

To examine the poles and zeros of $d(\mu)$ it is convenient to consider the complex μ plane divided into horizontal strips of width 2π. The poles of $d(\mu)$ fall at $\lambda_p + j2\pi q$. Although the eigenvalues λ_p may lie in any of these strips, values of q always exist to give one pole in the fundamental strip $0 \leqq \text{Im } \mu < 2\pi$ representative of each λ_p. Hence $d(\mu)$ has exactly n poles in each strip. It will be seen shortly that $d(\mu)$

also has n zeros in each strip so that a pole-zero constellation for $d(\mu)$ might be illustrated by Fig. 1.

The desired relation among the residues K_p is obtained by noting that

$$\int_{abcd} d(\mu)\, d\mu = \sum_{p=1}^{n} K_p \tag{47}$$

where the integral is taken around the rectangular contour *abcd* shown in Fig. 1 (or a congruent rectangle vertically displaced if a pole happens to fall at Im $\mu = 0$). The periodicity of $d(\mu)$ insures that the contributions to the integral from the horizontal sides *ab* and *cd* will cancel. The vertical sides *bc* and *da* are supposed to be displaced from the origin far enough to include all n poles in the rectangle so that (47) is valid. As their displacement approaches infinity the value of $d(\mu)$ approaches unity and the contributions from the vertical sides tend to cancel. Thus (47) implies

$$\sum_{p=1} K_p = 0. \tag{48}$$

This relation shows that (44) and (42) may be simplified to

$$K_\infty = 1 \tag{49}$$

and

$$d(\mu) = 1 + \frac{1}{2} \sum_{p=1}^{n} K_p \coth\left(\frac{\mu - \lambda_p}{2}\right). \tag{50}$$

It also allows one residue to be computed from the other $n - 1$, al-

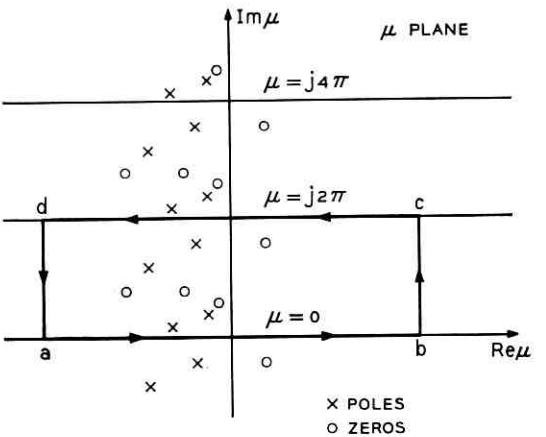


Fig. 1 — Pole-zero constellation.

though all n residues might be computed in practice and (48) used as a check for numerical accuracy.

Equation (50) expresses the characteristic determinant $d(\mu)$ in terms of the eigenvalues λ_p of the stationary part of the system and the residue determinants K_p. The characteristic exponents μ are thus by (38) and (50) the roots of the trigonometric equation

$$0 = 1 + \frac{1}{2}\sum_{p=1}^{n} K_p \coth\left(\frac{\mu - \lambda_p}{2}\right). \tag{51}$$

This trigonometric equation represents an explicit solution of the problem of finding characteristic exponents for an nth-order periodic system.

It is evident from (50) as well as from Fig. 1 and the periodicity of the function e^{μ} that the substitution

$$z = e^{\mu} \tag{52}$$

reduces (50) to a rational function in z. Zeros and infinities of z do not introduce superfluous poles or other singularities in this function because of (43). Thus, the poles and zeros of this rational function are mapped by (52) into the poles and zeros of $d(\mu)$ shown in Fig. 1. Any strip of vertical width 2π in the μ plane is mapped by (52) into the entire z plane so that the rational function of z has n poles in the z plane. The number of zeros of the rational function is also necessarily n. Hence, $d(\mu)$ has precisely n zeros in any horizontal strip of width 2π in the μ plane.

Because of the existence of well-developed techniques for polynomial manipulation, such as approximate solution methods, interpolation formulae, and stability criteria, it is practically convenient to utilize (50) and (51) in rational form. Accordingly let z be defined by (52) and x_p by

$$x_p = e^{\lambda_p}, \qquad p = 1,2, \cdots, n, \tag{53}$$

so that $d(\mu)$ is transformed to

$$D(z) = 1 + \frac{1}{2}\sum_{p=1}^{n} K_p \left(\frac{z + x_p}{z - x_p}\right) = d(\log z). \tag{54}$$

Further, let

$$g(z) = \prod_{p=1}^{n} (z - x_p) \tag{55}$$

be a characteristic polynomial defining the eigenvalues of the stationary part of $B(t)$. (This "characteristic polynomial" differs from the con-

ventional one in that its roots are e^{λ_p} rather than λ_p.) Equation (51), the characteristic equation, then becomes

$$0 = f(z) + g(z) \tag{56}$$

where

$$f(z) = \frac{1}{2} \sum_{p=1}^{n} K_p(z + x_p) \prod_{\substack{q=1 \\ q \neq p}}^{n} (z - x_q). \tag{57}$$

These equations demonstrate that the characteristic polynomial for the periodic system is obtained by adding a certain interpolating polynomial to the characteristic polynomial of the stationary part of the system. The behavior of the interpolating polynomial is prescribed at the roots of the stationary part.

The interpolating polynomial $f(z)$ has the n assigned values

$$f(x_p) = K_p x_p \prod_{\substack{q=1 \\ q \neq p}}^{n} (x_p - x_q); \tag{58}$$

because of the relation (48) among the residues K_p the polynomial (57) is identical with the Lagrangian interpolating polynomial

$$f(z) = \sum_{p=1}^{n} K_p x_p \prod_{\substack{q=1 \\ q \neq p}}^{n} (z - x_q). \tag{59}$$

Evidently the interpolating polynomial $f(z)$ is the unique polynomial of degree $n - 1$ having the assigned values (58). Thus $f(z) + g(z)$, the characteristic polynomial of the periodic system, is the unique monic polynomial of degree n having the n assigned values given in (58). This point of view may give some insight into stability questions. For example, the classical criteria of Routh and Hurwitz, and other results on bounds of zeros of sums of polynomials may be useful here.

If all the residues K_p vanish, as in the stationary case, the limiting values of the characteristic exponents obtained from (51) and (56) are $u = \lambda_p$, $p = 1,2, \cdots, n$. In cases of small variations where all $|K_p|$ are small the characteristic exponents differ very little from the eigenvalues of the stationary part. Asymptotically they may be calculated from any of the approximate equations

$$0 \approx 1 + \frac{1}{2} K_p \coth\left(\frac{\mu - \lambda_p}{2}\right) \tag{60}$$

$$z \approx x_p(1 - K_p) \tag{61}$$

or

$$\mu \approx \lambda_p - K_p . \tag{62}$$

Although perturbation type solutions such as (62) probably are more easily calculated by less complicated techniques, characteristic exponents obtained from (61) or (62) may be useful as starting values for solving (51) or (56) by numerical methods.

VIII. EXAMPLE

The following example illustrates the technique for finding characteristic exponents. A second-order case is chosen for convenience because some digital computer programs needed for the efficient evaluation of the residue determinants are not yet available. However, higher-order examples are not different in principle nor will they require inordinately longer computations.

The Mathieu equation

$$\frac{d^2y}{dz^2} + (3 - 4 \cos 2z)y = 0 \tag{63}$$

has the solution[12]

$$y = e^{j\beta z} \sum_{r=-\infty}^{\infty} c_{2r+1}\, e^{j(2r+1)z} \tag{64}$$

with

$$\beta = \pm 0.57943224 \cdots . \tag{65}$$

In vector form this equation is equivalent to

$$Y' = \pi \begin{bmatrix} 0 & 1 \\ -3 + 4 \cos 2\pi t & 0 \end{bmatrix} Y \tag{66}$$

with the identifications

$$Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad y_1 = y, \qquad z = \pi t. \tag{67}$$

Diagonalization by the transformation $PY = F$ where

$$P = \frac{1}{j2\sqrt{3}} \begin{bmatrix} j\sqrt{3} & 1 \\ -j\sqrt{3} & 1 \end{bmatrix} \tag{68}$$

yields (1), with

$$B(t) = \frac{\pi}{j\sqrt{3}}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}e^{-j2\pi t} + j\pi\sqrt{3}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix} + \frac{\pi}{j\sqrt{3}}\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}e^{j2\pi t}. \tag{69}$$

The determinant $d(\mu)$ has the form shown by (35), and the residue at $\mu = j\pi\sqrt{3}$ is approximately $K = j0.562096$, a result obtained from a 42nd-order approximant. From (56) and (57) the characteristic exponents are solutions of

$$\cosh \mu = \cos \pi\sqrt{3} - jK \sin \pi\sqrt{3}, \tag{70}$$

which yields, for $K \approx j0.562096$

$$\mu \approx \pm j0.42059\pi. \tag{71}$$

Corresponding correct values of μ from (64) and (65) are $\pm j0.42057\pi$. A somewhat longer computation would be required to produce a result accurate enough for certain purposes. Such a computation was not employed here because a more fundamentally sound computing technique for band-limited periodic variations as in (69) would exploit the form of the residue determinant and its large number of zero elements. Specifically, it is possible to program a determinant evaluation technique for such cases so that the computation time is asymptotically proportional to the order of the truncated determinant rather than to its cube. This possibility is discussed further in Appendix C.

IX. CONCLUSIONS

A method has been developed for analysis and calculation of solutions of nth-order linear periodic differential systems. The system description employed is a set of n simultaneous first-order linear differential equations. The method allows the determination of characteristic exponents from polynomial equations the coefficients of which are linear combinations of $n-1$ convergent infinite-order determinants. Approximate computation of the determinants is feasible for problems of finite order. In addition to characteristic exponents the complete solutions may also be computed if desired.

APPENDIX A

Convergence

The validity of the analysis presented here depends upon the convergence of the infinite processes employed. It must be shown that the

determinant $d(\mu)$ and the Fourier series for $Q(t)$ are convergent if the coefficient matrix $B(t)$ is suitably restricted. For this purpose (32) may be written as the infinite set of scalar equations

$$x_i + \sum_{j=-\infty}^{\infty} a_{ij}x_j = 0 \tag{72}$$

where a_{ij} and x_j are scalars, and the equations hold for all integral i. The coefficients a_{ij} actually are elements of the submatrices M_{kr}, and x_i elements of submatrices Q_r in (32). The determinant of coefficients of the scalar equations is

$$d(\mu) = | \delta_{ij} + a_{ij} | . \tag{73}$$

According to a theorem of St. Bobr[13] this determinant is absolutely convergent if

$$\sum_{i=-\infty}^{\infty} | a_{ii} | \tag{74}$$

and

$$\sum_{i=-\infty}^{\infty} \left[\sum_{\substack{j=-\infty \\ j\neq i}}^{\infty} | a_{ij} |^{p/(p-1)} \right]^{p-1} \tag{75}$$

converge for some value of p in the interval $1 < p \leqq 2$. (For $p = 2$, the case used here, the theorem was given by von Koch.) The expression in (74) obviously converges to zero since all a_{ii} in (72) are zero. Let the elements of the given matrix $B(t)$ be square integrable functions. Then Parseval's relation applies and the Fourier series coefficients for the matrix elements are surely square summable. Hence, the inside sum in (75) converges for $p = 2$. The outside sum also converges for $p = 2$, since its general term is asymptotically proportional to i^{-2} for large $| i |$ (as (33) and (34) indicate by their dependence on k). Of course, an exception occurs for values of μ given by Equation (39). The determinant $d(\mu)$ is singular at these points, but the convergence of the residue determinants K_p for simple poles is assured by St. Bobr's theorem. Thus, $d(\mu)$ converges absolutely and uniformly except for μ arbitrarily near $\lambda_p + j2\pi q$ and has poles at these values of μ.

Since the determinant $d(\mu)$ has zeros at any of the n characteristic values of μ within the strip $0 \leqq Im\mu < 2\pi$, the deletion of the zeroth equation ($i = 0$) from (72) and the transposition of $a_{i0}x_0$ in each equation produces a nonzero determinant of coefficients in the equations

$$x_i + \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} a_{ij}x_j = -a_{i0}x_0 = y_i , \qquad i \neq 0. \tag{76}$$

These equations with a_{ij} evaluated at a characteristic value of μ have as solutions the scalar quantities x_j needed to produce the matrices Q_p and thus the matrix $Q(t)$. That meaningful solutions to (76) exist for an arbitrary constant x_0 is shown by a theorem of L. W. Cohen.[14] This theorem (paraphrased) states that if (75) converges for the coefficients in (76), if the (convergent) determinant of (76) does not vanish, and if

$$\sum_{i=-\infty}^{\infty} | y_i |^p$$

converges, then the solutions exist, may be obtained by Cramer's rule (with infinite-order minor determinants), and have the property that

$$\sum_{i=-\infty}^{\infty} | x_i |^p$$

converges. Thus, if the elements of the given matrix $B(t)$ are square integrable functions, the coefficients a_{i0} are surely square summable, and the resulting trigonometric series for elements of $Q(t)$ have square summable coefficients. The Riesz-Fischer theorem[15] then states that the elements of $Q(t)$ are square integrable functions with Fourier coefficients given by the elements of Q_p and that the Fourier series for $Q(t)$ converges to $Q(t)$ in the mean. (Consequently there exists a sequence of partial sums of the Fourier series converging to $Q(t)$ "almost everywhere.")

In a more restricted case which might have more practical importance, it may be shown that if $B(t)$ is continuous so that the elements of B_p are $0(1/p^2)$, the solution matrix $Q(t)$ has the same property. Of course, the Fourier series for $Q(t)$ converges absolutely and uniformly in this case. This convergence condition and the more general one above demonstrate that the analysis technique is valid for a wide class of problems.

APPENDIX B

Multiple Poles of the Characteristic Determinant

If the matrix B_0, the stationary part of the coefficient matrix $B(t)$, has repeated eigenvalues, or if any of its eigenvalues differ by integral multiples of $j2\pi$, some denominators of rows of $d(\mu)$ are identical. In this case $d(\mu)$ has multiple poles, and the necessary analytical and computational procedures become more complicated. It is possible to treat the case of a single second-order pole of $d(\mu)$ by evaluation of $n - 1$

determinants as before, but greater multiplicities require considerably more extensive calculations.

When $d(\mu)$ has an m-fold pole at $\mu = \lambda_1$ the partial fractions expansion of $d(\mu)$ must contain the corresponding principal part of $d(\mu)$. The coefficients in the principal part involve derivatives of $(\mu - \lambda_1)^m d(\mu)$ evaluated at $\mu = \lambda_1$. These derivatives are more difficult to compute than the residue determinants of the simple case because they are linear combinations of most of the first minors of $d(\mu)$. The computation of such minors (not necessarily by direct methods) is also required if $Q(t)$ is to be determined (even when $d(\mu)$ has only simple poles). Appendix A shows this computation to be theoretically possible; it is equivalent to the inversion of a set of equations like (76). Nevertheless, the computation effort would be considerably greater than that required for computation of characteristic exponents when $d(\mu)$ has only simple poles.

When $d(\mu)$ has a single second-order pole, (48) may be utilized to make possible the calculation of characteristic exponents. It is convenient here to use the rational form of (54) for the infinite-order determinant $d(\mu)$. Let the repeated roots of B_0 be identified with λ_1, λ_2 and x_1, x_2 respectively. Define α_1 and α_2 by

$$\begin{aligned} \alpha_1 &= K_1(x_1 - x_2) \\ \alpha_2 &= K_2(x_2 - x_1) \end{aligned} \tag{77}$$

and allow x_1 to approach x_2. Substitution of (77) in (54) yields

$$\begin{aligned} D(z) = 1 + \frac{1}{2}\left(\frac{\alpha_1 - \alpha_2}{x_1 - x_2}\right) \frac{z^2 - x_1x_2}{(z - x_1)(z - x_2)} \\ + \frac{\alpha_1 + \alpha_2}{2(z - x_1)(z - x_2)} + \frac{1}{2}\sum_{p=3}^{n} K_p\left(\frac{z + x_p}{z - x_p}\right). \end{aligned} \tag{78}$$

Equation (48) may be written as

$$\frac{\alpha_1 - \alpha_2}{x_1 - x_2} + \sum_{p=3}^{n} K_p = 0 \tag{79}$$

so that

$$L + \sum_{p=3}^{n} \lim_{x_1 \to x_2} K_p = 0 \tag{80}$$

where

$$L = \lim_{x_1 \to x_2} \left(\frac{\alpha_1 - \alpha_2}{x_1 - x_2}\right) = \lim_{x_1 \to x_2} (K_1 + K_2) \tag{81}$$

is a finite limit. Evidently, as x_1 approaches x_2,

$$\lim_{x_1 \to x_2} D(z) = 1 + \frac{L}{2}\left(\frac{z + x_2}{z - x_2}\right) + \frac{\alpha_2}{(z - x_2)^2} + \frac{1}{2}\sum_{p=3}^{n} K_p \left(\frac{z + x_p}{z - x_p}\right). \quad (82)$$

The zeros of this limiting form of $D(z)$ correspond to the characteristic exponents in this case. The parameter α_2 may be determined by factoring $1/(\lambda_2 - \lambda_1)$ from the appropriate row of K_2 and computing the resulting determinant Δ, since

$$\alpha_2 = \Delta \cdot \lim_{x_1 \to x_2} \left(\frac{x_2 - x_1}{\lambda_1 - \lambda_2}\right) = \Delta. \quad (83)$$

The parameter L required in (82) may be computed from (80). Cases where two poles of $d(\mu)$ are almost coincident may be treated in a similar fashion, except that no limits are involved.

APPENDIX C

Approximate Computation of Residue Determinants

In practical cases where the number of terms in the Fourier series for $B(t)$ is limited, truncated approximants to the residue determinants may be evaluated by techniques that exploit the special form of these determinants. The form of a truncated residue determinant is illustrated by the scheme in Fig. 2, in which all elements outside of the shaded region are zero. Except for one submatrix near the center of the array the principal diagonal blocks represent nonsingular triangular sub-

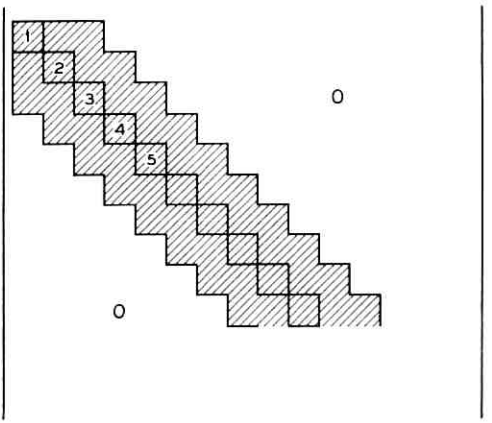


Fig. 2 — Form of truncated residue determinant.

matrices. To demonstrate the feasibility of computing truncated residue determinants of large order it will be shown that the computation time required for a reduction to triangular form is much smaller than for a general determinant of the same order. (The computation time required for a general determinant is asymptotically proportional to the cube of its order.)

To evaluate the determinant in Fig. 2 let zeros be produced below submatrix 1 by elementary operations with the rows passing through 1. Similar operations to produce zeros below 2 do not disturb the zeros already produced. Such operations may be continued in the usual manner to produce zeros below 3, 4, etc., until a triangular array of submatrices is realized. The number of arithmetic operations necessary in each step of zero production is essentially dependent only upon the order of the original system of equations and the number of terms in $B(t)$. Observation of Fig. 2 shows that the number of zero-producing steps for a truncated determinant of large order is asymptotically proportional to the order of the determinant. Thus, the computation time required for a reduction to triangular form is also asymptotically proportional to the order of the truncated determinant to be evaluated.

REFERENCES

1. Unger, H. G., Normal Modes and Mode Conversion in Helix Waveguide, B.S.T.J., **40,** 1961, pp. 255–280.
2. Coddington, E. A. and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
3. Darlington, S., An Introduction to Time-Variable Networks, Proc. of the Midwest Symposium on Circuit Analysis, Univ. of Illinois, 1955, pp. 5-1 to 5-25.
4. Kinariwala, B. K., Analysis of Time-Varying Networks, I.R.E. Conv. Record, **9,** Part 4, 1961, pp. 268–276.
5. Hill, G. W., Mean Motion of the Lunar Perigee, Acta Math., **8,** 1886, pp. 1–36.
6. Whittaker, E. T. and Watson, G. N., *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, Fourth Ed., 1927, Ch. 19.
7. Forsyth, A. R., *Theory of Differential Equations*, **4,** Cambridge Univ. Press, Cambridge, 1902, Chapters 8 and 9.
8. Riesz, F., *Les Systèmes d'Équations Linéaires a une Infinité d'Inconnues*, Gauthier-Villars, Paris, 1913, Chapters 2 and 6.
9. Wedderburn, J. H. M., Lectures on Matrices, Am. Math. Soc., Colloq. Publs., **17,** 1934.
10. Bellman, R., *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960, Ch. 11.
11. Knopp, K., *Theory of Functions*, Part II, Dover Publs., New York, 1947, Ch. 2.
12. Maclachlan, N. W., *Theory and Application of Mathieu Functions*, Oxford Univ. Press, Oxford, 1947.
13. Bobr, St., Eine Verallgemeinerung des v. Kochschen Satzes über die absolute Konvergenz der unendlichen Determinanten, Math. Zeit., **10,** 1921, pp. 1–11.
14. Cohen, L. W., A Note on a System of Equations with Infinitely Many Unknowns, Bull. Amer. Math. Soc., **36,** 1930, pp. 563–572.
15. Zygmund, A., *Trigonometrical Series*, Dover Publs., New York, 1955, Ch. 4.

Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III: The Dimension of the Space of Essentially Time- and Band-Limited Signals

By H. J. LANDAU and H. O. POLLAK

(Manuscript received February 23, 1962)

The purpose of this paper is to examine the mathematical truth in the engineering intuition that there are approximately 2WT independent signals φ_i of bandwidth W concentrated in an interval of length T. Roughly speaking, the result is true for the best choice of the φ_i (prolate spheroidal wave functions), but not for sampling functions (of the form sin t/t). Some typical conclusions are: Let f(t), of total energy 1, be band-limited to bandwidth W, and let

$$\int_{-T/2}^{T/2} | f^2(t) | \, dt = 1 - \epsilon_T^2 .$$

Then

$$\inf_{\{a_i\}} \int_{-\infty}^{\infty} \left| f(t) - \sum_{0}^{[2WT]+N} a_n \varphi_n \right|^2 dt < C\epsilon_T^2$$

is

(a) *true for all such f with N = 0, C = 12, if the φ_n are the prolate spheroidal wave functions;*

(b) *false for some such f for any finite constants N and C if the φ_n are sampling functions.*

I. INTRODUCTION AND SUMMARY OF RESULTS

Intuitive considerations based primarily on the sampling theorem have for a long time suggested that the space of signals "essentially" limited in time to the interval $| t | \leqq T/2$ and in frequency to $(-W,W)$ cycles is "essentially" $2WT$-dimensional. It is the object of the present

paper to investigate this problem thoroughly. The first step in the process is to see how the above statement may be made precise. The two main difficulties to be overcome in even *formulating* some mathematical problems in this area are contained in the two uses of "essentially" above: What shall we mean by "essentially" limited in time and frequency, and what can we mean by "essentially" $2WT$-dimensional?

Suppose that a function $f(t)$ is actually band-limited. It is then an analytic function of the complex variable t, and cannot vanish in $| t | > T/2$ without vanishing identically. We will therefore think of $f(t)$ as *approximately time-limited* to $| t | \leqq T/2$ if a large fraction of its energy is contained in that interval, that is, if

$$\frac{\int_{|t| \leqq T/2} | f(t) |^2 \, dt}{\int_{-\infty}^{\infty} | f(t) |^2 \, dt} = 1 - \epsilon_T^2 , \tag{0.1}$$

where ϵ_T will, in much of our thinking, be small; ϵ_T shall be used as a measure of the degree to which $f(t)$ fails to be concentrated on the interval $| t | \leqq T/2$. We will denote by $E(\epsilon_T)$ the set of band-limited functions $f(t)$ satisfying (0.1) with the further normalization for convenience that

$$\int_{-\infty}^{\infty} | f(t) |^2 \, dt = 1.$$

We should point out here that, by previous results,[1] T and ϵ_T are related: as ϵ_T becomes small, T must grow indefinitely.

We have now defined our set of functions; how can we speak precisely about its dimension? $E(\epsilon_T)$ is certainly not finite-dimensional for any $\epsilon_T > 0$, for there is no *finite* set of functions whose linear combinations exactly express each $f(t)$ in $E(\epsilon_T)$. We will, however, say that $E(\epsilon_T)$ is *approximately N-dimensional* if there exist N linearly independent functions $\varphi_0 , \cdots \varphi_{N-1}$ whose linear combinations approximate each $f(t)$ in $E(\epsilon_T)$ to within a small fraction of its energy, that is, if

$$\min_{\{a_i\}} \int_{-\infty}^{\infty} \left| f(t) - \sum_{0}^{N-1} a_i \varphi_i(t) \right|^2 dt < \delta_N^2 , \tag{0.2}$$

where we shall usually think of δ_N as small. Again, δ_N may be used as a measure of the degree to which $E(\epsilon_T)$ is N-dimensional.

In the above definition of the approximate dimension of $E(\epsilon_T)$, we have complete freedom in choosing the "basis" functions $\varphi_0 \cdots \varphi_{N-1}$

with which we will attempt to approximate $f(t)$. There are two different objectives we may have in choosing the φ_i. For real understanding of the dimension of $E(\epsilon_T)$ we must use the φ_i which *best approximate* $E(\epsilon_T)$, in the sense of making the error, represented by the left side of (0.2), as small as it can possibly be over the whole set $E(\epsilon_T)$. Alternatively, for practical purposes, we may wish to use the *simplest* available functions, and see how close we can come with them. Thus there is considerable interest in pursuing two lines of investigation:

(*i*) Let us first try to identify the best functions φ_i to use, that is the functions which achieve

$$\min_{\{\varphi_i\}_0^{N-1}} \max_{f \in E(\epsilon_T)} \min_{\{a_i\}_0^{N-1}} \int_{-\infty}^{\infty} \left| f(t) - \sum_0^{N-1} a_i \varphi_i(t) \right|^2 dt. \tag{0.3}$$

Once we have found these best functions, what is the relation between the number N of such functions, the measure of concentration ϵ_T, and the achievable degree of approximation δ_N?

(*ii*) If we pick for the φ's sampling functions, i.e., functions of the form $[\sin \pi(2Wt - r)]/[\pi(2Wt - r)]$, what is now the relation between N, ϵ_T, and δ_N?

It turns out that the answers to (*i*) and (*ii*) are rather different, that is, the degree of approximation achievable by sampling functions is in a very real sense poorer than the degree achievable by the *best* basis functions. And yet the solutions of the two problems are, as we shall see, remarkably intertwined.

In order to give a detailed picture of our results, it is necessary to summarize some of the previous work on time- and band-limiting which has appeared in Refs. 1 and 2.

The space $\mathfrak{L}^2$ of square-integrable functions on $(-\infty, \infty)$ forms a Hilbert space in which the inner product (f,g) is defined by

$$(f,g) = \int_{-\infty}^{\infty} f(t)\, \overline{g(t)}\, dt;$$

the norm squared of f, $\| f \|^2$, is defined by

$$\| f \|^2 = (f,f),$$

and is just the total energy. Two functions f and g are *orthogonal* if

$$(f,g) = 0.$$

To any closed subspace there corresponds a projection operation P, which assigns to every function its orthogonal projection onto the

subspace. Projections are characterized completely[3] by the properties

$$\begin{aligned} &P \text{ is self-adjoint, and} \\ &P^2 = P. \end{aligned} \tag{0.4}$$

We single out for consideration two projection operators on the space of square-integrable functions: time-limiting and band-limiting. *Time-limiting* a function f produces a function Df which is f restricted to $|t| \leqq T/2$:

$$Df \equiv \begin{cases} f & \text{if } |t| \leqq T/2 \\ 0 & \text{if } |t| > T/2 \end{cases}.$$

We shall write $D_T f$ if the specific interval is important to the discussion. *Band-limiting* a function f produces a function Bf whose Fourier transform agrees with the Fourier transform of f for $|\omega| \leqq 2\pi W$, and vanishes for $|\omega| > 2\pi W$. If

$$F(\omega) = \int_{-\infty}^{\infty} f(s)\, e^{-i\omega s}\, ds,$$

$$Bf = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega)\, e^{i\omega t}\, d\omega,$$

or, in terms of f directly,

$$Bf = \frac{1}{\pi} \int_{-\infty}^{\infty} f(s)\, \frac{\sin 2\pi W(t-s)}{t-s}\, ds.$$

The subspace of functions f in $\mathfrak{L}^2$ which are already time-limited, i.e. for which $Df = f$, will be called $\mathfrak{D}$, and similarly band-limited functions, for which $Bf = f$, the subspace $\mathfrak{B}$. The observation made previously that a band-limited function which vanishes for $|t| > T/2$ must vanish identically may now be phrased as

$$\mathfrak{B} \cap \mathfrak{D} = \{0\}.$$

A major result in Ref. 1 was that there is actually a non-zero minimum angle between the spaces $\mathfrak{B}$ and $\mathfrak{D}$.

A doubly orthogonal system of band-limited functions ψ_n was investigated in Refs. 1 and 2, and a number of properties were derived. The following are important to our development:

Given any $T > 0$ and any $W > 0$, we can find a countably infinite set of real functions $\psi_0(t)$, $\psi_1(t)$, $\psi_2(t)$, $\cdots$, and a set of real positive

numbers

$$\lambda_0 > \lambda_1 > \lambda_2 > \cdots ,$$

with the following properties:

(*i*) The $\psi_i(t)$ are band-limited, orthonormal on the real line, and complete in the space of square-integrable band-limited functions of bandwidth W cycles.

$$(\psi_i , \psi_j) = \begin{cases} 0, & i \neq j \\ & i, j = 0,1,2, \ldots . \\ 1, & i = j \end{cases}$$

(*ii*) In the interval $-T/2 \leqq t \leqq T/2$, the functions $D\psi_i(t)$ are orthogonal and complete in the space of square-integrable functions vanishing for $| t | > T/2$.

$$(D\psi_i , D\psi_j) = \begin{cases} 0, & i \neq j \\ & i, j = 0,1,2, \ldots . \\ \lambda_i , & i = j \end{cases}$$

(*iii*) For all values of t, real or complex,

$$\lambda_i \psi_i = BD\psi_i \left(= \int_{-T/2}^{T/2} \psi_i(s) \frac{\sin 2\pi W(t - s)}{\pi(t - s)} \, ds \right) .$$

We shall write $\lambda_i(T)$ if the specific interval is important to the discussion.

We are now in a position to give an account of our results. We repeat our basic definition:

$E(\epsilon_T)$ is the set of functions $f(t) \in \mathfrak{L}^2$ such that

(1) $f \in \mathfrak{B}$
(2) $\| f \| = 1$
(3) $\| Df \|^2 = 1 - \epsilon_T^2$.

Let us turn to the approximate dimension of $E(\epsilon_T)$. As we pointed out above, the basis $\{\varphi_i\}_0^N$ which we wish to use is the one which minimizes (0.3), that is, which minimizes

$$\max_{f \in E(\epsilon_T)} \min_{\{a_i\}_0^N} \left\| f - \sum_0^N a_i \varphi_i \right\|^2 .$$

It seems reasonable that the best basis, for any given N, should be the $(N + 1)$ linearly independent most concentrated band-limited functions,

and these are known, from previous work, to be $\psi_0, \cdots, \psi_N$. Although this seems to be harder to prove than one might expect, it is in fact true, and is the subject of

Theorem 1. For any fixed N, the functions $\psi_0, \cdots, \psi_N$ achieve the minimum in

$$\min_{\{\varphi_i\}_0^N} \max_{f \in E(\epsilon_T)} \min_{\{a_i\}_0^N} \left\| f - \sum_0^N a_i \varphi_i \right\|^2 .$$

Thus results on the approximation of $E(\epsilon_T)$ by linear combinations of a finite number of ψ_i are in fact best possible results on the approximate dimension of $E(\epsilon_T)$.

*Theorem 3. Let $f(t) \in E(\epsilon_T)$. Then**

$$\left\| f - \sum_0^{[2WT]} a_n \psi_n \right\|^2 \leqq C_1 \epsilon_T^2 ,$$

where the a_n are the Fourier coefficients of f in its expansion in the ψ's, and C_1 is independent of f, ϵ_T, and $2WT$, and may be taken as 12.

Theorem 3 shows that $[2WT] + 1$ of the best basis functions for $E(\epsilon_T)$ suffice to approximate a concentrated function to a degree proportional to the "unconcentrated part" ϵ_T^2 of the energy. We shall see that this is no longer the case when we use the simpler sampling functions.

In Theorem 3, as we have said, C_1 may be taken as 12. What does it take to make C_1 very close to 1, that is, to make the approximation almost as good as the concentration? First of all, it is important to see that roughly $2WT$ functions are *not* enough to do this, and this is the subject of

Theorem 5. For any $\epsilon_T^2 < 0.915$, there exists a function $f \in E(\epsilon_T)$ such that

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]-2} a_i \psi_i \right\|^2 \geqq C_2(\epsilon_T^2 - R(WT)),$$

where $C_2 > 1$ and $R(WT) \to 0$ as $WT \to \infty$. Here C_2 may be taken as $1/0.915$ and $R(WT)$ as $2\sqrt{2e}^{-\pi WT/2}$. (If $\epsilon_T^2 > 0.915$, the right side should be replaced by 1.)

By further analysis, this result may be strengthened so that it includes approximations by $[2WT] + N$ of the ψ_i functions, where N is any finite integer.

Theorem 8. For any given N and $\epsilon_T^2 < 0.916$, and for WT sufficiently

* $[x]$ means the largest integer $\leqq x$.

large, there will exist a function $f \in E(\epsilon_T)$ *such that*

$$\inf_{a_i} \left\| f - \sum_{0}^{[2WT]+N} a_i \psi_i \right\|^2 \geqq \frac{1}{0.916} (\epsilon_T^2 - 2\sqrt{2}\, e^{-\pi WT/2}).$$

(*If* $\epsilon_T^2 > 0.916$, *the right side should be replaced by 1.*)

Since, by Theorem 1, the ψ_i are the best approximating functions in $|t| \leqq T/2$, Theorems 7 and 8 hold, a fortiori, for any approximate basis $\{\varphi_i\}$.

What, then, does it take to bring the constant C of Theorem 3 arbitrarily close to 1? We do not know the best possible result, but there is considerable information in the following theorem, due to C. E. Shannon:

Theorem 4 (Shannon): Given any $\eta > 0$, *there exist constants* $C_3 = C_3(\eta)$ *and* $C_4 = C_4(\eta)$ *so that for* $f \in E(\epsilon_T)$,

$$\inf_{a_i} \left\| f - \sum_{0}^{[2WT]+C_3 \log^+ 2WT+C_4} a_i \psi_i \right\|^2 \leqq (1 + \eta)\epsilon_T^2 .*$$

Thus a number of functions boundedly more than $2WT$ cannot suffice for approximating $f \in E(\epsilon_T)$ to within $(1 + \eta)\ \epsilon_T^2$, but a logarithmically growing extra number of terms does.

Let us now turn to approximating $E(\epsilon_T)$ by sampling functions. The first result is that $[2WT] + 1$ sample functions will approximate f in energy roughly to within a constant times ϵ_T, that is, within a constant times the *square root* of the unconcentrated energy. The placement of the sample points depends on $2WT$, but of course not on the specific function.

Theorem 2. Let $f(t) \in E(\epsilon_T)$. *Then, if* $WT - [WT] \leqq \frac{1}{2}$,

$$\text{(a)} \qquad \left\| f - \sum_{|k| \leqq WT} f\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \right\|^2 \leqq \pi\epsilon_T + \epsilon_T^2 ,$$

and if $WT - [WT] > \frac{1}{2}$,

$$\text{(b)} \left\| f - \sum_{|k+\frac{1}{2}| \leqq WT} f\left(\frac{k + \frac{1}{2}}{2W}\right) \frac{\sin \pi(2Wt - k - \frac{1}{2})}{\pi(2Wt - k - \frac{1}{2})} \right\|^2 \leqq \pi\epsilon_T + \epsilon_T^2 .$$

An estimate valid for all WT *may be obtained by replacing* WT *in* (*a*) *by* $WT + 1$.

We note that the coefficients $f(k/2W)$ and $f(k + \frac{1}{2}/2W)$ are well-known to be the Fourier coefficients in the sampling series expansion, and hence the best constants to use.

* $\log^+ x = \max(\log x, 0)$.

This theorem is, in one sense, quite satisfactory because $\pi\epsilon_T + \epsilon_T^2$ does go to 0 as the unconcentrated part of the energy ϵ_T^2 goes to 0. On the other hand, $\pi\epsilon_T + \epsilon_T^2$ approaches 0 more slowly than ϵ_T^2 itself. That this estimate of the degree to which sampling functions approximate $E(\epsilon_T)$ cannot be too much improved is established in

Theorem 10. Let $f(t) \in E(\epsilon_T)$. Then an estimate of the form

$$\left\| f - \sum_{|k| \leqq WT+N} f\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \right\|^2 \leqq C\epsilon_T^2$$

cannot be valid independently of ϵ_T no matter how large the constants C and N are chosen.

Thus a sampling series approximation using ($2WT$ plus a constant) terms will not approximate every concentrated function to a degree proportional to the unconcentrated energy. As we have seen, this is in direct contrast to the theorem previously quoted for approximation with the best functions ψ_i. We also have the following negative result for approximation by sampling series to within $(1 + \eta)\epsilon_T^2$:

Theorem 11. For every $\beta < 1$, there exists $\delta > 0$, and ϵ_T such that

$$\left\| f - \sum_{|k| \leqq WT+(WT)^\beta} f\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \right\|^2 > (1 + \delta)\epsilon_T^2$$

for some $f \in E(\epsilon_T)$.

Once again, this is in direct contrast to the situation with the best functions ψ_i as given in Theorem 4.

We supposed near the beginning that $f(t)$ is actually band-limited. Suppose that it is only *almost* band-limited, that is, that

$$\frac{\int_{|\omega| \leqq 2\pi W} |F(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega} = 1 - \eta_W^2 . \tag{0.5}$$

It is interesting that our approximation theorems are stable in the sense that they continue to hold approximately for approximately band-limited functions. A sample is the following

Theorem 12. If $f(t) \in \mathfrak{L}^2$ with $\| f \| = 1$, and satisfies (0.1) and (0.5), then for some constants a_n we have

$$\left\| f - \sum_{0}^{[2WT]} a_n \psi_n \right\|^2 \leqq 12(\epsilon_T + \eta_W)^2 + \eta_W^2 .$$

An analogous result (Theorem 13) holds for a sampling approximation to f.

Before we proceed to the detailed exposition, let us mention one theorem, required for the proof of Theorem 10, which is of interest in its own right.

Theorem 9: When restricted to $t > 0$, the sample functions centered at the negative sample points are dense in $\mathfrak{L}^2(0, \infty)$, but those centered at the positive sample points are not dense in $\mathfrak{L}^2(0, \infty)$, nor even in $\mathfrak{B}$ restricted to $t > 0$. Specifically, given any square-integrable $f(t)$ we may find constants N and $a_n^{(N)}$ which make

$$\int_0^\infty \left| f(t) - \sum_{n=1}^{N} a_n^{(N)} \frac{\sin \pi(2Wt + n)}{\pi(2WT + n)} \right|^2 dt$$

as small as desired, but there exists a band-limited $g(t)$ for which

$$\int_0^\infty \left| g(t) - \sum_{n=1}^{N} b_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \right|^2 dt$$

cannot be made arbitrarily small regardless of the choice of N or b_n.

II. ACKNOWLEDGMENTS

The authors wish to acknowledge the many fruitful conversations with C. E. Shannon, D. Slepian, and L. J. Wallen which have contributed to the formulation of the present results.

III. DETAILED EXPOSITION

1. Given N functions $\varphi_0, \varphi_1, \cdots, \varphi_{N-1}$ in $\mathfrak{L}^2$, let us denote by S_φ^N the subspace spanned by them. The quantity $\min_{\{a_i\}} \| f - \sum_0^{N-1} a_i \varphi_i \|^2$ of (0.2) now represents the square of the distance $\rho(f, S_\varphi^N)$, measured in $\mathfrak{L}^2$, of f from S_φ^N. The number δ_N in (0.2) may therefore be taken to equal

$$\delta_N = \sup_{f \in E(\epsilon_T)} \rho(f, S_\varphi^N),$$

which, following the terminology of Ref. 4, we will call *the deflection of $E(\epsilon_T)$ from S_φ^N*.

We will first identify, for given T and N, that subspace of dimension N which best approximates $E(\epsilon_T)$, in the sense of minimizing this deflection.

Theorem 1: Let T be given. Then, for every N, the subspace spanned by the (orthonormal) functions $\psi_0, \cdots, \psi_{N-1}$ best approximates $E(\epsilon_T)$, in the sense that the deflection of $E(\epsilon_T)$ from that subspace is smaller than from any other subspace of dimension N.

Proof: We first compute the deflection of $E(\epsilon_T)$ from S_ψ^N. By definition, $f(t)$ is in $E(\epsilon_T)$ if and only if $f \in \mathfrak{B}$, with $\| f \|^2 = 1$ and $\| Df \|^2 = 1 - \epsilon_T^2$; thus, expanding f in the complete orthonormal system $\{\psi_i\}_0^\infty$, if and only if

$$f = \sum_0^\infty \alpha_i \psi_i, \quad \text{with} \quad \sum_0^\infty |\alpha_i|^2 = 1 \quad \text{and} \quad \sum_0^\infty \lambda_i |\alpha_i|^2 = 1 - \epsilon_T^2 .$$

Now by the orthonormality of the ψ_i,

$$\rho^2(f, S_\psi^N) = \min_{\{a_i\}} \left\| f - \sum_0^{N-1} a_i \psi_i \right\|^2 = \left\| \sum_N^\infty \alpha_i \psi_i \right\|^2 = \sum_N^\infty |\alpha_i|^2 .$$

To find the deflection of $E(\epsilon_T)$ from S_ψ^N we therefore compute

$$\sup_{f \in E(\epsilon_T)} \rho(f, S_\psi^N),$$

equivalently $[\sup \sum_N^\infty |\alpha_i|^2]^{\frac{1}{2}}$ subject to the conditions $\sum_0^\infty |\alpha_i|^2 = 1$ and $\sum_0^\infty \lambda_i |\alpha_i|^2 = 1 - \epsilon_T^2 \leqq \lambda_0$. We find

$$\text{deflection of } E(\epsilon_T) \text{ from } S_\psi^N = \begin{cases} 1, & 0 < 1 - \epsilon_T^2 \leqq \lambda_N \\ \left[\dfrac{\lambda_0 - (1 - \epsilon_T^2)}{\lambda_0 - \lambda_N} \right]^{\frac{1}{2}}, & \lambda_N < 1 - \epsilon_T^2 \leqq \lambda_0 \end{cases} . \tag{1.1}$$

Next suppose that $\varphi_0, \cdots, \varphi_{N-1}$ are any N given functions in $\mathfrak{L}^2$. By the Pythagorean theorem, the distance of $f \in \mathfrak{B}$ to any linear combination of the φ_i is no smaller than its distance to the same linear combination of the functions $B\varphi_i$, hence we may assume $\varphi_i \in \mathfrak{B}$. As before, let S_φ^N be the subspace spanned by $\varphi_0, \cdots, \varphi_{N-1}$, and denote by P_φ the operation of projecting orthogonally onto S_φ^N; explicitly, $P_\varphi f$ is the element of S_φ^N closest to f. In terms of P_φ, the quantity of interest in (0.2) can therefore be written simply as

$$\rho^2(f, S_\varphi^N) = \| f - P_\varphi f \|^2 = \| f \|^2 - \| P_\varphi f \|^2; \tag{1.2}$$

the last equality in (1.2) follows from the orthogonality of $P_\varphi f$ and $(f - P_\varphi f)$.

Now assign to every $f \in \mathfrak{B}$ the point in the $x - y$ plane whose x and y coordinates are $\| Df \|^2 / \| f \|^2$ and $[\| f \|^2 - \| P_\varphi f \|^2] / \| f \|^2$ respectively; denote by R_T the set of points so obtained. The significance of this map is that it sends every f in $E(\epsilon_T)$ into the line $x = 1 - \epsilon_T^2$, with y-coordinate equaling $\rho^2(f, S_\varphi^N)$; hence we see that

$$\text{deflection of } E(\epsilon_T) \text{ from } S_\psi^N = \left[\sup_{x = 1 - \epsilon_T^2} y \right]^{\frac{1}{2}} . \tag{1.3}$$

By previous results,[1] the x-coordinates of points in R_T satisfy $0 < x \leqq \lambda_0$; $x = \lambda_0$ is achieved only by the functions $k\psi_0(t)$, with k any constant. The y-coordinates of points in R_T satisfy $0 \leqq y \leqq 1$; $y = 1$ is achieved only by functions orthogonal to S_φ^N, equivalently to $\{\varphi_i\}_0^{N-1}$. Therefore, applying the Weyl-Courant lemma (Ref. 3, p. 238), we find

$$\sup_{y=1} x = \sup_{f \perp \{\varphi_i\}_0^{N-1}} \frac{\| Df \|^2}{\| f \|^2} \geqq \lambda_N .$$

Since there exist infinitely-dimensional subspaces of $\mathfrak{B}$ over which $\| Df \|^2 / \| f \|^2$ is arbitrarily small (for example those spanned by ψ_m, ψ_{m+1}, $\cdots$ for m sufficiently large), while S_φ^N is finite-dimensional, there are functions in those larger subspaces orthogonal to S_φ^N, and consequently $\inf_{y=1} x = 0$.

We show next that R_T is convex, equivalently that if P_1 and P_2 are two points in R_T, the line segment joining them is also contained in R_T. Let l be a line whose equation is $ax + by = c$. By definition of R_T, a function $f \in \mathfrak{B}$ will be sent on a point of l if and only if

$$\frac{a \| Df \|^2 + b[\| f \|^2 - \| P_\varphi f \|^2]}{\| f \|^2} = c,$$

equivalently, if and only if $a(Df,Df) - b(P_\varphi f, P_\varphi f) = (c - b)(f,f)$, or, using (0.4) and the fact that $f = Bf$, if and only if

$$([aBDB - bP_\varphi] f,f) = (c - b)\,(f,f). \tag{1.4}$$

An operator is completely continuous[3] if it transforms every bounded sequence (i.e. a sequence of functions $\{f_n\}$ for which $\| f_n \| \leqq k$ with some k) into a sequence which possesses a subsequence converging in $\mathfrak{L}^2$ norm. Since B is a projection, $\| Bf_n \| \leqq \| f_n \| \leqq k$. Writing $Bf_n(t)$ in terms of its Fourier transform $F_n(\omega)$ we obtain

$$Bf_n(t) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F_n(\omega)\, e^{i\omega t}\, d\omega,$$

whence $Bf_n(t)$ is an entire function of the complex variable t. Since a function and its Fourier transform have the same $\mathfrak{L}^2$ norm, Schwarz's inequality applied to this representation yields

$$| Bf_n(t) | \leqq c_1 e^{2\pi W |\operatorname{Im}\{t\}|} \| F_n \| \leqq c_1 k e^{2\pi W |\operatorname{Im}\{t\}|}, \tag{1.5}$$

so that the functions $Bf_n(t)$ are uniformly bounded on any compact set of the t-plane. Consequently (Ref. 5, p. 171), they form a normal family, and the sequence $Bf_n(t)$ possesses a subsequence $Bf_{n_k}(t)$ con-

verging uniformly on any compact set of the t-plane, in particular on the interval $|t| \leqq T/2$ of the real t-axis. Therefore, the functions $DBf_{n_k}(t)$ converge in $\mathfrak{L}^2$ norm as well, whence, since B is bounded, so do the functions $BDBf_{n_k}$. We have established the complete continuity of BDB. Since S_φ^N is finite-dimensional, the projection P_φ is completely continuous. By (0.4), both operators are self-adjoint. Consequently, the operator $A = aBDB - bP_\varphi$, which takes $\mathfrak{B}$ into itself, is also self-adjoint and completely continuous. Therefore[3] it has a set of orthonormal eigenfunctions $\theta_k(t) \in \mathfrak{B}$ with corresponding eigenvalues μ_k, and every function $f \in \mathfrak{B}$ has an expansion of the form

$$f = h_f + \sum_{k=0}^{\infty} \alpha_k \theta_k , \tag{1.6}$$

where $Ah_f = 0$ or, equivalently, h_f is orthogonal to all the θ_k. Using this representation, condition (1.4) becomes

$$\sum |\alpha_k|^2 \mu_k = (c - b)[\| h_f \|^2 + \sum |\alpha_k|^2] \quad \text{or}$$

$$\sum_1^\infty |\alpha_k|^2 (c - b - \mu_k) + (c - b) \| h_f \|^2 = 0. \tag{1.7}$$

We now argue that this set of functions is connected. For suppose that $f = h_f + \sum \alpha_k \theta_k$ and $g = h_g + \sum \beta_k \theta_k$ are each of the form (1.6) and satisfy (1.7). For every $0 \leqq u \leqq 1$ define, for $k = 0,1, \cdots$

$$\gamma_k^{(u)} = +\sqrt{u |\alpha_k|^2 + (1 - u) |\beta_k|^2}\, e^{i[u \arg \alpha_k + (1-u) \arg \beta_k]},$$

$$h_u = \frac{uh_f + (1 - u)h_g}{\| uh_f + (1 - u)h_g \|} \sqrt{u \| h_f \|^2 + (1 - u) \| h_g \|^2},$$

and set

$$r_u = h_u + \sum_0^\infty \gamma_k^{(u)} \theta_k .$$

We see that $Ah_u = 0$, since h_u is a linear combination of h_f and h_g, so that r_u is of the form (1.6); it is easily seen to satisfy (1.7). But as u varies between 1 and 0, the functions r_u trace a connected path in $\mathfrak{B}$ between f and g. Consequently, those functions in $\mathfrak{B}$ which map into the line l form a connected set in $\mathfrak{B}$. Since the map from $\mathfrak{B}$ onto R_T is continuous, it takes this connected set into a connected set, that is into a single segment of l. Thus, the intersection of R_T with any line l is a single segment, whence R_T is convex.

Combined with the information already derived about the points in

R_T, the convexity of R_T implies that

$$\sup_{x=1-\epsilon_T^2} y = 1, \qquad 0 < 1 - \epsilon_T^2 \leqq \lambda_N$$

$$\sup_{x=1-\epsilon_T^2} y \geqq \frac{\lambda_0 - 1 + \epsilon_T^2}{\lambda_0 - \lambda_N}, \qquad \lambda_N \leqq 1 - \epsilon_T^2 \leqq \lambda_0. \tag{1.8}$$

Combined with (1.3) and (1.1), (1.8) implies that deflection of $E(\epsilon_T)$ from $S_\varphi^N \geqq$ deflection of $E(\epsilon_T)$ from S_ψ^N. Theorem 1 is established.

We conclude from Theorem 1 that the quantity δ_N of (0.2), measuring the degree to which $E(\epsilon_T)$ is N-dimensional, may be taken to be equal to (1.1). Since, for $\lambda_N < 1 - \epsilon_T^2$,

$$\frac{\lambda_0 - (1 - \epsilon_T^2)}{\lambda_0 - \lambda_N} < \frac{\epsilon_T^2}{1 - \lambda_N},$$

and, for $\lambda_N \geqq 1 - \epsilon_T^2$,

$$1 \leqq \frac{\epsilon_T^2}{1 - \lambda_N},$$

we find

$$\delta_N < \frac{\epsilon_T}{\sqrt{\lambda_0 - \lambda_N}}. \tag{1.9}$$

Thus to establish an inequality of the form $\delta_k \leqq C\epsilon_T$ with C independent of T, it is sufficient to show that $\lambda_k(T)$ is bounded uniformly away from 1 independently of T. This will be done for $k = [2WT] + 1$ in Lemma 2, and for $k = [2WT] - N$, provided T is sufficiently large, in Theorem 8.1.

2. *Lemma 1. Let $f(s)$ be differentiable on $(-\infty, \infty)$. Then for any integers m and n, $m \leqq n$, and any $0 \leqq \frac{\alpha}{\beta} \leqq 1$,*

$$f(m) + \cdots + f(n) = \int_{m-\alpha}^{n+\beta} f(s)\,ds + (\tfrac{1}{2} - \beta)\,f(n + \beta) + (\tfrac{1}{2} - \alpha)\,f(m - \alpha) + \int_{m-\alpha}^{n+\beta} (s - [s] - \tfrac{1}{2})\,f'(s)\,ds.$$

Proof: The standard form of the Euler Summation Formula (Ref. 6, p. 539) gives

$$f(m) + f(m+1) + \cdots + f(n) = \int_m^n f(s)\,ds + \tfrac{1}{2}f(n) + \tfrac{1}{2}f(m) + \int_m^n (s - [s] - \tfrac{1}{2})f'(s)\,ds.$$

Our result then follows if

$$0 = \int_{m-\alpha}^{m} f(s)\,ds + (\tfrac{1}{2} - \alpha) f(m - \alpha) - \tfrac{1}{2} f(m) + \int_{m-\alpha}^{m} (s - [s] - \tfrac{1}{2}) f'(s)\,ds,$$

and if

$$0 = \int_{n}^{n+\beta} f(s)\,ds + (\tfrac{1}{2} - \beta) f(n + \beta) - \tfrac{1}{2} f(n) + \int_{n}^{n+\beta} (s - [s] - \tfrac{1}{2}) f'(s)\,ds.$$

Both follow immediately by partial integration on the last integrals, where $[s] = m - 1$ and n respectively. Lemma 1 is established.

We are now in a position to prove

Theorem 2. Let $g(t) \in E(\epsilon_T)$. Then if $WT - [WT] \leqq \frac{1}{2}$,

$$\text{(a)} \qquad \left\| g - \sum_{|k| \leqq WT} g\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \right\|^2 \leqq \pi\epsilon_T + \epsilon_T^2,$$

and if $WT - [WT] > \frac{1}{2}$,

$$\text{(b)} \qquad \left\| g - \sum_{|k+\frac{1}{2}| \leqq WT} g\left(\frac{k + \frac{1}{2}}{2W}\right) \frac{\sin \pi(2Wt - k - \frac{1}{2})}{\pi(2Wt - k - \frac{1}{2})} \right\|^2 \leqq \pi\epsilon_T + \epsilon_T^2.$$

An estimate valid for all WT may be obtained by replacing WT in (a) by $WT + 1$.

Proof: Without loss of generality, we assume $W = \frac{1}{2}$ for convenience. We apply Lemma 1 with $\beta = 0$ and f replaced by $|g|^2$. Then if $\alpha \leqq 1$,

$$|g^2(m)| + \cdots + |g^2(n)| = \int_{m-\alpha}^{n} |g^2(s)|\,ds + \tfrac{1}{2}|g^2(n)| + (\tfrac{1}{2} - \alpha)|g^2(m - \alpha)| + \int_{m-\alpha}^{n} (s - [s] - \tfrac{1}{2})\, 2\mathrm{Re}(g\bar{g}')\,ds.$$

It follows that if $\frac{1}{2} \leqq \alpha < 1$,

$$|g^2(m)| + \cdots + |g^2(n-1)| \leqq \int_{m-\alpha}^{n} |g^2(s)|\,ds + \int_{m-\alpha}^{n} (s - [s] - \tfrac{1}{2})\, 2\mathrm{Re}(g\bar{g}')\,ds. \tag{2.1}$$

If $0 \leqq \alpha < \frac{1}{2}$, we set $f(s) = |g^2(s+\frac{1}{2})|$ and $\alpha' = \alpha + \frac{1}{2}$. We have

$$|g^2(m+\tfrac{1}{2})| + \cdots + |g^2(n+\tfrac{1}{2})|$$

$$= \int_{m-\alpha}^{n} |g^2(s+\tfrac{1}{2})|\,ds + \tfrac{1}{2}|g^2(n+\tfrac{1}{2})|$$

$$+ (\tfrac{1}{2} - \alpha')|g^2(m+\tfrac{1}{2}-\alpha')|$$

$$+ \int_{m-\alpha'}^{n} (s - [s] - \tfrac{1}{2})\, 2\mathrm{Re}(g(s+\tfrac{1}{2})\bar{g}'(s+\tfrac{1}{2}))\,ds,$$

or

$$|g^2(m+\tfrac{1}{2})| + \cdots + |g^2(n+\tfrac{1}{2})|$$

$$= \int_{m-\alpha}^{n+\frac{1}{2}} |g^2(u)|\,du + \tfrac{1}{2}|g^2(n+\tfrac{1}{2})| - \alpha|g^2(m-\alpha)|$$

$$+ \int_{m-\alpha}^{n+\frac{1}{2}} (u - 1 - [u - \tfrac{1}{2}])\, 2\mathrm{Re}(g(u)\bar{g}'(u))\,du,$$

and

$$|g^2(m+\tfrac{1}{2})| + \cdots + |g^2(n-\tfrac{1}{2})| \leqq \int_{m-\alpha}^{n+\frac{1}{2}} |g^2(u)|\,du$$

$$+ \int_{m-\alpha}^{n+\frac{1}{2}} (u - 1 - [u - \tfrac{1}{2}])\, 2\mathrm{Re}(g(u)\bar{g}'(u))\,du. \tag{2.2}$$

If $\frac{1}{2} \leqq \alpha < 1$, we may apply (2.1) to $|g^2(t)|$ and $|g^2(-t)|$, and add the results. We obtain

$$\sum_{m \leqq |k| < n} |g^2(k)| \leqq \int_{m-\alpha \leqq |s| \leqq n} |g^2(s)|\,ds$$

$$+ \int_{m-\alpha \leqq |s| \leqq n} (s - [s] - \tfrac{1}{2})\, 2\mathrm{Re}(g\bar{g}')\,ds.$$

Now $|2(s - [s] - \frac{1}{2})| \leqq 1$; hence

$$\left| \int (s - [s] - \tfrac{1}{2})\, 2\mathrm{Re}(g\bar{g}')\,ds \right| \leqq \int |g|\,|g'|\,ds$$

$$\leqq \sqrt{\int |g|^2\,ds}\,\sqrt{\int |g'|^2\,ds},$$

by the Schwarz inequality. But

$$\int_{m-\alpha \leqq |s| \leqq n} |\, g' \,|^2\, ds < \int_{-\infty}^{\infty} |\, g' \,|^2\, ds$$

$$\leqq \pi^2 \int_{-\infty}^{\infty} |\, g \,|^2\, ds = \pi^2.$$

The last inequality holds because

$$g(t) = \int_{-\pi}^{\pi} G(x)\, e^{ixt}\, dx$$

$$g'(t) = \int_{-\pi}^{\pi} ixG(x)\, e^{ixt}\, dx$$

$$\int_{-\infty}^{\infty} |\, g'(t) \,|^2\, dt = 2\pi \int_{-\pi}^{\pi} x^2 \,|\, G(x) \,|^2\, dx$$

$$\leqq 2\pi \cdot \pi^2 \int_{-\pi}^{\pi} |\, G(x) \,|^2\, dx$$

$$= \pi^2 \int_{-\infty}^{\infty} |\, g(t) \,|^2\, dt.$$

Hence

$$\sum_{m \leqq |k| \leqq n} |\, g^2(k) \,| \leqq \int_{m-\alpha \leqq |s| \leqq n} |\, g^2(s) \,|\, ds + \pi \left[\int_{m-\alpha \leqq |s| \leqq n} |\, g^2(s) \,|\, ds \right]^{\frac{1}{2}}.$$

Now let $n \to \infty$; the preceding equation becomes

$$\sum_{m \leqq |k|} |\, g^2(k) \,| \leqq \epsilon_{2(m-\alpha)}^2 + \pi \epsilon_{2(m-\alpha)}.$$

Furthermore,

$$g(t) = \sum_{-\infty}^{\infty} g(k)\, \frac{\sin \pi(t-k)}{\pi(t-k)},$$

and the functions $\sin \pi(t-k)/\pi(t-k)$ are orthonormal. Hence,

$$\sum_{m \leqq |k|} g^2(k) = \left\| g(t) - \sum_{|k| < m} g(k)\, \frac{\sin \pi(t-k)}{\pi(t-k)} \right\|^2.$$

If we now set $m = [T/2] + 1$ and $m - \alpha = T/2$, then $\alpha \geqq \frac{1}{2}$ if

$$T/2 - [T/2] \leqq \tfrac{1}{2},$$

and we obtain (a).

Exactly the same argument, based on (2.2) rather than (2.1), gives the result (b) for the case $T/2 - [T/2] > \frac{1}{2}$.

If in (a), we use for $[T/2]$ the integer $m + 1$, then

$$\left\| g(t) - \sum_{|k| \leqq m+1} g(k) \frac{\sin \pi(t-k)}{\pi(t-k)} \right\|^2 \leqq \epsilon^2_{2(m+1)} + \pi\epsilon_{2(m+1)} ,$$

and, since ϵ_α is monotone decreasing in α, the last statement of Theorem 2 follows.

Corollary 2.1. Let $g(t) \in E(\epsilon_T)$, and let $W = \frac{1}{2}$ for simplicity. If, in addition, $g(k) = 0$, $|k| \leqq T/2$, when $T/2 - [T/2] \leqq \frac{1}{2}$, or if $g(k + \frac{1}{2}) = 0$, $|k + \frac{1}{2}| \leqq T/2$, when $T/2 - [T/2] > \frac{1}{2}$, then

$$\| Dg \|^2 \leqq \pi\epsilon_T .$$

Proof: This follows immediately from substitution into Theorem 2(a) and (b) of the additional conditions on $g(t)$.

Notice that the number of points at which g is required to vanish is $[T] + 1$, except if $T/2 - [T/2] = \frac{1}{2}$, when it is one less.

Lemma 2. With the normalization of $W = \frac{1}{2}$, for any $T > 0$

$$\lambda_{[T]+1}(T) \leqq 0.915.$$

Proof: Let us consider a function of the form

$$f = \sum_{n=0}^{[T]+1} a_n \psi_n(t). \tag{2.3}$$

The series contains $[T] + 2$ coefficients to be determined; it is therefore possible to make f vanish at the (at most) $[T] + 1$ integer or half-integer points α_k of Corollary 2.1 without having f vanish identically. More precisely, we wish

$$\sum_{m=0}^{[T]+1} a_n \psi_n(\alpha_k) = 0, \qquad k = 0,1, \cdots, [T].$$

The rank of the matrix $\{\psi_n(\alpha_k)\}$, $n = 0,1, \cdots [T] + 1$, $k = 0, \cdots, [T]$, is at most $[T]$, and hence there exists a solution vector $\{a_n\}$ not all of whose elements vanish. We may then pick the a_n so that $\sum |a_n|^2 = 1$. We have thus found a function of the form (2.3) and of total energy one, which vanishes at the $[T] + 1$ points of the Corollary 2.1.

We know for this function that

$$\int_{|t| \leqq T/2} |f|^2 dt = \sum_0^{[T]+1} |a_n|^2 \lambda_n ,$$

$$\int_{|t| > T/2} |f|^2 dt = 1 - \sum_0^{[T]+1} |a_n|^2 \lambda_n$$

$$= \sum_0^{[T]+1} (1 - \lambda_n) |a_n|^2.$$

Since the λ_n are decreasing in n, we have, remembering $\sum |a_n|^2 = 1$,

$$\lambda_{[T]+1} \leqq \sum_0^{[T]+1} |a_n|^2 \lambda_n$$

$$\leqq \pi \sqrt{\sum_0^{[T]+1} (1 - \lambda_n) |a_n|^2} \text{ by Corollary 2.1,}$$

$$\leqq \pi \sqrt{1 - \lambda_{[T]+1}}.$$

Therefore $\lambda_{[T]+1}$ is bounded from 1, and is, in fact, no larger than the root of the equation

$$x = \pi \sqrt{1 - x},$$

which is

$$\frac{-\pi^2 + \sqrt{\pi^4 + 4\pi^2}}{2} = 0.915.$$

Lemma 2 is established.

3. *Theorem 3. Let* $f(t) \in E(\epsilon_T)$. *Then*

$$\left\| f - \sum_0^{[2WT]+1} a_n \psi_n \right\|^2 \leqq 12\, \epsilon_T^2,$$

where the a_n *are the Fourier coefficients of* f *in its expansion in the functions* ψ_n.

Proof: The quantity defined in the theorem represents the square of the distance from $f \in E(\epsilon_T)$ to the subspace $S_\psi^{[2WT]+1}$ spanned by the functions ψ_n, with $0 \leqq n \leqq [2WT]$. Thus, by definition, it does not exceed $\delta^2_{[2WT]+1}$, the square of the deflection of $E(\epsilon_T)$ from $S_\psi^{[2WT]+1}$. Combining (1.9) and Lemma 2 now yields

$$\left\| f - \sum_0^{[2WT]} a_n \psi_n \right\|^2 \leqq \frac{\epsilon_T^2}{1 - \lambda_{[2WT]+1}} \leqq \frac{\epsilon_T^2}{1 - 0.916} \leqq 12\epsilon_T^2.$$

Theorem 3 is established.

4. *Theorem 4 (Shannon). Given any* $\eta > 0$, *there exist constants* $C_3 = C_3(\eta)$ *and* $C_4 = C_4(\eta)$ *so that for* $f \in E(\epsilon_T)$,

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]+C_3 \log^+ 2WT + C_4} a_i \psi_i \right\|^2 \leqq (1 + \eta)\epsilon_T^2.$$

Proof: Using properties (*ii*) and (*iii*) of the eigenfunctions ψ_i, and known results (Ref. 3, p. 242), we obtain

$$\rho_1(t - s) = \frac{\sin 2\pi W(t - s)}{\pi(t - s)} = \sum_0^\infty \psi_i(s)\psi_i(t).$$

Therefore

$$\rho_1(0) = \sum_0^\infty \psi_i^2(t), \quad \text{and}$$

$$\int_{-T/2}^{T/2} \rho_1(0)\, dt = 2WT = \sum_0^\infty \int_{-T/2}^{T/2} \psi_i^2(t)\, dt = \sum_0^\infty \lambda_i. \tag{4.1}$$

We now proceed to estimate $\sum_0^\infty \lambda_i^2$. The functions ψ_i satisfy the integral equation

$$\lambda_i \psi_i(t) = \int_{-T/2}^{T/2} \psi_i(s)\rho_1(t - s)\, ds.$$

Then

$$\lambda_i \int_{-T/2}^{T/2} \psi_i^2(t)\, dt = \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \rho_1(t - s)\psi_i(s)\psi_i(t)\, ds\, dt,$$

and if we sum on i, we obtain

$$\sum_{i=0}^\infty \lambda_i^2 = \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \rho_1^2(t - s)\, ds\, dt.$$

We now set
$s' = 2s/T$, $t' = 2t/T$, $c = \pi WT$, and $\rho(u) = \sin cu/(\pi u)$. Then

$$\begin{aligned}\sum_{i=0}^\infty \lambda_i^2 &= \int_{-1}^1 \int_{-1}^1 \rho^2(t' - s')\, ds'\, dt' \\ &= \int_{-1}^1 ds' \int_{-1-s'}^{1-s'} \rho^2(u)\, du.\end{aligned}$$

Integration by parts, and the substitution $cu = x$, give

$$\sum_{i=0}^\infty \lambda_i^2 = \frac{4c}{\pi^2} \int_0^{2c} \frac{\sin^2 x}{x^2}\, dx - \frac{2}{\pi^2} \int_0^{2c} \frac{\sin^2 x}{x}\, dx.$$

Asymptotically for large c, this is easily seen to equal

$$\frac{2c}{\pi} - \frac{1}{\pi^2} \log c + 0(1),$$

but we desire an actual lower bound. For $c \geqq \pi/8$,

$$\sum_{i=0}^\infty \lambda_i^2 = \frac{2c}{\pi} - \frac{4c}{\pi^2} \int_{2c}^\infty \frac{\sin^2 x}{x^2}\, dx - \frac{2}{\pi^2} \int_0^{3\pi/4} \frac{\sin^2 x}{x}\, dx - \frac{2}{\pi^2} \int_{3\pi/4}^{2c} \frac{\sin^2 x}{x}\, dx.$$

Therefore,

$$\sum_{i=0}^{\infty} \lambda_i^2 \geqq \frac{2c}{\pi} - \frac{4c}{\pi^2}\int_{2c}^{\infty} \frac{dx}{x^2} - \frac{2}{\pi^2}\int_0^{3\pi/4} x\,dx - \frac{2}{\pi^2}\int_{3\pi/4}^{2c} (\tfrac{1}{2} - \tfrac{1}{2}\cos 2x)\frac{dx}{x}$$

$$\geqq \frac{2c}{\pi} - \frac{2}{\pi^2} - \frac{9}{16} - \frac{1}{\pi^2}\log\frac{2c}{3\pi/4},$$

since

$$\int_{3\pi/4}^{2c} \frac{\cos 2x\,dx}{x} > 0 \quad \text{if} \quad c \geqq \pi/8$$

Thus

$$\sum_{0}^{\infty} \lambda_i^2 \geqq \frac{2c}{\pi} - \frac{1}{\pi^2}\log^+ c - 1 \tag{4.2}$$

for all c, since the inequality is trivially true for $c < \pi/8$.

Let us now introduce the following combinatorial problem. We consider infinite sequences of non-negative numbers μ_j such that

(a) $1 \geqq \mu_0 \geqq \mu_1 \geqq \cdots$,

(b) $\sum_0^{\infty} \mu_j = A$, a given positive constant,

(c) for a given integer $m \geqq A$, μ_m has a prescribed value,

and we seek to maximize $\sum_0^{\infty} \mu_j^2$ over all such sequences. Clearly the optimum $\{\mu_j\}$ will have $\mu_j = 0$ if $j > m$.

We claim that, with the possible exception of one μ_j, all the others in the optimum solution equal either 1 or μ_m. For suppose they do not, i.e. suppose $\{\mu_n\}$ takes on two values α and β such that $\mu_m < \alpha < \beta < 1$. If we now vary α and β between the limits μ_m and 1, keeping $\alpha + \beta$ a constant, and maximize $\alpha^2 + \beta^2$, we find an end-point maximum. In detail, if $\alpha + \beta = s$, then $\alpha^2 + \beta^2 = 2\,[(\alpha - s/2)^2 + s^2/4]$, which is maximized at an end-point value of α. Thus, the maximizing sequence $\{\mu_n\}_0^m$ can contain only one value which is neither 1 nor μ_m. This odd value is due to "breakage" in obtaining the exact total A. Let the maximizing sequence have k "1's", $(m - k)$ "μ_m's" and one value α, $\mu_m < \alpha < 1$. Then

$$k + (m - k)\,\mu_m + \alpha = A,$$

so that

$$k = \frac{A - \alpha - m\mu_m}{1 - \mu_m}.$$

Then

$$\sum_0^m \mu_j^2 = \frac{A - \alpha - m\mu_m}{1 - \mu_m} + \frac{m - A + \alpha}{1 - \mu_m}\mu_m^2 + \alpha^2 \tag{4.3}$$
$$= (A - \alpha)(1 + \mu_m) - m\mu_m + \alpha^2.$$

This is the maximum achievable value of $\sum_0^\infty \mu_j^2$ under the conditions (a), (b) and (c) above. But with $A = 2c/\pi$, and λ_m given, $m \geqq 2c/\pi$, the sequence of eigenvalues λ_j satisfies the above conditions. It therefore competes for the maximum, and hence

$$\frac{2c}{\pi} - \frac{1}{\pi^2}\log^+ c - 1 \leqq \sum_0^\infty \lambda_i^2 \leqq \frac{2c}{\pi}(1 + \lambda_m) - m\lambda_m .$$

Thus, for any $m \geqq 2c/\pi$,

$$\lambda_m \leqq \frac{\dfrac{\log^+ c}{\pi^2} + 1}{m - \dfrac{2c}{\pi}}.$$

For any given $\eta > 0$, if

$$m \geqq \frac{2c}{\pi} + \frac{12}{\eta}\left(\frac{\log^+ c}{\pi^2} + 1\right) \tag{4.4}$$

it follows that $\lambda_m \leqq \eta/12$. Then, by the reasoning of Theorem 3,

$$\inf_{a_i} \left\| f - \sum_0^m a_n\psi_n \right\|^2 \leqq \frac{\epsilon_T^2}{1 - \eta/12}.$$

If $\eta \leqq 11$, this implies

$$\inf_{a_i} \left\| f - \sum_0^m a_n\psi_n \right\|^2 \leqq (1 + \eta)\epsilon_T^2;$$

larger values of η are covered by Theorem 3. Theorem 4 is proved.

Note: If only small values of η are of interest, the "12" in (4.4) is of course unnecessarily large.

Lemma 3: With the normalization of $W = \frac{1}{2}$, we have for any $T > 1$,

$$\lambda_{[T]-1}(T) \geqq 0.085.$$

Proof: We begin, again, with Lemma 1. If we consider first the case $T/2 - [T/2] \geqq \frac{1}{2}$, we let $f(s) = |g^2(s)|$, and

(a) $m = 1$, $\alpha = \dfrac{1}{2}$, $n = \left[\dfrac{T}{2}\right]$, $\beta = \dfrac{T}{2} - \left[\dfrac{T}{2}\right]$;

(b) $m = -\left[\dfrac{T}{2}\right] - 1$, $\alpha = -\dfrac{T}{2} - \left[\dfrac{T}{2}\right]$, $n = 0$, $\beta = \dfrac{1}{2}$.

We obtain

$$| g^2(1) | + \cdots | g^2([T/2]) | \leqq \int_{\frac{1}{2}}^{T/2} | g^2(s) | \, ds + \int_{\frac{1}{2}}^{T/2} (s - [s] - \tfrac{1}{2}) \, 2\mathrm{Re}(g\bar{g}') \, ds$$

and

$$| g^2(-[T/2]) | + \cdots + | g^2(0) | \leqq \int_{-T/2}^{\frac{1}{2}} | g^2(s) | \, ds + \int_{-T/2}^{\frac{1}{2}} (s - [s] - \tfrac{1}{2}) \, 2\mathrm{Re}(g\bar{g}') \, ds.$$

Adding and applying the Schwarz inequality, we find, as in Theorem 2,

$$\sum_{|n| \leqq [T/2]} | g^2(n) | \leqq || Dg ||^2 + \pi || Dg || \, || g ||. \tag{4.5}$$

The Weyl-Courant lemma (Ref. 3, p. 238) asserts that

$$\lambda_n = \inf_{A_n} \sup_{\varphi \perp A_n} \frac{|| D\varphi ||^2}{|| \varphi ||^2},$$

where A_n ranges over all n-dimensional subspaces of $\mathfrak{B}$. If B_{n+1} is an $(n + 1)$-dimensional subspace of $\mathfrak{B}$, the orthogonal complement of every A_n must have at least one vector in common with B_{n+1}. Thus

$$\sup_{\varphi \perp A_n} \frac{|| D\varphi ||^2}{|| \varphi ||^2} \geqq \inf_{\varphi \in B_{n+1}} \frac{|| D\varphi ||^2}{|| \varphi ||^2},$$

and since the right-hand side of the inequality is independent of A_n, the Weyl-Courant lemma implies

$$\lambda_n \geqq \inf_{\varphi \in B_{n+1}} \frac{|| D\varphi ||^2}{|| \varphi ||^2}. \tag{4.6}$$

Now let $B_{[T]}$ be the subspace of $\mathfrak{B}$ spanned by the $[T]$ (orthonormal) functions $[\sin \pi(t - k)]/\pi(t - k)$, $| k | \leqq [T/2]$. For $g \in B_{[T]}$ we have

$$|| g ||^2 = \sum_{|n| \leqq [T/2]} | g(n) |^2,$$

since $g(n) = 0$ when $| n | > [T/2]$, so that (4.5) yields

$$1 \leqq \frac{\| Dg \|^2}{\| g \|^2} + \pi \frac{\| Dg \|}{\| g \|}.$$

Letting g vary in $B_{[T]}$, and using (4.6), we now find

$$1 \leqq \inf_{g \in B_{[T]}} \frac{\| Dg \|^2}{\| g \|^2} + \pi \frac{\| Dg \|}{\| g \|} = \inf_{g \in B_{[T]}} \frac{\| Dg \|^2}{\| g \|^2} + \pi \inf_{g \in B_{[T]}} \frac{\| Dg \|}{\| g \|}$$
$$\leqq \lambda_{[T]-1} + \pi \sqrt{\lambda_{[T]-1}},$$

whence

$$\lambda_{[T]-1} \geqq 0.085.$$

Similarly, if $T/2 - [T/2] < \frac{1}{2}$,

$$\sum_{|n+\frac{1}{2}| \leqq [T/2]} | g(n + \tfrac{1}{2}) |^2 \leqq \| Dg \|^2 + \pi \| Dg \| \| g \|,$$

and letting $B_{[T]}$ be the subspace of $\mathcal{B}$ spanned by the $[T]$ functions $[\sin \pi(t - k - \frac{1}{2})]/[\pi(t - k - \frac{1}{2})]$ with $| k + \frac{1}{2} | \leqq [T/2]$ we may apply the identical argument to find $\lambda_{[T]-1} \geqq 0.085$, as before. Lemma 3 is established.

Lemma 4: For any $WT > 0$,

$$\lambda_0 > 1 - 2\sqrt{2} e^{-\pi WT/2}.$$

Proof: For convenience, let $\Omega = 2\pi W$, and normalize so that $T = 2$. Consider the function $f(t)$ whose Fourier Transform $F(x)$ is given by

$$F(x) = \begin{cases} \dfrac{1}{(\Omega\pi)^{\frac{1}{4}}} e^{-x^2/2\Omega} & \text{if } | x | \leqq \Omega \\ 0 & \text{if } | x | > \Omega. \end{cases}$$

Then

$$\int_{-\infty}^{\infty} | f(t) |^2 \, dt = 2\pi \int_{-\Omega}^{\Omega} F^2(x) \, dx = 4\sqrt{\pi} \int_0^{\sqrt{\Omega}} e^{-u^2} \, du.$$

On the other hand,

$$f(t) = \int_{-\Omega}^{\Omega} F(x) \cos xt \, dx$$

$$= \frac{2}{(\pi\Omega)^{\frac{1}{4}}} \left[\sqrt{\frac{\pi\Omega}{2}} e^{-t^2\Omega/2} - \int_{\Omega}^{\infty} e^{-x^2/2\Omega} \cos xt \, dx \right] \tag{4.7}$$

$$\geqq \frac{2}{(\pi\Omega)^{\frac{1}{4}}} \left[\sqrt{\frac{\pi\Omega}{2}} e^{-t^2\Omega/2} - \int_{\Omega}^{\infty} e^{-x^2/2\Omega} \, dx \right].$$

It is easy to check that the expression in brackets is non-negative for $t = 1$, and hence for $|t| \leqq 1$. In fact,

$$\int_{\Omega}^{\infty} e^{-x^2/2\Omega}\, dx = \sqrt{\Omega} \int_{\sqrt{\Omega}}^{\infty} e^{-u^2/2}\, du,$$

and

$$G(\Omega) = \sqrt{\frac{\pi}{2}}\, e^{-\Omega/2} - \int_{\sqrt{\Omega}}^{\infty} e^{-u^2/2}\, du$$

is non-negative since it equals 0 at both $\Omega = 0$ and $\Omega = \infty$, and

$$G'(\Omega) = e^{-\Omega/2}\left(\frac{1}{2\sqrt{\Omega}} - \frac{\sqrt{\pi}}{2\sqrt{2}}\right)$$

is positive for $\Omega < 2/\pi$ and negative for $\Omega > 2/\pi$. Thus

$$\int_{-1}^{1} f^2(t)\, dt > 4\sqrt{\pi} \int_{0}^{\sqrt{\Omega}} e^{-u^2}\, du - 8\sqrt{2} \int_{\sqrt{\Omega}}^{\infty} e^{-u^2/2}\, du \int_{0}^{\sqrt{\Omega}} e^{-u^2/2}\, du.$$

Hence, from (4.7),

$$\frac{\int_{-1}^{1} f^2(t)\, dt}{\int_{-\infty}^{\infty} f^2(t)\, dt} > 1 - 2\sqrt{\frac{2}{\pi}} \int_{\sqrt{\Omega}}^{\infty} e^{-u^2/2}\, du \left[\frac{\int_{0}^{\sqrt{\Omega}} e^{-u^2/2}\, du}{\int_{0}^{\sqrt{\Omega}} e^{-u^2}\, du}\right].$$

But

$$\int_{\sqrt{\Omega}}^{\infty} e^{-u^2/2}\, du < \sqrt{\frac{\pi}{2}}\, e^{-\Omega/2} \text{ because } G(\Omega) \geqq 0,$$

and the expression in parentheses is bounded by $\sqrt{2}$. Thus

$$\frac{\int_{-1}^{1} f^2(t)\, dt}{\int_{-\infty}^{\infty} f^2(t)\, dt} > 1 - 2\sqrt{2}\, e^{-\Omega/2}.$$

But $f \in \mathfrak{B}$ by definition, and hence competes in the maximum problem[2] which defines λ_0. Hence

$$\lambda_0 = \max_{\mathfrak{B}} \frac{\int_{-1}^{1} |f(t)|^2\, dt}{\int_{-\infty}^{\infty} |f(t)|^2\, dt} > 1 - 2\sqrt{2}\, e^{-\Omega/2}.$$

Lemma 4 is established.

5. *Theorem 5. For any* $\epsilon_T^2 < 0.915$, *there exists a function* $f \in E(\epsilon_T)$ *such that*

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]-2} a_i \psi_i \right\|^2 \geqq \frac{1}{0.915} (\epsilon_T^2 - 2\sqrt{2}\, e^{-\pi WT/2}).$$

(*If* $\epsilon_T^2 \geqq 0.915$, *the right-hand side of the inequality should be replaced by 1.*)

Proof: Theorem 5 asserts the existence of a lower bound for the deflection of $E(\epsilon_T)$ from the subspace $S_\psi^{[2WT]-1}$, spanned by the functions ψ_k, with $0 \leqq k \leqq [2WT] - 2$. This deflection has already been calculated in (1.1) and is easily seen to be assumed by a function in $E(\epsilon_T)$. Thus there exists $f \in E(\epsilon_T)$ such that

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]-2} a_i \psi_i \right\|^2 = \min \left[1, \frac{\lambda_0 - (1 - \epsilon_T^2)}{\lambda_0 - \lambda_{[2WT]-1}} \right].$$

By Lemma 3, $\lambda_{[2WT]-1} \geqq 0.085$; this ensures that when $\epsilon_T^2 \geqq 0.915$ the smaller of the terms is 1. For other values of ϵ_T^2, since

$$\lambda_0 > 1 - 2\sqrt{2}\, e^{-\pi WT/2}$$

by Lemma 4, and $\lambda_0 < 1$, we find

$$\min \left(1, \frac{\lambda_0 - 1 + \epsilon_T^2}{\lambda_0 - \lambda_{[2WT]-1}} \right) \geqq \min \left[1, \frac{1}{0.915} (\epsilon_T^2 - 2\sqrt{2}\, e^{-\pi WT/2}) \right],$$

and now the second of the bracketed terms is the smaller. Theorem 5 is established.

6. In $\mathfrak{L}^2(-\infty, \infty)$ let D' denote the operation of projecting onto $[0, \infty]$, that is

$$D'f(t) = \begin{cases} f(t) & t \geqq 0 \\ 0 & t < 0. \end{cases}$$

Arguing as with DBD in the proof of Theorem 1 we see that $D'BD'$, which takes $\mathfrak{L}^2(0, \infty)$ into itself, is self-adjoint, positive, and bounded by 1 (though no longer completely continuous). It therefore has a spectrum[3] contained in the unit interval; we will show that its spectrum consists of all $0 \leqq \lambda \leqq 1$.

Theorem 6. The $\mathfrak{L}^2$ *spectrum of the operator*

$$D'BD'f = \frac{1}{\pi} \int_0^\infty \frac{\sin (x - y)}{x - y} f(y)\, dy, \qquad x \geqq 0,$$

consists of all $0 \leqq \lambda \leqq 1$.

Proof: Theorem 6 follows immediately as a special case of much more

general results of H. Widom[7] and M. Rosenblum (unpublished), which determine the spectra in $\mathfrak{L}^2$ of Wiener-Hopf equations with kernels whose Fourier transforms are bounded. We include a separate proof only because it is constructive.

By definition, λ is in the spectrum of an operator A if and only if for every $\epsilon > 0$ there exists φ_ϵ such that

$$\frac{\| A\varphi_\epsilon - \lambda\varphi_\epsilon \|}{\| \varphi_\epsilon \|} < \epsilon. \tag{6.1}$$

We will prove the theorem by constructing functions which satisfy (6.1) for any given $0 < \lambda < 1$. The spectrum being a closed set, it must then include all $0 \leqq \lambda \leqq 1$, but by the introductory remarks it is also contained in the closed unit interval, hence it consists of precisely the points $0 \leqq \lambda \leqq 1$.

Lemma 5: Let $\mu > 0$ be given. Then corresponding to any $\delta > 0$ there exists a function $H_\delta(z)$ satisfying

a. $H_\delta(z)$ *is analytic in* $|z| < 1$, *continuous in* $|z| \leqq 1$,

b. $H_\delta(0) = 0$,

c. $$\frac{\int_0^\pi |\mu H_\delta(e^{i\theta}) + H_\delta(e^{-i\theta})|^2 \sin\theta\, d\theta}{\int_0^\pi |H_\delta(e^{i\theta})|^2 \sin\theta\, d\theta} < \delta.$$

Proof: Suppose $0 < \alpha < \pi/2$. Denote in the z-plane by $P_1, P_2, P_3, P_4, P_5, P_6$ the points $1, e^{i\alpha}, -e^{-i\alpha}, -1, 0$, and i respectively, and let γ_1, γ_2 represent respectively the arcs P_1P_2, P_3P_4 of the unit circle. Let $w = P_0(z)$ be a conformal map of the upper half of the unit disc onto the region in the w-plane defined by $1 < |w| < q < \infty$, $\mathrm{Im}\{w\} > 0$, which takes the points P_1, P_2, P_3, P_4, onto $w = 1$, $w = q$, $w = -q$, $w = -1$ respectively. The required map exists as soon as q is chosen appropriately (for example, so as to make the extremal length of the family of curves joining γ_1 to γ_2 in the upper semicircle equal to the extremal length of the family of curves joining the two segments of the real axis in the image domain), and it defines q uniquely. Now by reflection, $P_0(z)$ is extendable across the diameter of the unit circle to a map of $|z| < 1$ onto the domain $1/q < |w| < q$, $\mathrm{Im}\{w\} > 0$, and satisfies

$$P_0(e^{i\theta})/q^2 = P_0(e^{-i\theta}) \qquad \alpha < \theta < \pi - \alpha$$

$$|P_0(e^{i\theta})| = q \qquad \alpha < \theta < \pi - \alpha$$

$$1/q < |P_0(e^{i\theta})|, |P_0(e^{-i\theta})| < q \qquad e^{i\theta} \in \gamma_1, \gamma_2.$$

Choose r so that $\mu = q^{-2r}$, and let $P(z) \equiv [qP_0(z)]^r$. Since $P_0(z)$ is bounded away from zero and infinity in $|z| \leqq 1$, the function $P(z)$ is also analytic in $|z| < 1$, continuous in $|z| \leqq 1$ (though no longer necessarily univalent) and satisfies

$$\mu P(e^{i\theta}) = P(e^{-i\theta}), \qquad \alpha < \theta < \pi - \alpha$$

$$|P(e^{i\theta})| = 1/\mu, \qquad \alpha < \theta < \pi - \alpha$$

$$m_\mu = \min(1/\mu, 1) < |P(e^{i\theta})|, |P(e^{-i\theta})| < \max(1/\mu, 1) = M_\mu,$$

$$e^{i\theta} \in \gamma_1, \gamma_2.$$

Next let $w = Q(z)$ map the region defined by $|z| < 1$, $\mathrm{Im}\{z\} > 0$, $\mathrm{Re}\{z\} > 0$ onto itself, taking the points P_5, P_1, P_2 onto $w = 0$, $w = 1$, and $w = i$ respectively. $Q(z)$ may be constructed from elementary maps and is given explicitly by

$$Q^2(z) \equiv \frac{\sqrt{\cos^2\alpha + \left(\frac{1+z^2}{1-z^2}\right)^2 \sin^2\alpha} - 1}{\sqrt{\cos^2\alpha + \left(\frac{1+z^2}{1-z^2}\right)^2 \sin^2\alpha} + 1}.$$

It may be extended by reflection to yield a map of $|z| < 1$ onto the domain in the w-plane formed by cutting the unit circle along the imaginary axis from $i[\sin\alpha/(1+\cos\alpha)]$ to i and from $-i[\sin\alpha/(1+\cos\alpha)]$ to $-i$. It satisfies

$$Q(0) = 0,$$

$$Q(e^{i\theta}) = -Q(e^{-i\theta}), \qquad \alpha < \theta < \pi - \alpha$$

$$|Q(e^{i\theta})| = |Q(e^{-i\theta})| = 1, \qquad e^{i\theta} \in \gamma_1, \gamma_2.$$

Now form $H(z) \equiv P(z)Q(z)$. We see that $H(z)$ satisfies conditions (a) and (b) of Lemma 5. Furthermore, by definition of H,

$$\mu H(e^{i\theta}) + H(e^{-i\theta}) = 0, \qquad \alpha < \theta < \pi - \alpha$$

$$|H(e^{i\theta})| = \frac{1}{\mu}|Q(e^{i\theta})|, \qquad \alpha < \theta < \pi - \alpha$$

$$m_\mu < |H(e^{i\theta})|, |H(e^{-i\theta})| < M_\mu, \qquad e^{i\theta} \in \gamma_1, \gamma_2.$$

Thus

$$\int_0^{\pi} |\mu H(e^{i\theta}) + H(e^{-i\theta})|^2 \sin\theta\, d\theta$$

$$= \int_0^{\alpha} + \int_{\pi-\alpha}^{\alpha} |\mu H(e^{i\theta}) + H(e^{-i\theta})|^2 \sin\theta\, d\theta \tag{6.2}$$

$$\leqq 2(\mu+1)^2 M_\mu^2 \int_0^{\alpha} \sin\theta\, d\theta = 2(\mu+1)^2 M_\mu^2 (1-\cos\alpha).$$

$$\int_0^{\pi} |H(e^{i\theta})|^2 \sin\theta\, d\theta > \int_{\alpha}^{\pi-\alpha} |H(e^{i\theta})|^2 \sin\theta\, d\theta$$

$$= \frac{1}{\mu^2}\int_{\alpha}^{\pi-\alpha} |Q(e^{i\theta})|^2 \sin\theta\, d\theta = \frac{2}{\mu^2}\int_{\alpha}^{\pi/2} |Q(e^{i\theta})|^2 \sin\theta\, d\theta. \tag{6.3}$$

Using the expression for $Q(z)$ we find, for $\alpha < \theta < \pi - \alpha$,

$$|Q(e^{i\theta})|^2 = \left| \frac{\sqrt{\cos^2\alpha - \dfrac{\cos^2\theta \sin^2\alpha}{\sin^2\theta}} - 1}{\sqrt{\cos^2\alpha - \dfrac{\cos^2\theta \sin^2\alpha}{\sin^2\theta}} + 1} \right| = \left| \frac{\sqrt{1 - \dfrac{\sin^2\alpha}{\sin^2\theta}} - 1}{\sqrt{1 - \dfrac{\sin^2\alpha}{\sin^2\theta}} + 1} \right|$$

$$= \frac{\sin^2\theta}{\sin^2\alpha}\left| 1 - \sqrt{1 - \frac{\sin^2\alpha}{\sin^2\theta}} \right|^2 \geqq \frac{1}{4}\frac{\sin^2\alpha}{\sin^2\theta}.$$

Introducing this into (6.3) yields

$$\int_0^{\pi} |H(e^{i\theta})|^2 \sin\theta\, d\theta > \frac{1}{2\mu^2}\sin^2\alpha \log\left(\frac{1+\cos\alpha}{\sin\alpha}\right),$$

whence, by (6.2),

$$\frac{\displaystyle\int_0^{\pi} |\mu H(e^{i\theta}) + H(e^{-i\theta})|^2 \sin\theta\, d\theta}{\displaystyle\int_0^{\pi} |H(e^{i\theta})|^2 \sin\theta\, d\theta} < \frac{K_\mu}{\log \csc\alpha},$$

where K_μ depends only on μ. Thus, if α is chosen sufficiently small, $H(z)$ satisfies the remaining condition (c). Lemma 5 is established.

We now pass to the construction of the functions φ_ϵ of (6.1). Given $0 < \lambda < 1$ and $\epsilon > 0$, set $0 < \mu = (1-\lambda)/\lambda$, choose δ so small that that $\sqrt{\delta}/(1-\sqrt{\delta}) < \epsilon$, and let $H_\delta(z) = H(z)$ be the function of Lemma 5 corresponding to δ.

Introduce the map $u + iv = w = \frac{1}{2}(z + 1/z)$, taking $|z| < 1$ onto the w-plane slit along the real axis from $w = -1$ to $w = 1$, and in

that region define $F(w) \equiv H(z)$. The function $F(w)$ is then analytic except on the slit. If $F_1(w)$ denotes $F(w)$ in the upper half-plane, $F_1(w)$ is continuous in the closed half plane $v \geqq 0$. If $v > 0$,

$$\int_{-\infty}^{\infty} |F_1(u + iv)|^2 \, du = \frac{1}{2} \int_{\Gamma_v} |H(z)|^2 \left| 1 - \frac{1}{z^2} \right| |dz|, \tag{6.4}$$

where Γ_v is the curve in the upper half of the unit circle defined in polar coordinates by $(r - 1/r) \sin \theta = 2v$. Since $H(0) = 0$, the function $[H(z)/z]\sqrt{1 - z^2}$ is analytic in $|z| < 1$, continuous in $|z| \leqq 1$, and by the maximum principle

$$\left| \frac{H(z)}{z} \sqrt{1 - z^2} \right| \leqq \sup_{|z|=1} \left| \frac{H(z)}{z} \sqrt{1 - z^2} \right| \leqq \sup_{|z|=1} |H(z)|.$$

By property (a) of Lemma 5, $H(z)$ is bounded in $|z| \leqq 1$, hence so is the integrand on the right-hand side of (6.4). Since the curves Γ_v have lengths bounded independently of v, it follows that

$$\int_{-\infty}^{\infty} |F_1(u + iv)|^2 \, du < c, \qquad v > 0.$$

Consequently, by a theorem of Paley-Wiener,[8] $F_1(w)$ coincides in $v \geqq 0$ with the Fourier transform of a function $\psi_1(t) \in \mathfrak{L}^2$ which vanishes for $t \geqq 0$. Letting $F_2(w)$ denote $F(w)$ in the lower half plane, the identical argument establishes that $F_2(w)$ coincides in $v \leqq 0$ with the Fourier transform of a function $\psi_2(t) \in \mathfrak{L}^2$ which vanishes for $t \leqq 0$. Let $\varphi_1(t) \equiv (1/\lambda)\psi_1(t)$, $\varphi_2(t) = -(1/\lambda)\psi_2(t)$, and $\varphi(t) = \varphi_1(t) + \varphi_2(t)$. Then with $\chi(u)$ the characteristic function of the interval $-1 \leqq u \leqq 1$, using the norm-preserving property of the Fourier transform, we have

$$\begin{aligned} \| BD'\varphi - \lambda\varphi \|^2 &= \| (B - \lambda)\varphi_2 - \lambda\varphi_1 \|^2 \\ &= \int_{-\infty}^{\infty} \left| [\chi(u) - \lambda] \frac{F_2(u)}{\lambda} + F_1(u) \right|^2 du \\ &= \int_{|u|>1} |F_1(u) - F_2(u)|^2 \, du \\ &\quad + \int_{-1}^{1} |\mu F_2(u) + F_1(u)|^2 \, du \\ &= \int_0^{\pi} |\mu H(e^{i\theta}) + H(e^{-i\theta})|^2 \sin\theta \, d\theta, \end{aligned} \tag{6.5}$$

while

$$\| \varphi \|^2 \geqq \| \varphi_2 \|^2 = \frac{1}{\lambda^2} \| \psi_2 \|^2$$

$$= \frac{1}{\lambda^2} \int_{-\infty}^{\infty} | F_2(u) |^2 \, du > \frac{1}{\lambda^2} \int_{-1}^{1} | F_2(u) |^2 \, du \tag{6.6}$$

$$= \frac{1}{\lambda^2} \int_0^{\pi} | H(e^{i\theta}) |^2 \sin \theta \, d\theta.$$

Thus, combining (6.5) and (6.6),

$$\frac{\| BD'\varphi - \lambda\varphi \|^2}{\| \varphi \|^2} < \lambda^2 \frac{\int_0^{\pi} | \mu H(e^{i\theta}) + H(e^{-i\theta}) |^2 \sin \theta \, d\theta}{\int_0^{\pi} | H(e^{i\theta}) |^2 \sin \theta \, d\theta} < \lambda^2 \delta. \tag{6.7}$$

From (6.7),

$$\| D'\varphi \| \geqq \| BD'\varphi \| \geqq \lambda(1 - \sqrt{\delta}) \| \varphi \|, \tag{6.8}$$

and

$$\| D'BD'\varphi - \lambda D'\varphi \| = \| D'(BD'\varphi - \lambda\varphi) \| \leqq \lambda\sqrt{\delta} \| \varphi \|. \tag{6.9}$$

Setting $\varphi_\epsilon = D'\varphi$ and combining (6.8) and (6.9) we obtain

$$\| D'BD'\varphi_\epsilon - \lambda\varphi_\epsilon \| / \| \varphi_\epsilon \| \leqq \sqrt{\delta}/(1 - \sqrt{\delta}) < \epsilon,$$

which is the required inequality (6.1). Theorem 6 is established.

7. *Theorem 7. Given any subinterval $0 < \alpha \leqq x \leqq \beta < 1$ of the unit interval, there exists T_0 such that for all $T > T_0$, the operator BD_TB has an eigenvalue contained in $[\alpha,\beta]$.*

Proof: Let $\lambda = \frac{1}{2}(\alpha + \beta)$ and choose ϵ so small that $3\epsilon/(\lambda - 3\epsilon) < (\beta - \alpha)/2$. Since $0 < \lambda < 1$, by Theorem 6 there exists a function $\varphi \in \mathfrak{L}^2$ such that

$$\frac{\| D'BD'\varphi - \lambda\varphi \|}{\| \varphi \|} < \epsilon.$$

Since φ and $D'BD'\varphi$ are fixed functions in $\mathfrak{L}^2$, there exists T_0 such that for each $T > T_0$

$$\| (D' - D_T)\varphi \|^2 = \int_T^{\infty} | \varphi(t) |^2 \, dt < \epsilon^2 \| \varphi \|^2$$

and

$$\| (D' - D_T)BD'\varphi \|^2 = \int_T^\infty | BD'\varphi(t) |^2 \, dt < \epsilon^2 \| \varphi \|^2.$$

Using the inequality $\| D_TB(D' - D_T)\varphi \| \leqq \| (D' - D_T)\varphi \|$ we then find

$$\begin{aligned} &\frac{\| D_TBD_T\varphi - \lambda\varphi \|}{\| \varphi \|} \\ &= \frac{\| D'BD'\varphi - \lambda\varphi - (D' - D_T)BD'\varphi - D_TB(D' - D_T)\varphi \|}{\| \varphi \|} \\ &\leqq \frac{\| D'BD'\varphi - \lambda\varphi \|}{\| \varphi \|} + \frac{\| (D' - D_T)BD'\varphi \|}{\| \varphi \|} \\ &\qquad + \frac{\| D_TB(D' - D_T)\varphi \|}{\| \varphi \|} \\ &\leqq 3\epsilon. \end{aligned} \tag{7.1}$$

Now from (7.1) we see

$$\| D_T \varphi \| \geqq \| D_TBD_T \varphi \| \geqq (\lambda - 3\epsilon) \| \varphi \| \tag{7.2}$$

and

$$\| D_TBD_T \varphi - \lambda D_T \varphi \| = \| D_T(D_TBD_T \varphi - \lambda\varphi) \| \leqq 3\epsilon \| \varphi \| \tag{7.3}$$

so that, combining (7.2) and (7.3),

$$\| D_TBD_T \varphi - \lambda D_T \varphi \| / \| D_T \varphi \| \leqq 3\epsilon/(\lambda - 3\epsilon). \tag{7.4}$$

Now by property *ii* of the functions ψ_i, we may expand $D_T\varphi$ in a series $D_T\varphi = \sum a_n\varphi_n$, where $\varphi_n \equiv D_T\psi_n/\sqrt{\lambda_n(T)}$. Inserting this into (7.4), and using the fact (*iii*) that the $\psi_n(t)$ (which depend also on T) are eigenfunctions of BD_TB, we find

$$\begin{aligned} \left(\frac{3\epsilon}{\lambda - 3\epsilon}\right)^2 \geqq \frac{\| D_TBD_T\varphi - \lambda D_T\varphi \|^2}{\| D_T\varphi \|^2} &= \frac{\| \sum a_n(\lambda_n(T) - \lambda)\varphi_n \|^2}{\| \sum a_n\varphi_n \|^2} \\ &= \frac{\sum | a_n |^2 \, | \lambda_n(T) - \lambda |^2}{\sum | a_n |^2} \\ &\geqq \inf_n | \lambda_n(T) - \lambda |^2. \end{aligned}$$

We conclude that for every $T > T_0$ there exists an eigenvalue $\lambda_n(T)$

of the operator BD_TB with $|\lambda_n(T) - \lambda| \leqq 3\epsilon/(\lambda - 3\epsilon) < (\beta - \alpha)/2$, or equivalently, since the $\lambda_n(T)$ are all real, that $\alpha < \lambda_n(T) < \beta$. Theorem 7 is established.

Corollary 7.1 The number of eigenvalues of the operator BD_TB contained in any subinterval J of the unit interval cannot remain bounded as $T \to \infty$.

Proof: Given any integer N, subdivide J into N disjoint intervals J_n. By Theorem 7, for all T sufficiently large each J_n will contain an eigenvalue of BD_TB, hence J will contain at least N such eigenvalues. Since N was arbitrary, Corollary 7.1 is established.

8. *Theorem 8. Let any integer N and $\epsilon_T^2 < 0.916$ be given. Then as soon as WT is sufficiently large, there will exist a function $f \in E(\epsilon_T)$ such that*

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]+N} a_i\psi_i \right\|^2 \geqq \frac{1}{0.916}\,(\epsilon_T^2 - 2\sqrt{2}\, e^{-\pi WT/2}).$$

(If $\epsilon_T^2 \geqq 0.916$, the right-hand side of the inequality should be replaced by 1.)

Proof: By Lemma 3, we have

$$\lambda_{[2WT]-1}\,(2WT) \geqq 0.085.$$

By Corollary 7.1, there exists a constant k_0, depending only on N, such that for all $WT > k_0$ the interval $0.084 \leqq x \leqq 0.085$ will contain at least $N + 2$ eigenvalues of BD_TB. Hence

$$\lambda_{[2WT]+N+1}\,(2WT) \geqq 0.084 \quad \text{for} \quad WT > k_0\,.$$

Now the proof of Theorem 5, applied without change to $\lambda_{[2WT]+N+1}$, establishes Theorem 8.

Theorem 8.1 Let ϵ_T and any integer N be given. Then as soon as T is sufficiently large

$$\inf_{a_i} \left\| f - \sum_0^{[2WT]-N} a_i\psi_i \right\|^2 \leqq 12\,\epsilon_T^2\,,$$

for all $f \in E(\epsilon_T)$.

Proof: According to Lemma 2,

$$\lambda_{[2WT]+1}\,(2WT) \leqq 0.915.$$

By Corollary 7.1, there exists a constant k_1, depending only on N, such that for all $WT > k_1$ the interval $0.915 \leqq x \leqq 0.916$ will contain at least $N + 1$ eigenvalues of BD_TB. Hence,

$$\lambda_{[2WT]-N}\,(2WT) \leqq 0.916 \quad \text{for} \quad WT > k_1\,.$$

Applying now the proof of Theorem 3 to $\lambda_{[2WT]-N}(2WT)$ establishes Theorem 8.1.

9. *Theorem 9. A. The restrictions to $t > 0$ of the functions*

$$[\sin \pi(2Wt - n)]/(2Wt - n),$$

for $n \leqq -1$, are dense in $\mathfrak{L}^2(0,\infty)$.

B. Their restrictions to $t < 0$ are not dense in $\mathfrak{L}^2(-\infty,0)$, nor even in $\mathfrak{B}$ restricted to $t < 0$.

Proof: Without loss of generality we may take $W = \frac{1}{2}$, to simplify notation. We begin with part A. Let

$$c(t) = \begin{cases} 1, & t \geqq 0 \\ 0, & t < 0. \end{cases}$$

The functions

$$\varphi_n(t) \equiv c(t)\frac{\sin \pi (t - n)}{t - n}$$

all lie in $\mathfrak{L}^2(0,\infty)$, so that their being dense in $\mathfrak{L}^2(0,\infty)$ is equivalent to the statement that $h(t) \equiv 0$ is the only function in $\mathfrak{L}^2(0,\infty)$ which is orthogonal to $\varphi_n(t)$, $n \leqq -1$ (Ref. 3, p. 72). We will prove A in this form.

Accordingly, suppose that $(h(t),\varphi_n(t)) = 0, n \leqq -1$. Using the Parseval theorem, and letting $\chi(u)$ be the characteristic function of the interval $|u| \leqq \pi$, we find

$$\begin{aligned} 0 = (h(t),\varphi_n(t)) &= \left[c(t)h(t), \frac{\sin \pi(t - n)}{t - n}\right] \\ &= (H(u),\chi(u)e^{inu}) \\ &= (\chi(u)H(u),e^{inu}), \qquad n \leqq -1, \end{aligned} \tag{9.1}$$

where $H(u)$ is the inverse Fourier transform of $c(t)h(t)$. The function $\chi(u)H(u)$ is in $\mathfrak{L}^2(-\pi,\pi)$ and may therefore be expanded there in a Fourier series $\chi(u)H(u) = \sum_{k=-\infty}^{\infty} a_k e^{iku}$. By (9.1) the coefficients a_k vanish for $k \leqq -1$, so that

$$\chi(u)H(u) = \sum_{k=0}^{\infty} a_k e^{iku}; \tag{9.2}$$

also

$$\sum_{k=0}^{\infty} |a_k|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(u)|^2\, du < \infty. \tag{9.3}$$

The function $H(u)$ may be continued analytically into the upper half of the $w = u + iv$ plane by its defining formula

$$H(w) = \int_0^\infty h(x)\, e^{iwx}\, dx\,,$$

from which, by Parseval's theorem,

$$\int_{-\infty}^{\infty} |\, H(u + iv)\,|^2\, du < A < \infty\,, \qquad v \geqq 0\,. \tag{9.4}$$

Set $G(w) \equiv \sum_{k=0}^{\infty} a_k e^{ikw}$; the function $G(w)$ is then also analytic in the upper half-plane $v > 0$, and is periodic there, with period 2π. We will now show that (9.2) implies $H(w) \equiv G(w)$ for $v > 0$, consequently that $H(w)$ is also periodic in $v > 0$ with period 2π. It then follows from (9.4) that $H(w) \equiv 0$, hence that $h(x) \equiv 0$, which was to be proved. We model our argument on one given by A. Beurling (unpublished).

Applying the Schwarz inequality to the defining expressions for $H(w)$ and $G(w)$ we find

$$|\, H(u + iv)\,|,\ |\, G(u + iv)\,| \leqq k/\sqrt{v},\quad 0 < v < 2. \tag{9.5}$$

Next set $F(w) \equiv H(w) - G(w)$ in $v > 0$.

Let $0 < \epsilon < \frac{1}{8}$, and in the w-plane denote by $P_1, P_\epsilon, P_2, Q_2, Q_\epsilon, Q_1$ the points $\pi, \pi + i\epsilon, \pi + i, -\pi + i, -\pi + i\epsilon, -\pi$ respectively. Let Γ, Γ_ϵ be the arcs made up of the line segments $P_1P_2 + P_2Q_2 + Q_2Q_1$ and $P_1P_\epsilon + P_\epsilon P_\epsilon + Q_\epsilon Q_1$ respectively. Let R_1, R_2 be the rectangles $|\, u\,| < \pi/2$, $\frac{1}{4} < v < \frac{1}{2}$ and $|\, u\,| < \pi/2$, $-\frac{1}{2} < v < -\frac{1}{4}$ respectively, and R a region which contains R_1 and R_2 and whose closure does not intersect Γ.

Form the function

$$J(w) = \int_\Gamma \frac{F(\zeta)\, d\zeta}{\zeta - w}\,.$$

By (9.5), $F(\zeta)$ is integrable on Γ, so that $J(w)$ is an analytic function of w for w off Γ, in particular for $w \in R$. Now we rewrite

$$J(w) = \int_{\Gamma - \Gamma_\epsilon} \frac{F(\zeta)\, d\zeta}{\zeta - w} + \int_{\Gamma_\epsilon} \frac{F(\zeta)\, d\zeta}{\zeta - w} \tag{9.6}$$

and estimate the second integral of (9.6). If $w \in R_1 \cup R_2$ and $\zeta \in \Gamma_\epsilon$ we see that $1/|\, \zeta - w\,| < B < \infty$. Consequently

$$\begin{aligned} \left| \int_{\Gamma_\epsilon} \frac{F(\zeta)\, d\zeta}{\zeta - w} = \int_{P_1P_\epsilon + Q_\epsilon Q_1} \frac{F(\zeta)\, d\zeta}{\zeta - w} + \int_{-\pi}^{\pi} \frac{F(u + i\epsilon)}{u + i\epsilon - w}\, du \right| \\ \leqq B \int_{P_1P_\epsilon + Q_\epsilon Q_1} |\, F(\zeta)\, d\zeta\,| + B \int_{-\pi}^{\pi} |\, F(u + i\epsilon)\,|\, du\,. \end{aligned} \tag{9.7}$$

By virtue of (9.5),

$$\lim_{\epsilon \to 0} \int_{P_1 P_\epsilon + Q_\epsilon Q_1} | F(\zeta)\, d\zeta | = 0. \tag{9.8}$$

Applying the Schwarz inequality to the remaining integral of (9.7), and using the definition of F and the triangle inequality in $\mathfrak{L}^2$ we find

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} | F(u + i\epsilon) |\, du &\leqq \left\{ \int_{-\pi}^{\pi} | F(u + i\epsilon) |^2\, du \right\}^{\frac{1}{2}} \\ &\leqq \left\{ \int_{-\pi}^{\pi} | H(u + i\epsilon) - H(u) |^2\, du \right\}^{\frac{1}{2}} \\ &\quad + \left\{ \int_{-\pi}^{\pi} | H(u) - G(u) |^2\, du \right\}^{\frac{1}{2}} \\ &\quad + \left\{ \int_{-\pi}^{\pi} | G(u) - G(u + i\epsilon) |^2\, du \right\}^{\frac{1}{2}}. \end{aligned} \tag{9.9}$$

By definition of $H(w)$,

$$H(u + i\epsilon) - H(u) = \int_0^\infty h(t)[e^{-\epsilon t} - 1]e^{iut}\, dt,$$

whence by Parseval's theorem

$$\begin{aligned} \int_{-\pi}^{\pi} | H(u + i\epsilon) - H(u) |^2\, du &\leqq \int_{-\infty}^{\infty} | H(u + i\epsilon) - H(u) |^2\, du \\ &= 2\pi \int_0^\infty | h(t) |^2 | e^{-\epsilon t} - 1 |^2\, dt. \end{aligned} \tag{9.10}$$

For each t, $\lim_{\epsilon \to 0} | h(t) |^2 | 1 - e^{-\epsilon t} |^2 = 0$, and $| h(t) |^2 | 1 - e^{-\epsilon t} |^2 \leqq 4 | h(t) |^2$, which by assumption is an integrable function. Consequently by the theorem on dominated convergence (Ref. 3, p. 37) applied to the last integral of (9.10),

$$\lim_{\epsilon \to 0^+} \int_{-\pi}^{\pi} | H(u + i\epsilon) - H(u) |^2\, du = 0. \tag{9.11}$$

Similarly from the definition of $G(w)$

$$G(u) - G(u + i\epsilon) = \sum_{k=0}^{\infty} a_k (1 - e^{-\epsilon k}) e^{iuk},$$

whence

$$\int_{-\pi}^{\pi} | G(u) - G(u + i\epsilon) |^2\, du = 2\pi \sum_{k=0}^{\infty} | a_k |^2 | 1 - e^{-\epsilon k} |^2,$$

so that using (9.3) and arguing as above

$$\lim_{\epsilon \to 0^+} \int_{-\pi}^{\pi} | G(u) - G(u + i\epsilon) |^2 \, du = 0. \tag{9.12}$$

Combining (9.12), (9.11), (9.2), (9.9), (9.8) and (9.7) we find that uniformly for $w \in R_1 \cup R_2$,

$$\lim_{\epsilon \to 0^+} \left| \int_{\Gamma_\epsilon} \frac{F(\zeta)\, d\zeta}{\zeta - w} \right| = 0. \tag{9.13}$$

Since $\Gamma - \Gamma_\epsilon$ forms the boundary of the rectangle $| u | \leqq \pi$, $\epsilon \leqq v \leqq 1$, in whose interior F is an analytic function, the first integral on the right-hand side of (9.6) is equal to $F(w)$ for $w \in R_1$ and to 0 for $w \in R_2$. From (9.13) it follows that $J(w)$, which is independent of ϵ, must itself coincide with $F(w)$ for $w \in R_1$ and with 0 for $w \in R_2$. But if $J(w) = 0$ in R_2, it must be identically 0 in its whole domain of analyticity, in particular in R, hence also in R_1. We conclude that $F(w) \equiv 0$ in R_1, hence in its whole domain of analyticity $v > 0$. Thus $H(w) \equiv G(w)$ in $v > 0$, whence, as we have already argued, part A of Theorem 9 follows.

We now pass to a proof of part B. We remark first that the restrictions of $\mathfrak{G}$ to $t < 0$ include the functions $[\sin \pi(2Wt - n)]/(2Wt - n)$, $n \geqq 1$, restricted to $t < 0$. Replacing t by $-t$, we see that, by part A, these are already dense in $\mathfrak{L}^2(-\infty, 0)$. Consequently to prove part B it is enough to establish its first assertion.

We argue by contradiction. Accordingly, suppose that the restrictions to $t < 0$ of the functions $[\sin \pi(2Wt - n)]/(2Wt - n)$, for $n \leqq -1$, are dense in $\mathfrak{L}^2(-\infty, 0)$. Then defining the function $g(t) \epsilon \mathfrak{L}^2(-\infty, 0)$ by

$$g(t) = \begin{cases} 1, & -1 \leqq t \leqq 0 \\ 0, & t < -1, \end{cases}$$

we could find a sequence of functions $f_n(t)$, each some linear combination of the $[\sin \pi(2Wt - n)]/(2Wt - n)$, $n \leqq -1$, such that $\{f_n(t)\}$ approaches $g(t)$ in $\mathfrak{L}^2(-\infty, 0)$, i.e. such that

$$\int_{-\infty}^{0} | g(t) - f_n(t) |^2 \, dt = \epsilon_n \to 0. \tag{9.14}$$

The triangle inequality in $\mathfrak{L}^2(-\infty, 0)$ applied to (9.14) yields

$$\int_{-\infty}^{0} | f_n(t) |^2 \, dt \leqq \left\{ \left[\int_{-\infty}^{0} | g(t) |^2 \, dt \right]^{\frac{1}{2}} + \sqrt{\epsilon_n} \right\}^2 = (1 + \sqrt{\epsilon_n})^2. \tag{9.15}$$

Now the functions $f_n(t)$ are all band-limited and $f_n(k) = 0$ for $k \geqq 0$. Thus by Ref. 9 there exists a constant C_1 such that

$$\int_0^\infty |f_n(t)|^2\,dt \leqq C_1 \int_{-\infty}^0 |f_n(t)|^2\,dt. \tag{9.16}$$

From (9.15) and (9.16) it follows that, as elements of $\mathfrak{L}^2(-\infty,\infty)$, the functions $f_n(t)$ have uniformly bounded norms as soon as $\epsilon_n < 1$. Applying (1.5), we conclude that the $f_n(t)$ are a uniformly bounded family of analytic functions in the strip $|\operatorname{Im}\{t\}| < 1$ of the complex t-plane, thus a normal family there (Ref. 5, p. 171). We may therefore extract from the sequence $\{f_n(t)\}$ a subsequence $f_{n_k}(t)$ converging (pointwise) in the whole strip, uniformly on any compact subset of the strip, to an analytic function $f(t)$; from (9.14),

$$f(t) = g(t), \qquad t < 0.$$

But $g(t)$ vanishes on an interval without vanishing identically, and so cannot coincide with an analytic function. We have reached a contradiction, and part B follows. Theorem 9 is established.

10. *Theorem 10*: *Let $f(t) \in E(\epsilon_T)$. Then an estimate of the form*

$$\min_{\{a_k\}} \left\| f - \sum_{|k| \leqq WT+N} a_k \frac{\sin \pi(2Wt-k)}{\pi(2Wt-k)} \right\|^2 \leqq C\epsilon_T{}^2$$

cannot be valid independently of ϵ_T, no matter how large the constants C and N are chosen.

Proof: Without loss of generality we may take $W = \frac{1}{2}$, to simplify notation.

Any function $f \in \mathfrak{B}$ has the (sampling series) expansion $f(t) = \sum_{-\infty}^{\infty} f(k)[\sin \pi(t-k)]/[\pi(t-k)]$. Since the functions

$$\varphi_k(t) = \frac{\sin \pi(t-k)}{\pi(t-k)} \text{ are orthonormal,}$$

$$\min_{\{a_k\}} \left\| f - \sum_{|k| \leqq (T/2)+N} a_k \varphi_k \right\|^2 = \sum_{|k| > (T/2)+N} |f(k)|^2. \tag{10.1}$$

Now consider the function $[\sin \pi(t-N-1)]/[\pi(t-N-1)]$ which is in $\mathfrak{B}$. By Theorem 9, we may approximate its restriction to $t > 0$ arbitrarily closely in $\mathfrak{L}^2(0,\infty)$ by finite linear combinations of the functions $[\sin \pi(t-n)]/[\pi(t-n)]$, $n \leqq -1$. That is, given $\eta > 0$, there exists constants $a_{-1},\ldots,a_{-m}$ (depending on η) such that

$$\int_0^\infty \left| \frac{\sin \pi(t-N-1)}{\pi(t-N-1)} - \sum_{k=-1}^{-m} a_k \frac{\sin \pi(t-k)}{\pi(t-k)} \right|^2 dt < \eta. \tag{10.2}$$

Let

$$\varphi_\eta(t) \equiv \frac{\sin \pi(t-N-1)}{\pi(t-N-1)} - \sum_{k=-1}^{-m} a_k \frac{\sin \pi(t-k)}{\pi(t-k)}; \tag{10.3}$$

the function $\varphi_\eta \in \mathfrak{B}$, and $\| \varphi_\eta \|^2 = 1 + \sum_{k=-1}^{-m} | a_k |^2 \geqq 1$. Since in particular φ_η is in $\mathfrak{L}^2(-\infty, \infty)$ we may choose an integer $T/2$ so large that

$$\int_{-\infty}^{-T} | \varphi_\eta(t) |^2 dt < \eta. \tag{10.4}$$

Now set

$$f(t) \equiv \frac{\varphi_\eta(t - T/2)}{\| \varphi_\eta \|};$$

We see that $f \in \mathfrak{B}$ and $\| f \| = 1$. Furthermore, by (10.2) and (10.4),

$$\int_{|t|>T/2} | f(t) |^2 dt = \frac{\int_{-\infty}^{-T} | \varphi_\eta(t) |^2 dt + \int_0^{\infty} | \varphi_\eta(t) |^2 dt}{\| \varphi_\eta \|^2} < \frac{2\eta}{\| \varphi_\eta \|^2},$$

so that $f \in E(\epsilon_T)$, with $\epsilon_T = (\sqrt{2\eta}/\| \varphi_\eta \|)$; we observe that ϵ_T can be made arbitrarily small by choosing η small, since $\| \varphi_\eta \| \geqq 1$. By definition

$$\sum_{|k|>(T/2)+N} | f(k) |^2 = \frac{1}{\| \varphi_\eta \|^2} \left[\sum_{k>N} | \varphi_\eta(k) |^2 + \sum_{k<-T-N} | \varphi_\eta(k) |^2 \right] \geqq \frac{1}{\| \varphi_\eta \|^2},$$

whence by (10.1)

$$\min_{\{a_k\}} \left\| f - \sum_{|k| \leqq (T/2)+N} a_k \varphi_k \right\|^2 \geqq \epsilon_T^2 \frac{1}{2\eta}.$$

Since η may be arbitrarily small, Theorem 10 follows.

11. *Theorem 11. For any $\beta < 1$, there exists $\delta > 0$ and ϵ_T such that*

$$\left\| f - \sum_{|k| \leqq WT+(WT)^\beta} f\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \right\|^2 > (1 + \delta)\epsilon_T^2,$$

for some $f \in E(\epsilon_T)$.

Proof: We again take $W = \frac{1}{2}$ without loss of generality. We follow a line of reasoning used in Ref. 9.

$$\text{Let } g(t) = \sin \pi t \sum_1^{\infty} \frac{1}{n^{1+\epsilon}} \frac{1}{t + n}.$$

Then

$$\frac{1}{2\epsilon} = \int_1^{\infty} \frac{dt}{t^{1+2\epsilon}} < \sum_1^{\infty} \frac{1}{n^{1+2\epsilon}} = \pi^2 \int_{-\infty}^{\infty} g^2(t)\, dt < \int_{\frac{1}{2}}^{\infty} \frac{dt}{t^{1+2\epsilon}} = \frac{2^{2\epsilon}}{2\epsilon}.$$

Now if $P, N > 0$, with $P > N + 1$, then

Proof: The decomposition

$$f = Bf + (f - Bf) \tag{12.1}$$

expresses f as the sum of its components in $\mathfrak{B}$ and orthogonal to $\mathfrak{B}$ respectively. The Pythagorean theorem then yields $1 = \| f \|^2 = \| Bf \|^2 + \| f - Bf \|^2$, whence $\| f - Bf \| = \eta_W$. Similarly, $\| f - Df \| = \epsilon_T$.

Let $g = Bf/\sqrt{1 - \eta_W^2}$, so that $g \in \mathfrak{B}$ and $\| g \| = 1$. We will apply Theorem 3 to g; to do so, we must estimate its degree of concentration. We first expand

$$\begin{aligned} \| Df - DBf \|^2 &= (Df - DBf, Df - DBf) \\ &= \| Df \|^2 + \| DBf \|^2 - 2\mathrm{Re}(Df, DBf). \end{aligned} \tag{12.2}$$

Moreover, since $\| Df - Bf \| \leqq \| f - Df \| + \| f - Bf \| = \epsilon_T + \eta_W$, we find

$$(\epsilon_T + \eta_W)^2 \geqq \| Df - Bf \|^2 = \| Df \|^2 + \| Bf \|^2 - 2\mathrm{Re}(Df, Bf). \tag{12.3}$$

Since D is a projection, $(Df, DBf) = (Df, Bf)$; hence subtracting (12.3) from (12.2),

$$\| DBf \|^2 - \| Bf \|^2 \geqq \| Df - DBf \|^2 - (\epsilon_T + \eta_W)^2,$$

or

$$\begin{aligned} \| Dg \|^2 = \frac{\| DBf \|^2}{1 - \eta_W^2} &\geqq \frac{\| Bf \|^2}{1 - \eta_W^2} - \frac{(\epsilon_T + \eta_W)^2}{1 - \eta_W^2} \\ &= 1 - \frac{(\epsilon_T + \eta_W)^2}{1 - \eta_W^2}. \end{aligned} \tag{12.4}$$

Consequently, by Theorem 3, there exist constants b_k such that

$$\| g - \sum_0^{[2WT]} b_k \psi_k \|^2 \leqq 12 \frac{(\epsilon_T + \eta_W)^2}{1 - \eta_W^2}. \tag{12.5}$$

Now from (12.1)

$$\frac{f}{\sqrt{1 - \eta_W^2}} - \sum_0^{[2WT]} b_k \psi_k = \left(g - \sum_0^{[2WT]} a_k \psi_k \right) + \left(\frac{f - Bf}{\sqrt{1 - \eta_W^2}} \right),$$

and the bracketed terms remain orthogonal. Thus, with

$$a_k = \sqrt{1 - \eta_W^2}\, b_k,$$

$$\| f - \sum_0^{[2WT]} a_k \psi_k \|^2 \leqq 12(\epsilon_T + \eta_W)^2 + \eta_W^2.$$

Theorem 12 is established.

We should point out that by letting $g = Df/\sqrt{1 - \epsilon_2^T}$ and working with the functions $D\psi_k$, the roles of ϵ and η may be interchanged, to yield the inequality

$$\| f - \sum_0^{[2WT]} c_k D\psi_k \|^2 \leqq 12(\epsilon_T + \eta_W)^2 + \epsilon_T^2 .$$

13. *Theorem 13*: *If* $f(t) \in \mathfrak{L}^2$ *with* $\| f \| = 1$, $\| Df \|^2 = 1 - \epsilon_T^2$, $\| Bf \|^2 = 1 - \eta_W^2$, *then for some constants* $c_k = c_k(f)$,

$$\| f - \sum_{|k| \leqq WT+1} c_k \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \|^2 \leqq (\epsilon_T + \eta_W)^2 + \eta_W^2 + \pi(\epsilon_T + \eta_W)\sqrt{1 - \eta_W^2}.$$

Proof: We proceed as in Theorem 12, up to (12.4) but now apply Theorem 2 instead of Theorem 3. Thus, for some constants b_k,

$$\| g - \sum_{|k| \leqq WT+1} b_k \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)} \|^2 \leqq \pi \frac{\epsilon_T + \eta_W}{\sqrt{1 - \eta_W^2}} + \frac{(\epsilon_T + \eta_W)^2}{1 - \eta_W^2}. \tag{13.1}$$

Replacing (12.5) by (13.1) and applying without change the rest of the proof of Theorem 12 establishes Theorem 13.

REFERENCES

1. Landau, H. J. and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—II, B.S.T.J., **40**, 1961, pp. 65–84.
2. Slepian, D. and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—I, B.S.T.J., **40**, 1961, pp. 43–64.
3. Riesz, F. and Sz-Nagy, B., *Functional Analysis*, Frederick Ungar, New York, N. Y., 1955.
4. Kolmogoroff, A., Annals of Mathematics, **37**, 1936, pp. 107–110.
5. Ahlfors, L. V., *Complex Analysis*, McGraw-Hill, New York, N. Y., 1953.
6. Knopp, K., *Theorie und Anwendung der Unendlichen Reihen*, Springer, Berlin, 1947.
7. Widom, H., Inversion of Toeplitz Matrices, Illinois Journal of Mathematics, 1960, pp. 88–99.
8. Paley, R. E. A. C. and Wiener, N., *Fourier Transforms in the Complex Domain*, Am. Math. Soc., 1934.
9. Pollak, H. O., Energy Distribution of Band-Limited Functions Whose Samples on a Half-Line Vanish, Journal of Math. Anal. and Appl., **2**, 1961, pp. 299–332.

A Method for Simplifying Boolean Functions

By A. H. SCHEINMAN

(Manuscript received March 6, 1962)

This article presents an iterative technique for simplifying Boolean functions. The method enables the user to obtain prime implicants by simple operations on a set of decimal numbers which describe the function. This technique may be used for functions of any number of variables.

At the present time, although several design aids have been introduced,[1,2,3] the synthesis of switching circuits remains a highly developed art, the theory being of only limited value to the circuit designer. In particular, that part of the design process involving the simplification of Boolean functions having large numbers of variables still presents a major problem.

Probably the best method currently available for the solution of such problems is the Quine-McCluskey Tabular Method, which consists of the exhaustive comparison of terms of the standard sum for adjacencies — terms which differ in only one variable. This technique, besides being a long, tedious one, determines all possible prime implicants of the given function. Thus a second problem is generated — selecting the essential prime implicants from among those found. This in itself is often a difficult procedure, for which specific methods have been developed.[4]

The method to be described in this paper is a simple, iterative technique for determining the prime implicants of a Boolean function. All of the essential prime implicants are found with this method, and in general, some or all of the nonessential prime implicants are automatically eliminated, materially simplifying the final search for the essential prime implicants.

Any Boolean function, say $f(x_1, \cdots, x_n)$ for example, may be expanded into the form

$$f(x_1, x_2, \cdots, x_n) = x_1 \cdot f(1, x_2, \cdots, x_n) + x_1' \cdot f(0, x_2, \cdots, x_n).$$

The above theorem is generally referred to as the expansion theorem[4] and enables one to expand any n-variable switching function about any one of the variables. As shown above the given function is said to be expanded about the variable x_1. Two new functions are thus formed, one multiplying x_1 and one multiplying x_1'. The two new switching functions are functions of $x_2, \cdots, x_n$ only, and are referred to as the residues of x_1.

Let us assume that the given function is specified in its canonical form as a sum of product terms,

$$f(x_1, x_2, \cdots, x_n) = \sum P_j$$

where P_j represents the general product term and the subscript j is the decimal equivalent of the associated product term, when the product terms are expressed in binary form with primed literals replaced by zeros and unprimed literals replaced by ones. This is often written as

$$f(x_1, x_2, \cdots, x_n) = \sum j, \qquad P \text{ being implied.}$$

Assume further that the variables are assigned binary weights in the order shown, with x_1 being assigned the highest weight and x_n the lowest. When the variable weights are assigned in this manner the expansion becomes a very simple process. In particular, if a specific series of decimal numbers characterizes the function, the residues may be formed as follows:

$$f(x_1, x_2, \cdots, x_n) = \sum j = x_1 R_1 + x_1' R_2.$$

R_2 is a sum consisting of the decimal numbers which are smaller than the weight of x_1 (the variable being expanded about), and R_1 is a sum consisting of the numbers which are greater than the weight of x_1, from each of which the weight of x_1 is subtracted.

Noting that the residues R_1 and R_2 are now each specified by summations, and the new sets of decimal numbers are interpreted as being functions of all succeeding variables, generically we may write:

$$R = f(x_2, x_3, \cdots, x_n).$$

Each of the residues can now be expanded about x_2, which is the highest weighted variable of the residues. Thus the expansion can be carried out in this manner using nothing more complicated than the subtraction process.

As an example, assume the following function:

$$f(A,B,C,D) = \sum(0,1,8,9,10) = AR_1 + A'R_2$$

where the order of the variables (A,B,C,D) indicates the relative binary weights, i.e., A is the highest weighted and D is the lowest weighted variable.

$$R_2 = f(B,C,D) = \sum 0,1 \quad \text{and}$$
$$R_1 = f(B,C,D) = \sum(8\text{-}8),\ (9\text{-}8),\ (10\text{-}8) = \sum 0,1,2$$

since the binary weight of $A = 8$.

This process may be repeated by expanding each of the residues about B. At this point let us examine R_1 and R_2.

There are five possibilities:

(a) $R_1 = R_2 = R$, therefore

$$f(A,B,C,D) = AR + A'R = R$$

indicating that A and A' are redundant

(b) $R_1 > R_2$, indicating that the decimal numbers representing R_2 form a subset of those representing R_1, then $R_1 = R_2 + R_b$, and

$$f(A,B,C,D) = A(R_2 + R_b) + A'R_2 = AR_2 + AR_b + A'R_2 = R_2 + AR_b.$$

(c) $R_2 > R_1$, where $R_2 = R_1 + R_a$

$$f(A,B,C,D) = AR_1 + A'(R_1 + R_a) =$$
$$AR_1 + A'R_1 + A'R_a = R_1 + A'R_a.$$

(d) The residues have no numbers in common, in which case

$$f(A,B,C,D) = AR_1 + A'R_2.$$

(e) Some of the numbers in each residue are the same,

$$R_1 = R_c + R_d, \qquad R_2 = R_c + R_e \quad \text{and}$$
$$f(A,B,C,D) = A(R_c + R_d) + A'(R_c + R_e)$$
$$= AR_c + AR_d + A'R_c + A'R_e$$
$$= R_c + AR_d + A'R_e.$$

Note that in this case the function may be written as

$$f(A,B,C,D) = R_c + A(R_c + R_d) + A'(R_c + R_e)$$

The summation R_c may be included with A and A' as a redundancy. This is significant in the method to be shown, because R_c may contribute to the simplification of residues resulting from subsequent

expansion. This is analogous to using a term of the standard sum in more than one subcube when using a Karnaugh map.[2]

The following method employs all of the above ideas and will be shown by means of an example.

Assume a function

$$f(A,B,C,D) = \sum 1,2,3,4,6,7,8,9,11,12,13,14.$$

Expansion about the variable A — the highest weighted variable — yields

$$f(A,B,C,D) = A'(\sum 1,2,3,4,6,7) + A(\sum 0,1,3,4,5,6)$$
$$= \sum 1,3,4,6 + A'(\sum 1,2,3,4,6,7) + A(\sum 0,1,3,4,5,6).$$

Where the summation $\sum 1,3,4,6$ is that part of the A residues which are the same, note that the numbers 1,3,4,6 are now redundant in the summations multiplying A and A'. The summation 1,3,4,6 actually represents those terms of the standard sum which differ only in the A variable.

There are now three summations, each of which is a function of B,C,D only. Each of these functions must now be expanded about the variable B. These operations are repeated exactly until the expansion about all variables is complete.

At this point the above illustration will be repeated, showing a mechanical technique to organize and simplify the procedure. Refer to Fig. 1.

Step 1. Arrange the decimal numbers representing the given function in a column.

Step 2. Divide the decimal numbers into two columnar groups, one headed with A' and one headed with A. The A' column contains the numbers of the original function which are smaller than 8 — the binary weight of A — and the A column contains the numbers which are equal to or greater than 8, first subtracting 8 from each.

Step 3. Include a third column, headed by a dash to indicate the redundancy of A and A', consisting of the numbers which are common to columns A and A'. Check the corresponding numbers in columns A and A' to record the fact that they are redundant. If any of the numbers in the dashed column have been *previously* checked in *both* the A and A' columns, they should also be checked in the dashed column.

Step 4. Examine each column. If any column consists of only checked numbers, eliminate the column entirely.

Each of the columns must now be expanded about B by repeating the above steps. The expansion of the function in the A' column is shown in

Fig. 1b. Since the weight of B is 4, all numbers less than 4 are placed in the B' column. These numbers are 1,2,3. Note that 1 and 3 must be checked since they were previously checked in the A' column. The numbers 4,6,7 are placed in the B column, first subtracting 4 from each, giving 0,2,3. The numbers 0,2 which correspond to 4,6 must be checked. A dashed column consisting of the numbers 2,3 is now included. Check the numbers 2,3 in both the B and B' columns. All of the numbers in the B and B' columns are now checked, hence both columns may be eliminated as shown.

Figure 2 illustrates the complete development. When the function is expanded about the final variable, note that the residues must be 0. At this point the prime implicants may be determined by simply tracing a path back to the start and reading the appropriate columnar headings.

Not all of the prime implicants obtained may be required to describe the function. In the example just shown, all of the prime implicants were essential, but another example will be shown in which this is not the case.

$$f = \sum 0,1,2,3,4,6,7,8,9,11,15$$

This function is simplified in Fig. 3, and four prime implicants are obtained, not all of which are essential. A simple method for determining

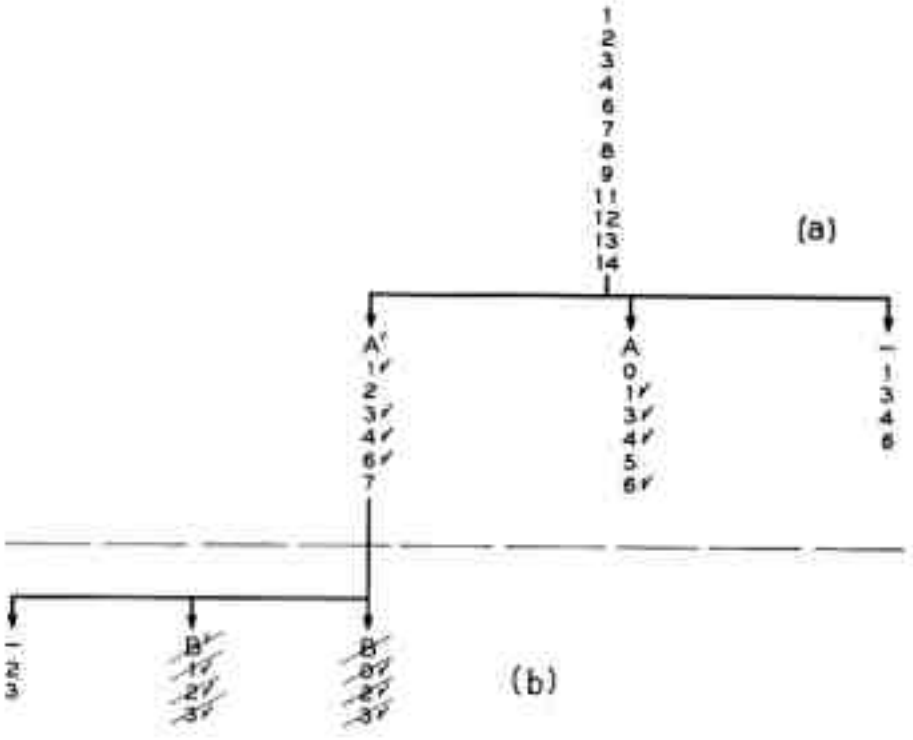


Fig. 1 — Example 1: $f(A,B,C,D) = \sum 1,2,3,4,6,7,8,9,11,12,13,14$; (a) expansion about A, (b) expansion about B.

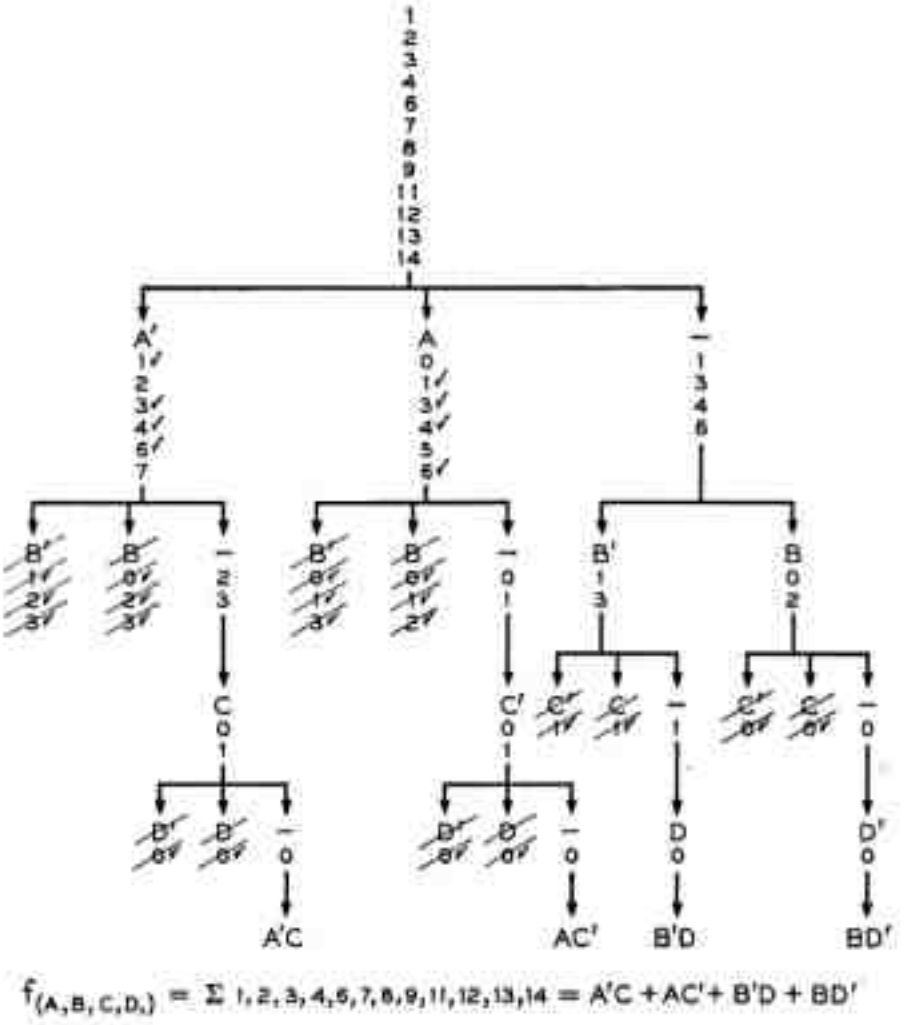


Fig. 2 — Example 1 completed.

the essential prime implicants is available in a prime implicant chart first proposed by Quine and later simplified by McCluskey. Such a chart is shown in Fig. 4 for the example of Fig. 3 (see Ref. 1).

The chart requires the establishment of columns, each of which represents one of the decimal numbers of the original function and is so headed. Each row represents one of the prime implicants and is thus identified.

Each prime implicant is a combination of 2^k of the decimal numbers of the original function; k may be any integer, including zero. It is an easy matter to find these numbers. One could, for example, trace backward from the final residue of the prime implicant (always zero) and if a column is passed through which is headed by an unprimed variable add the weight of the variable, if through a primed column the number remains unchanged. When tracing through a column headed by a dash

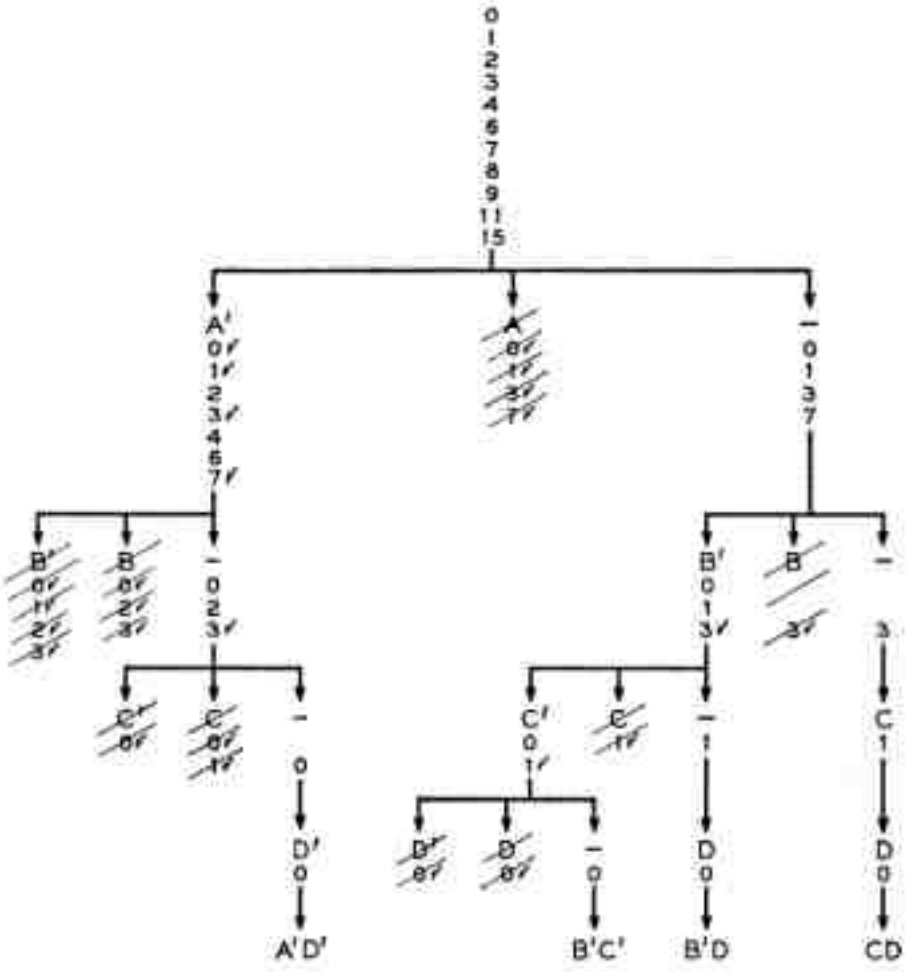


Fig. 3 — Example 2: $f(A,B,C,D) = \sum 0,1,2,3,4,6,7,8,9,11,15$; example yielding nonessential prime implicants.

each number becomes two numbers, one of which is the same as the number in the dashed column. The other number is obtained by adding the weight of the variable from whose expansion the dashed column resulted.

On each row of the chart, mark a cross under the decimal numbers

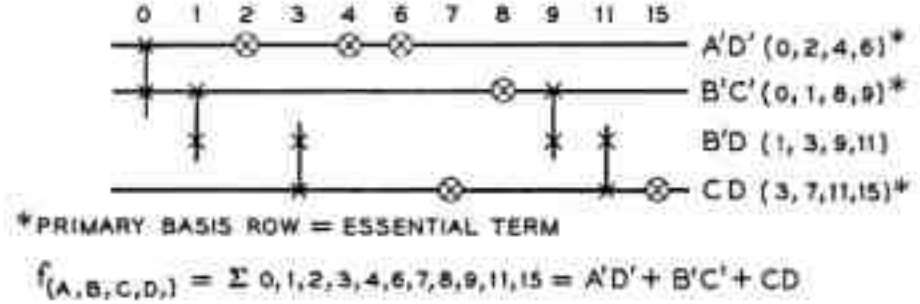


Fig. 4 — Chart method of obtaining essential prime implicants for Example 2.

associated with the terms contained in the prime implicant represented by that row. Then scan the columns and circle the crosses which stand alone in a column, and rule a line through each associated row. Such rows represent essential prime implicants. Now rule a vertical line through each cross in the ruled rows. If all other crosses in the chart are not thus ruled out by the vertical lines, additional prime implicants must be chosen. (See Ref. 1 for further discussion of prime implicant charts.) The chart of Fig. 4 shows that $A'D'$, $B'C'$, and CD are essential and the function may therefore be expressed by the following minimum sum

$$f(A,B,C,D) = A'D' + B'C' + CD.$$

If the function of Fig. 3 is simplified by other tabular methods, it will be noted that there are six possible prime implicants. This illustrates an important advantage of the method described here, and this is that while all essential prime implicants are obtained, some or all of the non-essential prime implicants may be eliminated automatically, materially simplifying the search for the essential terms by charting. The missing prime implicants are actually included in columns which were crossed out in the development because all elements were checked, indicating that the prime implicants if obtained would be redundant.

There are functions for which, if all possible prime implicants are found, charts would be produced which are cyclic in form. There are no immediately apparent choices of prime implicants which would yield a minimum sum. In such cases some initial choice must be made, and some cut and try is necessary. In fact, there will generally be more than one equally satisfactory solution, depending upon the initial choice.

When the method described in this paper is applied to functions which would normally produce a cyclical prime implicant chart, the chart obtained will not be cyclical. Some of the prime implicants will be eliminated as the work progresses, in effect making the initial choices automatically. This has two results. The chart is materially simplified and the solution is easily and automatically obtained. However if a particular initial choice would result in a more economical solution than another, then this method may or may not obtain the minimum sum, depending upon the weighting of the variables. The final solution in such cases must be regarded as an approximation to the minimum sum. The approximation will always be a close one.

If some of the decimal numbers specifying the given function are "don't care" terms, they must be checked initially and subsequently treated like any other checked number. The only real difference is noted when drawing a prime implicant chart, where columns corresponding to "don't care" terms would not be included (Refs. 1, 4).

A final example is included in Figure 5 to indicate which of the decimal numbers of the given function are combined to obtain any particular number in the expansion. The decimal numbers of the given function are shown parenthetically.

In conclusion the method described above offers the following advantages:

(a) The decimal numbers specifying the function may be operated on directly without any preliminary grouping.

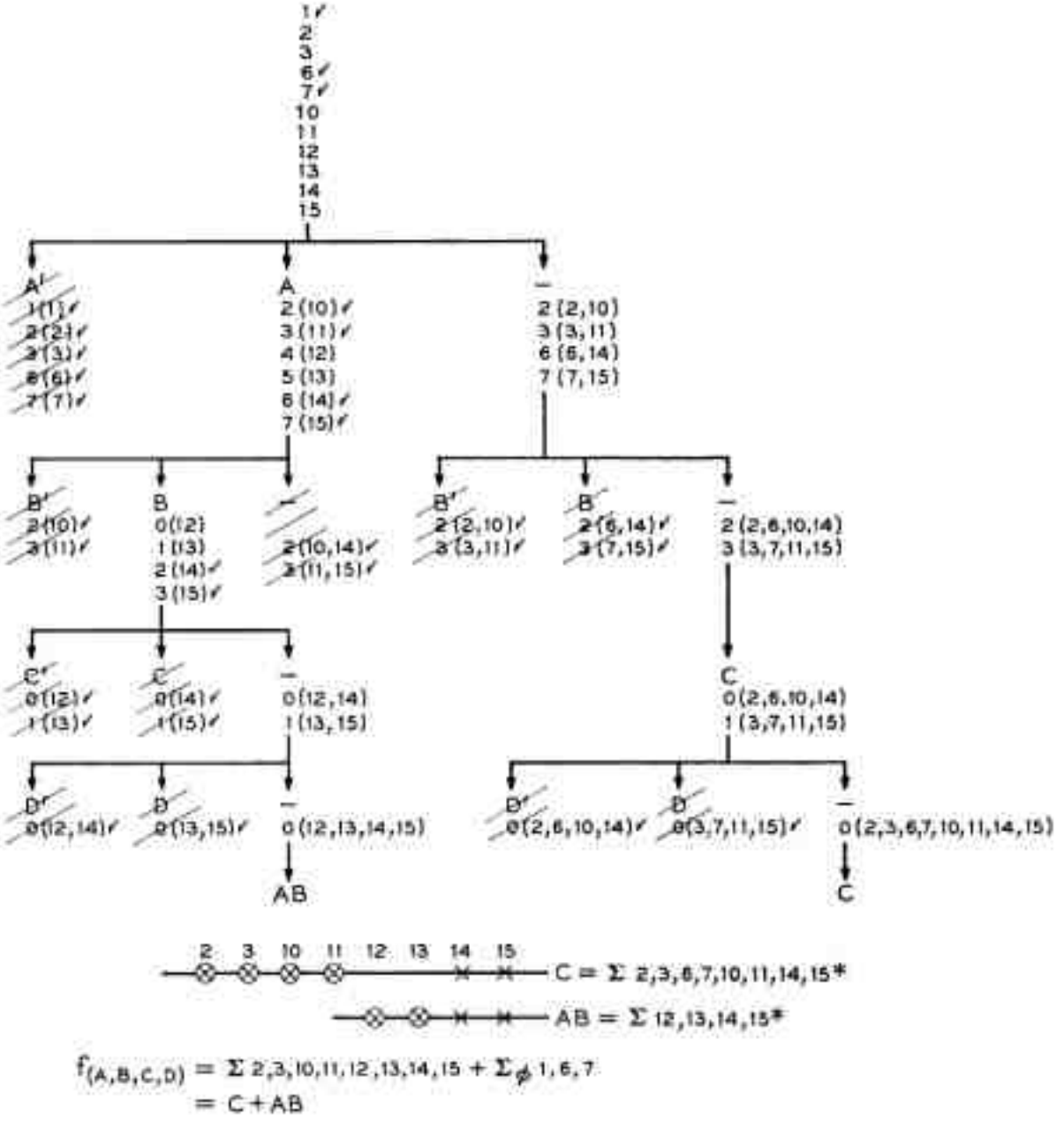


Fig. 5 — Example 3: $f(A,B,C,D) = \sum 1,2,3,6,7,10,11,12,13,14,15$; numbers in parentheses are decimal numbers which were combined to give numbers in expansion.

(b) Adjacencies (terms which differ in only one variable) are found in groups, making this method very rapid in use.

(c) The operations are simple and iterative, enabling functions having any number of variables to be simplified.

(d) Some nonessential prime implicants will in general be eliminated, simplifying the prime implicant chart.

REFERENCES

1. McCluskey, E., Jr., Minimization of Boolean Functions, B.S.T.J., **35**, 1956, pp. 1417–1445.
2. Karnaugh, M., The Map Method for Synthesis of Combinatorial Logic Circuits, A.I.E.E. Trans., Part 1: Communications and Electronics, **72**, 1953, pp. 593–599.
3. Hall, F. B., Boolean Prime Implicants by the Binary Sieve Method, A.I.E.E. Trans., Part 1: Communications and Electronics, 1962, pp. 709–713.
4. Caldwell, S. H., *Switching Circuits and Logical Design*, John Wiley & Sons Inc., N.Y., 1958.

Generalized Confocal Resonator Theory

By G. D. BOYD and H. KOGELNIK

(Manuscript received March 5, 1962)

The theory of the confocal resonator is extended to include the effect of unequal aperture size and unequal radii of curvature of the two reflectors. The latter is equivalent to a periodic sequence of lenses with unequal focal lengths. This treatment is in Cartesian coordinates as previously used. In an appendix the modes and resonant formulas are written in cylindrical coordinates.

The effect of unequal aperture size of the two reflectors is shown to produce mode patterns of unequal size on the two reflectors of a confocal resonator. The previous computations for diffraction losses are found to be applicable. Generalization of the theory to the case of reflectors of unequal curvature shows the existence of low-loss regions and high-loss regions as the reflector spacing is varied. One of the high diffraction loss regions occurs when the reflector spacing is between the two unequal radii of curvature. Such a region is interpretable in terms of instabilities in a periodic sequence of lenses of unequal focal length. An estimate of diffraction losses is obtained for the low-loss regions. The presence of a high diffraction loss region or unstable region should be of importance in the design of resonators or of a periodic sequence of lenses.

I. INTRODUCTION

The existence of modes in an open structure such as the confocal Fabry-Perot type resonator has been demonstrated by Boyd and Gordon[1] and by Fox and Li.[2] This resonator consists of two spherical reflectors separated by their common radius of curvature, as shown in Fig. 1. The reflectors were assumed to be of equal aperture and square[1] or circular[2] if viewed in the z-direction. Uniform reflectivity over the reflecting surface was postulated. Goubau and Schwering[3, 4] have also reported on this problem and have obtained similar results.

A mode may be defined as a field distribution that reproduces itself in spatial distribution and phase, though not in amplitude, as the wave bounces back and forth between the two reflectors. Because of losses

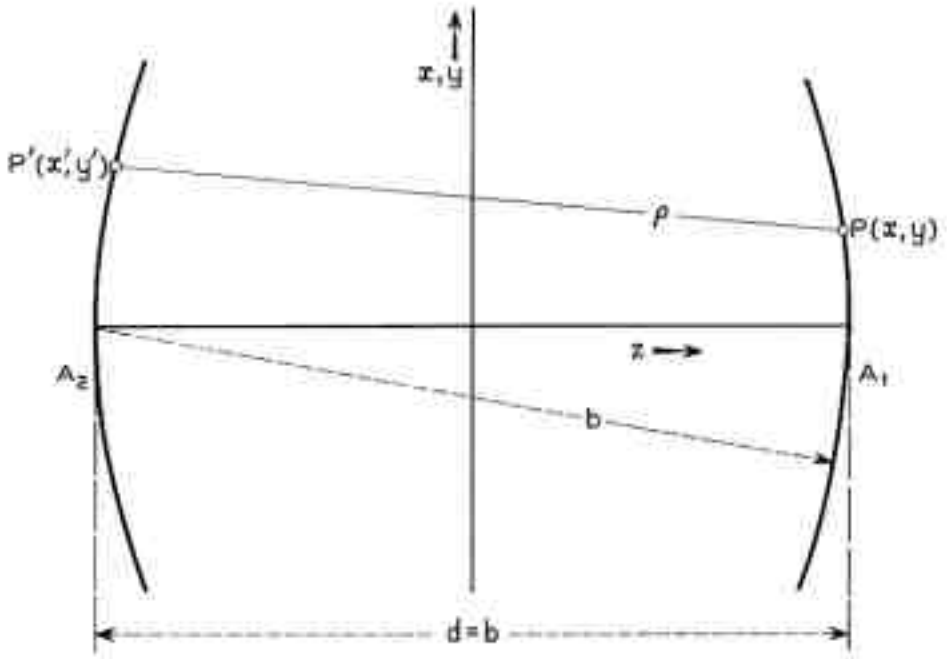


Fig. 1 — Confocal resonator with spherical reflectors.

due to diffraction and reflection, the reproduced pattern is reduced in intensity on each succeeding traversal of the resonator. The above-mentioned authors have shown that there is a set of modes which will reproduce themselves over the equal apertures A_1 and A_2 of the resonator.

Mathematically the modes of the confocal resonator form a complete orthogonal set of functions. For the confocal resonator these modes are highly degenerate in frequency; that is, many modes have the same resonance frequency. The degeneracy is split when the resonator is made nonconfocal by varying the plate spacing, though new degeneracies do appear at certain other spacings. Because of this frequency degeneracy, the modes of the confocal resonator are not unique unless the effects of loss are considered. Any linear combination of the degenerate frequency modes (the Hermite-Gaussian functions described in Ref. 1) is still a mode of the resonator.

When one includes the effect of diffraction losses due to finite apertures, the modes become unique, for then the eigenvalue degeneracy is split and each mode has its own characteristic rate of decay or Q. For the case of low diffraction losses, the eigenfunctions of the modes are still given with good approximation by the Hermite-Gaussian functions, which are exact only for the lossless case of infinite apertures. The frequency degeneracy is unaffected by the inclusion of diffraction losses.

Boyd and Gordon, in including diffraction losses, considered only the case where both reflectors, A_1 and A_2, are of *equal size*. This imposes a

certain symmetry on the system. If, however, the two reflectors are of different sizes, one might expect for the confocal resonator, with its high frequency degeneracy, stationary field configurations that are *asymmetric* in the z-direction. This may be understood by considering the set of degenerate Hermite-Gaussian modes which are resonant at a frequency given by $2q + m + n$ equal a constant. Combinations of these modes may be superimposed at one reflector to form various new field patterns. The field patterns on the two reflectors can now be different since the original modes with even and odd q change their relative phase by 180° in going from one reflector to the other. It is reasonable that the lowest-loss mode for an unequal aperture resonator will be such a combination that the field patterns will be asymmetrical. This also turns out to be true for all higher-order modes.

For the case of the nonconfocal resonator with spacing such that there are no frequency degeneracies, this asymmetry in the stationary field configurations is not possible. The field distribution is forced to be symmetrical between the two reflectors. The diffraction losses are then determined mainly by the smaller of the reflectors.

Resonators with reflectors of *different* radii of curvature are investigated also. A region of high diffraction loss is found for a range of separation of reflectors with unequal curvature near the confocal separation. This has some practical significance for resonators and transmission systems in that one must be sure to operate in only the low-loss region. Due to the possibility of slightly unequal radii of curvature in the fabrication of resonators, it is desirable to space the reflectors to obtain a nonconfocal condition. The existence of "stable regions" of low loss and "unstable regions" of high loss as the reflector spacing is varied is interpretable in terms of the stable and unstable regions of a periodic sequence of lenses of unequal focal length.

II. MODES IN A LOSSLESS CONFOCAL RESONATOR

It was pointed out in the introduction that the modes of the lossless confocal resonator are highly degenerate in frequency and thus not unique. Boyd and Gordon, in describing the modes of the lossless confocal resonator in terms of Hermite-Gaussian functions, considered only the symmetrical situation of identical field patterns and spot sizes over each aperture. Because of the high degeneracy of the lossless resonator, asymmetric field patterns between the two reflectors are just as possible. In this section the relation between asymmetric spot sizes is obtained. Only the introduction in the following section of unequal aper-

tures and the resulting diffraction losses will allow one to state which combination of asymmetric spot sizes is a unique mode of the system.

Boyd and Gordon have computed surfaces of constant phase within and without the confocal resonator. These surfaces have approximately a spherical shape. Any of these surfaces may be replaced by spherical reflectors to form a new resonating structure of arbitrary spacing and curvature. Except for the obvious special case, such a resonator was termed nonconfocal. Boyd and Gordon have shown that each confocal system of radius of curvature and separation equal to b generates a set of surfaces of constant phase of radius b' and separation d linked by the relation*

$$d^2 - 2db' + b^2 = 0. \tag{1}$$

For a given b and b' there are two possible reflector separations, d_1 and d_2 :

$$d_1 = b' + \sqrt{b'^2 - b^2},$$

and (2)

$$d_2 = b' - \sqrt{b'^2 - b^2}.$$

The field distribution of the modes of these nonconfocal systems is *symmetric* with respect to the system center (as are the fields of the generating confocal system). The fundamental modes of all these nonconfocal systems have a spot size of radius w_0 at the center of the resonator, given by

$$w_0 = \sqrt{\frac{b\lambda}{2\pi}}, \tag{3}$$

where λ is the wavelength. The spot size of the fundamental mode of the confocal system at the reflectors is $w_s = w_0\sqrt{2}$. In general, the spot size at a distance $d/2$ from the center is given by

$$w_s' = w_0 \sqrt{1 + \frac{d^2}{b^2}}. \tag{4}$$

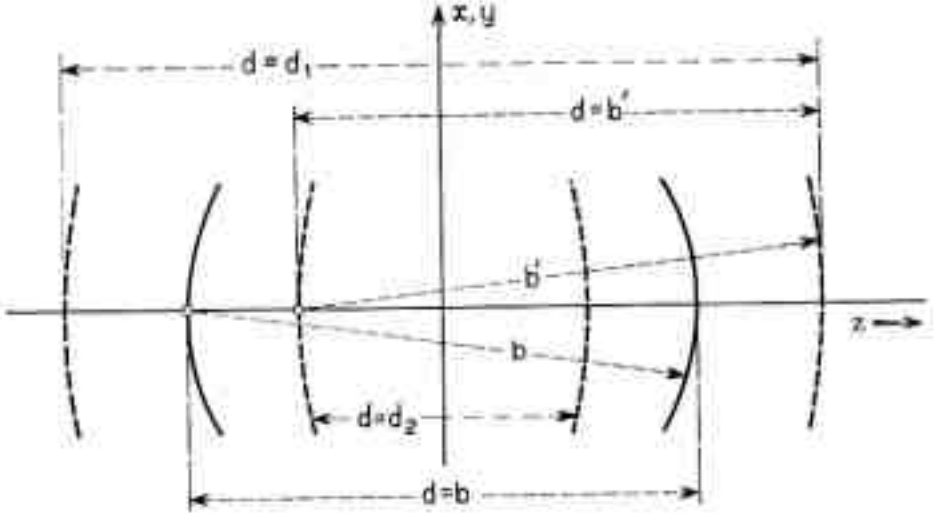
The surface of constant phase at the center ($z = 0$) is a plane, and the whole family of surfaces is symmetric with respect to it. So far we

* The notation used here is consistent with that of Boyd and Gordon but unfortunately not with that of Fox and Li, who use b for the spacing of the plane parallel resonator. We use d for the spacing of the reflectors, b for the confocal radius of curvature and thus its spacing, and b' for the radius of curvature of a surface of constant phase and for the radius of curvature of the reflectors of a nonconfocal resonator with identically curved reflectors. For the nonconfocal resonator with unequal radii of curvature we use b_1 and b_2 .

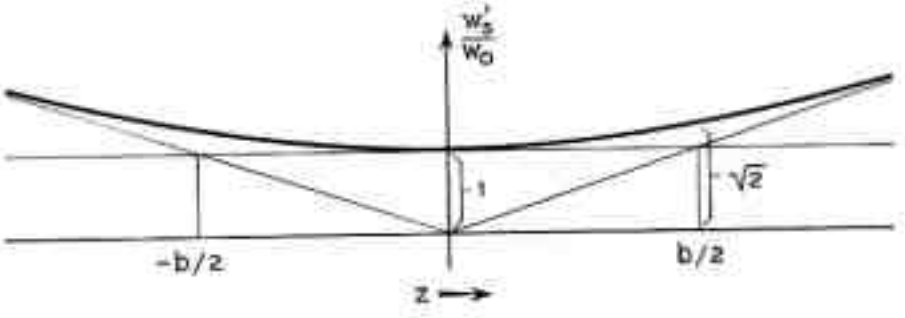
have arranged the reflector pair symmetrically to this plane, and symmetric mode configurations have resulted. It is possible, however, to construct a field configuration which is *asymmetric with respect to the reflector system* by placing one of the reflectors (with radius of curvature b') at $z = d_1/2$ and the other at $z = -d_2/2$. At these locations both surfaces of constant phase have the same radius of curvature b'. This is indicated in Fig. 2(a). The resulting reflector separation can be computed from (2) as

$$d = \tfrac{1}{2}(d_1 + d_2) = b'. \tag{5}$$

Since the spacing d equals the reflector radius of curvature b', it is apparent that a new confocal system has been formed. Apart from the surfaces of constant phase, consideration of spot sizes assures us that the



(a)



(b)

Fig. 2 — (a) Surfaces of constant phase including asymmetric confocal systems; (b) relative spot size $\frac{w_s'}{w_s} = \sqrt{1 + \frac{d^2}{b^2}}$.

modes obtained for this confocal system are, indeed, asymmetric. As indicated in Fig. 2(b) the spot size reaches its minimum value of w_0 at $z = 0$ and is not by any means in the center of the reflector system. The spot sizes at the new reflectors can be computed from (4) as

$$w_1 = w_0 \sqrt{1 + \frac{d_1^2}{b^2}}, \qquad w_2 = w_0 \sqrt{1 + \frac{d_2^2}{b^2}}. \tag{6}$$

This combined with (2) yields the relation

$$w_1 w_2 = \frac{\lambda b'}{\pi}, \tag{7}$$

where $\sqrt{\lambda b'/\pi}$ is the spot size at the reflectors, which we would expect for the symmetrical set of modes of a confocal system of spacing b'.

The resonance condition for this system is obtained via Boyd and Gordon's equation (20) after some computation as

$$\frac{4b'}{\lambda} = 2q + (1 + m + n), \tag{8}$$

where m, n, and q are the mode numbers as defined in Boyd and Gordon's work. By comparing their equation (14) with our result, we find that the resonance conditions for the symmetric and the asymmetric modes of the confocal system are identical, as expected.

By suitably choosing b for a given reflector curvature b', almost any ratio of reflector spot sizes w_1/w_2 can be obtained. Thus, for a given lossless confocal system, the confocal geometry allows an infinite number of sets of modes (characterized by the spot sizes at each aperture). It is the *finite* size and shape of the reflector that selects one particular set, as we shall see in the following section.

III. MODES OF A CONFOCAL RESONATOR WITH REFLECTOR SIZES UNEQUAL

Consider a confocal resonator. Assume that the reflectors A_1 and A_2 are, in general, of different sizes and/or shapes. With this asymmetry in mind, it is no longer reasonable to postulate that the field pattern on A_1 be reproduced on A_2 when looking for self-consistent field configurations. Instead, as a more generalized definition of a mode let us require that an energy distribution launched with a certain pattern on A_1 reproduce this pattern *on* A_1 after bouncing back from A_2. No condition on the pattern on A_2 is imposed.

To express this mathematically we use the approximations of Boyd and Gordon's paper and the scalar formulation of Huygens' principle. A wave leaving reflector A_1 with a field pattern $E(x, y)$ arrives at A_2 with a pattern $E'(x', y')$ given by

$$E'(x', y') = \frac{ik}{2\pi b} \int_{A_1} d\bar{x}\, d\bar{y}\, E(\bar{x}, \bar{y}) e^{-ik\rho}, \tag{9}$$

where b is the mirror separation, $k = 2\pi/\lambda$, and

$$\rho = b - \frac{1}{b}(\bar{x}x' + \bar{y}y'). \tag{10}$$

Most of the energy is reflected from A_2 and travels back to A_1. The radii of curvature of the reflectors are assumed very large compared to the wavelength λ, and we can therefore assume that laws for the reflection of plane waves apply locally. Then we find that the reflected wave leaves A_2 with the pattern $-E'(x', y')$. It will arrive at A_1 with a certain distribution pattern which we shall call $-\sigma_m^2 \sigma_n^2 E(x, y)$. At this point we have introduced the postulate that the field patterns be reproduced, except for the amplitude factor $\sigma_m^2 \sigma_n^2$, after one complete return trip. Again, the energy is bounced back and leaves reflector A_1 with a field distribution which can be expressed in terms of $E'(x', y')$ as

$$\sigma_m^2 \sigma_n^2 E(x, y) = \frac{ik}{2\pi b} \int_{A_2} dx'\, dy'\, E'(x', y') e^{-ik\rho'}, \tag{11}$$

with

$$\rho' = b - \frac{1}{b}(xx' + yy'). \tag{12}$$

Substituting (9) into (11) to eliminate E', then inserting the expressions (10) and (12) for ρ and ρ', and, finally, interchanging integrals, one obtains an integral equation

$$\sigma_m^2 \sigma_n^2 E(x, y) = -\frac{k^2 e^{-2ikb}}{4\pi^2 b^2} \int_{A_1} d\bar{x}\, d\bar{y}\, E(\bar{x}, \bar{y}) K(x, \bar{x}; y, \bar{y}), \tag{13}$$

with the kernel

$$K(x, \bar{x}; y, \bar{y}) = \int_{A_2} dx'\, dy' \exp\left(i\frac{k}{b}[x'(x + \bar{x}) + y'(y + \bar{y})]\right). \tag{14}$$

This is the fundamental integral equation that yields as its solution the modes of our system and their diffraction losses. The kernel $K(x, \bar{x};$

$y, \bar{y}$) depends on the shape and size of reflector A_2 and will be evaluated for some special cases in the following sections.

Integral equations for the modes of asymmetric nonconfocal systems can be derived on the basis of arguments similar to the ones used in this chapter, but solutions for them are in general not available.

IV. CONFOCAL RESONATORS WITH UNEQUAL SQUARE AND RECTANGULAR APERTURES

In this section a confocal resonator with two reflectors of finite but *unequal* size is considered. Let reflector A_1 extend from $-a_1$ to $+a_1$ in the x direction and from $-A_1$ to $+A_1$ in the y direction as shown in Fig. 3(a). Reflector A_2 is chosen to be of a rectangular shape $2a_2$ by $2A_2$ correspondingly.

For the above reflector dimensions, the kernel of integral equation (13) takes the form

$$K(x, \bar{x}; y, \bar{y}) = \int_{-a_2}^{a_2} dx' \int_{-A_2}^{A_2} dy' \exp\left(i\frac{k}{b}[x'(x+\bar{x}) + y'(y+\bar{y})]\right). \quad (15)$$

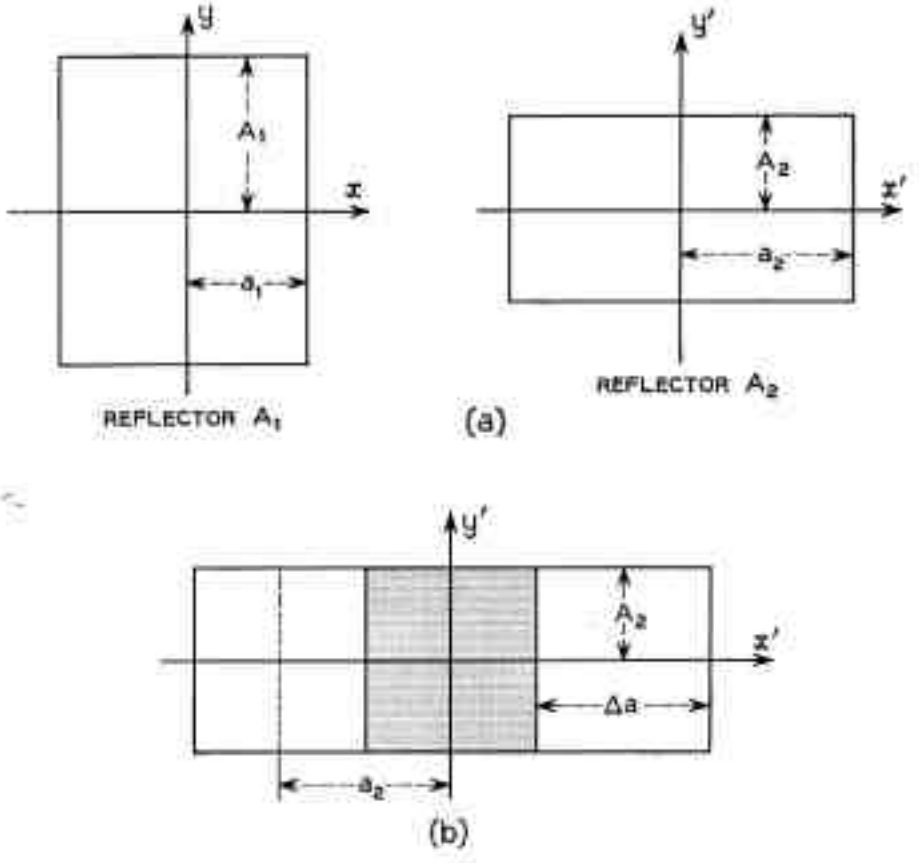


Fig. 3 — (a) Reflectors with rectangular aperture; (b) reflector A_2 blocked in center.

These integrals can be evaluated analytically and the kernel rewritten as

$$K(x, \bar{x}; y, \bar{y}) = \frac{4b^2}{k^2} \cdot \frac{\sin \frac{k}{b} a_2(x + \bar{x})}{(x + \bar{x})} \cdot \frac{\sin \frac{k}{b} A_2(y + \bar{y})}{(y + \bar{y})}. \tag{16}$$

Assume that the field pattern of a mode can be written in the form

$$E(x, y) = E_n f_m(x) g_n(y), \tag{17}$$

with E_n a constant amplitude factor, $f(x)$ a function of x only, and $g(y)$ a function of y only. Under these conditions integral equation (13) can be rearranged as

$$\begin{aligned} \sigma_m^2 \sigma_n^2 f_m(x) g_n(y) = -e^{-2ikb} \int_{-a_1}^{+a_1} d\bar{x}\, f_m(\bar{x}) \frac{\sin \frac{k}{b} a_2(x + \bar{x})}{\pi(x + \bar{x})} \\ \cdot \int_{-A1}^{A1} d\bar{y}\, g_n(\bar{y}) \frac{\sin \frac{k}{b} A_2(y + \bar{y})}{\pi(y + \bar{y})}. \end{aligned} \tag{18}$$

An integral relation satisfied by the angular prolate spheroidal wave functions $S_{0n}(c, s)$ is[5]

$$\frac{2c}{\pi} [R_{0n}^{(1)}(c, 1)]^2 S_{0n}(c, t) = \int_{-1}^{+1} ds \frac{\sin c(t - s)}{\pi(t - s)} S_{0n}(c, s), \tag{19}$$

where $R_{0n}^{(1)}(c, 1)$ is the radial prolate spheroidal wave function of nth order. Because

$$S_{0n}(c, s) = (-1)^n S_{0n}(c, -s), \tag{20}$$

it holds that

$$\frac{2c}{\pi} [R_{0n}^{(1)}(c, 1)]^2 S_{0n}(c, t) = (-1)^n \int_{-1}^{+1} ds \frac{\sin c(t + s)}{\pi(t + s)} S_{0n}(c, s). \tag{21}$$

We can compare this relation with integral equation (18) (with $c = (k/b)a_1a_2$ and $C = (k/b)A_1A_2$) and conclude that the field patterns of the TEM_{mnq} mode of the system are

$$\begin{aligned} &\text{on } A_1: \quad E(x, y) \propto S_{0m}\left(c, \frac{x}{a_1}\right) S_{0n}\left(C, \frac{y}{A_1}\right); \\ &\text{on } A_2: \quad E'(x', y') \propto S_{0m}\left(c, \frac{x'}{a_2}\right) S_{0n}\left(C, \frac{y'}{A_2}\right). \end{aligned} \tag{22}$$

As pointed out in Boyd and Gordon's work, the prolate spheroidal

wave functions S_{0m} can be approximated by orthogonal Hermite functions if the apertures are large enough. This enables one to derive an approximate formula for the dimensions of the "spot" of the fundamental, which indicates the location where the field of this mode has decreased by a factor e^{-1} with respect to its maximum. For rectangular reflectors these quantities will, in general, be different for the x and the y directions. One obtains

$$\begin{aligned} &\text{on reflector } A_1: \quad x_s = \sqrt{\frac{a_1}{a_2}}\sqrt{\frac{b\lambda}{\pi}}; \qquad y_s = \sqrt{\frac{A_1}{A_2}}\sqrt{\frac{b\lambda}{\pi}}; \\ &\text{on reflector } A_2: \quad x_s' = \sqrt{\frac{a_2}{a_1}}\sqrt{\frac{b\lambda}{\pi}}; \qquad y_s' = \sqrt{\frac{A_2}{A_1}}\sqrt{\frac{b\lambda}{\pi}}. \end{aligned} \tag{23}$$

Note that

$$x_s x_s' = y_s y_s' = w_s^2 = \frac{b\lambda}{\pi}, \tag{24}$$

which agrees with (7). From the above we see that, compared to a confocal resonator with equal apertures, the patterns of all modes on reflector A_1 are now magnified by a factor $\sqrt{a_1/a_2}$ in x direction and a factor $\sqrt{A_1/A_2}$ in y direction, if $a_1 > a_2$ and $A_1 > A_2$. The patterns on reflector A_2 are compressed correspondingly.

The center of the reflector system is no longer the position of maximum energy density. The position of maximum energy density will, in general, be different for the concentration in the x direction as compared to the concentration in the y direction. One computes displacements D_x and D_y of the positions of maximum concentration from the center in the direction of the smaller reflector:

$$D_x = \frac{b}{2}\cdot\frac{a_2^2 - a_1^2}{a_2^2 + a_1^2}; \qquad D_y = \frac{b}{2}\cdot\frac{A_2^2 - A_1^2}{A_2^2 + A_1^2}. \tag{25}$$

Comparison of (18) and (21) also yields the eigenvalues of the integral equation (20):

$$\sigma_m^2\sigma_n^2 = -(-1)^{n+m}e^{-2ikb}\frac{4cC}{\pi^2}[R_{0m}^{(1)}(c,1)\cdot R_{0n}^{(1)}(C,1)]^2. \tag{26}$$

This shows that

(i) the resonance condition

$$\frac{4b}{\lambda} = 2q + (1 + m + n) \tag{27}$$

is not changed by making the apertures of the reflectors of a confocal system unequal, and

(*ii*) the diffraction losses of a confocal system with *unequal* reflector apertures of dimensions a_1, A_1, a_2, and A_2 are equal to the diffraction losses of a confocal system with *equal* aperture dimensions a_e, A_e if $a_e^2 = a_1 a_2$ and $A_e^2 = A_1 A_2$.

V. CONFOCAL RESONATOR WITH ONE REFLECTOR PARTIALLY BLOCKED

In this section we would like to quickly sketch the analytical treatment of a confocal reflector system in which one reflector — in our case A_2 — is blacked out in the center as shown in Fig. 3(b). The other reflector dimensions are assumed to be the same as in the previous section. The effective shape of reflector A_2 is now that of two rectangles of width Δa, extending from $x' = -a_2 - (\Delta a/2)$ to $-a_2 + (\Delta a/2)$, and from $x' = a_2 - (\Delta a/2)$ to $a_2 + (\Delta a/2)$. If we insert the corresponding limits into (14), we obtain the kernel

$$K(x, \bar{x}; y, \bar{y}) = \frac{8b^2}{k^2(x + \bar{x})(y + \bar{y})} \sin\left[\frac{k\Delta a}{2b}(x + \bar{x})\right] \cdot \cos\left[\frac{ka_2}{b}(x + \bar{x})\right] \sin\left[\frac{k}{b} A_2(y + \bar{y})\right] \tag{28}$$

for the integral equation describing the system.

For very small reflector width $\Delta a \ll b\lambda/a$, the kernel is given with good approximation by

$$K(x, \bar{x}; y, \bar{y}) = \frac{4b\Delta a}{k(y + \bar{y})} \cos\frac{ka_2}{b}(x + \bar{x}) \sin\frac{kA_2}{b}(y + \bar{y}). \tag{29}$$

With this kernel, integral equation (13) can be separated into one equation containing functions of x and one containing functions of y only, as in the previous section. While the latter is the same as the integral equation treated in Section IV, the equation for $f(x)$ is of the form

$$\gamma f(x) = 2\Delta a \int_{-a_1}^{+a_1} d\bar{x}\, f(\bar{x}) \cos\frac{ka_2}{b}(x + \bar{x}). \tag{30}$$

Applying standard procedures, this integral equation can be solved elementarily. The solutions are

$$f(x) = \cos\frac{ka_2}{b} x, \tag{31}$$

with the eigenvalue

$$\gamma = 2\Delta a \left(a_1 + \frac{b}{2ka_2} \sin 2\frac{k}{b} a_1 a_2 \right), \tag{32}$$

and

$$f(x) = \sin \frac{ka_2}{b} x, \tag{33}$$

with the corresponding eigenvalue

$$\gamma = -2\Delta a \left(a_1 - \frac{b}{2ka_2} \sin \frac{2k}{b} a_1 a_2 \right). \tag{34}$$

The eigenvalues, of course, determine the diffraction losses and the resonance conditions for the reproducing patterns, as in the previous section. The resonance formula is

$$\frac{4d}{\lambda} = 2q + 1 + n \tag{35}$$

for the even cosine-function, and

$$\frac{4d}{\lambda} = 2q + 2 + n \tag{36}$$

for the odd sine-function.

One should note that the field distribution on reflector A_1 is simply the two-slit diffraction pattern one would expect from coherent excitation of the two narrow reflectors comprising A_2.

VI. RESONATORS WITH REFLECTORS OF UNEQUAL CURVATURE

To investigate resonator systems with concave reflectors of unequal radii of curvature, let us return to the consideration of a lossless system. From this model one can obtain information on spot sizes and resonance conditions. Diffraction losses will be estimated using the same approximation previously used by Boyd and Gordon for the nonconfocal resonator of equal curvature.

6.1 *Surfaces of Constant Phase*

Let reflector A_1 have a radius of curvature b_1, and reflector A_2 a radius of b_2. We shall base our argument on Boyd and Gordon's picture of surfaces of constant phase (Fig. 2), which we have already used in

Section II. In a set of surfaces, characterized by the confocal parameter b, reflector A_1 can be placed at the distances $\pm d_1/2$ from the center, and A_2 at $\pm d_2/2$ correspondingly. These distances can be computed from (2) as

$$\begin{aligned} d_1 &= b_1 \pm \sqrt{b_1^2 - b^2}, \\ d_2 &= b_2 \pm \sqrt{b_2^2 - b^2}. \end{aligned} \tag{37}$$

With given concave reflectors, therefore, four different resonator systems can be found which fit this particular set of surfaces of constant phase. The four different reflector separations $d = \frac{1}{2}(d_1 + d_2)$ are given by

$$2d = b_1 + b_2 \pm \sqrt{b_1^2 - b^2} \pm \sqrt{b_2^2 - b^2}. \tag{38}$$

To obtain various other resonator systems the parameter b can be varied. But the range of this variation is restricted, since only real valued distances have physical meaning in this context. If we assume that $b_2 > b_1$, it follows from (38) that b can be varied in the range from 0 to b_1. One can thus obtain reflector separations d in the range from 0 to b_1 and from b_2 to $b_1 + b_2$ as shown in Fig. 4. No information on resonators with reflector separations in the range from b_1 to b_2 can be obtained. It is of interest that the confocal system for reflectors of unequal curvature, with $d = \frac{1}{2}(b_1 + b_2)$, is just in this "unstable region."

Let us restrict our discussion to systems of given b_1, b_2, and d in the range covered by the picture of surfaces of constant phase. We can

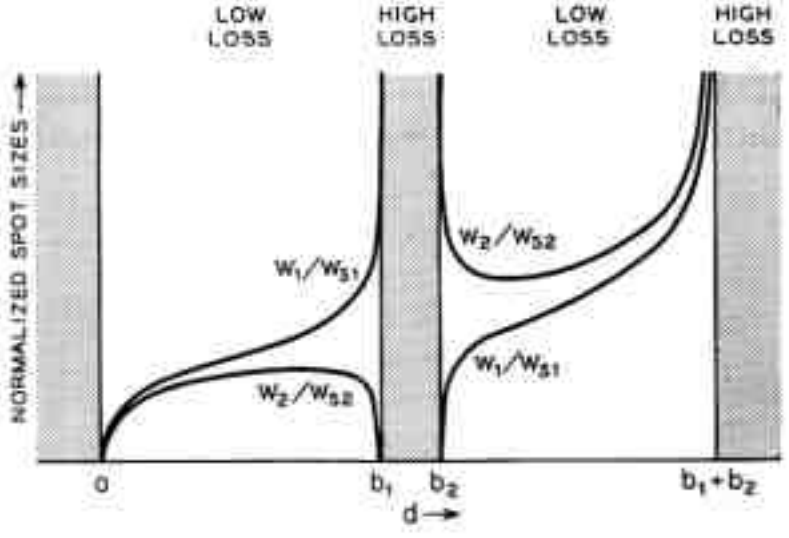


Fig. 4 — Spot sizes and high- and low-loss regions for a resonator with reflectors of unequal curvature and variable spacing.

inquire into the spot sizes w_1 on A_1 and w_2 on A_2. Combining (1), (4), (37), and (38) we obtain the relations

$$\left(\frac{w_1}{w_2}\right)^2 = \frac{b_1}{b_2}\cdot\frac{b_2 - d}{b_1 - d}, \tag{39}$$

$$(w_1 w_2)^2 = \left(\frac{\lambda}{\pi}\right)^2 \frac{b_1 b_2 d}{b_1 + b_2 - d}. \tag{40}$$

It also follows that the maximum concentration of energy occurs at the distance

$$\frac{d_1}{2} = d\,\frac{b_2 - d}{b_1 + b_2 - 2d} \tag{41}$$

from reflector A_1, and at

$$\frac{d_2}{2} = d\,\frac{b_1 - d}{b_1 + b_2 - 2d} \tag{42}$$

from A_2.

In Fig. 4 we have shown for a special case how the spot sizes w_1 and w_2 vary as a function of the reflector spacing d. In this figure the spot size on each reflector is normalized in terms of $w_{s1} = \sqrt{b_1\lambda/\pi}$ and $w_{s2} = \sqrt{b_2\lambda/\pi}$. These are the spot sizes at the reflectors of equal-radii confocal resonators with radii of b_1 and b_2 respectively.

Note that as d approaches b_1, the spot size w_2 on A_2 approaches zero, while the spot size w_1 on A_1 increases beyond limit. The corresponding effect occurs if d approaches b_2 from above. It should be remembered that the information obtained here can be applied usefully only so long as the spot sizes are somewhat smaller than the corresponding reflector dimensions.

The diffraction losses of the nonconfocal resonator of equal radii of curvature and aperture were previously estimated by Boyd and Gordon on the assumption that the diffraction loss is equal to that of its equivalent confocal resonator with reflector dimensions scaled up by the ratio of their spot sizes.

For the nonconfocal resonator of unequal radii of curvature and square apertures of sides $2a_1$ and $2a_2$ respectively, the equivalent Fresnel numbers at reflectors A_1 and A_2, which determine the diffraction losses at each reflector, are obtained from Boyd and Gordon's equation (29) as

$$\begin{aligned}\left(\frac{a^2}{b\lambda}\right)_1 &= \frac{a_1^2}{d_1\lambda}\left[2\frac{d_1}{b_1} - \left(\frac{d_1}{b_1}\right)^2\right]^{\frac{1}{2}},\\ \left(\frac{a^2}{b\lambda}\right)_2 &= \frac{a_2^2}{d_2\lambda}\left[2\frac{d_2}{b_2} - \left(\frac{d_2}{b_2}\right)^2\right]^{\frac{1}{2}},\end{aligned} \tag{43}$$

where d_1 and d_2 are determined by (41) and (42). The diffraction loss at each reflector is then obtainable from Fig. 3 of Boyd and Gordon as α_{D1} and α_{D2}.

The resonator Q is given by

$$Q = \frac{2\pi d}{\alpha\lambda}, \tag{44}$$

where

$$\alpha = \tfrac{1}{2}(\alpha_{D1} + \alpha_{D2}) + \alpha_R \tag{45}$$

and α_R represents the reflection loss per bounce at a reflector plus the single-pass scattering and absorption loss between the reflectors.

On the basis of this estimate, one concludes that the diffraction losses increase sharply if the separation d approaches an "unstable" region. No similar estimate of diffraction losses is available for the "unstable" regions. However, a ray optical analysis which we present in the next section shows the divergent nature of "unstable" resonator systems. This indicates relatively high diffraction losses.

With results obtained here and the help of Boyd and Gordon's equation (20), the resonance condition for the resonator with reflectors of different curvature can be computed as

$$\frac{2d}{\lambda} = q + \frac{1}{\pi}(1 + m + n)\cos^{-1}\sqrt{\left(1 - \frac{d}{b_1}\right)\left(1 - \frac{d}{b_2}\right)}. \tag{46}$$

To compare this with Boyd and Gordon's resonance formula for resonators with equal curvature b', we rewrite their equation (31) in terms of d and b':

$$\frac{2d}{\lambda} = q + \frac{1}{\pi}(1 + m + n)\cos^{-1}\left(1 - \frac{d}{b'}\right). \tag{47}$$

We have found this to be a very convenient form in which to rewrite their resonance formula (31). But due to well known relations between the trigonometric functions, various other formulations are possible. One of these formulations was given by J. R. Pierce.[6]

A half nonconfocal resonator may be formed by a plane reflector and a spherical reflector of radius of curvature b_1 and spacing d between the reflectors with $d < b_1 < \infty$. The resonant condition may be obtained from (46) by letting $b_2 \to \infty$. The result is given by

$$\frac{2d}{\lambda} = q + \frac{1}{2\pi}(1 + m + n)\cos^{-1}\left(1 - \frac{2d}{b_1}\right). \tag{48}$$

6.2 *Equivalent Sequence of Lenses*

Let us call the regions $b_1 + b_2 > d > b_2$ and $0 < d < b_1$ "stable" or "low loss," and the regions $d > b_1 + b_2$ and $b_1 < d < b_2$ "unstable" or "high loss." We can understand these stable and unstable regions of a resonator system with reflectors of unequal curvature from another point of view if we replace the resonator by an equivalent sequence of lenses. These lenses are spaced at distances d and have focal lengths of $f_1 = b_1/2$ and $f_2 = b_2/2$ respectively. Lens systems of this type have been used in periodic focusing of long electron beams and instabilities have been observed.

Stability investigations of sequences of lenses of equal focal length are readily available.[7] These systems are stable if

$$0 < \frac{L}{f} < 4, \tag{49}$$

where L is the lens spacing and f the focal length.

A pair of lenses of focal lengths f_1 and f_2 spaced at the distance d can be replaced by an equivalent optical system[8] of a focal length f given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}. \tag{50}$$

The system's principal planes are found to be spaced at distances $h_1 = d(f/f_2)$ and $h_2 = d(f/f_1)$ from the corresponding lenses (see Fig. 5).

If we substitute such a thick lens for each pair of unequal lenses of our system, we obtain a sequence of equal optical systems of focal length f. If, furthermore, we define as their "effective" spacing

$$L = d + h_1 + h_2, \tag{51}$$

the arguments of Pierce's treatment[7] are applicable to our case.

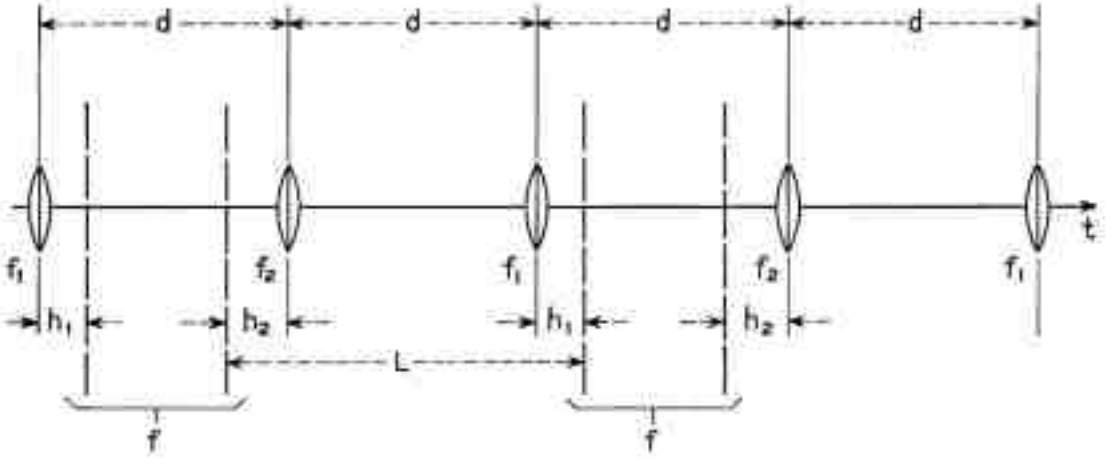


Fig. 5 — Sequence of lenses of alternating focal length

Combining equations we obtain

$$\frac{L}{f} = d\left\{\frac{2}{f_1} + \frac{2}{f_2} - \frac{d}{f_1 f_2}\right\}. \tag{52}$$

From (49) and (52) one can show that the boundaries of the stable regions are given by

$$\left(\frac{d}{b_1} - 1\right)\left(\frac{d}{b_2} - 1\right) \leqq 1, \tag{53}$$

and

$$\left(\frac{d}{b_1} - 1\right)\left(\frac{d}{b_2} - 1\right) \geqq 0. \tag{54}$$

These relations define the stable and unstable regions in agreement with the preceding discussion, but they are also valid for negative values of b_1 and b_2. If one allows for convex reflectors, one can also obtain this somewhat generalized result from the picture of surfaces of constant phase.

6.3 *Stability Diagram*

A. G. Fox and T. Li have suggested a two-dimensional diagram of the stable and unstable regions which is very instructive. Several choices of coordinates are possible. In Fig. 6 we have plotted d/b_1 and d/b_2 as coordinates. In this diagram the boundary lines described by (54) appear as straight lines, and the curve represented by (53) as a hyperbola, as shown. For confocal systems, i.e., systems with coinciding reflector foci, we have $2d = b_1 + b_2$, which may be written:

$$\left(\frac{d}{b_1} - \frac{1}{2}\right)\left(\frac{d}{b_2} - \frac{1}{2}\right) = \frac{1}{4}. \tag{55}$$

In our diagram, therefore, these systems are represented by points on another hyperbola and fall within the high-loss region. A transition from a "stable" to an "unstable" region means an extremely sharp increase of diffraction losses for reasonably large Fresnel numbers.

The confocal system with reflectors of equal curvature is represented by a rather singular point in our diagram. We see that certain deviations from the ideal dimensions $d = b_1 = b_2$ will greatly increase the system losses. This should be taken into account when designing maser resonators or optical transmission systems, and it may be advisable to choose points of operation at a safe distance from the unstable region. The degenerate frequency characteristics of the confocal system can be

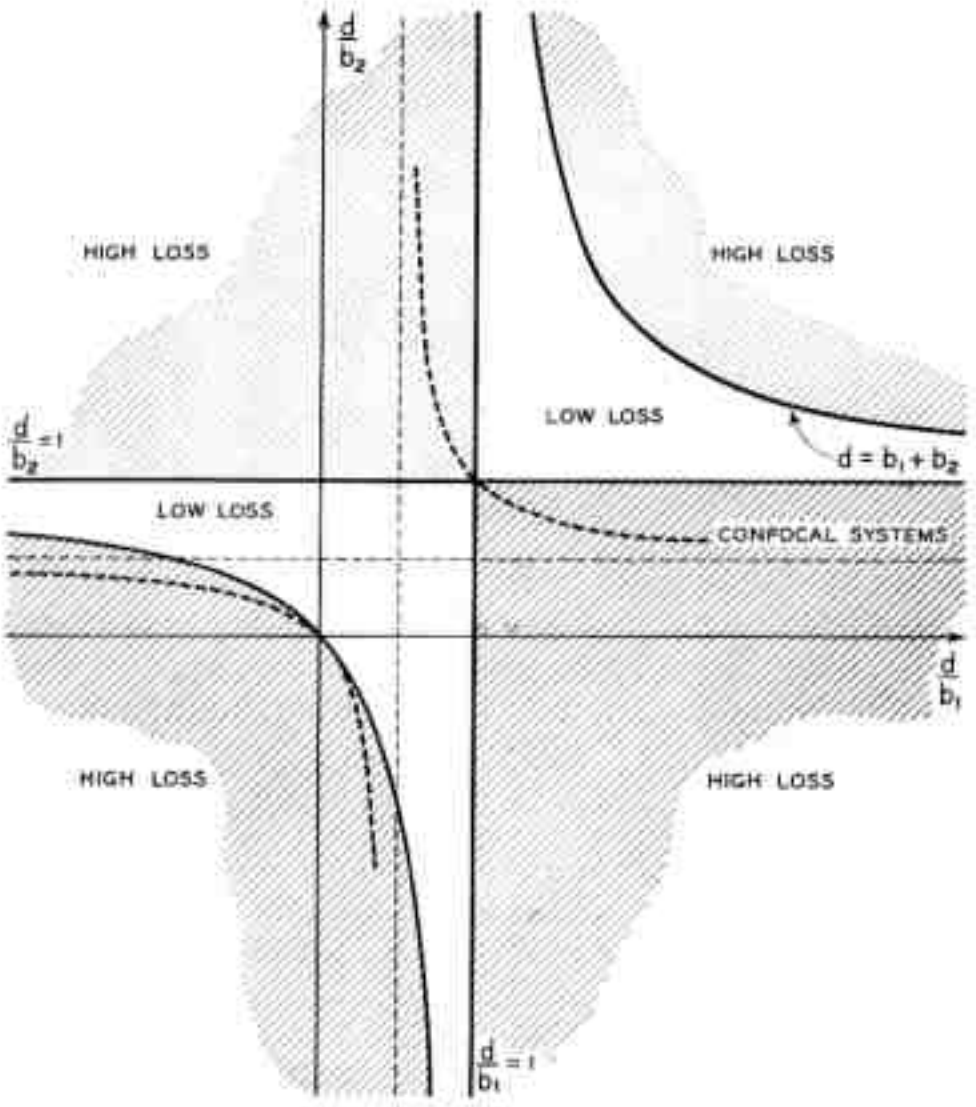


Fig. 6 — Two-dimensional diagram of stable and unstable regions.

obtained, if desired, by using a concave reflector ($b_1 = 2d$) and a flat one ($b_2 = \infty$) spaced at the distance $d = b_1/2$. This system is in a stable region.

6.4 *Equivalent Systems*

One may ask whether operation in one of the two low-loss regions is to be preferred to the other. We have found that the two regions are absolutely equivalent as far as diffraction losses, spot sizes, and resonance conditions are concerned. For a given reflector spacing d, we can find a corresponding pair of resonator systems, one operating in the lower stable region with reflector radii b_{1L} and $b_{2L} < d$, and one in the upper region with reflector radii b_{1U} and $b_{2U} > d$. Spot sizes, diffraction losses,

and frequency response of both systems are the same if the conditions

$$\frac{1}{b_{1L}} + \frac{1}{b_{1U}} = \frac{2}{d}$$

and (56)

$$\frac{1}{b_{2L}} + \frac{1}{b_{2U}} = \frac{2}{d}$$

hold. In this case the mode patterns of one resonator correspond to the complex conjugates of the mode patterns of the other.

Some of these conclusions can be drawn on the basis of (39), (40), and (46) alone. To prove the correspondence of diffraction losses and mode patterns, however, the integral equations of the resonator systems have to be investigated using a procedure suggested by T. Li for the analysis of nonconfocal resonators with equal reflectors. Examination of the integral equations shows that unstable systems also are equivalent under the conditions given above.

VII. SUMMARY

The effects of unequal reflector apertures on the modes of a confocal resonator have been discussed. It was found that unequal apertures at the two reflectors have a large effect in determining the mode patterns. The resonant condition, however, is not changed by an asymmetry of this kind. Boyd and Gordon's picture of surfaces of constant phase does not contain *nonconfocal* systems (of equal curvature) with similar asymmetries in spot size. This is because lossless nonconfocal resonators are not, except for special cases, degenerate in frequency. This shows that the mode patterns of nonconfocal systems are not significantly changed if the reflectors are of unequal size but are larger than the spot size at the reflectors.

For resonators formed of two reflectors of unequal curvature, unstable regions of high loss are shown to exist in that an equivalent sequence of lenses becomes defocusing. The true confocal resonator is on the border of such unstable regions, though in fact it has minimum diffraction losses. Unfortunate deviations from the dimensions of the ideal confocal resonator can produce a system of high loss. The implications for the design of resonators for gaseous or solid optical masers or of long distance optical transmission systems are that the "equal" radii of curvature should be made slightly larger (or smaller) than the reflector spacing. One can also choose to simulate a confocal resonator with one curved

and one flat reflector spaced at half the radius of curvature. This system is stable.

VIII. ACKNOWLEDGMENTS

Stimulating discussions with and constructive criticism from J. P. Gordon, W. W. Rigrod, A. G. Fox and T. Li are sincerely appreciated.

APPENDIX

Cylindrical Coordinates

In this paper and in the paper of Boyd and Gordon[1] the mathematical analyses were based on a system of Cartesian coordinates. Resonators with reflectors of square or rectangular aperture have been investigated. For large apertures the authors were able to obtain approximations for the properties of the resonator modes on which many arguments of this paper are based. They showed that the mode patterns are describable in terms of Hermite-Gaussian functions.

For certain classes of problems, for instance if it is desired to obtain diffraction losses for circular apertures, it is preferable to use a cylindrical system of coordinates. Approximate solutions for the modes of resonators with reflectors of large circular apertures can be obtained from the work of Goubau and Schwering.[3] The results of these authors are presented in terms of hybrid waves, but by suitably combining two hybrid waves one obtains modes which for our purposes can be regarded as linearly polarized TEM waves. Goubau and Schwering show that the mode patterns are describable in terms of associated Laguerre-Gaussian functions. Fox and Li[9] have given asymptotic solutions for the mode patterns in terms of Sonine's polynomials, which can be shown to be equivalent to the above results.

One can obtain asymptotic solutions for the modes of the confocal resonator in cylindrical coordinates by making a scalar wave approximation and using Huygens' principle. This leads to an integral equation for the modes of the confocal resonator with reflector spacing and curvature b in cylindrical coordinates, which has been given in Appendix C of Fox and Li.[2] For an infinitely large aperture their result may be written in the form

$$\chi\{\sqrt{t}\,R_{pl}(t)\} = \exp\left[-ikb + i\frac{\pi}{2}(l+1)\right] \cdot \int_0^\infty \sqrt{t'}R_{pl}(t')\cdot J_l(tt')\cdot\sqrt{tt'}\,dt', \tag{57}$$

where $t = r\sqrt{(k/b)}$ and r is the radial distance in the plane of the aperture. As the solution of the above equation the function $R_{pl}(t)$ describes the radial dependence of the modes and the angular dependence is $e^{+il\varphi}$. The resonance condition for the individual modes is obtained from the eigenvalue χ of (57).

From Magnus and Oberhettinger[10] one observes that the associated Laguerre-Gaussian function is self-reciprocal under the Hankel transformation. Thus the solution of (57) is given by

$$R_{pl}(t) \propto t^l L_p^l(t^2) \cdot e^{-t^2/2} \tag{58}$$

where $L_p^l(t)$ is the associated Laguerre polynomial. The associated eigenvalue is found to be

$$\chi = \exp\left[-ikb + i\frac{\pi}{2}(2p + l + 1)\right] \tag{59}$$

which leads to the resonance condition for the confocal resonator with q as the longitudinal mode number:

$$\frac{4b}{\lambda} = 2q + 2p + l + 1. \tag{60}$$

The field distribution of the modes inside and outside the confocal resonator can be derived from the mode patterns on the reflectors by using Huygens' principle, as in Boyd and Gordon's paper. The field distribution can of course be obtained from Goubau and Schwering's work. Comparing Goubau and Schwering's equation (5a) with Boyd and Gordon's equation (20), one finds that the surfaces of constant phase of a Cartesian TEM_{mnq} mode are identical with the surfaces of constant phase of a cylindrical TEM_{plq} mode if

$$m + n = 2p + l. \tag{61}$$

The fields and therefore the spot size of the fundamental TEM_{00q} Cartesian mode and the fields of the fundamental cylindrical mode are identical throughout the resonator.

The resonance conditions and spot sizes of the modes of resonators with large circular apertures can therefore be deduced for nonconfocal resonators and resonators with reflectors of unequal radii of curvature, in exactly the same fashion as has been done for square apertures.

We do not propose to present this derivation again, but list as a reference some characteristic properties of the modes of resonators with large circular apertures, together with properties of the modes of resonators with large square apertures. It may be worth repeating that an aperture is considered "large" for a particular mode if the mode's

energy, as calculated from the approximations below, is well concentrated within the aperture. Only under this condition are the mode's characteristics reasonably well described by the formulae listed below.

A.1 *Approximations for Resonators with Large Circular Apertures.*

A system of cylindrical coordinates (r, φ, z) is used, where the z-axis coincides with the resonator's optical axis. The corresponding modes are designated TEM_{pl_q}.

A.1.1 *Nonconfocal Resonators with Reflectors of Equal Radius of Curvature b' and a Reflector Spacing d.*

At the reflectors the spot size w_s' of the fundamental TEM_{00_q} mode is given by

$$w_s' = \sqrt{\frac{\lambda b'}{\pi}} \left(\frac{d}{2b' - d}\right)^{\frac{1}{2}}. \tag{62}$$

The relative field distribution (mode pattern) at the reflectors ($z = \pm d/2$) of a TEM_{pl_q} mode is given by

$$\frac{E(r, \varphi, \pm d/2)}{E_0} = \left(\frac{r}{w_s'}\sqrt{2}\right)^l \cdot L_p^{\,l}\left(2\frac{r^2}{w_s'^2}\right) \cdot e^{-r^2/w_s'^2} \cos l\varphi \tag{63}$$

where $L_p^{\,l}$ are the associated Laguerre polynomials. The mode resonates at a wavelength given by

$$\frac{2d}{\lambda} = q + \frac{1}{\pi}(2p + l + 1) \cos^{-1}\left(1 - \frac{d}{b'}\right). \tag{64}$$

A.1.2 *Resonators with Reflectors of Unequal Radii of Curvature b_1 and b_2 and a Reflector Spacing d.*

The spots of the fundamental mode are in general of different size on the two reflectors. We have a spot size w_1 on the reflector with radius of curvature b_1 and vice versa. In (39) and (40) these quantities have been expressed in terms of λ, d, b_1, and b_2. The TEM_{pl_q} mode patterns on the reflectors are obtained from equation (63) by substituting for w_s' the corresponding w_1 or w_2.

The resonance condition is

$$\frac{2d}{\lambda} = q + \frac{1}{\pi}(2p + l + 1) \cos^{-1} \sqrt{\left(1 - \frac{d}{b_1}\right)\left(1 - \frac{d}{b_2}\right)}. \tag{65}$$

A.2 *Approximations for Resonators with Large Square Apertures.*

A system of Cartesian coordinates (x, y, z) is used with the z-axis coinciding with the resonator axis. The corresponding modes are designated TEM_{mnq}.

A.2.1 *Nonconfocal Resonators of Equal Radius of Curvature b' and a Reflector Spacing d.*

The spot size w_s' of the fundamental mode at the reflectors is again given by (62). The mode pattern of a TEM_{mnq} mode at the reflectors is given by

$$\frac{E(x, y, \pm d/2)}{E_0} = H_m\left(\frac{x\sqrt{2}}{w_s'}\right) \cdot H_n\left(\frac{y\sqrt{2}}{w_s'}\right) \cdot e^{-(x^2+y^2)/w_s'^2} \qquad (66)$$

where the H_m are the Hermitian polynomials. The resonance condition for this mode is given by (47).

A.2.2 *Resonators with Reflectors of Unequal Radii of Curvature b_1 and b_2 and a Reflector Spacing d.*

The spot sizes w_1 and w_2 of the fundamental mode at the two reflectors are the same as those discussed in Section A.1.2. The mode patterns of the TEM_{mnq} mode at the corresponding reflectors are obtained by substituting w_1 or w_2 for w_s' in (66). The resonance condition is given by (46).

REFERENCES

1. Boyd, G. D., and Gordon, J. P., B.S.T.J., **40,** 1961, pp. 489–508; and *Advances in Quantum Electronics,* edited by J. R. Singer, Columbia University Press, 1961, pp. 318–327.
2. Fox, A. G., and Li, Tingye, B.S.T.J., **40,** 1961, pp. 453–488; and *Advances in Quantum Electronics,* edited by J. R. Singer, Columbia University Press, 1961, pp. 308–317.
3. Goubau, G., and Schwering, F., Trans. I.R.E., **AP-9,** 1961, p. 248.
4. Schwering, F., Archiv der Elektrischen Übertragung, **15,** 1961, pp. 555–564.
5. Slepian, D., and Pollak, H. O., B.S.T.J., **40,** 1961, pp. 43–63.
6. Pierce, J. R., Proc. National Academy of Sciences, **47,** 1961, pp. 1808–1813.
7. Pierce, J. R., *Theory and Design of Electron Beams,* D. Van Nostrand and Company, New York, 1954.
8. Jenkins, F. A., and White, H. E., *Fundamentals of Optics,* 3rd Ed., McGraw-Hill Book Company, New York, 1957.
9. Private communication.
10. Magnus, W., and Oberhettinger, F., *Function of Mathematical Physics,* Chelsea Publishing Company, New York, 1954 p. 137.

A Unidirectional Traveling-Wave Optical Maser*

By J. E. GEUSIC and H. E. D. SCOVIL

(Manuscript received March 20, 1962)

The basic ideas leading to a unidirectional traveling-wave optical maser are presented. Experimental data on the performance of pulsed ruby amplifying sections and high density PbO glass Faraday rotation isolators are given. Feasibility tests on a two-section device have been made and are in agreement with predictions. Some remarks are made concerning image definition, channel capacity, noise and pump power requirements.

I. INTRODUCTION

Net gain at optical frequencies was demonstrated by the successful operation of maser oscillators.[1,2] The first direct measurements of gain at optical frequencies were made by Javan, Bennett and Balik[3] for the helium-neon gas maser and by Kisliuk and Boyle[4] in a ruby solid state maser. The optical amplifiers described by these investigators were low-gain devices. Since at optical frequencies the maser is the only available amplifier which at present preserves amplitude and phase information, it may well have to provide extremely high stable gains of between 30–60 db in many applications, as for example in an optical communications system.

It is well known that to realize a stable high-gain amplifier it is essential to make the amplifier nonreciprocal. This has been accomplished in the case of the microwave traveling-wave maser[5] but has not previously been reported at optical frequencies. The objective of this paper is to discuss the basic principles which are necessary for realizing a nonreciprocal optical amplifier and to report the successful operation of a pulsed unidirectional traveling-wave optical maser (TWOM) using ruby.

Also the image amplifying ability of the TWOM is discussed and demonstrated.

* This work was supported in part by the U. S. Army Signal Corps under Contract DA-36-039-sc-87340.

II. BASIC PRINCIPLES

Consider a multimode transmission line containing active maser material as shown in Fig. 1. Over the length shown, assume that the single-pass power gain is G_0, that it is reciprocal, and that at each end the power reflection coefficient is r. If $G_0 r < 1$, the amplifier is stable. For a ruby optical maser with an air interface, $r \approx 0.07$; hence such a device will exhibit a stable gain if $G < 14$. If an $r \approx 0.07$ is accepted as typical (it may be made much smaller by the use of antireflection coatings), then it is evident that a simple optical maser with a gain of $\approx$6 db is a very stable amplifier with very little regeneration. On the other hand if a gain of 30 db is required, then r must be <0.001. This is extremely difficult to achieve, and it is evident that a stable gain of 60 db is essentially impossible from such a device.

The optical traveling-wave maser (TWOM) shown in Fig. 2 is a device capable of achieving extremely high stable gains. The device consists of a succession of amplifying sections, each of which has a gain of 6 db for the typical figures given above. These sections are separated by nonreciprocal elements or isolators so that power is easily transmitted in the direction of the arrows but strongly attenuated in the reverse direction. In fact the reverse loss in db of the isolator is chosen to exceed twice the single-pass gain in db of an amplifying section.

A simple way to realize isolation is to use the optical equivalent of a microwave Faraday rotation isolator. The main differences are the frequency, the material used, and the fact that our transmission line can support many modes. One may use numerous materials for the nonreciprocal rotator. Although transparent ferromagnetics, ferrimagnetics or anti-ferromagnetics may be considered, many other classes of materials could also be used; for instance it is not even necessary to use a material with a permanent magnetic dipole moment. Diamagnetics may be employed since they also exhibit Faraday rotation. In fact, nonreciprocal rotation was first observed by Faraday in a diamagnetic glass.

In the microwave Faraday rotation isolator the plane of polarization is determined by the rectangular waveguide. In the optical counterpart a polarizing medium can be used to fulfill the same function. In addition to having nonreciprocal rotation and defining the plane of polarization,

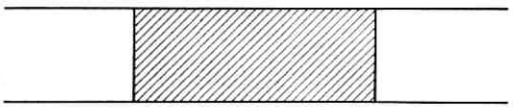


Fig. 1 — Transmission line containing an active maser material.

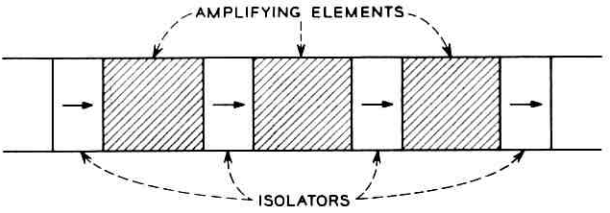


Fig. 2 — Schematic diagram of a unidirectional traveling-wave optical maser.

we need to be able to absorb waves with the unwanted polarization. In the microwave equivalent this is done by using a sheet of loss film placed so that its plane is parallel to the electric field of the wave which we wish to absorb. In the optical version this absorption may be combined in the polarizing medium itself if we use a material which has dichroic characteristics. The construction of an isolator is now straightforward and is shown in Fig. 3. In Fig. 3 a wave enters from the left-hand side with its plane of polarization defined by the dichroic polarizer as shown in the end view. The plane of polarization is rotated by 45° in the clockwise direction by the Faraday medium and passes through the right-hand polarizer. However, a wave entering the right-hand polarizer has its plane rotated 45° in the clockwise direction and hits the left-hand polarizer with its plane of polarization in the low transmittance or absorbing direction of the polarizer. The direction of the axially applied magnetic field with respect to the forward direction must be chosen in accordance with the sign of the Verdet constant of the Faraday medium used.

A TWOM may be built as shown in Fig. 4. In the figure the active medium is illuminated with pump power in one of several known ways. A two-section maser is shown; however, additional sections can be added.

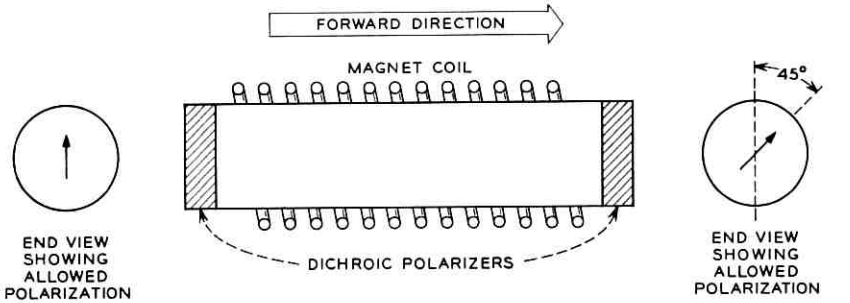


Fig. 3 — An optical Faraday rotation isolator.

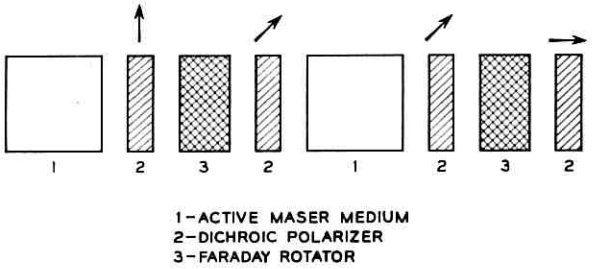


Fig. 4 — Elements in a two-stage TWOM.

For best operation, of course, all interfaces would have antireflection coatings. In the TWOM, alternate polarizers can be eliminated for simplicity.

An obvious application of the TWOM is as a power amplifier for an optical maser oscillator. Another application may be to obtain very high peak optical powers. It is well known that because of its energy storage the peak power of a maser may greatly exceed its average power. In fact some of the pulsed characteristics of a TWM have been analyzed.[6] The interaction of high optical power levels with matter will certainly allow many new and interesting phenomena to be studied. A focused beam with these peak intensities should have application to microwelding or micro-cutting of materials. Finally the TWOM may find application to amplify received signals in an optical relay system.

This concludes the discussion of the general ideas relating to a unidirectional traveling-wave optical maser. The concepts apply to either a CW amplifier or a pulsed amplifier. It now remains to discuss in more detail the theory and design of the individual amplifying sections and the optical isolators which have been used and tested in a pulsed ruby TWOM.

III. THE AMPLIFYING SECTION

Consider two levels 1 and 2 as shown in Fig. 5 with populations N_1 and N_2 and degeneracies g_1 and g_2 respectively. Also assume that the sample interacts at the transition $1 \rightleftarrows 2$ with a beam incident in a direction θ and φ having an energy density per unit frequency range per unit solid angle given by $\rho(\nu,\theta,\varphi,\mathbf{P})$. Here ν is the frequency of the radiation, $\mathbf{P}$ is a vector which defines an independent state of polarization of the radiation and θ and φ are the polar and azimuthal angles of a spherical coordinate system fixed in our sample. For this situation

Condon and Shortly[7] give the net rate of emission for transitions $1 \rightleftarrows 2$ as

$$\frac{d\rho}{dt}(\nu,\theta,\varphi,\mathbf{P}) = W(\theta,\varphi,\mathbf{P}) \left\{ \frac{\rho(\nu,\theta,\mathbf{P})v^3}{\nu^2} (g_1N_2 - g_2N_1) + h\nu g_1N_2 \right\} g(\nu) \quad (1)$$

where $W(\theta,\varphi,\mathbf{P})$ is the probability per unit time per unit solid angle of a spontaneous transition from $2 \rightarrow 1$ with emission in a direction θ,φ and polarization $\mathbf{P}$.

$g(\nu)$ is the normalized line shape for the transition.

N_1 and N_2 are the number of atoms per unit volume in levels 1 and 2, and

v is the velocity of wave propagation in the medium.

Equation (1) gives both the stimulated emission and the spontaneous emission per unit solid angle in a direction θ and φ. Equation (1) takes into account the anisotropy of the transition probability.

In order to find the gain or loss through the medium in a direction θ,φ at the $1 \rightleftarrows 2$ transition, the portion of (1) due to stimulated emission is integrated in the following manner,

$$\int_{\rho_0}^{\rho} \frac{d\rho(\nu,\theta,\varphi,\mathbf{P})}{\rho(\nu,\theta,\varphi,\mathbf{P})} = \int_0^{l(\theta,\varphi)/v} \frac{W(\theta,\varphi,\mathbf{P})v^3}{\nu^2} (g_1N_2 - g_2N_1)g(\nu)\, dt$$

where $\rho_0(\nu,\theta,\varphi,\mathbf{P})$ is the incident energy density and $l(\theta,\varphi)$ is the length of the medium in the direction considered. Hence the gain in db is given by

$$G_{\text{db}}(\nu,\theta,\varphi,\mathbf{P}) = (10 \log e) \left[W(\theta,\varphi,\mathbf{P}) \frac{\lambda_0^2 l(\theta,\varphi)}{\epsilon} (g_1N_2 - g_2N_1)g(\nu) \right]$$

where λ_0 is the free-space wavelength and ϵ the dielectric constant at

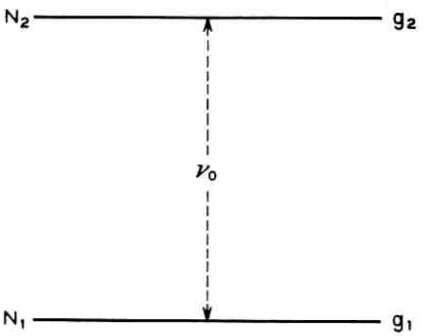


Fig. 5 — Two typical energy levels.

λ_0. For a Gaussian line shape, (approximately true for the ruby R_1 line at room temperature), the gain is given by

$$G_{\text{db}}(\nu,\theta,\varphi,\mathbf{P}) = (10 \log e)\left[W(\theta,\varphi,\mathbf{P})\frac{\lambda_0^2 l(\theta,\varphi)}{\epsilon}(g_1N_2 - g_2N_1)\right]\frac{2}{\Delta\nu} \cdot\sqrt{\frac{\ln 2}{\pi}}\exp -\frac{(\nu-\nu_0)^2}{(\Delta\nu)^2}4\ln 2 \tag{2}$$

where $\Delta\nu$ is the linewidth and ν_0 is the frequency at the center of the line. Using (2), the bandwidth B over which the gain is within 3 db of the peak gain is given by

$$B = B_M\sqrt{\frac{\log\dfrac{G_{\text{db max}}}{G_{\text{db max}}-3}}{\log 2}} \tag{3}$$

where $B_M = \Delta\nu$ (cf. equation (25) of Ref. 5). If the transition considered is isotropic (i.e. $W(\theta,\varphi,\mathbf{P}) = W$ = constant) then the gain given by (2) becomes

$$G_{\text{db}}(\nu,\theta,\varphi,\mathbf{P}) = (10 \log e)\left[\frac{\lambda_0^2 l(\theta,\varphi)}{8\pi\epsilon\tau}\right]\left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{2}{\Delta\nu}\sqrt{\frac{\ln 2}{\pi}} \cdot\exp -\frac{(\nu-\nu_0)^2}{(\Delta\nu)^2}4\ln 2$$

where τ is the spontaneous emission lifetime for the transition and is equal to $(8\pi g_1 W)^{-1}$. From (2) or this last expression it is observed that the medium exhibits gain if $(g_1N_2 - g_2N_1) > 0$ and loss if $(g_1N_2 - g_2N_1) < 0$.

In the case of ruby at room temperature, inversion at the R_1 line (14,400 cm^{-1}) is obtained by maser operation as shown in Fig 6. Here we have lumped the blue and green bands as one level; this is of course an obvious oversimplification. Also the ground state is treated as a single level with degeneracy $g_1 = 4$. At room temperature this is a valid assumption since the linewidth of the R_1 and R_2 transition is large compared to the separation of the two zero field levels in the ground state. At lower temperatures the linewidths of the R_1 and R_2 transitions are small, so that the two zero-field levels of the ground state are resolved; then it is necessary to consider the ground state as two levels each with degeneracy 2. For ruby the time spent in the pumping states is negligible compared with the normal lifetime of the two metastable states shown, and the efficiency for atoms getting to the metastable states upon pumping into the green or blue bands is near 100 per cent. Based on this,

and assuming that pumping power is supplied for times short compared to the normal or stimulated emission lifetimes of the two metastable states, we can write

$$\frac{dN_1}{dt} \approx -fPN_1 \tag{4}$$

where P is the incident pumping power and f is a pumping efficiency factor, which depends on the pump transition probability and the properties of the electromagnetic structure used in the pumping process. Since we assume that the pump power P is supplied for a short time, (4) can be integrated to obtain

$$\begin{aligned} N_1 &= N \exp -f \int_0^t P dt \\ &= Ne^{-fE} \end{aligned} \tag{5}$$

where N is the total number of Cr^{3+} ions/cc in the ruby sample and E is the total energy absorbed by the material at the pumping transition. Now since it was assumed that the time spent in the pumping levels is negligible compared to the spontaneous lifetimes of the metastable

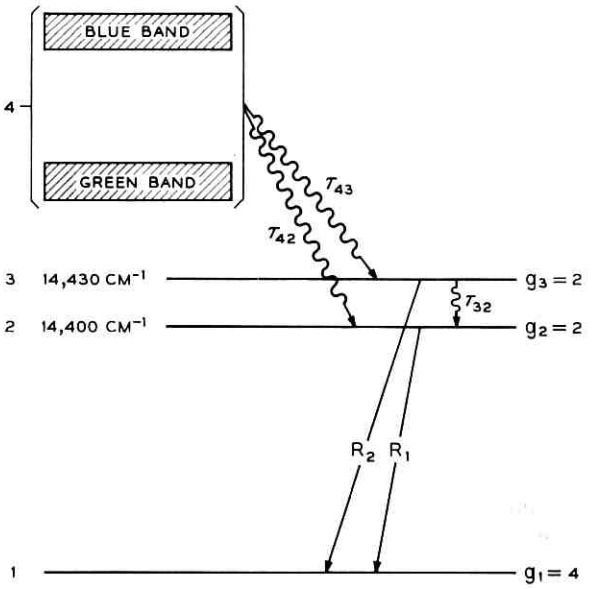


Fig. 6 — Optical energy levels of Cr^{3+} in Al_2O_3 .

states, the following relation holds,

$$N_1 + N_2 + N_3 \approx N \tag{6}$$

and $N_3 = \gamma N_2$ where γ is determined by τ_{42}, τ_{43} and τ_{23}. Using (2), the power gain G of the ruby material at the center of the ruby R_1 line when inverted is

$$G_{\text{db}} = A\left[N_2 - \frac{g_2}{g_1} N_1\right] \tag{7}$$

where

$$A = (10 \log e)\left[g_1 W(\theta,\varphi,\mathbf{P}) \frac{\lambda_0^2 l(\theta,\varphi)}{\epsilon}\right] \frac{2}{\Delta\nu} \sqrt{\frac{\ln 2}{\pi}}.$$

Upon substitution of (5) and (6) into (7) we obtain

$$\frac{G_{\text{db}}}{(L_{\text{db}})_{\max}} = \frac{1}{\frac{g_2}{g_1}(1+\gamma)}\left[1 - \left\{1 + \frac{g_2}{g_1}(1+\gamma)\right\} e^{-fE}\right] \tag{8}$$

where $(L_{\text{db}})_{\max}$ is equal numerically to the loss in db at the center of the line if all ions are in the ground state. To a good approximation this is equal to the loss through the unpumped material for a temperature T such that $kT \ll h\nu_{R_1}$, which is the case for the ruby R_1 line for temperatures 300°K and lower. Since (8) is derived for the case where the gain is measured in a time T_{meas} after the pump power has been supplied, which is short compared to τ_{21} and τ_{31} but long compared to τ_{43} and τ_{42}, the value of γ can be simply given for two limiting cases:

(a) $\tau_{23} < \tau_{42}$ and $\tau_{43} < T_{\text{meas}}$

$$\gamma = \exp h(\nu_{R_1} - \nu_{R_2})/KT.$$

At room temperature for ruby $\gamma = 0.865$

(b) τ_{42} and $\tau_{43} < T_{\text{meas}} < \tau_{23}$

$$\gamma = \frac{\tau_{42}}{\tau_{43}}.$$

At present no accurate data are available for ruby on the times τ_{23}, τ_{42} and τ_{43}; however, the measurements of Wieder[8] give upper limits for these times which indicate that case (a) is satisfied at room temperature. Also the gain measurements to be presented are consistent with the assumption $\gamma = 0.865$.

To determine what values of fE could be achieved and what gains

could be obtained for a pulsed ruby amplifying stage, gain measurements were carried out on an amplifying test section at room temperature. This test section consisted of a 3-inch long, 0.250-inch diameter, c-axis oriented (0.065 per cent by weight of Cr_2O_3 in Al_2O_3) ruby rod and a G.E. FT-91 Xenon flash tube located at the foci of a 3-inch long elliptical cylinder, as shown in Fig. 7. The gains were measured as is shown schematically in Fig. 8. The beam from a ruby oscillator was used as the signal source, and the gain or loss through the amplifying section was determined by measuring the ratio of the signal at the photomultiplier with and without the amplifier in the oscillator path. The measured loss through the unpumped amplifier at the peak of the R_1 line was 12 db. Because the ends of the ruby were not coated with antireflection layers, the measured gains had to be corrected to obtain the single-pass gain.

The expression for the numerical gain G of a reciprocal amplifier with feedback is given by

$$G = \frac{(1 - r)^2 G_0}{1 + r^2 G_0^2 - 2rG_0 \cos \varphi} \tag{9}$$

where r is the power reflection coefficient at each end and φ is the relative phase shift of a wave making one complete, round-trip traversal of the amplifier. Since the length of the ruby amplifier is not known to an

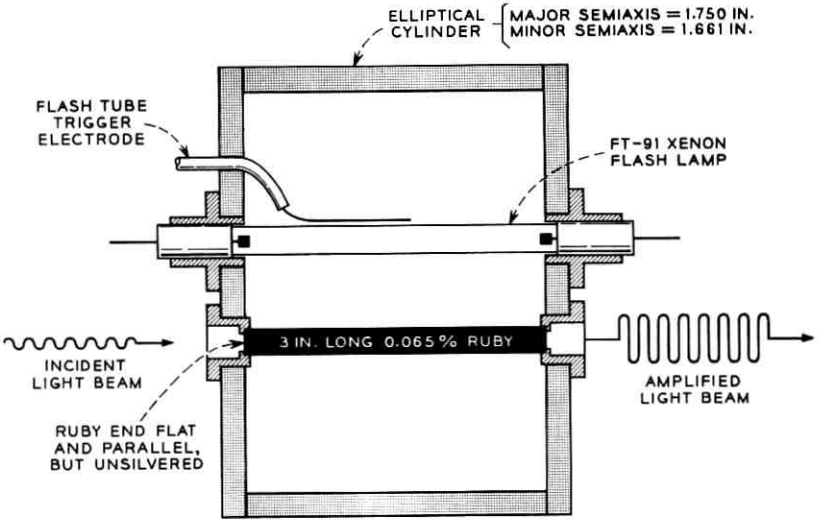


Fig. 7 — Ruby optical amplifier section.

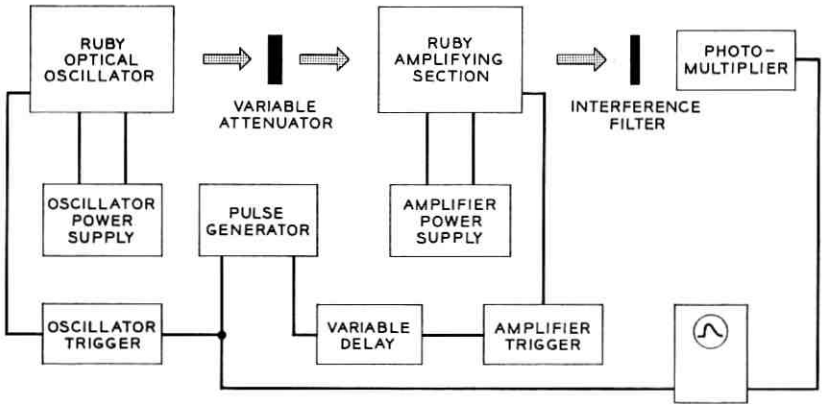


Fig. 8 — Schematic description of optical gain measurements.

accuracy better than a wavelength, the value of φ to be used in (9) to obtain the single-pass gain G_0 is not known. In the gain measurements an interference filter was used which had a passband larger than the ruby amplifying linewidth. Further, the measured bandwidth of a ruby oscillator is 1-2 kmc, whereas the separation of Fabry-Perot modes in our amplifying rod is

$$\Delta\nu \approx \frac{c}{2\sqrt{\epsilon}\, l} \approx 0.83 \text{ kmc.}$$

Hence it is reasonable to correct the data by assuming the proper gain formula to use in determining G_0 is one where G has been averaged over all possible values of φ. For the case where G is averaged we obtain

$$\bar{G} = \frac{(1 - r)^2 G_0}{(1 - G_0^2 r^2)}. \tag{10}$$

The experimental data corrected in this manner for the 3-inch long amplifying section are plotted in Fig. 9 along with the theoretical curve given by (8) for $\alpha = 0.865$, $g_1 = 4$ and $g_2 = 2$. Also plotted are data obtained on a 1-inch long ruby where the measured gains were smaller and the regeneration corrections needed to obtain the single-pass gain were less important. It is seen from Fig. 9 that our measured gain variation versus the input pumping parameter fE is in good agreement with the theory developed.

It remains to explain how the correlation between the values of fE on the abscissa of Fig. 9 and the measured light intensity at the green

and blue bands was obtained. For this purpose, the relative intensity at the green and blue bands versus electrical energy input to the flash lamp was measured photoelectrically. The absolute relation of this relative intensity to fE was then determined by observing at what value of light intensity unity gain (0 db) was achieved with the amplifier in place.

The maximum measured net gains for the amplifiers tested depended rather critically on the FT-91 tubes used; however, by selecting FT-91 flash tubes and operating them at input energies of 250 joules (which is approximately twice their rating) net gains of 10 db and 6 db were obtained respectively for two different amplifiers. All the measurements were made at room temperature. Higher gains per unit length could have easily been achieved by cooling, for it is known from the measurements of Schawlow and Devlin[9] that the linewidth of the ruby R_1 line

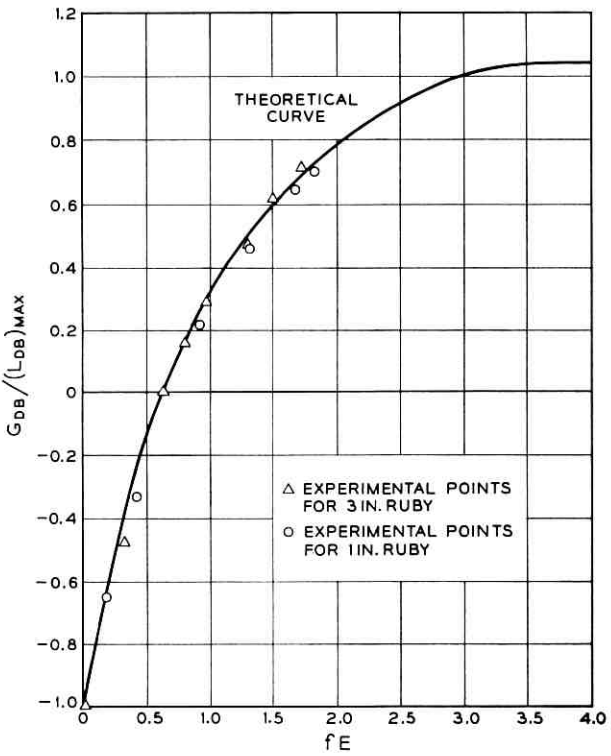


Fig. 9 — Gain versus input pump energy curve for a pulsed ruby maser at room temperature. $(L_{db})_{max}$ is 12 db and 4 db for the 3-inch and 1-inch ruby, respectively.

decreases quite rapidly with decreasing temperature between 300 and 80°K.

IV. THE OPTICAL ISOLATOR

To complete the discussion of the individual elements in the TWOM, it remains to discuss the basic principles involved in the design of an optical isolator and to report some performance data.

It is well known that the plane of polarization of a light beam, when passing through a Faraday medium, will be rotated by a magnetic field H applied to the material in a direction parallel to the direction of light propagation. The angular rotation θ of the plane of polarization is related to the strength of the magnetic field and the path length l in the medium by the expression

$$\theta = V \cdot H \cdot l \tag{11}$$

where the angle θ is chosen in the positive screw sense along the applied magnetic field. The constant of proportionality V is known as the Verdet constant and is a function of the material and the wavelength of light used. The Verdet constant is related to the material properties in the following manner

$$V = \pi \nu_0 (N_+ - N_-)/Hc$$

where N_+ and N_- are the refractive indices for right- and left-handed circularly polarized light of frequency ν_0. In the case of ferromagnetics and antiferromagnetics, large Verdet constants have been measured. The large Verdet constant measured in each of these materials is due to the fact that the ions which produce the rotation see, in addition to the applied field, an internal field which is of the order of magnitude of the exchange field. Ferromagnetics would then be especially attractive. For the visible region of the optical spectrum, however, for ferromagnetics known to the authors, this large rotation is accompanied by large attenuation per unit length of material. This is because the absorption bands which give rise to the rotation in the ferromagnetic material usually extend from the ultraviolet well into the visible. However, ferromagnetic materials should be useful rotators in the infrared — i.e., beyond about 1μ.

The problem of building a good isolator in the visible (particularly at the R_1 line) depends upon finding a material which has a Verdet constant large enough to produce 45° rotation with reasonable fields and lengths, and which has low attenuation in the length of material which must

be used. For the ruby R_1 line at 14,400 cm^{-1}, known ferromagnetics are at present unsatisfactory. However, two diamagnetic materials appear to be good choices. One of these, ZnS(β), has a Verdet constant of 0.22 minute/cm/gauss at the R_1 line in ruby and is essentially transparent at this wavelength. For ZnS(β) a rotation of 45° requires a magnetic field of 3100 gauss for a 4-cm length. The second material, high-density PbO glass, also is attractive since it has a Verdet constant of 0.09 minute/cm/gauss at the R_1 line and an attenuation of 0.08 db/cm. Because of the commercial availability of the PbO glass from Corning Glass Works, this material was chosen for use in the optical isolator. In both the ZnS(β) and the PbO glass it is the Zn^{++} and Pb^{++} ions which are probably responsible for the rotation. In the PbO glass the absorption band which produces the rotation rises sharply at 4000 A, and extends to shorter wavelengths; this band is probably composed of the $^1S_0$ to $^3P_0$, $^3P_1$, $^3P_2$ transitions in the free-ion notation. Since the configuration which gives rise to $^3P_0$, $^3P_1$ and $^3P_2$ states in the free ion contains bonding electrons in the solid state, the transition observed beyond 4000 A should more properly be referred to as a charge transfer band. This tentative assignment seems supported by measurements made on Pb^{++} in CaO by Ewles as discussed by McClure.[10] The Verdet constant and the attenuation per unit length have been measured for Corning #8363 high PbO content glass as a function of wavelength and are shown in Fig. 10. The data in Fig. 10 indicate that PbO glass is a useful Faraday rotator for optical isolators in the wavelength range 5000–7000 A.

To define the plane of polarization in the isolator and to provide reverse loss, as was discussed earlier, dichroic polarizers such as Tourmaline and Polaroid sheet or crystal polarizers such as Nicol prisms and Glan-Thomson prisms can be used. At the R_1 line, type HN-38 Polaroid sheet is a good choice, having a transmittance of 0.86 for the parallel orientation and a transmittance of approximately 0.01 for the perpendicular orientation.

An optical isolator was constructed in the manner as depicted in Fig. 3. A 4-inch long, $\frac{1}{2}$-inch diameter PbO glass rod was used as the rotator, and Type HN-38 Polaroid polarizers were used to define the plane of polarization and to provide reverse loss. A water-cooled solenoid provided the necessary field to produce 45° rotation. The reverse loss of the isolator was measured at the R_1 (14,400 cm^{-1}) line and was found to be approximately 16 db. The measured insertion loss of the isolator was 3 db. In this isolator, optimized operation has not been realized; however, measurements on the individual components were made to determine the best performance that can be achieved with a more carefully con-

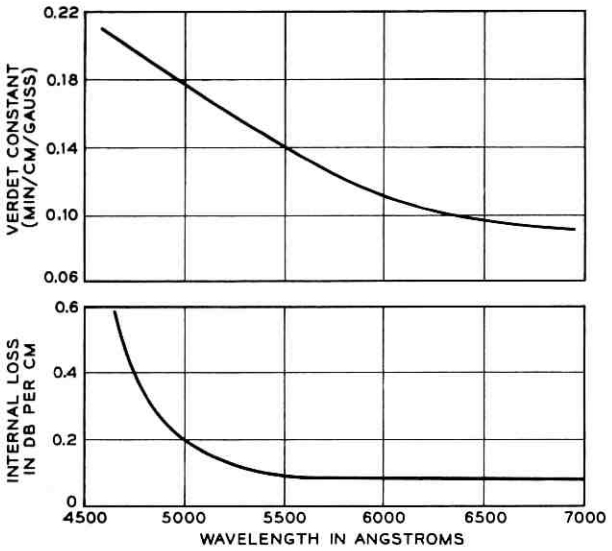


Fig. 10 — Variation of the Verdet constant and the internal loss of PbO glass versus wavelength.

structed isolator. These measurements on the individual components indicate that an isolator with an insertion loss of 1 db and a reverse loss of 15 to 17 db could be realized if antireflection coatings are used on all isolator surfaces.

V. OPTICAL TRAVELING-WAVE MASER PERFORMANCE

To test the feasibility of a high-gain (30–60 db) optical amplifier a test TWOM was assembled, consisting of two amplifiers separated by an isolator section. The two individual amplifiers had net gains of 10 db and 6 db respectively, and the isolator had an insertion loss of 3 db. A net gain of 13 db was expected. The measured net gain was 12.2 db; the discrepancy of 0.8 was probably due to the lack of perfect alignment. The oscillogram presentation of the gain of the amplifier is shown in Fig. 11. The top trace is the signal due to the oscillator alone as observed with the photomultiplier, with the TWOM out of the beam path. The middle trace shows the signal after amplification by the first amplifying stage with the isolator and second amplifier out of the beam path. Taking into account the changes made in the gain of the oscilloscope, the second trace indicates a net gain of 10 db through the first amplifier. The bottom trace shows the signal after passing through the

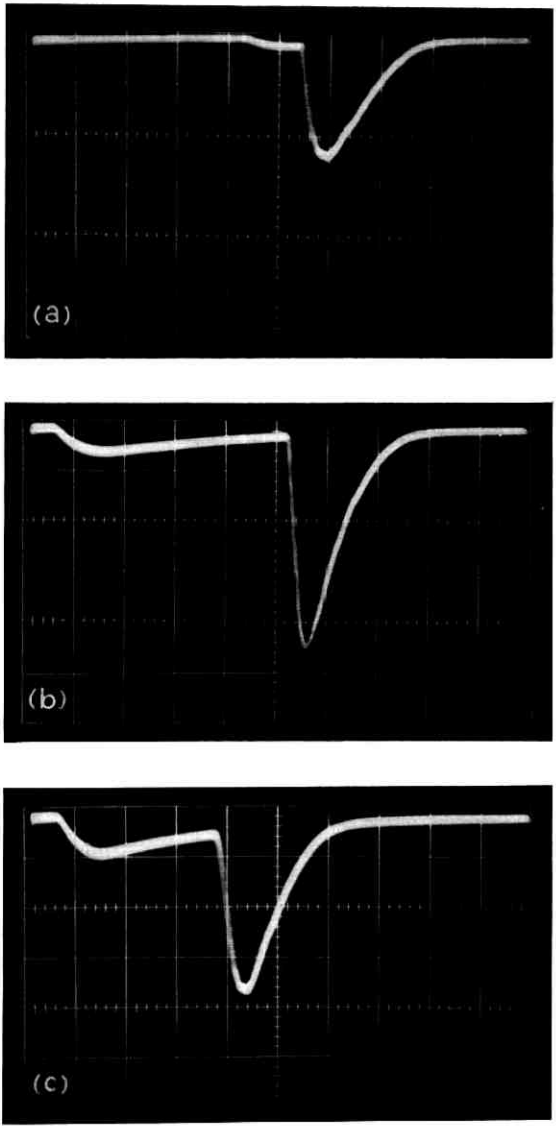


Fig. 11 — Oscilloscope photographs showing gain in the TWOM: (a) upper trace shows the signal at the photomultiplier due to the oscillator alone with the TWOM removed from the beam. Scope gain is 2 volts/cm. (b) Middle trace is the received photomultiplier signal with the first amplifier in the beam. Scope gain is 10 volts/cm. (c) The lower trace shows the oscillator signal as amplified by the entire TWOM. Scope gain is 10 volts/cm with 3 db of optical attenuation added. The sweep time is 100 microseconds/cm in all three traces.

entire TWOM. For this measurement 3 db of optical attenuation was introduced, and if this is taken into account the net gain of the TWOM was found to be 12.2 db.

In making all the gain measurements reported, a time constant was used so that individual oscillator spikes were not observed. A measurement of the gain with a shorter time constant did not contradict the measured peak gains observed (over the oscillator duration) with the longer time constant and also did not reveal any new spikes that were not present in the oscillator. In all the measurements the firing of the oscillator flash tube was delayed with respect to the firing of the amplifier pumping tubes; this was done so that the amplifier could build up to full gain before the probing signal was sent through the amplifier. In Fig. 11 this delay was approximately 400 microseconds. There was about a ± 50-microsecond jitter in firing the oscillator, and this accounts for the difference in position of the signal in the three traces shown in Fig. 11. The total physical length of the test TWOM was approximately 20 inches, with the isolator solenoid presently being the most space-consuming element. A shorter physical length is possible with a more careful and sophisticated design.

Inasmuch as the diameter of the ruby rods used in the amplifiers is 0.250 inch, the TWOM that has been described is a multimode amplifier capable of handling approximately 10^8 spatial modes. The fact that it can support 10^8 spatial modes suggests that image information can be sent through the amplifier. That is to say, the TWOM can be considered to be a limited aperture, infinite focal length lens with gain. To show that an image could be sent through a TWOM and amplified, a projection slide having on it a number of dark lines was placed before the input of the TWOM (in this case the second amplifier was removed for simplicity) and illuminated with the oscillator beam. The slide was then viewed with a lens and a camera as shown in Fig. 12. In Fig. 13 are shown

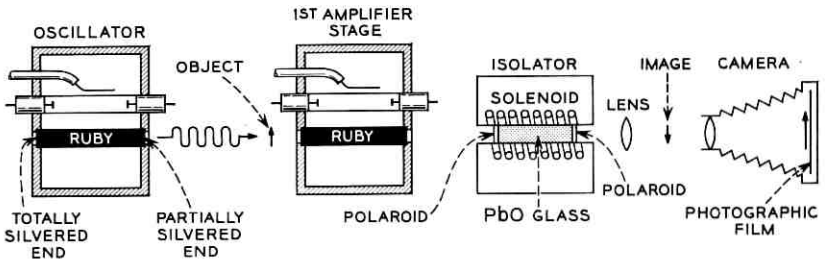


Fig. 12 — Schematic description of the image amplification experiment.

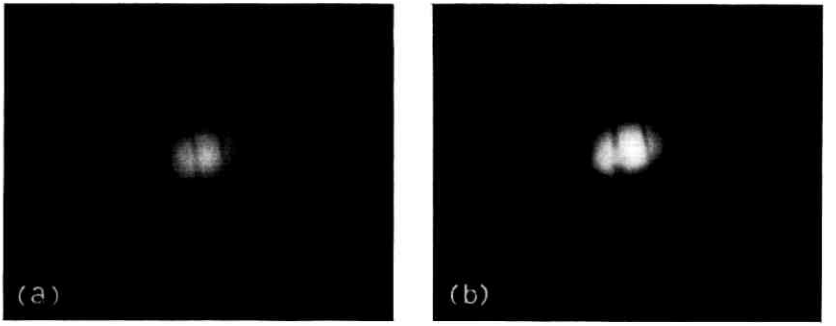


Fig. 13 — Photographs showing image amplification in the TWOM: (a) This photograph shows the observed image of two lines on the object slide as observed with unity gain. (b) This photograph shows the image when observed with the TWOM at a gain of approximately 5 db.

two pictures of this slide as taken through the TWOM. Only two bars on the slide were illuminated by the oscillator. The left photograph was made with the amplifier operated at unity gain, and the right photograph was made with an amplifier gain of approximately 5 db. It is seen that there is greater contrast in the amplified image than in the unamplified image. A quantitative measurement of the photographic negative with a densitometer showed a gain of approximately 4 db in the amplified image over the unamplified image. Although admittedly the observed image amplification is quite crude, it is nevertheless felt that the image amplification properties of a TWOM have been shown and that this property of the TWOM may be of importance in a communication system.

VI. IMAGE DEFINITION, CHANNEL CAPACITY AND NOISE

The capacity of a communication system depends upon the total number of unit phase-space cells which the system can handle. The usual communication system has a spatial width of only one cell (it is single mode), and information is conveyed only in the longitudinal direction — i.e., the frequency and time domain. At optical frequencies however, it may be convenient to use many cells in transverse space. A discussion is therefore given of the channel capacity of a multimode image-amplifying maser.

We define channel capacity as the product of the number of cells in transverse space and the bandwidth. The latter has been discussed earlier; it remains to consider the former.

Evidently if we are dealing with an infinitely long cylindrical amplifier

with perfectly reflecting walls — i.e., a multimode waveguide with gain — then the number of modes for a unidirectional, single-polarization amplifier is $\pi A/\lambda^2$, where A is the cross-sectional area and the channel capacity C is

$$C = \frac{\pi A}{\lambda^2} B$$

where the bandwidth B is given by (3) for a Gaussian line shape.

In practice, the amplifier is not infinitely long, and in fact the device described here has a length $< (A/\lambda)$. Under these conditions one is concerned with the near field region of the amplifier aperture, and image transmission is possible (i.e., some modes can be transmitted from input to output independently of the boundary conditions).

Consider an idealized, unidirectional, single-polarization, cylindrical amplifier as depicted in Fig. 14 of length l and cross sectional area A.

We consider first the image space. A transmitted image can occupy in transverse (i.e., x,y,P_x,P_y) space an area equal to $A\Omega$ where Ω is the angular aperture A/l^2 and $P = (h\nu/c)$ is the photon momentum. The aperture A and its associated diffraction solid angle $\omega = \lambda^2/A$ define the transverse dimensions of a unit cell.[11] Consequently the total number of image cells available is

$$\frac{\Omega}{\omega} B\Delta t = \frac{A^2}{l^2\lambda^2} B\Delta t \tag{12}$$

where Δt is the time over which the observation is made. Equation (12) may be interpreted to mean: (*i*) that Ω/ω identifiable image points can be amplified (this also follows from classical diffraction considerations) and (*ii*) that the image channel capacity $\Delta C = (\Omega/\omega)B$. The device whose operation has been described here has a $\Delta C \approx 3 \times 10^{15}$ cps.

One can classify cells into groups according to the behavior of their corresponding rays of geometric optics. Those used for transmitting an image consist of $A^2/l^2\lambda^2$ rays which pass directly from the input to the

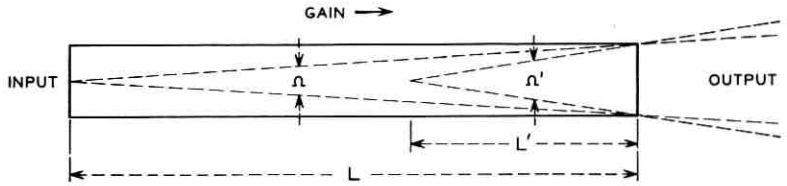


Fig. 14 — An idealized image amplifier.

output. The next group consists of $3(A/l^2\lambda^2)$ rays which undergo a single internal reflection at the wall before reaching the output and fall within the solid angle 4Ω minus the image core Ω. Successive groups undergo an increasing number of internal reflections. The reason for the A^2 dependence in (12) is now clear — the area and consequently the total number of cells which the transmission line can support is proportional to A, and out of this total number, that fraction which can pass directly from input to output is also proportional to A. If the length of the amplifier is greater than A/λ, then all rays are reflected from the walls and, in this limit, one is concerned with waveguide propagation.

In principle, and to a certain extent in practice, the "off axis" rays are available for transmitting information. The internal reflections will result in image distortion, but cylindrical symmetry is preserved and spatial coding can be used. At the present time, however, the channel capacity using only the image core is embarrassingly large and there seems little advantage in using the "off axis" rays.

The limiting noise of a maser (or for that matter any coherent amplifier) corresponds to one noise photon per unit cell of phase space, (noise number[12] $N = 1$). Optical masers can come very close to this limit and in fact the ruby amplifier described in this paper has a calculated $N \approx 2$. The number of noise photons associated with the signal is consequently $N\ \Delta C$.

In the frequency domain it is well known that there is an optimum bandwidth for a given signal and that excess bandwidth gives excess noise. Similarly in the transverse domain there is an optimum spatial bandwidth; in particular the angular aperture Ω used should not exceed that required for the necessary definition, as this would add excess noise. If the amplifier has excess angular aperture, the optics should be arranged so that the transmitted image and the following receiver occupy only the necessary solid angle.

In practice, additional noise is produced at the output, and proper precautions should be taken to eliminate this excess noise. Consider first only the noise associated with image rays. A cross section along the amplifier, shown dotted in Fig. 14, may be considered as the input of the remaining part of the amplifier with a new angular aperture $\Omega' = A^2/l'$ which is larger than Ω. We have therefore an amplifier whose bandwidth (spatial) increases (to a limit $\Omega' = 2\pi$) as we travel along the device. Excess noise will appear at the output in this larger solid angle but will of course see less gain than that in the useful angle Ω. A detector, insensitive to angle, placed at the output will see this noise, which can be severe in some designs. This excess noise, which is not intrinsic to the

signal band, can perhaps most easily be eliminated by focussing the input plane on the detector and using an aperture stop which passes only the focused image.

Even though one may not intend to use the "off axis" rays they may, unless special precautions are taken, be available at the output. Commonly used solid-state maser materials have large indices of refraction and, consequently, internal reflections may occur for large angles of incidence (for ruby the critical angle is $\approx 60°$). It is evident, therefore, that an appreciable fraction of the spontaneously emitted photons can be amplified and reach the output. These can be separated from those in the signal image core by the method mentioned above.

One should differentiate between presently available image intensifiers and maser image amplifiers. Unlike the maser, the former are not coherent amplifiers and are perhaps better described as image quantum detectors. The quantum detector preserves the intensity of the signal but not its phase (and in practice it is difficult to preserve the momentum). The maser preserves both amplitude and phase and easily preserves the frequency and direction of the signal. It is a direct consequence of uncertainty that a coherent amplifier has a noise number $N \geqq 1$. The detector which preserves only one variable can in principle be noise free, although this does not mean that it is more "sensitive" than a maser in some systems applications.[13]

At wavelengths of $< 1\mu$ the black body temperature of an object must be $> 10^4$°K before the phase-space density can exceed unity. As a consequence, the phase of the radiation is indeterminate for objects illuminated by usual incoherent light sources. There appears little object in using a coherent amplifier with such an image, and in fact its noise can be a serious disadvantage.

If an object is illuminated by a maser oscillator, however, then densities $\gg 1$ can be obtained, and in fact densities > 1 are necessary in a communication system in order to make full use of channel capacity. In such systems a coherent image amplifier which can also easily preserve the frequency, phase and direction of a signal may be of value.

VII. PUMP POWER REQUIREMENTS

One of the principal practical difficulties associated with optical masers is that of providing the pump. Minimum pump power and minimum pump color temperature are mutually exclusive, and in practice a compromise must be reached. In this section we are concerned with the effects of signal circuit design on the pump requirement of a CW

amplifier. It is clear that with the use of materials whose quantum efficiency is near unity (ruby has a Q.E. ≈ 0.9) pump photons are required only to replace those photons dissipated at the signal transition. In practice, signal circuit design can affect the pump requirements by several orders of magnitude.

The total number of photons emitted at the signal transition from an uncertainty-limited ($N = 1$) maser amplifier is

$$N_T = \sum \left(\frac{c^3}{8\pi\nu^2} \frac{G_0 - 1}{\ln G_0} \frac{n}{\tau} + G_0 D \right). \tag{13}$$

Here G_0 is the numerical gain for a given cell and D the dynamic range (equal to the maximum number of photons per cell in the input signal). n is the density of particles per unit volume per spectral interval, and from the previous equations n/τ is related to G_0. The first term is due to spontaneous emission and amplified spontaneous emission, and the second is due to the amplified signal. Equation (13) may conveniently be regrouped as

$$N_T = \sum_{\substack{\text{All Signal} \\ \text{Cells } \Delta C}} \left(\frac{c^3}{8\pi\nu^2} \frac{G_0 - 1}{\ln G_0} \frac{n}{\tau} + G_0 D \right) + \sum_{\text{The Remainder}} \frac{c^3}{8\pi\nu^2} \frac{G_0 - 1}{\ln G_0} \frac{n}{\tau}. \tag{14}$$

The first term is determined entirely by the specifications placed on the amplifier; i.e., it has a channel capacity ΔC, a gain G_0, and a dynamic range D over this band. The first part accounts for the noise from an uncertainty-limited amplifier. The second part is due to the signal. When $G_0 \& D \gg 1$, the first term becomes $G_0 \Delta C D$. The second term is the sum, over all other cells into which the particles can radiate, of the spontaneously emitted and amplified spontaneously emitted photons. This term, which is concerned with cells outside the signal band, represents wasted energy. If ΔC contains only a small fraction of the total number of cells, this term may be approximated as $K(N/\tau)$ where N is the total number of particles at the signal transition and K is the average of $(G_0 - 1)/(\ln G_0)$. Evidently if the gain for these excess cells is unity, then the second term is just N/τ, the total spontaneous emission.

In a practical amplifier it seems likely that the specifications and the design will be such that (14) will be represented to a good approximation by

$$N_T = G_0 \Delta C D + K N/\tau. \tag{15}$$

Certainly it is likely that a $G_0 \& D \gg 1$ will be required and hence the

first part is a good approximation. If $G_0 \& D \gg 1$, then an insignificant amount of pump power will be saved by designing so that ΔC occupies an appreciable fraction of the total phase space and further, since such a design would appreciably increase the pump color temperature requirements, it seems likely that the fraction will be kept small. Under these circumstances, the second part is also a good approximation. A signal circuit efficiency η_s may then be defined as

$$\eta_s = G_0 \Delta C D / (G_0 \Delta C D + K N/\tau). \tag{16}$$

The preceding discussion is now applied to a single-mode transmission system — i.e., one where the information is to be sent only in the longitudinal domain. An obvious design for an amplifier in this application might be a single-mode, single-polarization waveguide TWM using fibre optics. A material with a quantum efficiency near unity, an inversion near 100 per cent, and a signal linewidth just sufficient to give the required bandwidth might be used. As far as the material and signal structure is concerned, such a design has 100 per cent photon efficiency; i.e., every photon absorbed from the pump is available for amplification at the signal frequency and in the signal mode. (To be strictly correct, the efficiency is 25 per cent if it is a single-polarization, unidirectional amplifier.) The ratio of stimulated to spontaneous emission is $\approx [(G_0 - 1)/\ln G_0]$ for high gain in the absence of a signal. If the requirement was for a gain of 30 db then this ratio is $\approx$20 db. Further since one-half of the stimulated emission photons will be produced in the last 3 db (10 per cent of the length) of the amplifier, the rate at which photons are produced at the output section is $\approx 10^3$ times that from spontaneous emission alone. This means that the idler relaxation time must be $< 10^{-3}$ of the spontaneous emission lifetime and that the pump density must be 10^3 greater than that required to produce the same inversion in the absence of gain. So far we have considered only the contribution from amplified noise. If in addition we require that the device be able to amplify a signal $D \times$ noise, the above figures of 10^{-3} and 10^3 become 10^{-5} and 10^5, respectively, for $D = 20$ db. These perhaps impossible requirements on the idler relaxation time and pump color temperature have arisen because the device is too efficient.

A less efficient alternative design is now considered for the above application. Consider an image amplifier of the type depicted in Fig. 14. Since we are concerned with only one transverse mode, we need the definition for only one image point — i.e., $l \approx A/\lambda$.

Equation (15) is a good approximation in this case. The first term, which represents essential energy, is

$$G_0 \Delta CD = G_0 D \Delta\nu \sqrt{\frac{\log \frac{G_{\text{db max}}}{G_{\text{db max}} - 3}}{\log 2}}.$$

The second term, which represents wasted energy, is given by

$$K \frac{N}{\tau} = K \frac{n}{\tau} \times \text{volume} = K l^2 \lambda \frac{n}{\tau}.$$

Also, if W is isotropic and we have complete inversion, (2) gives

$$l = \frac{G_{\text{db}} \epsilon \Delta\nu}{0.16\lambda^2} \frac{g_2}{g_1} \frac{\tau}{n}$$

and hence

$$K \frac{N}{\tau} = K \times \frac{G_{\text{db}}^2 \epsilon^2 \Delta\nu^2}{0.0256\lambda^3} \left(\frac{g_2}{g_1}\right)^2 \frac{\tau}{n}.$$

As in the last example we assume the amplifier is required to have a gain of 30 db and a dynamic range of 20 db. If 0.05 per cent ruby is used at room temperature, if complete inversion is obtained and if further we assume that, contrary to the device described in this paper, the isolators use a negligible fraction of the total length, then

$$\Delta C \approx 10^{11} \text{ cps}$$

giving

$$G_0 \Delta CD \approx 10^{16} \text{ photons per sec}$$

also

$$l \approx 16 \text{ cm}$$

giving

$$K \frac{N}{\tau} \approx 0.85 \times 10^{20} K \text{ photons per sec.}$$

From (16) the structure efficiency is

$$\eta_s \approx \frac{1}{K} 1.2 \times 10^{-4}.$$

Since the signal is capable of interacting with almost all of the spins, the signal photon density has negligible effect upon the required pump color temperature with this value for η_s. It is essential to keep K small since it has a direct effect on the required color temperature as well as

on the required pump power. The index of refraction of most solid-state maser materials is so large that unless special precautions are taken to prevent internal reflections, a large fraction of the off-axis rays will see the full amplifier gain. This could easily lead to a value for $K \approx 100$ for a 30-db amplifier with a corresponding two order-of-magnitude increase in the required pump power and energy density. An estimate indicates that it should be possible to keep $K < 2$ with correspondingly little effect on the power and density of the pump.

The two amplifiers just described represent design extremes. The single-mode waveguide amplifier has a minimum pump power but an excessively high pump density requirement (in this example 5 orders of magnitude above that for spontaneous emission alone). Conversely the single-cell image amplifier has a minimum pump density but an excessive pump power requirement (in this example 4 orders of magnitude above that intrinsic to the specifications). A reasonable compromise objective might be for $\eta_s \approx 50$ per cent, in which case both the power and density requirements would both be increased only by a factor of 2 from the minimum.

The single-cell image amplifier just described is very inefficient because only a small fraction of the total phase space is available to the useful image core. A somewhat different geometry can improve this. As an example, one could break up the amplifier into say 10 sections, each with a gain of 3 db. Each section is now capable of amplifying one image point if $l/10 \approx A/\lambda$; i.e., A and hence $K(N/\tau)$ can be reduced a factor of 10 with a corresponding increase in η_s. Of course with this decrease in A, the diffraction angle is larger and unity magnification lenses must be placed between sections to reform the image at the input plane of the next section. The extent to which this subdivision can be continued is in practice limited by the fact that electronic gain per section decreases with the number of sections while the circuit loss per section will be substantially constant, leading to an eventual decrease in the net gain of the entire device.

This section has been concerned primarily with amplifiers for use in a single-mode transmission system, and the information has been conveyed in the longitudinal domain. It has been shown that pump density and power can vary over wide ranges by signal circuit design. In principle, analogous arguments apply to an image amplifier where information is to be conveyed primarily in the transverse domain: in practice, the problem is primarily one of increasing the efficiency and would appear to depend largely on obtaining narrow linewidth materials. Narrow-frequency band filters are probably also necessary to obtain high effi-

ciency. It is perhaps worthwhile noting that filters of the Fabry-Perot type are narrow in entire momentum space and as such inhibit image transmission.

VIII. RECIRCULATION

Since the amplifiers described here are traveling-wave devices, they may be folded to achieve a more desirable geometry. An extreme case of folding is to send successive signal passes through what is otherwise a single section; i.e., an amplifier is designed so that its transverse capacity exceeds that required for the signal, which may then be recirculated in this otherwise excess image space.

If, for example, one is concerned, as in the last section, with an amplifier for a single-mode transmission system, then each pass must occupy a region $\Delta x\ \Delta y\ \Delta P_x\ \Delta P_y \approx h^2$. Further it is essential to make efficient use of this transverse space if η_s is not to deteriorate. It may be necessary to provide image guard bands to inhibit feedback: the resulting decrease in η_s may in practice be more than compensated by an increase in the pump circuit efficiency with this geometry.

The recirculating optics can take many forms. Rectilinear reflections can produce the simple single-cell image amplifier described in the last section in compact form. The addition of lenses can produce the folded equivalent of the third type of amplifier described in the last section. If a Faraday rotation isolator is used, an optically active material such as quartz in the recirculating path can restore the polarization. The number of passes is limited by the losses in the optics and by the amount of feedback which can be tolerated.

IX. TRANSIENT BEHAVIOR

The fact that a TWM can store energy and release it on demand is well known. This property has been of only academic interest at microwave frequencies but may be one of the most important characteristics of an optical device. The transient behavior of a TWM has been investigated by Schulz-DuBois[6] and may be summarized as follows. If a strong signal in the form of a step function is applied, then the leading edge will see the full amplifier gain and, in the process, will drain off active particles so that successive portions of the signal experience less gain. As the signal travels along the amplifier, it assumes somewhat the shape of a shock wave where the sharp leading edge sees the full gain, and the integrated energy in the pulse is essentially equal to the original stored energy in the preceding portion of the amplifier. The output pulse

width is limited by the rise time of the amplifier. This theory is valid as long as the material characteristics at the signal frequency can be described in terms of the rate equations.

The rise time of a ruby TWOM at room temperature is $\approx 2 \times 10^{-12}$ sec. By using the pulse-sharpening property of the device, it may be possible to produce light pulses shorter than any available by conventional electronics.

It is easy to design a TWOM of cross section less than 1 cm^2 whose stored energy is greater than one joule. With such a device, rise time is unlikely to limit the peak power; instead, the limitation will probably be material breakdown. In the present design the Polaroid is believed to be the most susceptible component but could be replaced by, say, Glan-Thomson or Brewster angle polarizers. We expect that peak powers of 10^7 watts/cm^2 can be achieved quite easily without focusing and that considerably higher power densities are realizable.

Once the limiting power density and cross-sectional area are reached, the beam can be divided into parallel channels forming a phased array. If phase stabilities approaching that of the microwave TWM can be achieved, several divisions will be possible.

X. CONCLUSION

The feasibility of an image-amplifying optical traveling wave maser has been demonstrated. Since each section is short circuit stable, they may be cascaded to produce any desired gain.

The experimental data on the gain as a function of pump power are in good agreement with theory.

The loss and Verdet constant of high density PbO glass has been measured and is such that in combination with dichroic polarizers it leads to a satisfactory isolator. Materials with higher Verdet constants would be preferable.

The theory shows that, in order to build an amplifier which is low noise and which has a minimum pump requirement, one must use high quality optics which are diffraction limited.

It is believed that very high peak powers and very short pulses can be produced with the device.

An effective channel capacity $\approx 10^{15}$ cps has been obtained and one is faced with the rather unusual problem of reducing the bandwidth of an amplifier.

XI. ACKNOWLEDGMENTS

The authors wish to acknowledge the assistance of Mr. H. Marcos in making the measurements on the optical isolator. We also are in-

debted to our colleagues Dr. E. O. Schulz-DuBois and Dr. J. G. Skinner for critically reading the manuscript and making valuable suggestions.

REFERENCES

1. Maiman, T. H., Nature, **187,** 1960, p. 493.
2. Javan, A., Bennett, W. R., Jr., and Herriott, D. R., Phys. Rev. Letters, **6,** 1961, p. 106.
3. Javan, A., Bennett, W. R., Jr., and Ballik, E. A., private communication.
4. Kisliuk, P. P., and Boyle, W. S., Proc. I.R.E., **49,** 1961, pp. 1635–1639.
5. DeGrasse, R. W., Schulz-DuBois, E. O., and Scovil, H. E. D., B.S.T.J., **38,** 1959, pp. 305–334.
6. Schulz-DuBois, E. O., Microwave Solid-State Devices, *11th Interim Report*, Army Signal Corps Contract DA-36-039-sc-73224.
7. Condon, E. U., and Shortley, G. H., *The Theory of Atomic Spectra*, Cambridge University Press, London and New York, 1951.
8. Wieder, I., and Sarles, L., *Advances in Quantum Electronics*, edited by J. R. Singer, Columbia University Press, New York, 1961.
9. Schawlow, A. L., *Advances in Quantum Electronics*, edited by J. R. Singer, Columbia University Press, New York, 1961.
10. McClure, D. S., *Advances in Solid State Physics*, **9,** edited by F. Seitz and D. Turnbull, Academic Press, Inc., New York, 1959.
11. Dirac, P. A. M., *Quantum Mechanics*, 3rd edition, Oxford University Press, 1947, p. 238.
12. Weber, J., Masers, Rev. Mod. Phys., **31,** 1959, p. 681.
13. Gordon, J. P., to be published in Proc. I.R.E.

Further Analysis of Errors Reported in "Capabilities of the Telephone Network for Data Transmission"

By ROBERT MORRIS

(Manuscript received October 23, 1961)

The recorded error data from a field testing program reported on by Alexander, Gryb, and Nast have been further analyzed. New methods of analysis have given more information on the causes and nature of errors experienced by data in the switched telephone plant. The results obtained will enable workers in the field of error control to use the field test data more effectively. The ideas presented will be useful to designers of future field tests.

TABLE OF CONTENTS

I. Introduction 1399
II. The Task Force Testing Program 1400
III. The Effect of Dropouts 1401
IV. The Effects of Noise 1402
V. Methods of Determining Bit Phase 1403
VI. Methods of Finding Short Dropouts 1408
VII. Analysis of the Dropouts 1408
VIII. Analysis of Errors due to Noise 1409
IX. The Case of Call No. 2390 1409
X. Some Special Results 1410
XI. The Role of the Test Word 1411
XII. Relation to Error-Control Schemes 1412
XIII. Conclusions 1413
XIV. Acknowledgments 1414

I. INTRODUCTION

The purpose of this paper is to describe certain characteristics of errors affecting digital data in the switched telephone network. The present results are based on an analysis of error data recorded in a field testing program conducted by the Data Transmission Evaluation Task Force of the Bell System. These data are summarized in a previous paper by Alexander, Gryb, and Nast[1] which sets forth numerous conclusions about the telephone plant, based on the test program. The present paper reports the results of a sequence of statistical investiga-

tions which have made it possible to classify most of the errors into several populations. The effect of the test equipment on the time-distribution of errors is also discussed.

II. THE TASK FORCE TESTING PROGRAM

This section contains a description of those aspects of the test program which bear on the present discussion. Test calls were made by the Task Force over a wide variety of circuits. Both local and long distance calls were made over the switched telephone network and the performance of the circuits was recorded. Data concerning 1010 test calls were gathered and used for analysis. The test calls were of three categories: 10-minute calls at 600 bits/second, 10-minute calls at 1200 bits/second, and 30-minute calls at 1200 bits/second. Each of these categories was further subdivided into "Exchange Calls" which used no long distance switching facilities, "Short Haul Long Distance Calls" which were made over distances of less than 400 miles, and "Long Haul Long Distance Calls" which were made over distances of between 400 and 3000 miles.

The source of the transmitted data was a word generator which produced a sequence of marks and spaces repeating with a period of thirty bits. This word generator was used to drive an FM modulator. (The terms "mark" and "space" designate the two states of the FM channel. The convention here is that the lower of the two frequencies used is called mark and the higher is called space.) At the receiving end of the circuit, the FM signal was demodulated and compared with the output of another word generator identical to that at the transmitting end. When the output of the word generator at the receiving end agreed with the received signal, the bit was accepted as correct; when they differed, an error was noted. The two generators were kept in step by electronic clocks. The sequence of marks and spaces produced by the word generators was as follows:

SSSSMSSSSMMSMSSMMMMSMMMMSSMSSM.

The sequence of bits received correctly and bits received in error was recorded serially on a magnetic tape with two channels, as follows: The first channel contained only clock pulses, one pulse per bit. The second channel contained a pulse opposite the corresponding pulse in the first channel every time an error was noted, and was blank whenever no error occurred. The relative bit phase of the word generator, i.e., the position in the 30-bit word where the call began, was not recorded for any of the calls. The methods described below made it possible in most cases

to determine whether a particular error changed a mark into a space or a space into a mark, and to determine which, if any, of the thirty positions in the test-generator word were most susceptible to errors.

The demodulator contained what is known as a zero-crossing detector, which counted the number of times the received signal crossed a neighborhood of zero. This detector delivered an output which was proportional to the number of zero crossings per unit time. A threshold was established and when fewer zero crossings were detected, a mark was scored; when more zero crossings were detected, a space was scored. Thus when no signal at all was received, the absence of signal was interpreted as a mark.

With a detector of this type, impulse noise would be more likely to add zero crossings to the received signal than to subtract them if the impulse noise were of higher frequency than the signal frequencies used in the tests. On the other hand, disturbances of lower frequency would tend to change spaces into marks by subtracting zero crossings. As it turned out, most errors caused by noise were mark-to-space errors, but the fact most important to our methods of analysis was that a single type of disturbance during a call caused one of the two types of error predominately. By a "single type of disturbance" we mean any statistically recognizable pattern of errors.

Much of the remainder of the discussion of this test program will depend on three of the properties of the system mentioned above, namely:

(1) a cyclic 30-bit word generator was used,
(2) any particular kind of disturbance on the line caused errors asymmetrically, i.e., either a majority of mark-to-space errors or a majority of space-to-mark errors,
(3) loss of signal caused all bits to be received as marks.

During some of the calls in the test program, the circuit was lost and a new attempt was made. On other occasions, the clock which kept the receiving word generator in phase got out of step, sometimes because of loss of signal. In either case, the error tape was erased and a new error recording was started from the beginning. Some of the calls encountered one or both of the above difficulties. Only the calls which were successfully completed were reported.

III. THE EFFECT OF DROPOUTS

Rather early in the study it became obvious that, in some of the test calls, errors were caused not by noise but by loss or serious attenuation ("dropout") of the received signal, often for very short periods of time.

When this occurred, all bits were received as marks and errors were recorded only when the receiving word generator produced a space. When the receiving word generator produced a mark, the bit was recorded as being received correctly even though the mark that was received was completely independent of what was transmitted. The pattern of errors recorded during such a period was:

$$111101111001011000010000110110,$$

where a 0 stands for a bit accepted as correct, and a 1 stands for a recorded error. Thus, when the line was open or for some other reason no signal was being received during the tests, 16 errors were recorded for every 30 bits transmitted. The error rate observed during a dropout is evidently dependent on the proportion of spaces transmitted, which in many systems of error control differs widely from 16 out of 30; for instance, in the so-called 2/8 code, 2 marks and 6 spaces are sent in every block of 8 bits.

Computer programs were written by the author which could detect the pattern shown above in those cases where the dropout extended over more than about twelve bits. A dropout occurring in the first nine bit-positions of the test word would have caused eight errors, and, since the computer program was adjusted so as to find dropouts causing more than five errors, this dropout would be found. A dropout which covered the sixteenth through the twenty-fourth bit-positions would not be found, since it would have caused only one error and could not be distinguished from other error-producing effects. Because of the distinctive pattern of errors, the bit phase of the word generator could easily be determined in calls with long dropouts.

IV. THE EFFECTS OF NOISE

During some test calls, the nature of the disturbances on the transmission facilities was such that there were many more mark-to-space errors than space-to-mark errors. On other test calls, the reverse was true. A computer program was written by the author which made it possible to determine which calls were of these two types and to decide the bit phase of the word generator in those cases. The exact methods by which this was done are described in Section V.

In some calls, there were ten to twenty times as many mark-to-space errors as there were space-to-mark errors. In these calls, errors did not fall on all of the mark positions with even approximately equal frequency. Some mark positions were more vulnerable to errors than others, al-

though the vulnerable positions were not always the same in different calls. In certain calls, the average number of errors on the four exposed mark positions (a single mark between spaces) exceeded by a factor of ten the average number of errors on the remaining mark positions.

V. METHODS OF DETERMINING BIT PHASE

Three different methods were used by the author to determine the bit phase of the word generator in the test calls. None of these was effective on calls with less than about fifty errors; in fact they occasionally failed on calls with as many as several hundred errors. However, the bit phase was determined for enough calls to account for somewhat more than 65 per cent of all the errors reported in the test program.

The first method was by far the simplest. The error data were visually scanned for the characteristic pattern of a dropout. In only a few calls were there long enough dropouts for this method to produce results.

The second method used the IBM 7090 computer to divide the sequence of good bits and bits in error into consecutive blocks of thirty bits each. The positions in these blocks were numbered from 1 to 30 and the number of errors corresponding to each numbered position was totaled over all of the blocks in the call. This produced a sequence of thirty numbers such as the one in Fig. 1, which is the result actually obtained for one of the test calls, Call No. 2330; this was a 30-minute Exchange Call at 1200 bits/second. This call had a total of 678 errors, and if the errors fell randomly at each bit position, then more than half of the thirty numbers in Fig. 1 would be expected to lie in the range 19–26 inclusive. This is far from being the case, since only three of them do so. About half of the numbers are surprisingly small and most of the rest are surprisingly large. There is a wide gap between the two collections of numbers. When the relative bit phase of the word generator in this call is assumed to be that of Fig. 2, which is correct, the numbers of errors corresponding to the fourteen bit-positions of the test generator word which are marks are:

25, 31, 20, 85, 27, 40, 28, 79, 46, 55, 43, 22, 66, 44

and the numbers of errors corresponding to the sixteen bit-positions which are spaces are:

9, 1, 11, 2, 3, 8, 2, 2, 6, 4, 7, 1, 1, 6, 4, 0.

It is difficult to ignore the fact that this choice of the relative bit phase of the word generator divides the thirty numbers in Fig. 1 into two dis-

9	25	31	20	85	1	27	40	28	79	11	2	46	3	8	55	2	2	6	4	43	7	1	1	6	22	66	4	44	0

Fig. 1 — Output of computer program for call No. 2330.

9	25	31	20	85	1	27	40	28	79	11	2	46	3	8	55	2	2	6	4	43	7	1	1	6	22	66	4	44	0
S	M	M	M	M	S	M	M	M	M	S	S	M	S	S	M	S	S	S	S	M	S	S	S	S	M	M	S	M	S

Fig. 2 — Correct choice of word generator phase.

9	25	31	20	85	1	27	40	28	79	11	2	46	3	8	55	2	2	6	4	43	7	1	1	6	22	66	4	44	0
M	S	S	S	S	M	S	S	S	S	M	M	S	M	S	S	M	M	M	M	S	M	M	M	M	S	S	M	S	S

Fig. 3 — Incorrect choice of word generator phase.

tinct populations; the smallest number in the first list is 20 and the largest number in the second list is 11. Another (but erroneous) choice of the bit phase of the word generator is represented in Fig. 3. With this choice, the numbers of errors corresponding to marks are:

9, 1, 11, 2, 3, 2, 2, 6, 4, 7, 1, 1, 6, 4

and the numbers of errors corresponding to spaces are:

25, 31, 20, 85, 27, 40, 28, 79, 46, 8, 55, 43, 22, 66, 44, 0.

Two of the numbers in the second list (the 8 and the 0) seem to stand out as belonging more naturally to the first list.

Because of the pattern in Fig. 1 of four large numbers, then one small, then four large, followed after an interval of six bits by four small, then one large, then four small, the only choices of word-generator phase which make sense are those shown in Figs. 2 and 3. The choice between them is not difficult in this case. This pattern is repeated in dozens of calls; often it shows up even more clearly than in the example given here.

Calls in which there were numerous dropouts too short to be recognized by eye had an excess of space-to-mark errors. These calls were found and analyzed by the same method, but in this case high numbers corresponded to spaces.

These two methods failed to determine the bit phase in many calls which had very large numbers of errors. It was suspected that some of these calls had both short dropouts and, in addition, enough mark-to-space errors so that neither of the two kinds of error was in large excess. To handle this situation, a computer program was written which produced sequences of thirty numbers as above, but this time the first line included only the single errors, the second line included only the double errors, etc. Two examples of this presentation are given in Figs. 4 and 5. Now, errors caused by noise most commonly occurred one by one, and errors during a dropout occurred primarily in two's and four's as can be seen by reference to the test-generator word. Therefore the first line of this presentation could be scanned for any great preponderance of mark-to-space errors caused by noise, and the other lines could be scanned for evidence of short dropouts, which were also made more visible by this presentation. Many of the remaining calls succumbed to this double-barrelled attack. Fig. 4 gives an example of a call in which the errors were mostly due to dropouts, and Fig. 5 gives an example of a call in which the errors were mostly isolated mark-to-space errors.

3	1	1	2	1	2	0	1	3	2	2	(14)	2	1	3	4	2	1	2	(14)	2	1	0	3	3	1	3	2	4	3
0	0	0	1	1	0	0	2	2	1	0	0	0	(11)	1	0	0	0	0	1	0	0	1	0	(13)	0	1	(10)	1	2
0	2	1	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
(8)	0	0	0	0	(8)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	S	S	S	M	S	S	S	S	M	M	S	M	S	S	M	M	M	M	S	M	M	M	M	S	S	M	S	S	M

Fig. 4 — Output of second computer program for call No. 2249. The top line of this presentation contains the total number of isolated errors corresponding to each bit-position in the test word. The second line contains the total number of double errors corresponding to each position, with an entry made only once for each double error at the first of the two bit-positions it covers. That is, a double error on the first and second bit-positions would contribute 1 to the count in the first entry of the second line. Third and following lines are similar displays of triple errors, etc. In this, the circled entries constitute strong evidence of drop-outs.

2	3	2	1	31	0	4	1	0	21	6	0	24	0	1	17	7	4	11	0	11	7	4	6	0	0	22	0	1	26
0	0	0	2	2	0	0	0	1	5	1	0	1	0	0	6	1	5	1	1	9	1	2	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	1	2	0	0	0	0	0	1	1	0	0	0	5	1	0	0	0	0	1	0	0	1
0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	S	S	S	M	S	S	S	S	M	M	S	M	S	S	M	M	M	M	S	M	M	M	M	S	S	M	S	S	M

Fig. 5 — Output of second computer program for call No. 1607. In this call there is no evidence of any dropout, but the first line indicates many more isolated errors on the marks than on the spaces.

VI. METHODS OF FINDING SHORT DROPOUTS

Even though the last method mentioned was able to reveal short dropouts in a call, it gave no indication of their exact lengths or where they were located in the call. Another computer program was written which went bit-by-bit through each call of known bit phase, and counted up how many marks had been received out of the previous thirty bits. If no errors were recorded in the previous thirty bits this number was at all times 14. If mark-to-space errors were occurring, the number was less than 14. But, if a dropout occurred, even over a few bits, the number rose dramatically and made it clear where the dropout probably began and ended. By this method, combined with visual inspection of the original data, every dropout which caused more than five errors in any call of the test program was found, except for the remote possibility that during a dropout noise simulated the transmitted word.

Another interesting phenomenon was discovered by means of these computations; namely, that often there was a dropout during which a few spaces would be received, but there was no correlation between these spaces and the spaces actually sent by the transmitter. This indicates that on occasion, even when the line was open, there was enough noise received to simulate a space, so that some bits which were recorded as correct in the field trials were actually the result of two potential errors in the same bit-position whose effects cancelled. Short periods of time during which the transmitted signal was absent, but during which impulse noise caused several spaces to be received, are not classified as dropouts, since the resulting error patterns are not distinctive.

VII. ANALYSIS OF THE DROPOUTS

In the test calls, a total of 58 dropouts extended over more than 12 bits, and these were analyzed. Of these, none were in Exchange Calls, 3 were in Short Haul Calls, contributing 110 errors, and the remaining 55 were in Long Haul Calls, contributing 3400 errors. The longest dropout was 1129 bits long (approximately 1 second). It is estimated that dropouts lasting 12 bits or less contributed at least 1500 additional errors.

In addition to those discussed above, many dropouts occurred during the tests and were not reported because these dropouts caused a loss of synchronization between the transmitting and receiving word generators or were long enough to cause the test personnel to terminate the call. The recording of errors in these cases was abandoned and a new recording started.

It is difficult, therefore, on the basis of the reported data, to specify quantitatively the contribution of dropouts to the total error rate. This contribution was negligible in the case of Exchange Calls and Short Haul Calls. In Long Haul Calls, dropouts were a major source of error. Even after the omission of records of many dropouts, as explained above, about 10 per cent of all Long Haul Calls had one or more dropouts, and these dropouts contributed over 20 per cent of the errors in Long Haul Calls.

Since there were no dropouts in the final data on Exchange Calls, and only negligible ones for Short Haul Calls, and since a very large number of calls were made in these categories, it seems safe to assume that the causes for dropouts lay in long-distance transmission or switching facilities, or at least that these facilities formed an essential link in the chain of causes leading to a dropout.

VIII. ANALYSIS OF ERRORS DUE TO NOISE

There were 36 calls in which the bit phase of the word generator was established and in which there were no errors that could be attributed to dropouts. In these calls, the number of mark-to-space errors was slightly more than four times as great as the number of space-to-mark errors. This statement, though true, is somewhat misleading, because this was just the kind of call in which the bit phase was easy to determine. It is certain that none of the remaining calls with more than about 50 errors has any such discrepancy between the two kinds of error, or else its bit phase would have been determined. What can be deduced is that some transmission paths favored this kind of error very much.

In some calls, an abnormally high number of errors fell on those bit positions which corresponded to single marks between spaces; in others, an abnormally high number fell on the positions corresponding to marks immediately preceding spaces. These effects may have been caused by conditions on the line, such as high delay distortion; or they may have been caused by characteristics of the test equipment, as illustrated in Section IX.

IX. THE CASE OF CALL NO. 2390

Call No. 2390 was a 30-minute Long Haul Call at 1200 bits/second with 4565 errors. This call had more errors than any other call in the test program. In spite of the large number of errors, the three methods described did not reveal the bit phase of the word generator in this call. Therefore, the pattern of errors in the call was subjected to close scru-

tiny. It was discovered that the sequence of good bits and bits in error at the end of the call was precisely that which would be recorded if the receiving word generator were one bit out of step with the transmitting word generator. Somewhat further back in the call the pattern corresponded to the two generators being two bits out of step. This was traced back through the call step by step until finally a place was reached where the two word generators were out of step by fifteen bits. By this time, several thousand errors had been accounted for. The search was not carried beyond this point.

Evidently the bulk of the errors in Call No. 2390 were caused by a failure of the terminal equipment to keep in step. These 4565 errors make up 12 per cent of the errors observed during the entire testing program and 17 per cent of those observed in Long Haul Calls.

It does not seem justified to let these pseudo-errors contribute to the reported error rate on the telephone network. And it is certain that this call should not be used in a study of error control or of the burst structure of errors, since not only the pattern of errors but the errors themselves were caused by the nature of the test equipment and not by the characteristics of the telephone plant.

X. SOME SPECIAL RESULTS

During Call No. 1568, a Long Haul Call at 1200 bits/second, the field test personnel noted that four bursts of errors which caused a total of about 400 errors coincided with audible multifrequency keypulses. A direct inspection of these errors failed to reveal much regularity in the pattern of errors within the bursts. After the bit phase of the word generator in this call was determined, the sequence of received marks and spaces was inspected. During the four keypulses, the received signal conformed to the following four patterns, respectively:

M M M S S S M M M S S S M M M S S S
M M M M M S M M M M M S M M M M M S
M M S M M S M M S M M S M M S M M S
M M M M S S M M M M S S M M M M S S

with occasional errors superimposed, presumably arising from noise. These patterns have periods of 6, 6, 3, and 6 bits, respectively, which correspond to frequencies of 200, 200, 400, and 200 cps. These frequencies may be related to the fact that the two tones transmitted simultaneously during an MF pulse are always separated by a multiple of 200 cps.

In parts of some calls, remarkably many errors were separated by exactly 57 good bits. On investigation, these all turned out to be calls

at 1200 bits/second which terminated in a few particular step-by-step central offices. Most probably, these errors were caused by some switching mechanism in the telephone plant which, when it operates, produces a 21-cps disturbance.

In Call No. 1420, a Long Haul Call at 1200 bits/second, there were four long dropouts and their lengths were 301, 302, 329, and 337 bits. The closeness of lengths of these four dropouts suggests that all of them had the same cause.

In Call No. 2429, a Long Haul Call at 1200 bits/second, there were eight long dropouts and their lengths were 33, 33, 67, 67, 94, 95, 97, and 126 bits. All of these lengths are close to being multiples of 32 bits, again suggesting a common source of the dropouts.

XI. THE ROLE OF THE TEST WORD

The choice of a cyclic word generator as the source of the transmitted data did not affect the measurement of error rate significantly. However, its characteristics were strongly reflected in the final error recordings because of the considerable contribution of dropouts to the total error rate. The fact that the length of the test word was 30, a number with many small factors, again did not affect the measurement of error rate; but if the data were used in simulation experiments — for example, to test an error-detecting code whose length was a divisor of the number 30 — the results could be biased by this choice of length of test word.

After the bit phase of many of the calls was revealed by analysis of the error recordings, it was possible to resolve the following questions:

(1) whether any particular error changed a mark into a space or the reverse,
(2) which, if any, of the 30 bit-positions were most susceptible to errors,
(3) to what extent errors were caused by dropouts, and
(4) the contribution of terminal-equipment malfunction to the observed error rate.

Using the bit phase to study the distribution of errors among the 30 bit-positions of the test word, it was discovered that the most vulnerable positions were different in different calls. It would be instructive to be able to compare these differences in vulnerability with the measured characteristics of the transmission medium. There were a number of calls in the series for which almost all of the errors fell on a single bit-position in the test word. It is not known either why this occurred or which bit position was affected, since it is impossible to determine the bit phase for such calls.

XII. RELATION TO ERROR-CONTROL SCHEMES

One of the results of this study is the knowledge that with the modulation scheme used in the testing program, the mark-to-space and space-to-mark errors were not evenly mixed with each other. For a long period of time the mark-to-space variety was more abundant by a large factor, owing to noise; at other times the reverse was true, because of dropouts, interfering tones, and the like. An error-detecting code which counts the number of marks per block is much more efficient in this situation than when mark-to-space errors and space-to-mark errors are well mixed.

For example, in the 2/8 code, which has been suggested for error detection, bits are transmitted in blocks of eight, and in every block there are exactly two marks and six spaces. A received block with more or less than two marks is detected as an error. This scheme will always detect an error condition if a single error occurs in a block. If a block contains two errors (which is likely because of the burst character of the errors) then an error condition will be detected if both errors are in the same direction but not if they are in opposite directions. Assuming that the two kinds of error are well mixed and that sent marks are four times as likely to be in error as sent spaces, we find that a double error in a block will be detected only 39 per cent of the time. The truth is that, because of the lack of mixing, a double error in a block would be detected over 90 per cent of the time by this code. During a dropout the situation is even more striking. A message using this code would experience a bit-error rate of 75 per cent during a dropout but *every one of these errors would be detected*. The relative advantage of this code and of other codes of this type would go unnoticed if the codes were evaluated by a direct simulation using the field test data.

A useful technique in studying error-control methods is the construction of mathematical models of the occurrence of errors. Several such models have been constructed using the field test data,[2,3] and have been used to study the effectiveness of error-control procedures.[4] To test the validity of such models and to evaluate the parameters used in the models, it is necessary to study the manner in which errors are distributed in time, e.g., to what extent they occur in bursts. One way of doing this is to determine the distribution of lengths of runs of consecutive errors and the distribution of lengths of runs of good bits between errors.[2] Because of the structure of the test generator word, every run of consecutive errors in the field test data during a dropout had length 1, 2, or 4 and every run of consecutive good bits during a dropout had length 1, 2, or 4. These distributions of run lengths depended only on the choice of test-generator word. Also, repetitive patterns of errors in Long Haul

Calls were often produced by causes other than dropouts. For instance, the single event described in Section IX contributed 17 per cent of all the errors in these calls, and occasions (described in Section X) when the signal was overridden by tones contributed about another 10 per cent. Thus, including dropouts, about half the errors in Long Haul Calls occurred in repetitive patterns, and these patterns reflected the nature of the test equipment and the modulation scheme rather than that of the telephone plant.

Also, when short spikes of impulse noise caused bits to be received as spaces regardless of what was transmitted, errors were recorded in the field test data only when the occurrence of a spike of noise coincided with a transmitted mark. Those which coincided with transmitted spaces were not recorded as errors. Therefore, the mean length of runs of good bits in the data is about twice the mean distance between spikes of impulse noise. If the mark-counting code described above were tested against these data, it would appear to experience twice as many errors as it would have experienced if it had been used on a real-life channel.

XIII. CONCLUSIONS

With no knowledge of the bit phase of the word generator, the only error statistics which could be reliably deduced from the field test data were the error rates themselves. With the recovery of the bit phase, the proportion of mark-to-space errors could be determined, dropouts were revealed, and certain malfunctions of the test equipment were discovered. Now a considerable amount of additional information is available about the causes and nature of errors experienced by data in the switched telephone network. This information makes the existing data more useful for simulation and analysis.

The analysis reported in this paper confirms some results of Ref. 1 and modifies some others. It is unrelated to most of the material discussed in Ref. 1. Specifically:

(1) The subject matter of this paper has no bearing on Figures 1–25 of Ref. 1.

(2) The reported error rates and the distributions of error rates by classes of calls are essentially as indicated in Figures 26–29 of Ref. 1.

(3) The time distribution of errors over large blocks (e.g., more than 50–100 bits), as shown in the right-hand portions of Figures 30, 31, 32, 34, 36, and 38 of Ref. 1, is not changed.

(4) The validity of the time distribution of errors over very short blocks, (e.g., 2–10 bits), as represented in Figures 33, 35, 37, 39

and the left-hand portions of Figures 30, 31, 32, 34, 36, and 38 of Ref. 1, has not been confirmed.

(5) The problem mentioned in (4) throws doubt on the data used to construct Figures 40–43 of Ref. 1.

The results whose validity is questioned in (4) and (5) follow from arbitrary choices made in designing the field tests. The effects of such choices are considerable and hard to evaluate. It is therefore difficult to assess the nature and extent of the changes that should be made in the figures named in (4) and (5).

Because line dropouts (and the loss of word-generator synchronization in one call) contributed significantly to the error rate, the choice of the test word influenced the time distribution of errors, as noted in (4) above. This must be considered in evaluating schemes for error correction and detection.

In most of the calls with high error rates not caused by dropouts, the errors had a distinct tendency to change marks into spaces. This also must be considered in evaluating error-control schemes.

It must also be borne in mind that the property of receiving all bits as marks in the absence of signal and the property that impulse noise usually changes marks into spaces are *not* themselves characteristics of the telephone plant, but are dependent on the particular devices used for modulation and demodulation (Ref. 1, pp. 435–6). The important point is that the methods and point of view of this paper can be applied to data obtained with any system of modulation. One cannot hope to devise a good method of controlling errors without a deep knowledge of the nature of the errors and of how the terminal equipment is affected by them.

XIV. ACKNOWLEDGMENTS

The author is indebted to R. M. Gryb and G. J. McAllister for their cooperation in making the field test data available, and to Eric Wolman for technical advice.

REFERENCES

1. Alexander, A. A., Gryb, R. M., and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39,** 1960, p. 431.
2. Gilbert, E. N., Capacity of a Burst-Noise Channel, B.S.T.J., **39,** 1960, p. 1253.
3. Mertz, P., Model of Error Burst Structure in Data Transmission, Proc. Nat. Elect. Conf., **16,** 1960, p. 232.
4. Bennett, W. R., and Froehlich, F. E., Some Results on the Effectiveness of Error-Control Procedures in Digital Data Transmission, Trans. I.R.E., **PGCS-9,** March, 1961.

Grade of Service of Direct Traffic Mixed with Store-and-Forward Traffic*

By JOSEPH OTTERMAN†

(Manuscript received November 28, 1961)

The dual use of trunks for both direct and store-and-forward (S/F) traffic makes high trunk efficiency possible. The resultant trunk savings are important in communication systems in which long-haul trunks contribute heavily to the cost of the system. This paper reports work on the computation of trunking tables that could be used to engineer trunk requirements for prescribed loads and distributions of direct and S/F traffic.

A method of computation and some specific results in terms of grade of service of direct traffic and traffic capacity for S/F traffic are presented. The numerical results are for two to forty-eight trunks. The results apply to the case of exponentially-distributed holding times of both the direct and the S/F traffic.

I. INTRODUCTION

Direct traffic is a user-to-user service (generally voice) requiring a connection to be established promptly on demand. Store-and-forward (S/F) traffic, on the other hand, is stored at or near the originator's location and is later sent to its destination, either directly or through further intermediate storage. If S/F traffic is sent only when the direct traffic load is light, large amounts of S/F traffic can be accommodated with very slight degradation of direct traffic service. This can be especially important in a long-haul communication system where the trunk group cross sections (number of trunks in a group) are small. Such trunk groups are notoriously inefficient if used only for direct traffic, but they have a substantial ability to handle additional S/F traffic.

* This work has been carried out under U. S. Army Signal Corps Contract DA-36-039-SC-78806.

† I.T.T Federal Laboratories, Nutley, N. J., (work performed for Bell Telephone Laboratories). Present affiliation, General Electric Company.

The basic operating method analyzed in this paper is as follows. The occupancy of the trunks in a trunk group is monitored, and whenever the occupancy drops below a certain level, the S/F traffic is allowed access to the idle trunks. The sending of an S/F message is *not* interrupted in order to service arrivals of direct traffic, but when transmission of an S/F message is begun, a specified number of trunks is always held in reserve for direct traffic. Under the foregoing rules, the present analysis establishes (*i*) the grade of service of direct traffic and (*ii*) the amount of S/F traffic that can be accommodated. Direct traffic congestion discipline is assumed to be governed by the lost-calls-cleared assumption,* which means that direct calls which encounter an all-trunks busy condition do not reappear in the busy hour.

The statistical nature of traffic is first summarized. The problem of dual use of trunks is formulated as a probabilistic net representing a Markov chain. It is assumed that a queue of S/F traffic exists at all times. Using cut-sets in the state graph, this probabilistic net is analyzed to derive expressions for the steady-state probabilities of trunk occupancy. From the expressions of trunk occupancy for a given load of direct traffic, the grade of service of the direct traffic and the amount of the S/F traffic that can be accommodated are determined. A glossary of symbols is given in the appendix.

II. MATHEMATICAL ANALYSIS

2.1 *Statistical Properties of Traffic*

The direct calls are assumed to be generated individually and collectively at random. The expected number of arrivals per hour is denoted as n. The expected number of arrivals during a fraction of an hour $d\tau$ is simply given by $n\,d\tau$. When $d\tau$ is short enough so that multiple events are improbable, the probability of an arrival during $d\tau$ is equal to the expected number of arrivals: i.e., equal to $n\,d\tau$.

In the following discussion it will be more convenient to measure time in fractions, dt of the average holding time of the direct traffic, which is denoted T. Then the probability of an arrival during a short time, dt, is therefore:

$$\text{Probability of an arrival during } dt = nT\,dt = a\,dt. \qquad (1)$$

* The assumption of lost-calls-cleared is especially appropriate when an alternate route is provided for calls that encounter the all-trunks-busy condition. When an alternate route is provided, the probability of loss should be interpreted as the portion of calls that overflow, seeking the alternate route.

The product nT, which is denoted a, denotes the offered direct traffic measured in erlangs. It represents the expected number of calls in progress on a fully served basis. Equation (1) indicates still another significance of a: it is the expected number of arrivals during one holding time.

The holding times of both the direct and the S/F traffic are assumed to be exponentially distributed (with average T for the direct and t_0 for S/F traffic). Under the assumption of exponential holding times, the probability of termination of a call with average holding time T during a differential interval $d\tau$ is given by $d\tau/T$ plus terms negligible in the limit $d\tau \to 0$. In fractions dt of the holding time T, this probability is simply dt. When x calls are in progress, the probability of the termination on one of these calls during a differential interval dt is given by:

$$\text{Probability of a termination out of } x \text{ calls} = x\,dt. \tag{2}$$

Similarly, the probability of a termination out of z S/F messages in transmission is $rz\,dt$, where r is the ratio T/t_0.

The probability of a termination during dt depends on the present status of the trunk group: i.e., the number of calls in progress. Neither the probability of an arrival (1), nor the probability of a termination (2), depends upon the past history of the calls. This demonstrates the fact that under the above conditions the trunk occupancy is a Markov process.

The assumption of lost-calls-cleared is used; that is, a direct call that encounters a condition of all-trunks-busy is cleared from that trunk group and does not reappear in the busy hour. In the absence of S/F traffic, the trunk occupancy is a relatively simple Markov process, which is shown in the flow diagram in Fig. 1. Statistical equilibrium considerations lead to the following formula for the probability G_x of exactly x trunks being busy:

$$G_x = \frac{\dfrac{a^x}{x!}}{\displaystyle\sum_{y=0}^{c} \frac{a^y}{y!}}. \tag{3}$$

The probability of loss (probability of all-trunks-busy) is the well known Erlang B formula:[2]

$$B(c,a) = \frac{\dfrac{a^c}{c!}}{\displaystyle\sum_{y=0}^{c} \frac{a^y}{y!}}. \tag{4}$$

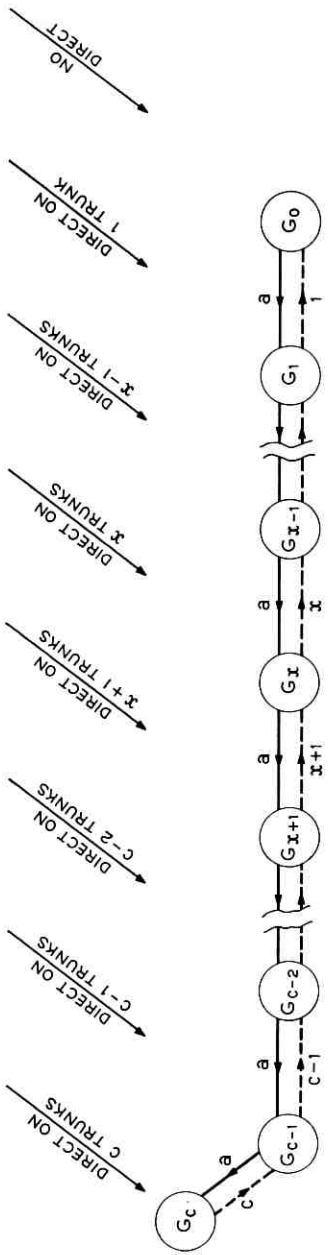


Fig. 1 — Markov chain for trunk occupancy under lost-calls-cleared assumption, no S/F traffic.

2.2 *Trunk Occupancy by Direct and S/F Traffic on a Lost-Calls-Cleared Basis*

2.2.1 *One Trunk in Reserve, $s = 1$*

The queue of S/F traffic is given access to the trunk group only when the occupancy is such that one trunk will remain idle after transmission of an S/F message is initiated. An S/F queue is assumed to exist at all times, which means that whenever two trunks in the trunk group become idle, one of them is taken for S/F transmission. When the reserve trunk becomes busy with direct traffic, S/F transmission is stopped on the trunk on which the next (in time sequence) termination of an S/F message occurs. This last-mentioned trunk then becomes a reserve trunk.

The occupancy of trunks can be regarded as a Markov process, which is shown in the flow diagram of Fig. 2. The circles in the bottom row represent states in which one trunk is available to accommodate possible arrivals of direct traffic. Conversely, the circles in the top row represent all-trunks-busy conditions. Trunk occupancy by S/F traffic is represented by the index z, which, in Fig. 2, increases from left to right. S_z represents the steady-state probability of z trunks busy with S/F traffic (in the bottom row): i.e., $(c - z - 1)$ trunks busy with direct traffic and one trunk in reserve. R_z represents the steady-state probability of z trunks busy with S/F traffic (in the top row): i.e., $(c - z)$ trunks busy with direct traffic and no trunks idle.

In the steady state, the transition probabilities out of S_z during dt are: $a\ dt\ S_z$ through a new arrival (transition into R_z) and $(c - z - 1)$ $dt\ S_z$ through a termination of one of $(c - z - 1)$ direct calls in progress (into S_{z+1} since, as soon as two trunks are recognized as idle, one of them is seized for S/F transmission).

Transition probabilities into S_z are: $(c - z)dtR_z$ through a termination of one of $(c - z)$ direct calls in progress in state R_z (in the time interval after an arrival of a direct call and before the next termination, in time, of an S/F message); $(z + 1)r\ dtR_{z+1}$ through a termination of one of the $(z + 1)$ S/F messages in progress in state R_{z+1}; and $(c - z)$ $dt\ S_{z-1}$ through a termination of one of the $(c - z)$ direct calls in progress in state S_{z-1}. Equating the transition probabilities out of S_z to transition probabilities into S_z results in the following equation:

$$(a + c - z - 1)S_z = (c - z)S_{z-1} + (z + 1)r\,R_{z+1} + (c - z)R_z\,. \quad (5)$$

The transition probabilities out of R_z during dt are: $(c - z)\ dt\ R_z$

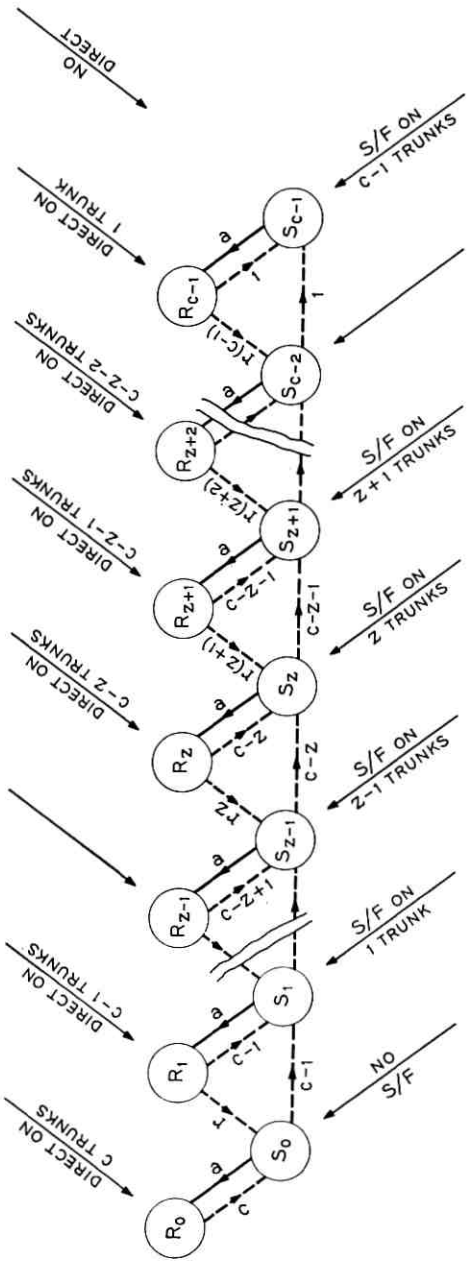


Fig. 2 — Markov chain for trunk-occupancy under lost-calls-cleared assumption, dual use of trunks, one trunk in reserve.

through a termination of one of $(c - z)$ direct calls in progress (into S_z) and $zr\,dt\,R_z$ through a termination of one of z S/F messages in progress (into S_{z-1}). The transition probability into R_z is $adt\,S_z$ through an arrival of a direct call.

It should be pointed out here that if a new, direct arrival occurs when the chain is in state R_z, one call is lost (or takes another route), but transition to another state does not occur under the assumption of lost-calls-cleared.

Equating the transition probabilities into R_z to transition probabilities out of R_z results in the following equation:

$$aS_z = (c - z + zr)R_z = [c + z(r - 1)]R_z . \tag{6}$$

Subtracting (6) from (5) results in

$$(z + 1)rR_{z+1} = (c - z - 1)S_z - (c - z)S_{z-1} + zr\,R_z . \tag{7}$$

This relation is satisfied if the following holds:

$$(z + 1)rR_{z+1} = (c - z - 1)S_z \tag{8}$$

and therefore,

$$zrR_z = (c - z)S_{z-1} . \tag{9}$$

Actually, the recurrence relation of (8) and (9) can be observed directly by equating the transition probabilities through the appropriate cut-set in Fig. 2.

We will express first R_z and S_z, $0 \leqq z \leqq (c - 1)$, in terms of R_0, the probability that all trunks are busy with direct traffic. Combining (6) and (8) shows that

$$ar(z + 1)R_{z+1} = (c - z - 1)[c + z(r - 1)]R_z \tag{10}$$

and therefore,

$$arzR_z = (c - z)[c + (z - 1)(r - 1)]R_{z-1} . \tag{11}$$

Iteration of (11) results in the following expression for R_z:

$$R_z = \frac{[c + 0(r - 1)][c + 1(r - 1)][c + 2(r - 1)] \cdots [c + (z - 1)(r - 1)]}{z!\,r^z} \cdot \frac{(c - 1)(c - 2) \cdots (c - z)}{a^z} R_0 . \tag{12}$$

From (12) and (6),

$$S_z = \frac{[c + 0(r-1)][c + 1(r-1)][c + 2(r-1)] \cdots [c + z(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^{z+1}} R_0. \tag{13}$$

Now, R_0 can be determined for a given load a by noting that the sum of probabilities R_z and S_z adds up to unity:

$$\sum_{z=0}^{c-1} R_z + \sum_{z=0}^{c-1} S_z = 1. \tag{14}$$

Introducing R_z and S_z from (12) and (13), respectively,

$$\frac{1}{R_0} = \sum_{z=0}^{c-1} \frac{[c + 0(r-1)][c + 1(r-1)][c + 2(r-1)] \cdots [c + (z-1)(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^z}$$

$$+ \sum_{z=0}^{c-1} \frac{[c + 0(r-1)][c + 1(r-1)][c + 2(r-1)] \cdots [c + z(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^{z+1}} \tag{15}$$

$$= \sum_{z=0}^{c-1} \{a + c + z(r-1)\} \cdot \frac{[c + 0(r-1)][c + 1(r-1)] \cdots [c + (z-1)(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^{z+1}}.$$

The probability of loss corresponds to the probability that all trunks are busy: i.e., the sum of the probabilities R_z. This can be referred to as the dual-use-of-trunks formula, $D(c,a,r,1)$ for c trunks, a erlangs of offered direct load, ratio r of holding times, and one trunk in reserve.

$$D(c,a,r,1) = \sum_{z=0}^{c-1} R_z$$

$$= \frac{\displaystyle\sum_{z=0}^{c-1} \frac{[c + 0(r-1)][c + 1(r-1)] \cdots [c + (z-1)(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^z}}{\displaystyle\sum_{z=0}^{c-1} [a + c + z(r-1)]} \cdot \frac{[c + 0(r-1)][c + 1(r-1)] \cdots [c + (z-1)(r-1)]}{z!\, r^z} \cdot \frac{(c-1)(c-2)\cdots(c-z)}{a^{z+1}} \tag{16}$$

The amount of S/F traffic in erlangs that can be accommodated is denoted b. It is given by

$$b = \sum_{z=0}^{c-1} zS_z + \sum_{z=0}^{c-1} zR_z. \tag{17}$$

Equations (16) and (17) constitute the results sought.

The load b can be determined by a simpler formula when the grade of service $D(c,a,r,1)$ has been calculated. This formula can be arrived at by the following reasoning. The load in terms of trunks occupied is $c - 1$ at all times, plus one additional trunk with probability $D(c,a,r,1)$. The erlang load in terms of traffic carried is b for the S/F traffic and $a[1 - D(c,a,r,1)]$ for the direct traffic. [The portion $aD(c,a,r,1)$ is cleared under the assumption of lost-calls-cleared.] Equating the two results yields the following equation,

$$b + a[1 - D(c,a,r,1)] = c - 1 + D(c,a,r,1) \tag{18}$$

and therefore

$$\begin{aligned} b &= c - 1 - a[1 - D(c,a,r,1)] + D(c,a,r,1) \\ &= c - (a + 1)[1 - D(c,a,r,1)]. \end{aligned} \tag{19}$$

It should be pointed out that for $r \to \infty$ the probabilities $R_z \to 0$ as $1/r$ (except for R_0). In this limiting case the probability of loss tends to $B(c,a)$ with the difference—i.e., the impairment due to the S/F traf-

fic—being of the order of $1/r$. This can be written

$$\lim_{r\to\infty} [D(c,a,r,1) - B(c,a)] = 0\left(\frac{1}{r}\right). \tag{20}$$

The probabilities R_0, S_0, S_1, $\cdots$ S_{c-1} tend to become, respectively, G_c, G_{c-1}, G_{c-2} $\cdots$ G_0.

2.2.2 *Two Trunks in Reserve, $s = 2$*

In this case, the queue of S/F traffic is denied access to the trunk group unless two trunks will remain idle after transmission of an S/F message is initiated. A queue of S/F traffic is again assumed to exist at all times, which means that as soon as three trunks in the trunk group become idle, one of them is taken for transmission of an S/F message. When one (or two) of the reserve trunks becomes busy with the direct traffic, S/F transmission is stopped on the trunk on which the next (in time sequence) termination of an S/F message occurs. The flow diagram of the process is shown in Fig. 3. In states S, two trunks are idle; in states W, one trunk is idle; in states R, all trunks are busy. Thus, in state S_z, z trunks are busy with the S/F traffic and $(c - z - 2)$ with the direct traffic; in state W_z, z trunks are busy with the S/F traffic and $(c - z - 1)$ with the direct traffic; in state R_z, z trunks are busy with the S/F traffic and $(c - z)$ with the direct traffic. The transition probability coefficients are as given in the Fig. 3.

Three basic equations are presented now, which state that the net transition flow through three cut-sets is zero under the steady-state conditions. The three cut-sets are indicated in Fig. 3 by lines of dots.

$$aS_z + (z + 1)rR_{z+1} = (c - z - 1)W_z + (c - z - 1)S_{z-1} \tag{21}$$

$$(c - z - 1)S_{z-1} = zrR_z + zrW_z \tag{22}$$

$$(c - z + zr)R_z = [c + z(r - 1)]R_z = aW_z. \tag{23}$$

On the basis of (23), (22) can be rewritten as follows:

$$\begin{aligned}(c - z - 1)S_{z-1} &= zr\left[1 + \frac{c + z(r - 1)}{a}\right]R_z \\ &= zr\left[1 + \frac{a}{c + z(r - 1)}\right]W_z\end{aligned} \tag{24}$$

and similarly,

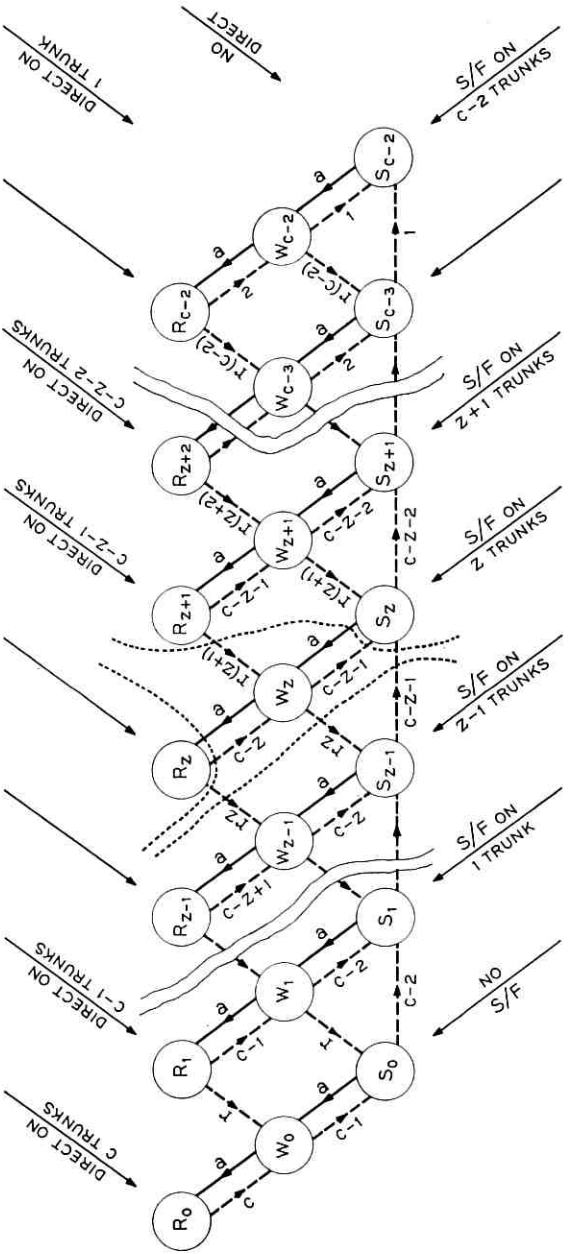


Fig. 3 — Markov chain for trunk occupancy under lost-calls-cleared assumption, dual use of trunks, two trunks in reserve.

$$(c - z - 2)S_z = (z + 1)r\left[1 + \frac{c + (z + 1)(r - 1)}{a}\right]R_{z+1}$$
$$= (z + 1)r\left[1 + \frac{a}{c + (z + 1)(r - 1)}\right]W_{z+1}. \tag{25}$$

Introducing the relation from (25) into the left side of (21) and the relation from (24) into the right side of (21), we obtain

$$a\left[1 + \frac{c - z - 2}{a + c + (z + 1)(r - 1)}\right]S_z$$
$$= \left[(c - z - 1) + zr\left(1 + \frac{a}{c + z(r - 1)}\right)\right]W_z \tag{26}$$
$$= \left[c + z(r - 1) - 1 + \frac{zra}{c + z(r - 1)}\right]W_z.$$

We assume R_0 is known and proceed to derive the other probabilities in the following manner: W_z is determined from R_z through the use of (23), which is rewritten in a convenient form as (27). S_z is determined from W_z through the use of (26), which is rewritten in a convenient form as (28). R_{z+1} is determined from S_z through the use of (25), which is rewritten in a convenient form as (29):

$$W_z = \frac{[c + z(r - 1)]}{a}R_z \tag{27}$$

$$S_z = \frac{\left[c + z(r - 1) - 1 + \dfrac{zra}{c + z(r - 1)}\right]}{a\left[1 + \dfrac{c - z - 2}{a + c + (z + 1)(r - 1)}\right]}W_z \tag{28}$$

$$R_{z+1} = \frac{(c - z - 2)a}{(z + 1)r[a + c + (z + 1)(r - 1)]}S_z. \tag{29}$$

We obtain, using (27), for $z = 0$,

$$W_0 = \frac{[c + 0(r - 1)]}{a}R_0. \tag{30}$$

Using (28), for $z = 0$,

$$S_0 = \frac{\left[c + 0(r-1) - 1 + \dfrac{0ra}{c + 0(r-1)}\right]}{a\left[1 + \dfrac{c - 0 - 2}{a + c + 1(r-1)}\right]} W_0$$

$$= \frac{\left[c + 0(r-1) - 1 + \dfrac{0ra}{c + 0(r-1)}\right]}{\left[1 + \dfrac{c - 0 - 2}{a + c + 1(r-1)}\right]} \frac{[c + 0(r-1)]}{a^2} R_0. \tag{31}$$

Using (29), for $z = 0$,

$$R_1 = \frac{(c - 0 - 2)a}{1r[a + c + 1(r-1)]} S_0$$

$$= \frac{\left[c + 0(r-1) - 1 + \dfrac{0ra}{c + 0(r-1)}\right]}{\left[1 + \dfrac{c - 0 - 2}{a + c + 1(r-1)}\right]} \frac{[c + 0(r-1)]}{a} \cdot \frac{c - 0 - 2}{1r[a + c + 1(r-1)]} R_0. \tag{32}$$

Using (27), for $z = 1$,

$$W_1 = \frac{[c + 1(r-1)]}{a} R_1. \tag{33}$$

Using (28), for $z = 1$,

$$S_1 = \frac{c + 1(r-1) - 1 + \dfrac{1ra}{c + 1(r-1)}}{a\left[1 + \dfrac{c - 1 - 2}{a + c + 2(r-1)}\right]} W_1. \tag{34}$$

Finally

$$R_z = \prod_{k=0}^{z-1} \frac{c + k(r-1)}{a} \prod_{k=0}^{z-1} \frac{\left[c + k(r-1) - 1 + \dfrac{kra}{c + k(r-1)}\right]}{a\left[1 + \dfrac{c - k - 2}{a + c + (k+1)(r-1)}\right]} \cdot \prod_{k=0}^{z-1} \frac{(c - k - 2)a}{(k+1)r[c + (k+1)(r-1)]} R_0 \tag{35}$$

$$W_z = \prod_{k=0}^{z} \frac{c + k(r-1)}{a} \prod_{k=0}^{z-1} \frac{\left[c + k(r-1) - 1 + \dfrac{kra}{c + k(r-1)}\right]}{a\left[1 + \dfrac{c-k-2}{a + c + (k+1)(r-1)}\right]} \cdot \prod_{k=0}^{z-1} \frac{(c-k-2)a}{(k+1)r[c + (k+1)(r-1)]} R_0 \tag{36}$$

$$S_z = \prod_{k=0}^{z} \frac{c + k(r-1)}{a} \cdot \prod_{k=0}^{z} \frac{\left[c + k(r-1) - 1 + \dfrac{kra}{c + k(r-1)}\right]}{a\left[1 + \dfrac{c-k-2}{a + c + (k+1)(r-1)}\right]} \cdot \prod_{k=0}^{z-1} \frac{(c-k-2)a}{(k+1)r[c + (k+1)(r-1)]} R_0 . \tag{37}$$

Now R_0 can be determined, since

$$\sum_{z=0}^{c-2} R_z + \sum_{z=0}^{c-2} W_z + \sum_{z=0}^{c-2} S_z = 1. \tag{38}$$

The probability of loss $D(c,a,r,2)$ is equal to the probability that all trunks are busy

$$D(c,a,r,2) = \sum_{z=0}^{c-2} R_z . \tag{39}$$

When values for R_z obtainable from (35), (36), (37), and (38) are used in (39), it can be regarded as the dual-use-of-trunks formula $D(c,a,r,2)$ for c trunks, a erlangs of offered direct load, ratio of holding times r, and two trunks in reserve.

The amount b of the S/F traffic in erlangs that can be accommodated is given by

$$b = \sum_{z=0}^{c-2} zR_z + \sum_{z=0}^{c-2} zW_z + \sum_{z=0}^{c-2} zS_z . \tag{40}$$

It should be pointed out that for $r \to \infty$, the probabilities $R_z \to 0$ as $1/r^2$ (except for R_0) and the probabilities $W_z \to 0$ as $1/r$ (except for W_0). In this limiting case, the probability of loss tends to $B(c,a)$ with the difference — i.e., the impairment due to S/F traffic — being of the order $1/r^2$:

$$\lim_{r \to \infty} [D(c,a,r,2) - B(c,a)] = 0\left(\frac{1}{r^2}\right). \tag{41}$$

The probabilities R_0, W_0, S_0, S_1, $\cdots$ S_{c-2} tend to become, respectively,

G_c , G_{c-1} , G_{c-2} , G_{c-3} $\cdots$ G_0 . The load b becomes

$$b = c - 2 - a[1 - D(c,a,r,2)] + 2G_c + G_{c-1} . \qquad (42)$$

This can be seen by the following reasoning. The load in terms of trunks occupied is $(c - 2)$ trunks at all times, one additional trunk with the probability G_{c-1} and two additional trunks with the probability G_c . The erlang load in terms of traffic carried is b for the S/F and $a[1 - D(c,a,r,2)]$ for the direct traffic. Equating the two results yields (42).

III. RESULTS AND CONCLUSIONS

The recurrence relations for trunk occupancy probabilities as derived above were programmed on an IBM 704 computer by Miss B. Berman. Two types of programs were written:

i. The grade of service of direct traffic and the amount of S/F traffic accommodated are calculated for a given amount a of direct traffic, a given number of trunks in a trunk group, and either one or two trunks in reserve.

ii. The amount a of direct traffic and the amount of S/F traffic are calculated for a given grade of service assigned to direct traffic, a given number of trunks in a trunk group, and either one or two trunks in reserve. This computer program is similar to (*i*) except that it involves successive approximations in the amount of direct traffic to obtain the desired grade of service.

These computer programs were used for trunk groups ranging in size from 2 to 48 trunks. Special attention was given to one ratio of the average holding time of direct traffic to the average holding time of S/F traffic, $r = T/t_0 = 300/3.6 = 83.3$. This ratio was selected to correspond to the average holding times which were originally expected in the UNICOM system, for which the study was initially done. For the direct traffic, $T = 5$ minutes $= 300$ seconds, and for the S/F traffic, $t_0 = 3.6$ seconds. This last figure was obtained by assuming 150 words for the average message length, 42-bits-per-word digital coding, and 2400-bits-per-second transmission speed. Approximately one second is allowed in the total for the average time required to establish the connection. Thus, on the average, for one fully loaded trunk, 1000 S/F messages per hour is the equivalent of only 12 direct calls.

Calculations in the first type of program above were made for offered loads of direct traffic determined by the formula $B(c,a) = 0.05$. The resulting grade of service of direct traffic and the amount of S/F traffic carried for $r = 83.3$ are presented in Table I. (The grade of service is plotted as a solid line in Figs. 4 and 5.) It can be seen that substantial

TABLE I — $D(c,a,\ 83.3,s)$ AND b FOR $B(c,a)\ =\ 0.05$

c	a	One Trunk in Reserve		Two Trunks in Reserve	
		$D(c,a,83.3,1)$	b	$D(c,a,83.3,2)$	b
2	0.38132	0.05255	0.69127	—	—
3	0.89940	0.05477	1.20463	0.05003	0.41676
4	1.52462	0.05670	1.61853	0.05008	0.79092
5	2.21848	0.05842	1.96954	0.05014	1.11664
6	2.96033	0.05996	2.27714	0.05021	1.40371
7	3.73782	0.06137	2.55294	0.05028	1.66049
8	4.54297	0.06267	2.80439	0.05036	1.89308
9	5.37025	0.06387	3.03662	0.05044	2.10599
10	6.21572	0.06500	3.25327	0.05052	2.30257
12	7.95007	0.06704	3.64998	0.05069	2.65643
16	11.54361	0.07053	4.34111	0.05105	3.25047
20	15.24928	0.07344	4.94398	0.05140	3.74232
24	19.03073	0.07592	5.49001	0.05176	4.16546
30	24.80184	0.07908	6.23847	0.05229	4.71144
36	30.65736	0.08173	6.92986	0.05279	5.18284
48	42.53693	0.08598	8.20641	0.05374	5.98058

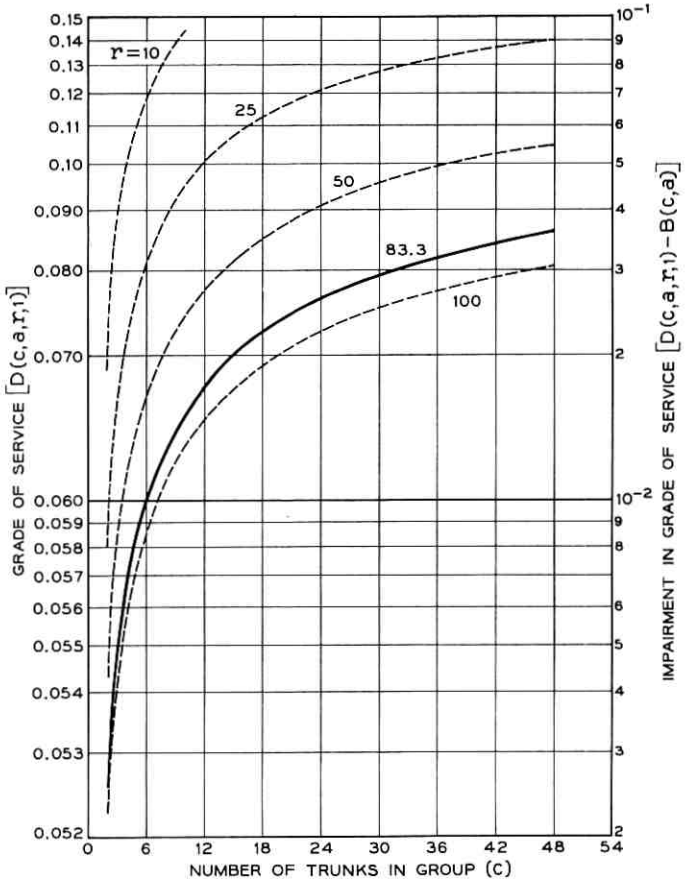


Fig. 4 — Grade of service $D(c,a,r,1)$ with a given by $B(c,a)\ =\ 0.05$, vs number of trunks c.

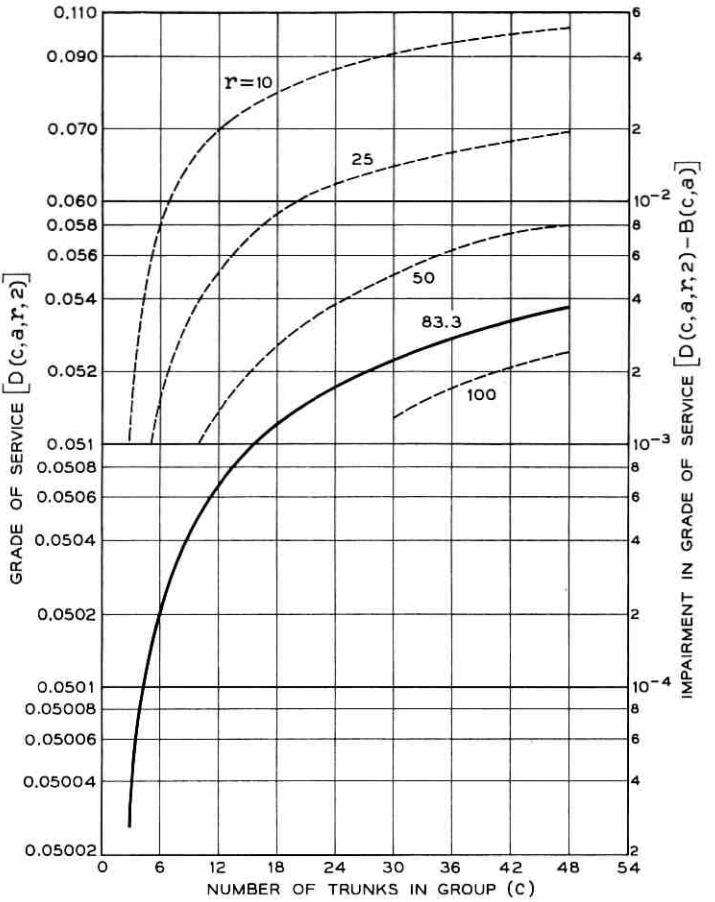


Fig. 5 — Grade of service $D(c,a,r,2)$ with a given by $B(c,a) = 0.05$, vs number of trunks c.

amounts of S/F traffic can be carried. When one trunk is held in reserve, the amount ranges from 0.7 erlang (700 messages per hour) in the case of a trunk group with 2 trunks, to 8.2 erlangs (8200 messages per hour) in the case of a trunk group with 48 trunks. When two trunks are held in reserve, the amount ranges from 0.42 erlang (420 messages per hour) in the case of a trunk group with 3 trunks, to 6.0 erlangs (6000 messages per hour) in the case of a trunk group with 48 trunks.

The impairment in grade of service to direct traffic — i.e., the difference between the grade of service under the dual use of trunks and the

grade of service when no S/F traffic is sent (for the same amount of direct traffic) — is rather small when 1 trunk is held in reserve and practically insignificant when 2 trunks are held in reserve. The impairment increases from 0.003 for the 2-trunk group to 0.036 for the 48-trunk group when 1 trunk is held in reserve, and from 0.00003 for the 3-trunk group to 0.0037 for the 48-trunk group when 2 trunks are held in reserve.

Similar calculations were carried out for different ratios r of holding times, ranging from 100 to 1. The grade of service to the direct traffic is plotted, with r as a parameter, as dashed lines in Figs. 4 and 5. The impairment increases rapidly with decreasing r. If the impairment were plotted as a function of r, it can be seen that for $r > 10$ the impairment obeys closely the asymptotic functional relation $1/r$ of (20) for $s = 1$ and $1/r^2$ of (41) for $s = 2$. For lower r, the impairment increases more slowly than $1/r$ for $s = 1$, or $1/r^2$ for $s = 2$. The capacity b for carrying S/F traffic increases slowly with the increase of probability of loss, which is indicated for $s = 1$ by (19).

Calculations in the second type of program were used to obtain the offered amount of direct traffic, which will result in grade of service 0.05 under the dual use of trunks: i.e., load a in erlangs determined by $D(c,a,83.3,1) = 0.05$ and by $D(c,a,83.3,2) = 0.05$. The amount b of S/F traffic was computed at the same time. These results are presented in Table II and Fig. 6. Comparing dual use of trunks having 1 trunk in reserve with use of trunks for direct traffic only at the same grade of service, it can be seen that for the 2-trunk group, 0.7 erlangs of S/F

TABLE II — a AND b FOR $D(c,a,83.3,s) = 0.05$

c	One Trunk in Reserve $D(c,a,83.3,1) = 0.05$		Two Trunks in Reserve $D(c,a,83.3,2) = 0.05$	
	a	b	a	b
2	0.36953	0.69894	—	—
3	0.85954	1.23344	0.89912	0.41686
4	1.44427	1.67795	1.52364	0.79143
5	2.08803	2.06637	2.21629	1.11790
6	2.77203	2.41657	2.95642	1.40608
7	3.48519	2.73907	3.73167	1.66431
8	4.22041	3.04060	4.53409	1.89868
9	4.97292	3.32572	5.35812	2.11369
10	5.73925	3.59771	6.19983	2.31268
12	7.30392	4.11127	7.92525	2.67221
16	10.52058	5.05545	11.49555	3.28063
20	13.80962	5.93086	15.17145	3.79025
24	17.14272	6.76441	18.91730	4.23396
30	22.19328	7.96639	24.62530	4.81501
36	27.28129	9.13277	30.40758	5.32546
48	37.51678	11.40906	42.11484	6.21013

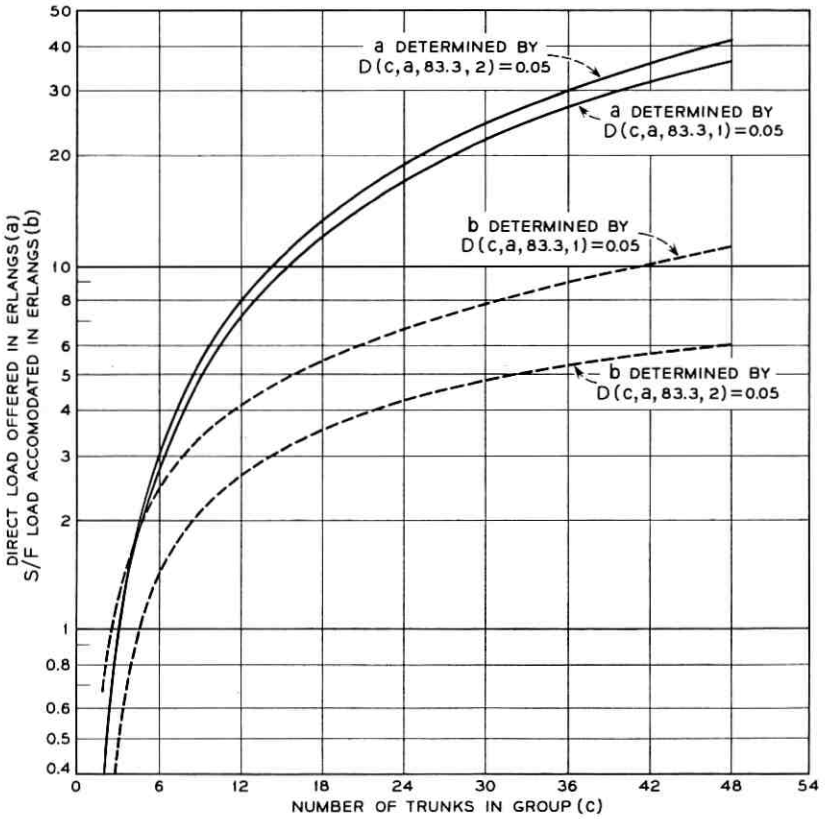


Fig. 6 — Load of S/F traffic b and offered load of direct traffic a defined by $D(c,a,83.3,1) = 0.05$ and by $D(c,a,83.3,2) = 0.05$, vs number of trunks c.

traffic (700 messages per hour) can be sent at the cost of decreasing the offered direct load from 0.381 erlang to 0.369 erlang: i.e. by 0.012 erlang (a fraction 0.14 of the average holding time of a direct call). For the 48-trunk group, 11.4 erlangs of S/F traffic (11,400 messages per hour) can be sent at the cost of decreasing the offered direct load from 42.5 erlangs to 37.5 erlangs: i.e. by 5.0 erlangs (60 direct calls). Making a similar comparison when two trunks are held in reserve, it will be noted that if the offered load a determined by $B(c,a) = 0.05$ were to be plotted in Fig. 6, it would virtually coincide with the plotted offered load a determined by $D(c,a,83.3,2) = 0.05$. Thus, the decrease in direct capacity is negligible. In Table III an example is presented of trunk occupancy probabilities which result from a calculation in the second type of pro-

TABLE III — TRUNK OCCUPANCY PROBABILITIES

z	One Trunk in Reserve $D(5,a,83.3,1) = 0.05$		z	Two Trunks in Reserve $D(5,a,83.3,2) = 0.05$		
	R_z	S_z		R_z	W_z	S_z
0	0.03912	0.09368	0	0.04977	0.11229	0.19610
1	0.00450	0.18807	1	0.00018	0.00688	0.27160
2	0.00338	0.27508	2	0.00004	0.00322	0.24702
3	0.00220	0.26559	3	0.00001	0.00098	0.11190
4	0.00080	0.12758				
Total	0.05000	0.95000		0.05000	0.12337	0.82663

gram; probabilities R_z and S_z for $D(5,a,83.3,1)$ and the probabilities R_z, W_z, and S_z for $D(5,a,83.3,2)$ are given.

These calculations were repeated for different values of r ranging from 100 to 1. The capacity a for carrying direct traffic determined by $D(c,a,r,1) = 0.05$ is plotted in Fig. 7 and by $D(c,a,r,2) = 0.05$ in Fig. 8.

The following conclusions can be drawn from the numerical results. The dual use of long-haul trunks for both direct and S/F traffic can be economically attractive. The calculations indicate that when one or two trunks are held in reserve for possible arrivals of direct traffic, it is possible to send large amounts of S/F traffic on the same trunk groups as direct traffic with little or negligible impairment of the direct traffic, provided the parameters of group size and the number of trunks held in reserve are properly selected. The dual use of trunks is especially attractive if the average holding time of the S/F traffic is short compared with the average holding time of the direct traffic. When the holding time of the S/F traffic approaches the average holding time of the direct traffic, the capacity for carrying direct traffic deteriorates sharply, unless two trunks (or more, for large trunk-groups) are held in reserve.

The potential economies are believed to be of greatest importance in systems with small trunk groups. A manipulation of the data (1 trunk in reserve) will serve to point out that the gain in trunk group efficiency is very substantial in the smaller groups and decreases as the number of trunks increases. The trunk efficiency is the load carried divided by the number of trunks. The load carried is $[1 - B(c,a)]a = 0.95a$ (from Table I) when no S/F traffic is present, and

$$[1 - D(c,a,r,1)]a + b = 0.95a + b$$

(from Table II) under the dual use of trunks. The 2-trunk group increases in efficiency from 18 to about 52 per cent, the 10-trunk group from 59

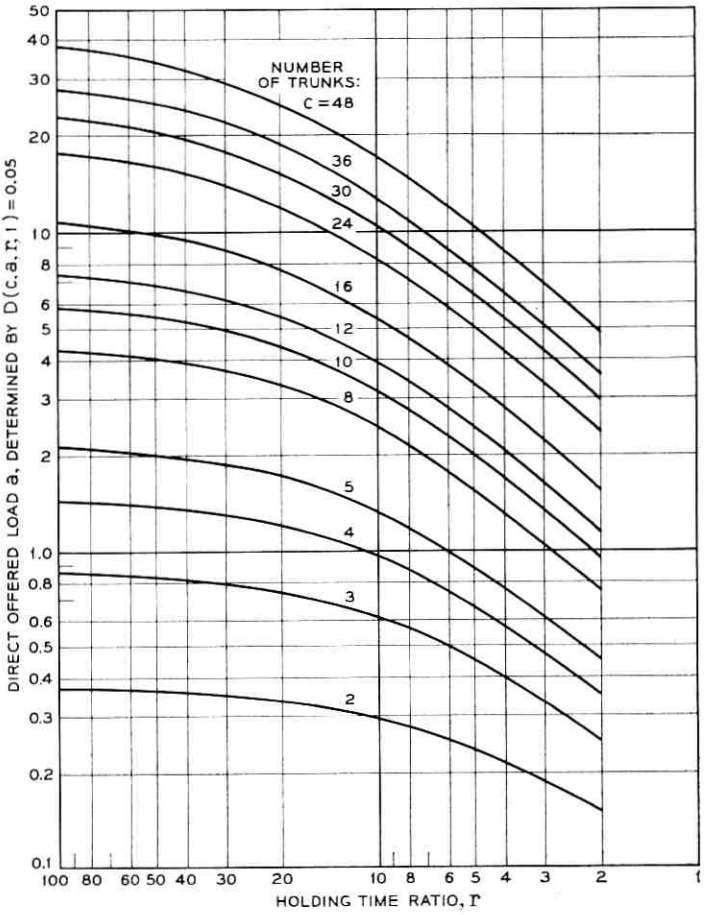


Fig. 7 — Offered load of direct traffic a defined by $D(c,a,r,1) = 0.05$, vs ratio of holding times r.

to about 90 per cent, and the 48-trunk group from 84 to about 98 per cent. These advances are significant in terms of possible savings in long-haul transmission plant.

This method of operating circuit groups at such a high level of occupancy remains to be evaluated in terms of the grade of service of the S/F user. The grade of service of the S/F traffic has not been discussed here.

The analysis given here was confined to a single trunk group carrying

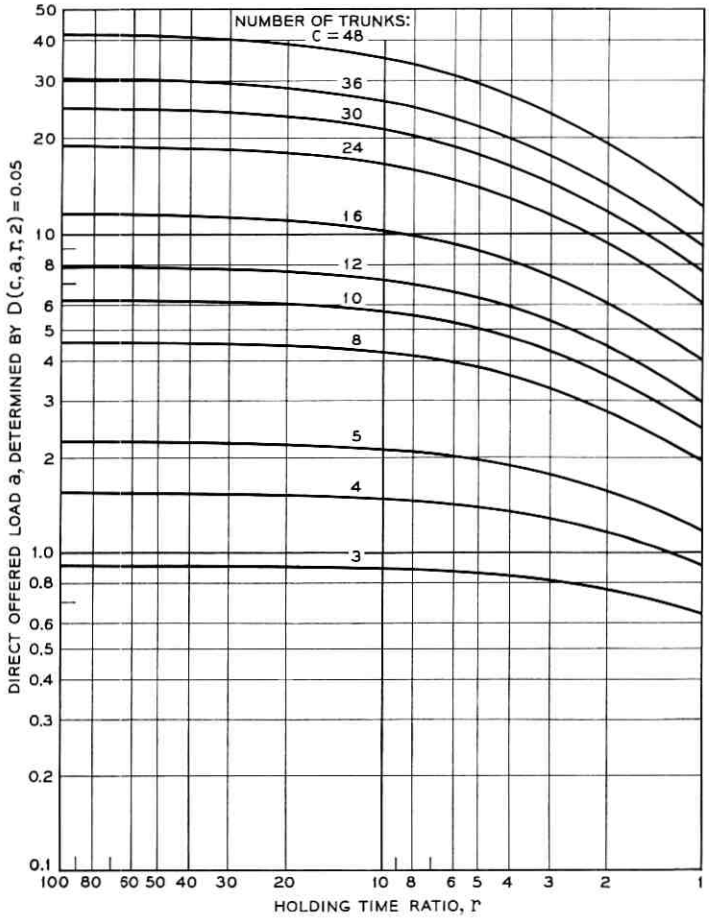


Fig. 8 — Offered load of direct traffic a defined by $D(c,a,r,2) = 0.05$, vs ratio of holding times r.

traffic between two points. The problems of alternate routing and through-switching of traffic are yet to be explored.

IV. ACKNOWLEDGMENT

The author wishes to acknowledge helpful discussions with F. Assadourian, R. C. Pfarrer, J. H. Weber and especially H. W. Townsend. The preparation of data by L. A. Gimpelson and the meticulous computer programming by Miss Beverly Berman are very much appreciated.

APPENDIX

Glossary of Symbols

a	Offered amount of direct traffic, in erlangs, during busy hour.
b	Amount of S/F traffic per busy hour that can be accommodated, in erlangs.
c	Number of trunks in a trunk group.
s	Number of trunks held in reserve.
t_0	Average holding time of S/F traffic.
T	Average holding time of direct traffic.
r	Ratio of the average holding time of direct traffic to the average holding time of S/F traffic, T/t_0.
$B(c,a)$	Grade of service to direct traffic on lost-calls-cleared basis, with c trunks in the trunk-group and an offered load a.
$D(c,a,r,s)$	Grade of service to direct traffic on lost-calls-cleared basis, with c trunks in the trunk-group, offered load a of direct traffic, ratio of holding times r, and s trunks in reserve for possible arrivals of direct traffic.
G_x	Probability of exactly x trunks being occupied by direct traffic.
n	Expected number of arrivals of direct traffic during the busy hour.
R_z	Probability of, or state of, all-trunks-busy, with S/F traffic on z trunks.
S_z	Probability of, or state of, specified number of trunks (one when $s = 1$, two when $s = 2$) in reserve, with S/F traffic on z trunks.
W_z	Probability of, or state of, one trunk in reserve, with S/F traffic on z trunks (under the operating method where $s = 2$).
x	Number of trunks occupied by direct traffic.
z	Number of trunks occupied by S/F traffic.

REFERENCES

1. Feller, William, *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, 1959, pp. 411–412.
2. Brockmoyer, et al, *The Life and Works of A. K. Erlang*, Copenhagen Telephone Company, Denmark, 1948.

Over-All Characteristics of a TASI System

By J. M. FRASER, D. B. BULLOCK, and N. G. LONG

(Manuscript received September 19, 1961)

TASI (Time Assignment Speech Interpolation) has been in service on transatlantic submarine cable channels since mid-1960. Measurement of service quality on one TASI system (White Plains–London) indicates that system performance equals or exceeds the original engineering objectives in all but a few cases. Field modifications now being made should bring these exceptions into closer agreement with objectives.

A companion paper[4] discusses in detail the design considerations for TASI speech detectors and describes subjective tests made to determine the maximum permissible loading of TASI circuits without impairment of service.

TASI, an abbreviation of *T*ime *A*ssignment *S*peech *I*nterpolation, is a high-speed switching and transmission system which uses the idle time in telephone calls to interpolate additional talkers.[1,2] In a normal telephone conversation each subscriber speaks less than half of the time. The remainder of the time is composed of listening, gaps between words and syllables, and pauses while the operator or subscriber leaves the line. Measurements on working transatlantic channels, Fig. 1, show that a TASI speech detector with a sensitivity of −40 dbm is operated by speech from one talker on the average about 40 per cent of the time the circuit is busy at the switchboard. Since long distance circuits use separate facilities for the two directions of transmission, each one-way channel is, on the average, free about 60 per cent of the time.

In order to take advantage of this free time to interpolate additional conversations, a considerable group of channels must be available. An attempt to interpolate two independent conversations on a single channel would result in a large percentage of the speech being lost, since the probability of both talkers speaking at the same time is high. However, with a large group of channels serving a larger group of talkers, the variations in demand become much smaller. Even with 74 talkers on 37 channels, the percentage of speech lost (freeze-out fraction) is reduced to a point where there is no noticeable effect on continuity of conversation.

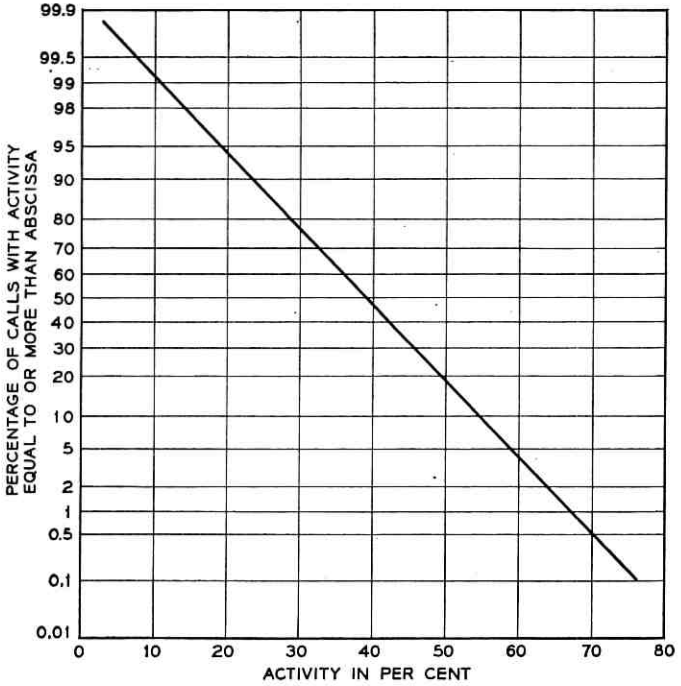


Fig. 1 — Circuit activity.

The increase in channel capacity with TASI is illustrated in Fig. 2, which gives the TASI advantage (ratio of switchboard positions, or trunks, to channels) for a range of activities and number of channels. A freeze-out fraction of 0.5 per cent has been assumed for each curve, since this amount of speech loss has been found from tests to have a negligible effect on transmission quality.

It will be noted in Fig. 2 that a TASI advantage of at least two can be obtained on a 37 channel group as long as the average activity is not significantly greater than 40 per cent. TASI is designed to use 36 channels for speech interpolation and one additional channel as a control channel for transmitting disconnect and error checking signals. This fits the needs of present day submarine cable systems, since 37 is close to the maximum number of channels that can be made available for TASI out of the total 48 channels derived by submarine cable type channel equipment[3] employing 3-kc filter spacing. The remainder of the channels are required for special services such as program material and certain types of data which are ordinarily not transmitted through TASI.

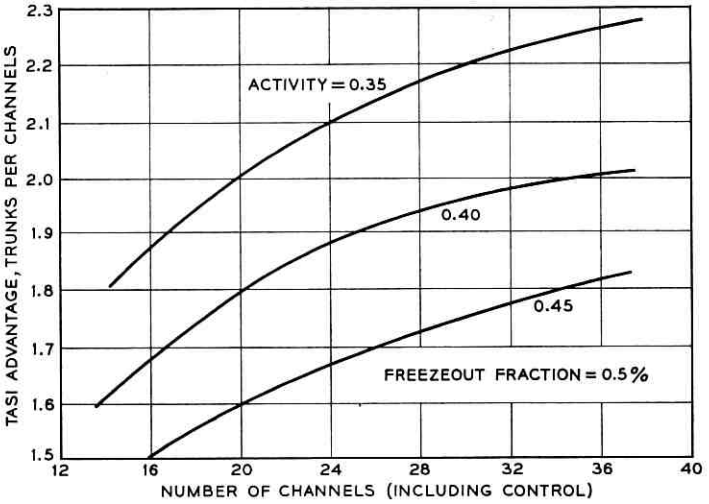


Fig. 2 — TASI advantage.

The first TASI system was put in service in June, 1960 on the transatlantic cable system between White Plains and London (TAT-1) and the second followed a few months later on the cable between New York and Paris (TAT-2). TASI is well suited to submarine cable application for several reasons. Although TASI requires considerable expensive terminal equipment, it is an economical means of doubling the number of conversations that can be handled on expensive submarine cable facilities. In addition, TASI is easier to apply to long submarine cables than to the land plant with its many branching points and alternate routes.

The principal purpose of this paper is to describe the application of TASI to the White Plains–London submarine cable system. The channel requirements that must be met for TASI operation are detailed along with measurements of the combined system characteristics, such as noise, bandwidth, etc. Measurements of the amount of speech lost in an actual working system are compared with earlier theoretical computations.

I. DESCRIPTION OF OVER-ALL SYSTEM

The message channels on the land portions of the TAT-1 system employ standard 4-kc channel spacing, and the undersea channels are spaced at 3-kc intervals. The land and undersea-type channels are interconnected at voice frequencies at Sydney Mines, Nova Scotia and Oban,

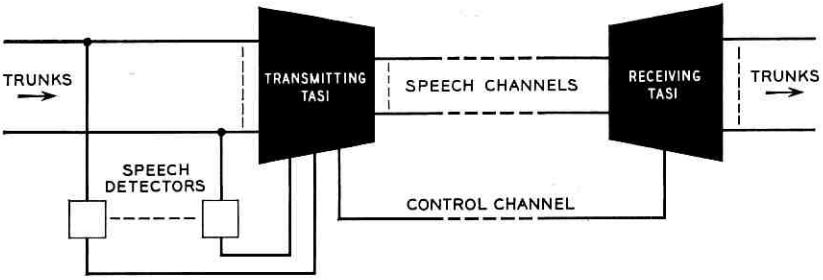


Fig. 3 — TASI equipment — one direction of transmission.

Scotland. A typical message circuit between the TASI terminals at White Plains and London extends from about 275 to 3150 cps.

Fig. 3 shows a block diagram of the TASI equipment for one direction of transmission; an independent TASI is used in the opposite direction of transmission. Presence of speech on a trunk causes the speech detector to operate, initiating a request for a channel. The transmitting common control equipment selects an idle channel, if one exists, and assigns it to the requesting trunk. Before the talker is connected to the channel, a "connect" signal is sent over the assigned channel specifying the trunk to be connected to that channel at the distant receiving terminal. During the time required to connect talker and listener the initial part of the talker's speech is clipped. In order to minimize clipping the signaling time has been made as short as possible, 17 ms, consistent with reliable signaling and quiet switching. The signal information consists of a single burst of 4 tones out of a possible 14, ranging in frequency from 615 to 2419 cps. Once a talker is assigned a channel he does not lose the connection as long as he continues talking. When he stops talking he may still retain the connection unless he has to be disconnected to provide a channel for another talker. A similar burst of 4 tones out of 15 (615 to 2501 cps) is used to disconnect the talker and the listener, but this signal is sent over a separate control channel. During periods when no disconnect signals are being used, the same type of code signals are used to send information over the control channel as to the trunk-channel connections existing at the transmitting end. This connection-checking information overrides any earlier information and determines the connection made at the receiver. In addition, a comparison at the receiver between existing and overriding information is used to detect bad channels.

As shown in Fig. 2, the number of trunks which can be served by TASI depends upon the number of channels available. To prevent excessive

speech loss when the connecting channels fail, TASI has been designed to automatically remove bad channels from service; trunks are then removed until the proper trunk-channel ratio is reached.

In addition to the provisions for automatically reducing the number of connected trunks and channels, TASI contains audible and visible alarms to identify internal failures. In the event of a major failure in TASI the terminals automatically switch themselves out at both ends, reducing the number of connected trunks to the number of available channels. TASI is also switched out automatically if both the regular and alternate control channels fail. If only the regular control channel fails, the disconnect and error checking signals are automatically switched to the alternate control channel and TASI will continue to operate with only a momentary interruption.

When TASI is switched out, the voice-frequency amplifiers associated with TASI are also switched out. The schematic relationship of these voice-frequency amplifiers and other transmission equipment is shown for one terminal in Fig. 4 along with typical operating level points. The combination of VF amplifiers, TASI, and appropriate attenuation pads provides a zero-loss device and also provides optimum transmission levels to TASI.

The echo suppressor shown in Fig. 4 performs the usual function of preventing echoes, generated at points of impedance mismatch, from reaching a subscriber's ear and interfering with normal conversation.

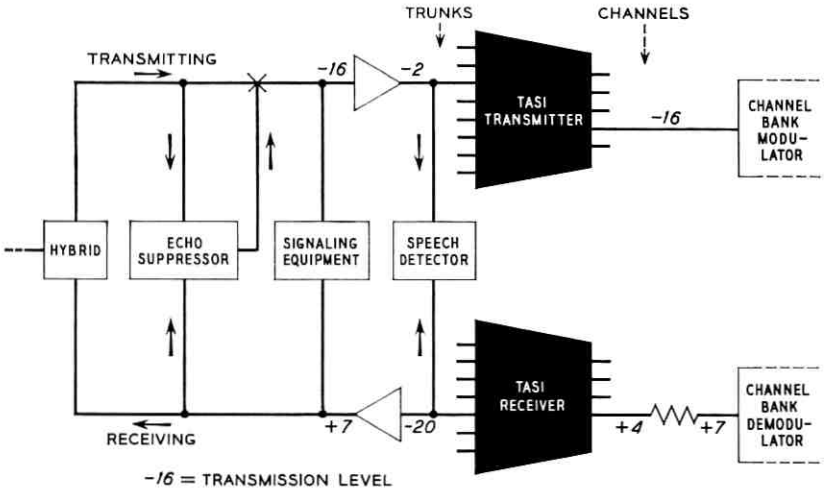


Fig. 4 — TASI and associated equipment.

On TASI circuits, these suppressors must be of the receiving end split type at each end of the circuit to prevent the distant talker's echo from operating the speech detector. The location of a split suppressor is shown schematically in Fig. 4. Speech received from the distant party reaches the echo suppressor and operates it, causing a large loss to be inserted in the transmitting path before the speech detector. This large loss prevents most of the echoes capable of operating the speech detector from reaching the speech detector, but the initial part of the echo may get through because the existing echo suppressors were designed with a slow operate time to minimize operation on line noise. On the other hand, the speech detectors must be relatively fast operating to minimize initial clipping of speech. The result is that the operate time of the echo suppressors is about 7 ms longer than that of the speech detectors. During this interval the speech detector may be operated by echoes before the echo suppressor can operate. To minimize this difficulty, and still use existing echo suppressors, the speech detectors are equipped with fast-acting circuits which reduce the sensitivity of the speech detectors as much as 13 db depending on the energy present on the receiving side of the trunk. This echo protector function of the speech detector reduces the probability of operating during the initial part of the spurt; the large loss inserted later by the echo suppressors prevents operation during the remainder of the spurt, which usually contains the higher energy.

In addition to the equipment shown in Fig. 4, compandors can be applied to noisy channels as necessary to meet noise objectives. Tests have indicated that TASI can signal satisfactorily through Bell System 1A-type compandors or their British equivalent.

1.1 *Toll Signaling and Supervision*

Because TASI is a time sharing device, there are problems involved in transmitting supervisory and dialing pulses. TASI can work satisfactorily with the present ringdown manual arrangement, but it is obvious that the usual method of continuous supervision by means of a steady tone during the idle time cannot be used. Likewise, dial pulses cannot compete for a TASI channel on the same basis as a talker, because TASI clipping would cause signaling errors. A burst signaling system is required.

II. ENGINEERING OBJECTIVES FOR TASI

In order that TASI could operate over existing telephone facilities and would fit in with existing performance standards, certain engineering

TABLE I — ENGINEERING OBJECTIVES FOR TASI

Capacity	At least 72 message trunks to be operated over 37, 3-kc spaced cable channels. If the number of available channels is less than 37, the number of trunks to be provided by TASI will be less, as illustrated in Fig. 2. If the total busy-hour speech activity is increased above about 40%, the maximum number of trunks to be provided will be less.
Speech quality	With the TASI system fully loaded as defined above, the degradation to speech quality due to TASI should not exceed about 1 db. When the number of talkers equals the number of available channels, the TASI degradation should be close to 0 db.
Signaling errors	On the average, during the busy hour, no more than 0.01% of the talkspurts transmitted should be lost because of signaling errors if the transmission medium meets the objectives noted in Table II. Assuming that the average activity is 40%, this means about one talkspurt lost in thirty average 10-minute calls.
Reliability	The reliability objective is that the amount of time trunks are removed from service because of TASI failure shall be less than 0.1% of the total time.
Frequency response	The TASI transmitter and receiver connected back to back should pass a band of 200–3500 cps. The average variation from flatness of all the channels should be within ±0.5 db over this frequency range. In addition the standard deviation of the variations from the average should not exceed 0.2 db.
Net loss	The net loss at 1000 cps through the TASI equipment alone should be adjustable to within 0.15 db of 0 db and should stay within ±0.15 db of the adjusted value for at least one month.
Circuit	The noise generated by TASI in the transmission path should not exceed about 12 dba as measured at the zero level points.
Crosstalk	To provide adequate crosstalk performance, an equal level coupling loss of 70 db should be obtained between talking paths in TASI. This applies to both near-end and far-end crosstalk.

objectives were set up to guide the planning and development of TASI. They were considered as reasonable goals rather than rigid requirements. These objectives are listed in Table I.

III. CHARACTERISTICS OF CHANNELS FOR TASI

Because of the high-speed signaling used in TASI and because subscribers are switched rapidly between channels, the transmission requirements of the channels connecting TASI terminals are somewhat tighter than required for the usual telephone message service. The characteristics of importance to TASI are listed in Table II.

TABLE II — REQUIRED TRANSMISSION CHARACTERISTIC OF CONNECTING CHANNELS FOR TASI

Minimum bandwidth	300 cps–2900 cps (10-db cutoff frequencies).
Flatness of band 565 to 2550 cps	Difference between maximum and minimum loss should not exceed 2.5 db.
1-kc net loss value of any channel	Not more than ±3 db from the nominal value.
Envelope delay distortion 565 to 2550 cps	Not greater than 2 ms.
Flat delay	Maximum difference between channels should be no more than 10–15 ms at 1000 cps. (The control channel should be one of the fastest.)
RMS noise	Without compandors, 38 dba at zero transmission level (38 dba 0). With compandors, noise on line ahead of compandors should not exceed 51 dba 0. Difference in output noise between channels should not exceed 6 db.
Crosstalk	Equal level crosstalk loss on all channels should be at least 60 db.
Frequency stability (565 to 2550 cps)	No frequency shifted more than 2 cps.
Working levels for TASI (excluding pads or amplifiers outside of TASI)	Transmitting terminal input, −2 with respect to zero transmission level point. Receiving terminal input, +4 with respect to zero transmission level point.

IV. MEASURED TRANSMISSION CHARACTERISTICS

After TASI was installed on TAT-1 extensive measurements and tests were made to determine how closely TASI came to meeting the engineering objectives and how TASI operated in the environment of the telephone plant. The results of the tests and measurements are given briefly in the following paragraphs.

4.1 *Frequency Response*

Fig. 5 shows the frequency response of a TAT-1 message channel, with and without TASI. Figure 5 applies to most connection channels, although a few channels differ significantly due to channel bank and pilot filters. The sharp cutoff frequencies of the channel, about 275 and 3150 cps, are principally due to the 3-kc submarine cable terminal equipment. It can be seen that TASI does not affect the frequency response of the system appreciably.

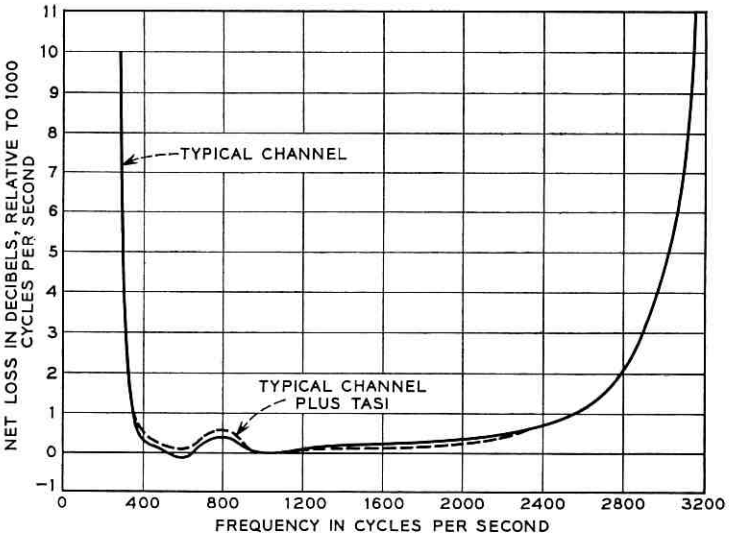


Fig. 5 — Frequency response — TAT-1.

4.2 *Net Loss Variations*

The net loss variations on TAT-1 are kept to a minimum through the use of pilot-controlled automatic gain controls. An associated alarm reporting system informs all important stations automatically if the pilot levels deviate beyond prescribed limits.

Fig. 6 shows a distribution of the difference in net loss of successive talkspurts experienced by a typical White Plains–London subscriber talking through TASI during the busy hour. These changes are due to differences in the net losses of the channels and the various paths through TASI. Only 8 per cent of the changes were greater than about 3 db, which is just noticeable. There were no changes greater than about 5 db. During periods of light traffic, the number of switches between channels, and the number of net loss changes, will decrease and cease entirely if traffic is very light.

4.3 *Envelope Delay Distortion*

A fixed delay equalizer was added to each of the TAT-1 message channels to reduce the delay distortion. In addition, the voice-frequency interconnections between land and undersea channels were arranged to avoid combinations having excessive delay distortion. The median

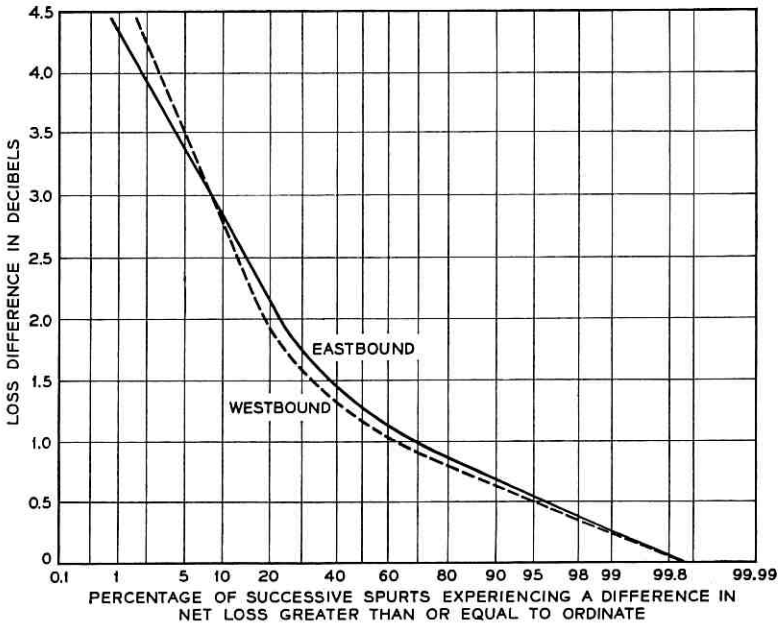


Fig. 6 — Net loss differences.

value of delay distortion of the resulting White Plains–London message channels is under 0.5 ms, with no channel being worse than 1.6 ms, within the TASI signaling band (565 to 2550 c/s).

Fig. 7 shows the delay distortion of a typical message channel before equalization, and after the addition of equalizers plus TASI. Channels located in the frequency spectrum near the cutoffs of group connector filters and pilot-frequency filters have delay distortion characteristics significantly different from the example on Fig. 7, particularly at the lower and upper edges of the passband.

4.4 *Noise*

TASI contributes very little to the over-all system noise; the average of the TAT-1 channels is about 36 dba at zero transmission level (abbreviated "36 dba 0"), and the average noise generated by TASI plus amplifiers is less than 12 dba 0. Since noise is increasing slowly on the transatlantic circuits caused by the change in cable characteristics with time, compandors have been installed on some channels. This will not influence the operation of TASI.

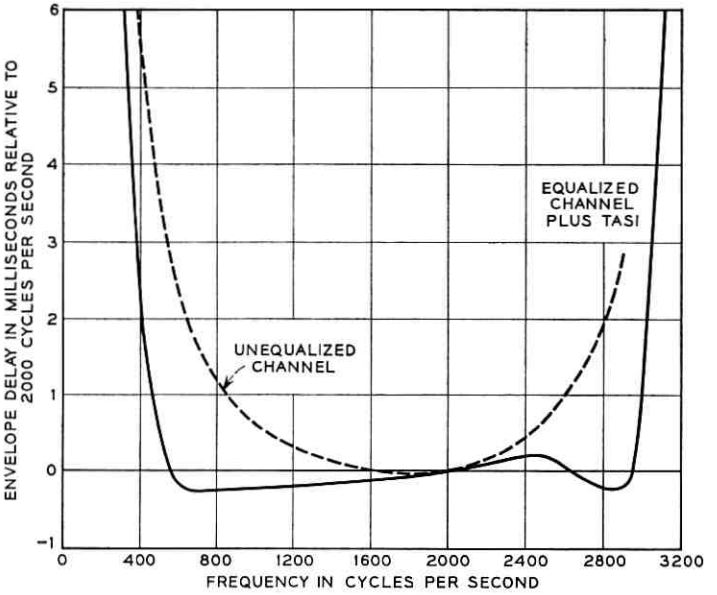


Fig. 7 — Envelope delay distortion typical TAT-1 Channel.

When a channel is disconnected during a conversation, the listener may notice a change in the background noise, because he no longer hears the channel noise of the whole system but instead hears only the noise of the trunk connecting him to TASI. To prevent the subscriber from feeling that he has lost his circuit due to the sudden noise change between successive connections, the receiving TASI terminal transmits random noise of about 33 dba 0 toward the listener whenever his trunk is not connected to a channel.

V. SPEECH CLIPPING

In addition to the measurements described above, measurements were made of the amount of speech clipped from subscribers making calls through the TAT-1 TASI system during the busy hour. The measured values were compared with computed values which were based on measured speech activities and talkspurt lengths.

The computations took into account the two major components of speech clipping in TASI, which are:

1. Signaling clipping, which is the time lost (17 ms) while a new connection is established;

2. Freeze-out, which is the time lost because no channels are available.

The length and frequency of these clips will vary from call-to-call due to loading, speech habits, or statistical chance.

In addition to signaling and freeze-out, some clipping is also caused by (a) speech-detector response time and threshold; (b) disconnection delays caused by control channel crowding during heavy loading. However, tests[4] have shown that the TASI speech detector introduces negligible speech impairment, and as described later it is estimated that control channel overload contributes very little clipping.

The results of the computations for various trunk-channel combinations are shown in Fig. 8 for the median call, and in Fig. 9 for the worst 1 per cent case. TASI can operate with a maximum of 36 speech channels plus one for control, but because of the demands for special services, 37 channels are not available on all systems for TASI. Results are shown, therefore, assuming different numbers of channels are available to TASI.

The right-hand scale in Figs. 8 and 9 gives the estimated db impairment corresponding to the speech loss shown by the left-hand scale. The upper solid curves assume that during the busy hours the circuits are carrying calls 100 per cent of the time (A = 1.0); the lower dotted curves assume that over the busy hours the average loading of the circuits is 0.85.

As shown in Fig. 8, for example, a subscriber in a 74 trunk, 36 + 1-channel fully-loaded system will have 1.1 per cent or less of his speech clipped during 50 per cent of his calls; Fig. 9 shows that this same subscriber will lose 2.9 per cent or less of his speech during 1 per cent of his

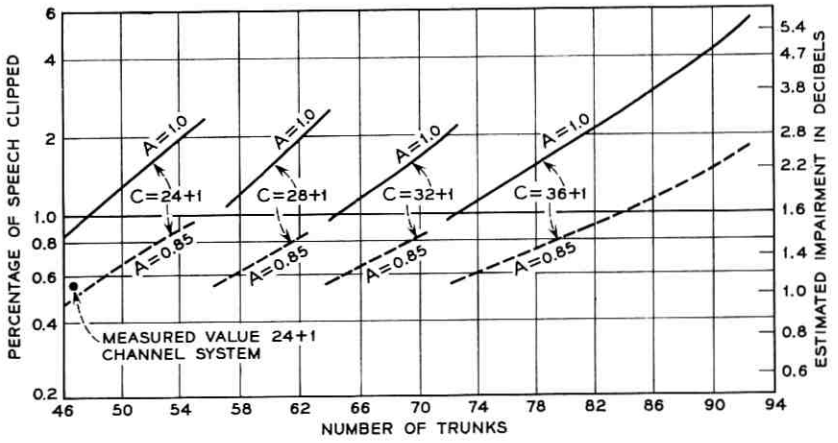


Fig. 8 — Median speech clipping.

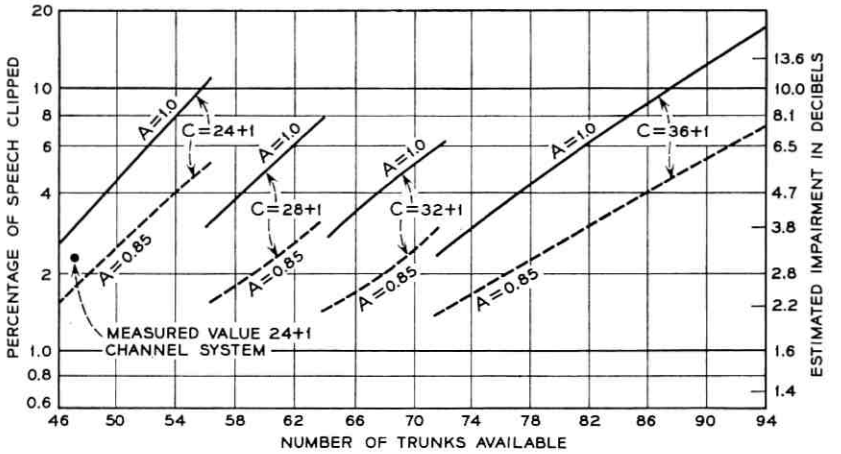


Fig. 9 — Upper 1 per cent speech clipping.

calls. In a TASI system with an average loading of 0.85, the median clipping would be 0.6 per cent or less, and the upper 1 per cent clipping would be 1.6 per cent or less. The loading factor varies with season, world events, etc. In order to be conservative, and to provide good quality speech even during peak periods, the recommended trunk-channel ratios shown in Table III assume 100 per cent loading of all trunks.

TABLE III — RECOMMENDED OPERATING CONDITIONS

Number of connection channels available to TASI	Normal max. no. of talker trunks	Approximate percentage of speech clipped	
		Median-%	1-%
24 + 1	47	1%	3%
28 + 1	56	1	3
32 + 1	65	1	3
36 + 1	74	1	3

The dots in Figs. 8 and 9 for 47 trunks on 24 + 1 channels represent measurements made during the busy hours on the TAT-1 TASI system. Since the results lie between the A = 1 and A = 0.85 computed values, the indications are that the loading of this system, during the busy hours, is between 0.85 and 1.0. Direct measurements of circuit usage were made and confirm that the average lies in this range.

Fig. 10 shows the complete distribution of the measured speech loss during subscriber calls caused by signaling clipping and freeze-out on TAT-1. Both computed and measured curves of the two components

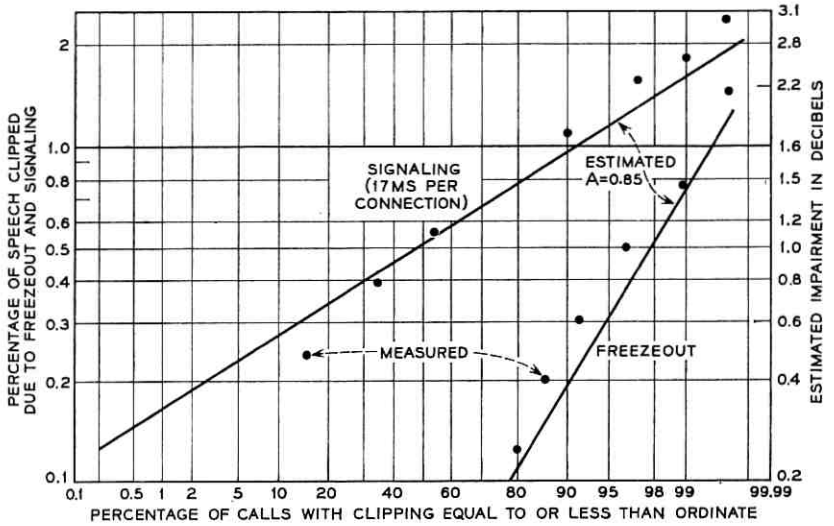


Fig. 10 — Estimated and measured clipping distributions on a 24 + 1 channel 47 trunk TASI system.

are approximately log normal. The good agreement shown in Fig. 10 between computed values (assuming A = 0.85) and the measured values indicates that with respect to freeze-out fraction the operating TASI behaves statistically as expected.

VI. CLIP LENGTHS

Another important characteristic of TASI clipping is the distribution of the lengths of clips a talker experiences. The length of clip a talker receives whenever he receives a new connection is composed of two major parts; the constant signaling clip (17 ms) and the freeze-out, which may vary from 0 ms to 500 ms (upper limit for all practical purposes), depending upon the instantaneous load. The computed distribution of clip lengths a subscriber may experience in a 47/24 + 1 TASI is shown in Fig. 11 for A = 0.85 and A = 1 together with measured values on the New York–London TASI system. The measured 1 per cent clip length in this system was about 60 ms. Here again the measured values lie between A = 0.85 and A = 1, indicating the actual loading was between these two values. While the computed distributions are for a 47/24 + 1 TASI installation, the distributions apply fairly well to all the trunk-channel ratios shown in Table III.

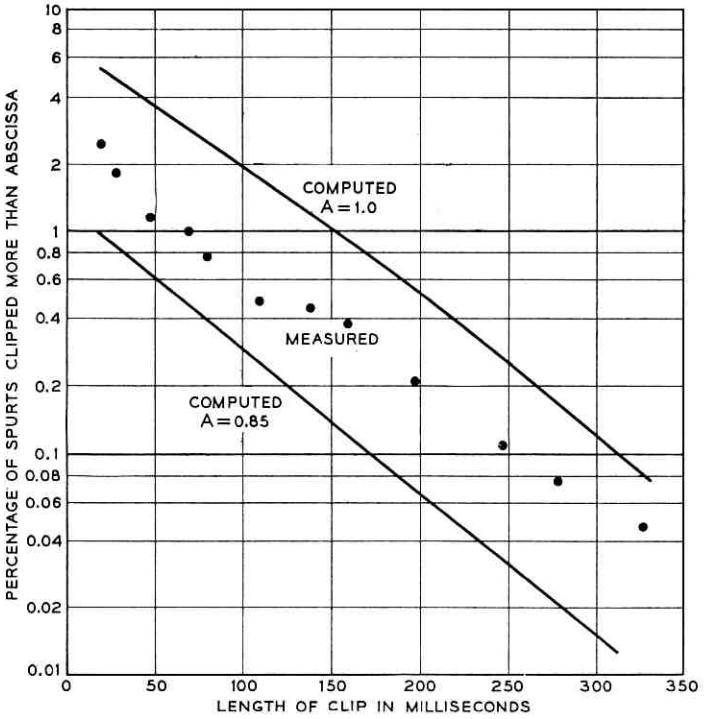


Fig. 11 — Measured and computed clip length distributions for a 24 + 1 channel 47 trunk TASI.

VII. CONTROL CHANNEL CAPACITY

As mentioned earlier, control channel crowding under heavy loading conditions can delay disconnections and in some cases increase clipping by delaying subsequent connections.

In a 74 talker, 36 + 1 channel TASI system in which each spurt requires a new connection, it is estimated that disconnects are delayed, on the average, about 4 ms each. This results in an estimated increase in average freeze-out fraction of less than 0.2 per cent. Since a working system switches less often than every spurt — a fact confirmed by measurement — the actual effect of control channel crowding is believed less than the above estimate indicates. Although in some rare cases bunching could seriously effect individual talkers, it appears that control channel overload contributes negligible clipping.

VIII. CONCLUSIONS

TASI has been operated successfully on submarine cable systems to provide approximately twice as many good quality message trunks as

existed before TASI. The measurements made on an installed TASI system have shown that TASI has come close to meeting its engineering objectives and that field modifications now going on will bring closer agreement between performance and objectives. The close agreement between computations of speech clipping and the measured values show that TASI theory is well understood and that TASI in the field is conforming closely to theory.

REFERENCES

1. Bullington, K., and Fraser, J. M., Engineering Aspects of TASI, B.S.T.J., **38,** 1959, pp. 353–364.
2. O'Neill, E. F., TASI—Time Assignment Speech Interpolation, Bell Laboratories Record, **37,** 1959, pp. 83–87.
3. Tucker, R. S., Sixteen-Channel Banks for Submarine Cables, Bell Laboratories Record, **38,** 1960, pp. 248–252.
4. Miedema, H., and Schachtman, M. G., TASI Quality — Effect of Speech Detectors and Interpolation, this issue, pp. 1455–1473.

TASI Quality — Effect of Speech Detectors and Interpolation

By H. MIEDEMA and M. G. SCHACHTMAN

(Manuscript received September 19, 1961)

This article describes tests made to select design parameters for the speech detectors in the TASI system. Results of subjective tests carried out to determine maximum permissible loading of TASI circuits during busy hours are also described. Finally, conclusions drawn from observations on a working TASI system are given. These observations indicate that TASI is a more satisfactory method of increasing transatlantic cable capacity than alternate methods, such as the use of 2-kc channel banks.

I. Introduction

TASI (*T*ime *A*ssignment *S*peech *I*nterpolation) is a new component in the telephone system that can approximately double the message capacity of existing long submarine cables. With TASI many calls share the same facilities, each requiring an available channel only when speech is transmitted. In order to recognize that speech is being transmitted by the subscriber, a highly sensitive speech detector is required. To assign the speech to an idle channel and to connect the proper talker and listener at each end requires a rapid switching system. A description of the switching system and other related matters can be found in other sources.[1,2,3] This paper deals with: (a) the work carried out to select the parameters of a speech detector satisfactory for TASI operation and (b) the results of subjective tests made to determine the approximate effect of the type of speech clipping that can occur in a fully loaded TASI during busy traffic periods.

Some speech is lost whenever the number of individuals talking or starting to talk in one direction on TASI circuits exceeds the number of available channels. The amount of lost speech must be kept small so that the transmission quality is not affected appreciably. On the average, less than 0.5 per cent of the total speech is lost due to interpolation, as long as the number of calls in progress is held to no more than twice

the number of channels. The effect of this loss on transmission quality is practically negligible.

In addition to the speech lost through interpolation, some speech is lost during the time required to connect a talker and his listener at the other end to the assigned channel. In TASI the switching time (17 milliseconds) has been kept as short as possible, consistent with reliable signaling. During the busy hours a subscriber will be switched about every second talkspurt; however, during occasional periods of peak load, the subscriber may be switched almost every talkspurt. Consequently, 17 milliseconds will be clipped from a large number of talkspurts during the busy hours. In order to minimize the amount of speech lost due to switching time, TASI has been designed so that a subscriber loses his channel only when it is not needed by that subscriber and when it is required by another talker.

A third possible source of lost speech results from the operate time of the speech detector. Whenever a talker has to be reconnected to a channel, the speech detector must recognize that speech is present and initiate the proper action. The interval between the time that the speech starts and the speech detector reacts adds to the amount of lost speech. The operate time of the speech detector can be kept small compared with the clipping caused by interpolation and connect signaling, but it should not be made so fast that the detector operates too often on noise.

In order to determine suitable speech-detector characteristics, subjective laboratory tests and field measurements were made on working transatlantic circuits. Since the demand for transatlantic circuits exceeded the capacity of the existing cable facilities, the schedule for developing TASI was of necessity very short for such a complex system. The tight schedule limited the type and length of test to the minimum needed for reasonable assurance that a satisfactory speech detector could be built. Consequently, a straightforward voltage threshold detector was chosen instead of a more complicated type. The first part of this paper describes the test results that led to the selection of the speech-detector characteristics. The second part of this paper describes the results of subjective tests to determine the impairment in speech quality caused by various amounts of lost speech. This work was needed as a guide to the maximum number of circuits that can be assigned to TASI without affecting speech quality adversely.

II. SPEECH DETECTOR CHARACTERISTICS

The ideal speech detector for TASI should operate only when speech is present and should not operate when noise and extraneous signals

are present. A practical detector must represent a compromise between ideal operation on speech signals and ideal rejection of noise signals. In addition, the activity, or percentage of the total time that a detector is operated, must be minimized. The parameters of the detector were chosen to insure that it

i. operates when very low levels of speech are present,

ii. operates a minimum amount of time on line noise, and

iii. minimizes the number of times a subscriber must be switched, consistent with allowing twice as many calls as there are channels available.

The speech-detector parameters are interrelated and one parameter could not be selected without considering the effect on all other parameters. In order to choose the best combination, a series of subjective tests were made in the laboratory to find the speech-detector characteristics that provided good results under simulated plant conditions. Later, field measurements were made during a large number of transatlantic telephone calls to determine the performance of several possible speech detectors under actual plant conditions for a wide variety of talkers. The field tests measured speech activities, talkspurt lengths, and number of talkspurts. The combined results of the laboratory subjective tests and field measurements led to the choice of speech-detector characteristics shown in Table I. Each of these parameters will be treated separately.

When the speech detector is made too sensitive, the detector operates on noise and thereby reduces the possible TASI advantage. Conversely, when the speech detector is not sensitive enough, part of the first syllable is lost before the detector is operated. A sensitivity of the TASI speech detector of −40 dbm at zero transmission level when combined with an adequate speech detector hangover (slow release time) results in satisfactory speech quality with volumes as low as −31 vu and also results in minimum false operations due to noise. As shown on Fig. 1, a −31 vu talker has a lower volume than almost all talkers on transatlantic calls.

TABLE I — TASI SPEECH DETECTOR CHARACTERISTICS

1000-cps sensitivity	−40dbm at zero transmission level point*
Frequency range	500–3000 cps
Operate time	5 milliseconds
Hangover (release time)	240 milliseconds "deferred"
Echo suppression	Maximum 13 db

* The zero transmission level point is a point to which all level points in a toll system can be referred. It is analogous to citing altitude by referring to height above sea level. The zero level point is at the transmitting toll switchboard of the system under consideration.

If the speech detector responded only to speech power within a very narrow frequency band, the problem of noise activity would be greatly reduced. However, for a speech detector with a fixed sensitivity to recognize initial consonants from many talkers, it is desirable to have a reasonably wide frequency band. Tests have shown that a bandwidth of approximately 500 to 3000 cps is suitable for telephone speech. If the bandwidth were extended below 500 cps or above 3000 cps, noise operations would be increased and the initial speech power seen by the detector would not be increased sufficiently to permit an offsetting decrease in sensitivity.

To operate the speech detector the power on the line must remain above −40 dbm for about 5 milliseconds. Laboratory experiments have shown that when the operate time is made as fast as possible, the detector sensitivity for equal quality speech can be decreased to about −37 dbm. However, the faster operate time results in increased operation of the speech detector by noise spikes which increases the activity even with

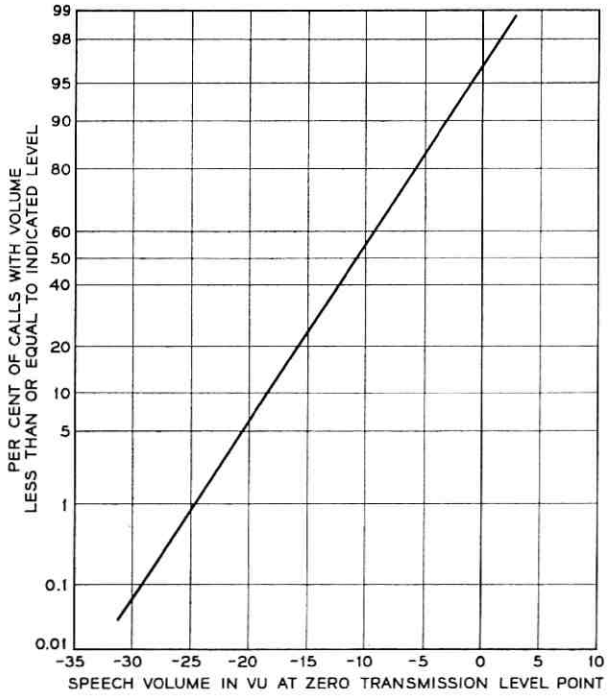


Fig. 1 — Subscriber speech volume distribution on transatlantic telephone circuits.

the lower sensitivity. Alternatively, an increase in operate time from 5 to 10 milliseconds requires an additional 3 db increase in sensitivity (−43 dbm) with no significant reduction in noise activity.

The sensitivity threshold of −40 dbm used for the detector is substantially above the threshold of hearing; hence, noticeable dropouts may occur when the speech power is below the sensitivity threshold. The ear is particularly sensitive to the loss of weak syllables as well as clips within words and closely connected phrases. Consequently, the speech detector should not release until the speech power has remained below the threshold for a period of time that is comparable to the time of one additional syllable. The hangover required for satisfactory transmission of low speech volumes varies with sensitivity and amounts to about 240 milliseconds for a −40 dbm detector. A much shorter hangover results in a higher TASI switching rate which increases the amount of speech lost by connect signaling clipping; a longer hangover results in increased speech activity and higher interpolation speech loss. Listening tests have shown that good transmission quality for weak talkers can be obtained with the combinations of sensitivity and hangover shown in Fig. 2.

Peaks of noise higher than −40 dbm may operate the speech detector and hence add to the circuit activity. Most of the noise peaks last less

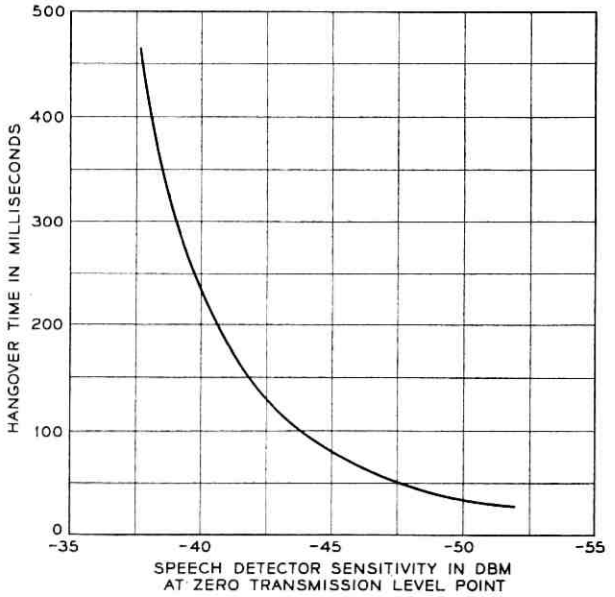


Fig. 2 — Variation of hangover with sensitivity for constant speech quality.

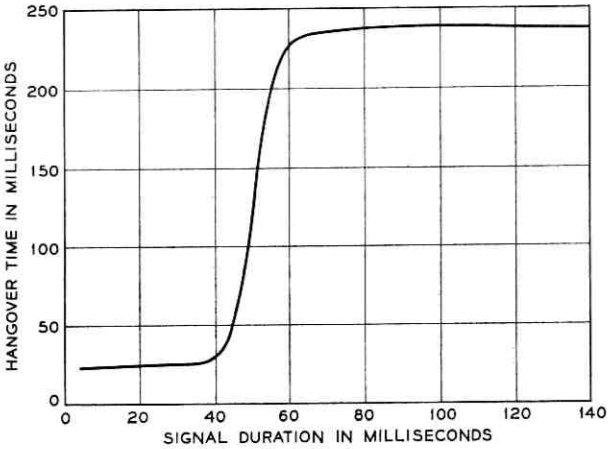


Fig. 3 — Speech-detector hangover vs signal duration.

than 5 milliseconds and are virtually eliminated by the 5-millisecond operate time of the speech detector. Once the speech detector is operated, however, the circuit cannot be released until the hangover time has elapsed. The effects of those noise peaks, which last from 5 to about 50 milliseconds, are minimized by the use of the deferred hangover characteristic shown in Fig. 3. The minimum hangover time is about 25 milliseconds. Noise peaks lasting substantially longer than 50 milliseconds are indistinguishable from speech syllables and operate the circuit to the extent of the full hangover of 240 milliseconds. The combination of 5-millisecond operate time, −40 dbm sensitivity, 240 milliseconds deferred hangover, and 500 to 3000 cps frequency range is about optimum for the expected telephone speech and noise levels.

In the preceding sections only the operation of the speech detector by normal speech and noise incoming to the TASI system has been considered. In the actual telephone plant, another group of unwanted signals, called echoes, can result in false operations of the detector. Transatlantic cable circuits have a one-way delay of about 40 milliseconds and this amount is sufficient to require echo suppressors in the four-wire part of the plant to prevent subscribers from hearing the echo returned from the two-wire part. The echo suppressors have an operate time of about 12 milliseconds to minimize false operation by noise. At times, the slowly operating suppressor will permit small bursts of echo to get past the suppressor at the beginning of talkspurts from the distant terminal. The distant listener is unaware of these short echoes which are therefore

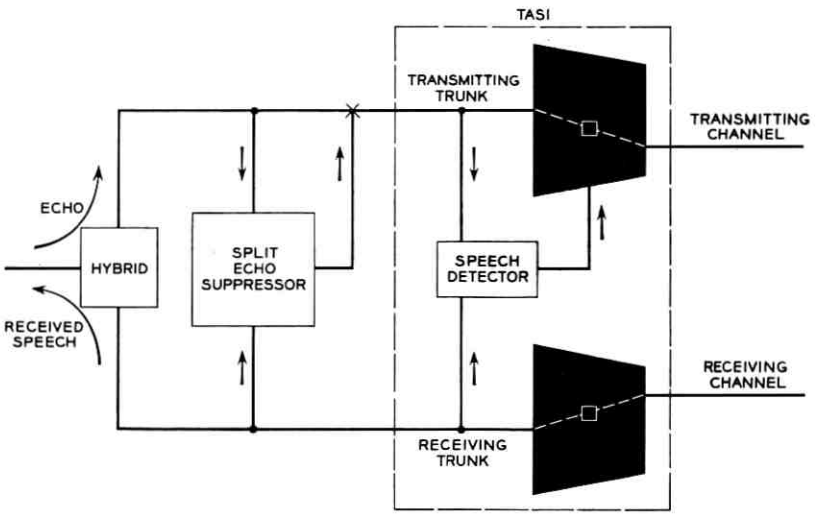


Fig. 4 — Simplified diagram of TASI terminal

of no consequence in non-TASI circuits. However, on TASI-equipped circuits, these short echoes can operate the faster acting (5 milliseconds) speech detector and thereby use valuable channel time. The echo path is shown in Fig. 4 for an individual trunk in which the TASI terminal is represented in simplified form.

To prevent these short echoes from operating the detector, an echo protecting circuit was added to the speech detector and connected to the receiving trunk output. When the distant talker is active, his speech reduces the sensitivity of the near-end speech detector at a uniform rate up to a maximum sensitivity decrease of 13 db as shown in Fig. 5. This value was chosen after consideration of the existing return losses in the plant and of the characteristics of the echo suppressors. The objective was the elimination of nearly all echo operation under all operating conditions. Measurements on many calls have confirmed that echo operation of the speech detector is negligible.

III. LABORATORY TESTS AND FIELD MEASUREMENTS ON SPEECH DETECTORS

From the foregoing description of the interaction of speech-detector parameters, it is apparent that there are several combinations which will result in a speech detector capable of recognizing the presence of speech and acting on this signal with hardly any effect on the speech quality.

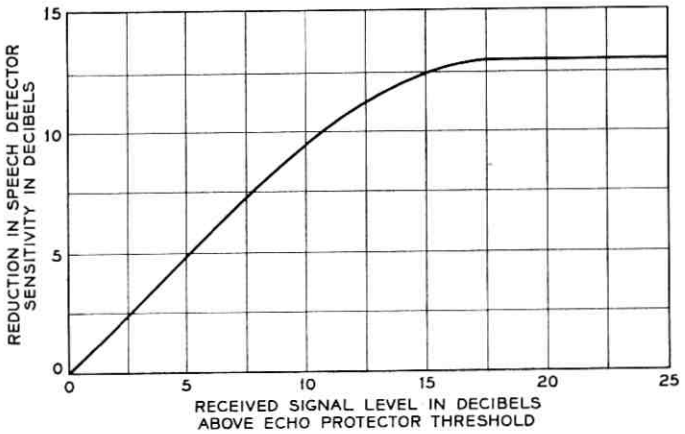


Fig. 5 — Effect of echo protection circuit.

Listening tests in the laboratory were made to determine the speech-detector sensitivity required for each type of detector over a range of speech levels. These tests were made using tape recordings of a number of voices, both male and female.

Speech from the tape recorder operated the speech detector, which in turn operated a gate to connect the observers to the tape recorder. The arrangement was such that observers could hear the recorded speech only when the detector was active. The observers were selected from Bell Telephone Laboratories personnel, both technical and clerical. They were asked to determine the minimum sensitivity required for acceptable speech quality. The rapid deterioration of speech quality when the speech-detector sensitivity was reduced below a certain minimum made this adjustment critical for each observer and resulted in a reproducible relation between speech-detector sensitivity and speech level. While each observer had a well defined tolerance level, the variation of the results among individuals was large. The sensitivity selected as satisfactory for a given speech detector was the value that satisfied 50 per cent of all observers for the minimum speech level of -31 vu at the zero transmission level point.

After determining several combinations of speech-detector characteristics that resulted in equal speech quality, various speech detectors were connected across transatlantic trunks to determine for each call the number of operations and the activity. The average activity of transatlantic subscribers, determined from measurements on many calls, determines to a large extent the possible TASI advantage (ratio of trunks

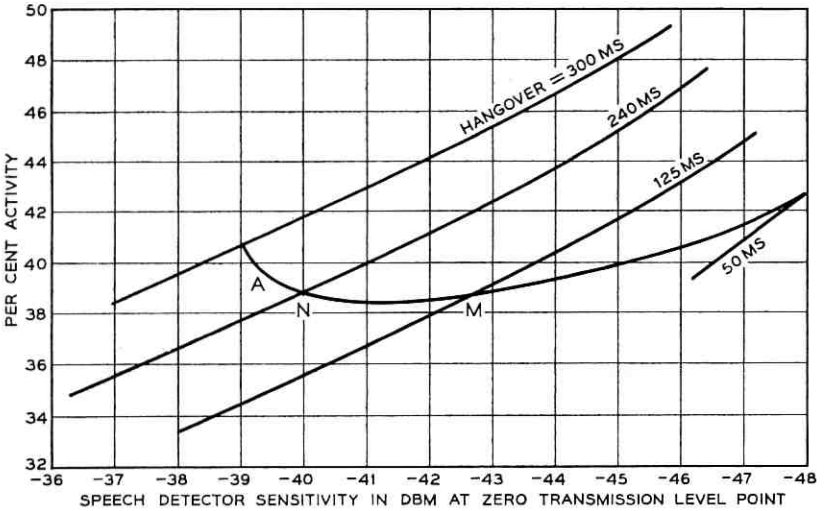


Fig. 6 — Average speech-detector activity during busy hour on transatlantic telephone circuits.

to channels). The number of operations affects how often a subscriber is disconnected from and reconnected to the channels. The results of the field measurements are shown by the four sloping lines on Fig. 6. On the same figure, curve A indicates the locus of points for constant and acceptable quality from Fig. 2. This locus permits the selection of the detector that provides acceptable quality and minimum activity. It will be noted that the point of minimum activity on curve A is not very sharply defined. Two speech detectors, points N (240 milliseconds hangover) and M (125 milliseconds hangover), were selected for further study. The average activity and talkspurt length resulting from each of these detectors were obtained from the measurements on transatlantic calls and are summarized in Table II.

It will be noted in Table II that the activity from the 240-millisecond detector is higher than for the 125-millisecond detector. However, the talkspurt length for the customer using the 240-millisecond detector is almost twice as long as for the 125-millisecond detector. As a guide in the choice between these detectors, computations were made to determine the total amount of speech a subscriber would lose during the busy hour if one or the other detector were used in TASI. The computed speech loss (interpolation plus switching) associated with each detector is shown on Fig. 7. The total speech lost is slightly less for the 240-millisecond hangover detector. Although use of the 240-millisecond de-

TABLE II — MEASURED ACTIVITIES AND TALKSPURT LENGTHS FOR TRANSATLANTIC SPEECH ORIGINATING IN U.S.

	125-millisecond Detector	240-millisecond Detector
Activity		
Operator	23%	24%
Subscriber	47%	48%
*Average during busy hour	38%	39%
Duration of average detector operation		
Operator	0.4 second	0.7 second
Subscriber	0.8 second	1.3 second
*Average during busy hour	0.6 second	1.1 second

* The results for operators and subscribers are combined in the ratio of 0.375 to 0.625, respectively. This ratio was found to be approximately the division of operator and subscriber time on transatlantic trunks during the busy hour.

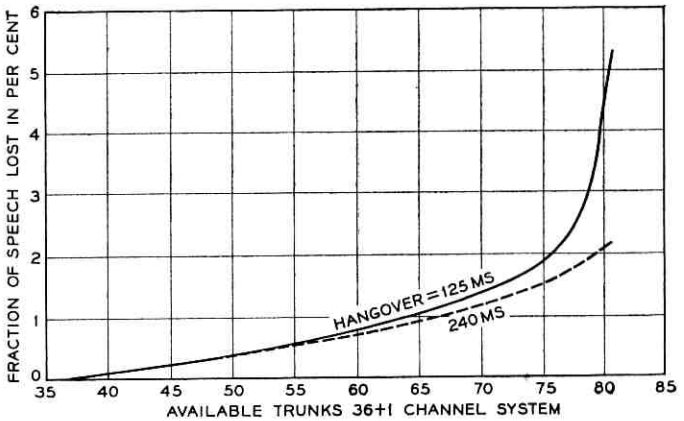


Fig. 7 — Subscriber speech losses caused by TASI signaling and interpolation.

tector results in a higher activity than the 125-millisecond detector and a slight increase in interpolation clipping, these disadvantages are more than offset by the fewer switching operations.

IV. SUBJECTIVE EVALUATION OF CLIPPING

The preceding tests showed that a satisfactory speech detector could be built whose impairment would be negligible. Additional tests were needed to determine how much clipping of all types could be tolerated without a significant impairment in speech quality. These quantitative

tests were needed in order to specify the maximum number of circuits that could be used with TASI.

Under normal operation, TASI clipping consists of many segments which are short compared with an average syllable length. During the busiest hours of the day, about one-half of the talkspurts will experience some clipping but in most cases, it will be limited to the 17-millisecond switching clip. Only one talkspurt in 10,000 will be clipped longer than about 0.4 second. Since the listener rarely misses a syllable, the result is a slight abruptness in speech which has a negligible effect on intelligibility.

The initial series of subjective tests to evaluate clipping in TASI was conducted before an actual TASI terminal became available. Preselected clips of a fixed length were introduced at the beginning of every Kth talkspurt, where K was 1, 2, $\cdots$, etc. By eliminating the randomness of TASI, the testing time was shortened and the data could be analyzed to discover whether the most significant factor affecting the impairment of TASI was the total speech lost or the pattern of clip occurrence. In this series of tests, the percentage of total speech time lost by clipping was varied from 0 to 6 per cent, which encompassed the expected range of TASI clipping. A group of observers placed normal business calls over circuits in which the amount of clipping would be controlled and measured. The observers were asked to base their replies solely upon *quality* considerations and to rate the calls as either "good," "fair," or "poor." The testing methods are described more fully in the Appendix.

The results of the clipping tests are shown on Fig. 8. Each point represents the percentage of calls rated "good" out of groups of more than 75 test calls for each condition. The left-hand scale indicates the amount of impairment found to elicit the corresponding percentages of "good" responses in a previous study of other degradations.[4] The curve was an approximate fit with a second-order polynomial using the least squares method.

Impairment in db has the following meaning: if a telephone subscriber is given the choice between the transmission system under test and a reference transmission system, how much loss could be inserted in the reference path for the subscriber to rate the circuits as equivalent? The loss inserted is defined to be the impairment, and is usually expressed in db. This impairment has been determined in the past for such things as noisy systems, restricted bandwidth, etc., but has not been determined directly in this study. The rendering of the present data into db of impairment rests on the assumption that "percentage good" means the same thing in this and the Coolidge-Reier study. This assumption is

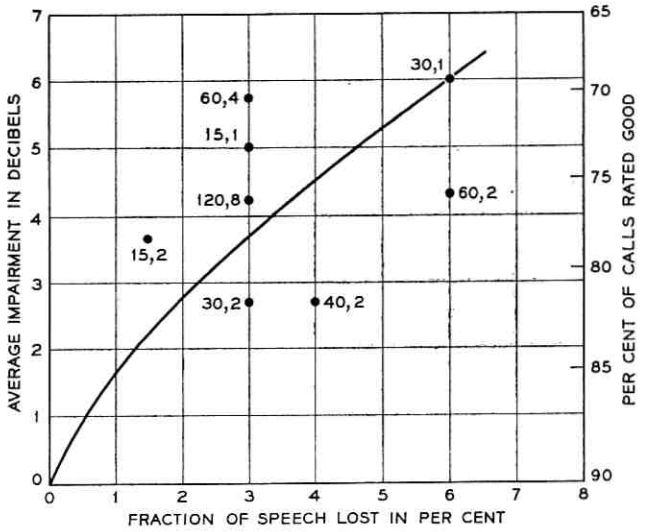


Fig. 8 — Subjective impairment vs fraction of speech lost.
Note: designation 30,2 means a 30 ms clip every second talkspurt.

unlikely to hold with great precision, so the impairment figures given here (and the implied comparability with other impairments) must not be considered precise quantities. This reservation also applies to similar conversions referred to later in this paper including those involving intelligibility figures. The use of such impairments provides a common denominator for comparing the effect of these different forms of speech degradation.

After the initial series of tests had been completed, another series of tests expanded the measurements to still greater amounts of clipping, in order to evaluate the loss in intelligibility that might occur with unusually high TASI loadings. The second series of tests used articulation methods and measured the degree of *successful communication* rather than simply quality as before. By this time a working TASI terminal was available so the occurrence and duration of the clipping were essentially the same as in normal TASI operation. The tests used "phonetically balanced" words and both the articulation in percentage of correct words and the percentage of lost speech were measured. The results of three articulation tests for widely different amounts of lost speech are shown by the dots on Fig. 9. The upper dotted curve is based on the three articulation tests, while the lower dotted curve has been taken from Fig. 8. The combined result is indicated by the solid line. The translation

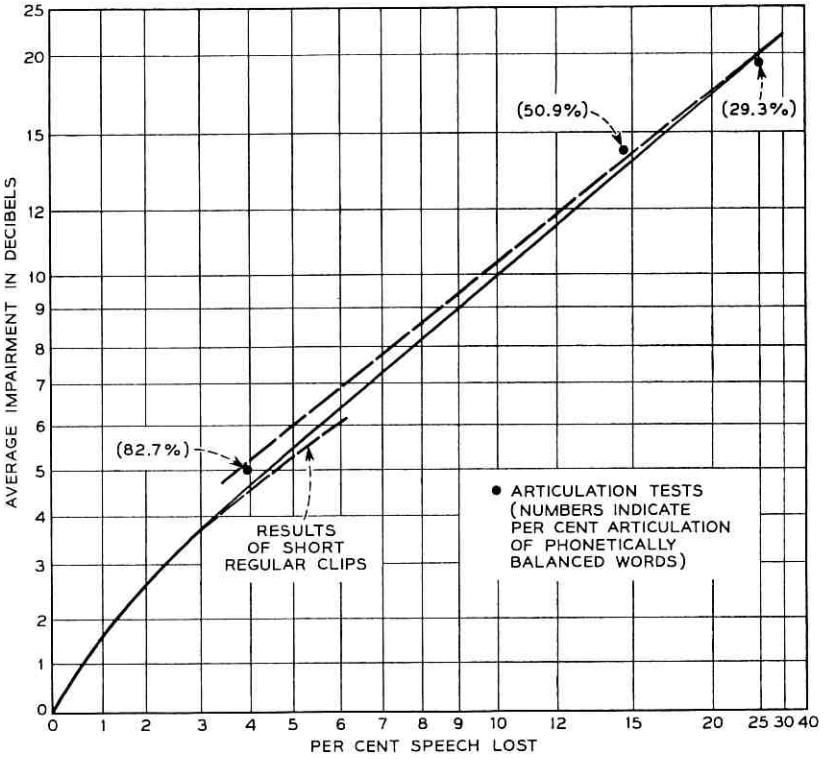


Fig. 9 — Subjective impairment vs fraction of speech lost.

from the percentage of correct words to db impairment made use of previously published results.[5] The testing methods are described more fully in the Appendix.

The foregoing tests indicated that the impairment is determined for the most part by the percentage of speech lost, at least for the case of normal TASI clipping which has certain definite characteristics: the clips are of short duration, they occur only at the beginning of talkspurts and they are irregularly distributed throughout a call. Hence, these results do not necessarily apply to the case where long clips of many syllables occur very infrequently.

The results on Fig. 9 can be applied to a statistical analysis of TASI clipping to obtain an estimated impairment for an actual TASI system. Fig. 10 shows the probable busy-hour impairment for a typical transatlantic TASI system, which uses 36 interpolation channels plus one

channel for disconnect and error checking signals. The estimates on Fig. 10 are based on the amount of lost speech which has been computed from appropriate distributions of subscriber activities and talkspurt lengths as a function of the number of trunks available for service.

The median curve indicates that in 50 per cent of the calls, the impairment caused by TASI will be less than 2 db as long as the number of active circuits is no more than 74 trunks on 36 + 1 channels. However, the impairment amounts to about 4 db for one call in a hundred and to about 6 db for one call in 10,000. For comparison purposes, the approximate impairments of 2-kc and 3-kc underseas channel banks are also indicated on Fig. 10. It will be noted that on nearly every call, a 2:1 TASI advantage causes less busy-hour impairment than would have been obtained through the use of 2-kc channel banks. In addition, the TASI impairment decreases rapidly as the number of talkers decreases during non-peak periods, while the impairment caused by narrow channels is independent of the traffic load.

V. DETERMINATION OF THE GRADE OF SERVICE

The foregoing results are presented in terms of db impairment to permit an easy comparison between TASI and alternate methods of

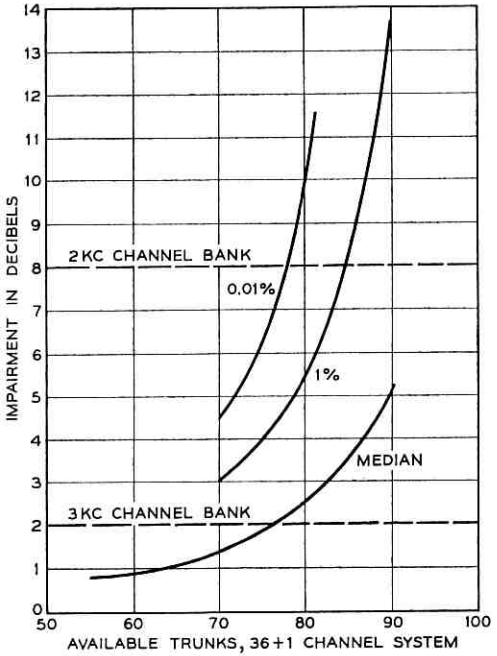


Fig. 10 — Busy-hour subscriber impairment due to TASI.

obtaining additional channels. It is also of interest from an operating standpoint to estimate the over-all grade of service on transatlantic calls when the additional practical factors such as variations in net loss, bandwidth and talker volumes are included. For more accurate results, the effect of each component should be investigated under all possible conditions of the other factors such as noise, volume bandwidth, etc., in order to look for possible interaction effects. However, in order to find the small effects looked for in this study, many samples would be required for each combination, and the amount of work involved would be almost prohibitive. Past experience indicates that a rough estimate of the over-all impairment can be obtained by adding together the individual impairments in db and then expressing the end result in the "good," "fair," and "poor" ratings.

On this basis the estimated grade of service on transatlantic calls with TASI applied to 36 3-kc interpolation channels is shown on Fig. 11 for loading conditions that are equal to or greater than the normal operating condition of 74 trunks on 36 + 1 channels. For comparison the estimated grade of service is given for 3-kc underseas channel banks used without TASI as well as the desired objective for all long-distance circuits. While this objective is not fully met at present, it is expected that the TASI grade of service under normal operation will be improved and will approach the desired objective as the result of the current program to improve local plant transmission. Figures 10 and 11 also indicate that during emergency peak periods, TASI can provide a greater number of higher quality circuits than can be realized with 2-kc channel bank equipment.

VI. SUBSEQUENT OBSERVATIONS OF TASI IN SERVICE

After TASI systems were put in service on transatlantic telephone cables, service observations were made on several TASI circuits as well as on a reference non-TASI circuit. The results indicated that the percentage of calls rated "good" by a qualified service observer was practically the same for TASI and non-TASI trunks.

The grade of service for greater than normal TASI loading was measured when additional trunks were utilized for emergency service at the time of a break in one of the transatlantic cables. One transatlantic system was utilized with TASI to carry traffic normally carried by both transatlantic systems. Table III gives the grade of service for normal loading and for two conditions of greater than normal loading. With the exception of a somewhat greater "poor" rating for 90 trunks, the observations on operating circuits correspond to those predicted in Fig. 11.

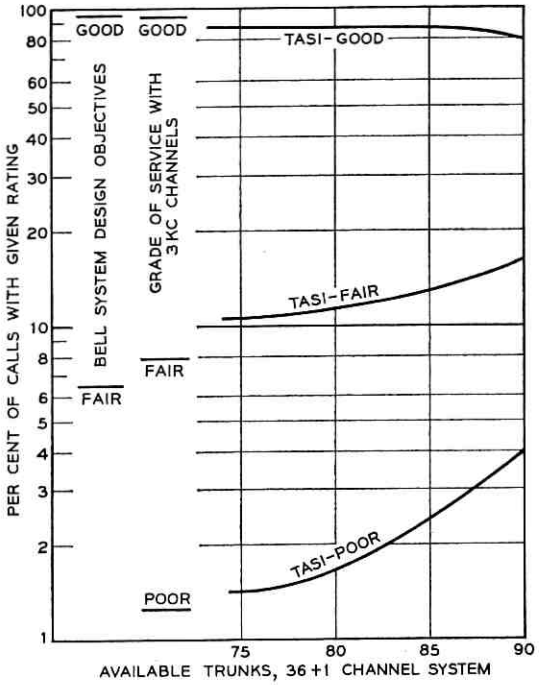


Fig. 11 — Grade of service — TASI plus 3-kc channels.

TABLE III — TASI GRADE OF SERVICE FOR NORMAL AND PEAK PERIOD LOADING AS DETERMINED BY TELEPHONE COMPANY SERVICE OBSERVERS

Trunks on 36 + 1 channels	Per cent Rating "good"	Per cent Rating "fair"	Per cent Rating "poor"
74	90	9	1
84	87	11	2
90	74	16	10

VII. CONCLUSION

The data from these service observations combined with the other results presented in this paper give assurance that most telephone users making calls through a TASI system with normal load will not be aware that their conversation is being interpolated. This holds even during the busy period when any degradation that does exist is at a maximum. During emergency conditions when service is partially disrupted, TASI

provides additional circuits with only a moderate decrease in grade of service. This is felt to be preferable to long service delays.

VIII. ACKNOWLEDGMENT

The authors are indebted to K. Bullington and J. M. Fraser for their helpful suggestions and comments. Messrs. A. E. Donkin, L. R. Tower and K. F. Trofatter provided valuable help in constructing and operating specialized testing and measuring equipment.

APPENDIX

Method of Conducting the Tests

A.1 *Initial Series of Conversational Tests*

The random nature of speech and clipping, and the dependence of the subjective ratings on the environment in which the clipping occurs indicated that the tests of clipping should be conducted under the most representative conditions. To implement this requirement, observers at Bell Telephone Laboratories engaged in normal telephone conversations in which the TASI clipping was introduced and the bandwidth and net loss were those of a typical transatlantic circuit. These calls were from the desks of the observers and consisted of normal business traffic that corresponded well with typical traffic on transatlantic calls. Following each call they rated the quality of the circuit as "good," "fair," or "poor." A total of 45 hours of test calls made by 25 talkers was analyzed.

The results of the tests were analyzed and yielded the percentage of all test calls rated "good" in over-all quality of transmission. The rating of per cent "good" for the various clipping conditions was converted to db impairment by the previously mentioned relation between subjective ratings and received volume. This step assumes that the impairment is roughly equivalent to the effect of a decrease in received volume on a transatlantic circuit.

The individual data points are plotted in Fig 8 and the resulting curve is repeated as the lower portion of the curve of Fig. 9. The percentage of speech lost is determined by the length of the clip, the frequency of occurrence (every Kth talkspurt) and the average talkspurt length. For these tests, the 125-millisecond hangover detector was used which results in an average talkspurt length of about 0.5 second. Thus, for 30 milliseconds, $K = 2$, a 30 millisecond clip occurs on every second talkspurt of length 0.5 second, and this condition results in a percentage speech loss of 3 per cent. As indicated in Fig. 8, speech losses of 3 per

cent and 6 per cent resulted from more than one test condition. These different conditions were used to determine whether the manner in which a given percentage speech loss occurs results in a significant variation in the impairment. Such a significant variation is not apparent from the results. Rather, the spread that occurred can be attributed more to variability in the testing than to the different ways of producing clipping. While each point represents more than 75 test calls rated by observers, the estimated confidence interval is only ± 2 db because only 3 separate ratings were possible. The solid curve has been used for TASI engineering but any use for other purposes should take into account the testing methods and the variations that are inherent in such subjective tests.

A.2 *Articulation Tests*

Articulation tests were conducted to obtain information about the effect of clipping when the amount of lost speech was great enough to affect comprehension. Single word articulation tests were chosen to provide the severest type of test of TASI clipping.

These tests were conducted with a list of one hundred "phonetically balanced"[6] words, which have been so chosen that all speech sounds are represented according to their frequency of occurrence in normal speech. The list of words was recorded using four different voices which were combined so that twenty-five words from the list were recorded by each voice. Two male and two female voices were used.

The recorded lists of words were played through a TASI terminal which was artificially loaded to insure the desired amount of TASI clipping. The clipped speech was played back in the laboratory through standard telephone handsets. Six technical and nontechnical observers were used for each test. The speech levels at the receivers and the receiver noise were adjusted to simulate an average transatlantic connection. The observers wrote down the words as they heard them and their responses were compared with the original lists to determine the phonetically balanced word articulation in per cent.

The approximate impairment in db corresponding to the per cent PB word articulation can be obtained from previous information that relates phonetically balanced word articulation to syllable articulation and syllable articulation to volume above threshold. By this means the articulation results for the conditions tested are related to the corresponding volumes, and the variation in these volumes from the reference condition yields the impairment in db.

REFERENCES

1. Bullington, K., and Fraser, J. M., Engineering Aspects of TASI, B.S.T.J., **38,** 1959, pp. 353–364.
2. O'Neill, E. F., TASI — Time Assignment Speech Interpolation, Bell Laboratories Record, **37,** 1959, pp. 83–87.
3. Fraser, J. M., Bullock D. B., and Long, N. G., Over-all Characteristics of a TASI System, this issue, pp. 1439–1454.
4. Coolidge, O. H., and Reier, G. C., An Appraisal of Received Telephone Speech Volume, B.S.T.J., **38,** 1959, pp. 877–897.
5. Richardson, E. G., *Technical Aspects of Sound*, Vol. I, Elsevier Publishing Company, Amsterdam, Houston, London, New York, 1953, pp. 280–282.
6. American Standard Method for Measurement of Monosyllabic Word Intelligibility, ASA No. S3.2, 1960.

Comment on "Discrimination against Unwanted Orders in the Fabry-Perot Resonator"

(Manuscript received May 18, 1962)

In the above paper,[1] Kleinman and Kisliuk state that "Fox and Li have investigated these configurations and the corresponding frequencies and losses for interferometers consisting of perfectly reflecting plates in air. In the usual laboratory interferometer the Fox and Li modes cannot be resolved because of insufficient reflectivity of the plates. Therefore the role played by these modes in optical masers is not settled."

Unfortunately these statements might be interpreted to mean that there is doubt as to the validity of the normal mode concept applied to maser interferometers.

We should like to correct the impression that the analysis of Fox and Li[2] was limited to perfectly reflecting mirrors. As a matter of fact, the reflectivity of the mirrors is completely unimportant in determining the normal modes, providing only that it is uniform over the mirrors.

It is quite true that in most solid state masers the inhomogeneities of the medium appear to create so much chaos in the radiation fields that correlation with a simple theoretical picture is often hard to demonstrate. However, gas masers appear to behave in a reasonably ideal way, and both the near-field and far-field radiation patterns for these masers appear to confirm the normal mode picture. In the case of such a maser equipped with plane mirrors, Herriott[3] has observed 1.3 mc beats which correspond well with the expected difference frequency between the dominant (even-symmetric) mode and the lowest order odd-symmetric mode.

An even more striking confirmation is seen in Herriott's pictures[3] of the light distribution across the plane mirrors of his helium-neon maser. These show fairly symmetrical multi-lobed distributions, which are at least qualitatively what one would expect for low-order transverse modes.

Finally the very beautiful pictures of Kogelnik and Rigrod[4] have demonstrated convincingly the existence of higher order modes in a helium-neon maser with concave mirrors.

With regard to passive interferometers, E. H. Scheibe has reported[5]

that in 1955 a "spurious" resonance was observed in a parallel plate resonator at 9.4 kmc. This turns out to have been the lowest order interferometer mode. Scheibe states that his value of measured Q agrees well with the loss curves of Fox and Li and with a curve given by Goubau and Christian. Christian and Goubau[6] have given a number of measured values for diffraction loss in a parallel-plate resonator over a range of values in N and have shown that these all agree closely with the theoretical loss curve given by Fox and Li for the dominant mode. Good evidence for higher order modes exists in a report by Culshaw[7] on a millimeter wave interferometer in which small subsidiary resonances (Fig. 6 of Culshaw) appeared at slightly greater reflector separations than the main resonances. The observed separations agree within a few per cent with what would be predicted from the results of Fox and Li for a TEM_{03} mode.

These findings leave very little doubt that the iterative normal mode picture does apply to laboratory interferometers, either with loss or with gain.

A. G. FOX
TINGYE LI
D. A. KLEINMAN
P. P. KISLIUK

REFERENCES

1. Kleinman, D. A., and Kisliuk, P. P., Discrimination Against Unwanted Orders in the Fabry-Perot Resonator, B.S.T.J., **41**, March, 1962, pp. 453–462.
2. Fox, A. G., and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., **40**, March, 1961 pp. 453–488.
3. Herriott, D. R., Optical Properties of a Continuous Helium-Neon Optical Maser, J. Opt. Soc. Amer., January, 1962, p. 31.
4. Kogelnik, H., and Rigrod, W. W., Visual Display of Isolated Optical-Resonator Modes, Proc. I.R.E. Letters, February, 1962, p. 220.
5. Scheibe, E. H., Measurements on Resonators Formed From Circular Plane and Confocal Paraboloidal Mirrors, Proc. I.R.E., June, 1961, p. 1079.
6. Christian, J. R., and Goubau, G., Some Measurements on an Iris Beam Waveguide, Proc. I.R.E., November, 1961, p. 1679.
7. Culshaw, W., The Fabry-Perot Interferometer at Millimeter Wavelengths, a report of the Telecommunications Research Establishment, G. Malvern, England, January, 1953.

Contributors to This Issue

Vaclav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. Member American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, Mind Association, Phi Beta Kappa.

Gary D. Boyd, B.S., 1954, M.S., 1955 and Ph.D., 1959, California Institute of Technology; Bell Telephone Laboratories, 1959—. He is engaged in optical maser research. Member American Physical Society, I.R.E.

Donald B. Bullock, B.S., 1951, University of California; Bell Telephone Laboratories, 1951—. His early assignments included development work on coaxial cable carrier and government communications equipment. Later, he did systems engineering on submarine telephone cables and, more recently, TASI. At present he is engaged in engineering of satellite communications systems. Member A.I.E.E., I.R.E., Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Phi Beta Kappa.

John M. Fraser, B.S. in E.E., 1945, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1934—. Prior to World War II, he was concerned with the evaluation of subjective factors affecting the transmission performance of telephone systems. During the war he was chiefly concerned with the design and evaluation of communication systems for the military. Later he was engaged in transmission engineering work on the transatlantic telephone cable and on other carrier systems. More recently he has been responsible for an engineering group on TASI. He is now working on new submarine cable systems and in investigating the usefulness of new devices such as optical masers and Vocoders. Senior member I.R.E.; member Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

JOSEPH E. GEUSIC, B.S., 1953, Lehigh University; M.S., 1955 and Ph.D., 1958, Ohio State University; Bell Telephone Laboratories, 1958—. He was engaged in research and development work on the solid state maser. At present he is engaged in research on optical masers. Member Sigma Xi, Pi Mu Epsilon.

HERWIG KOGELNIK, Dipl.-Ing., 1955, Dr. techn., 1958, Technische Hochschule Wien, Austria; D. Phil., 1960, Oxford University, England; Bell Telephone Laboratories, 1961—. He is engaged in optical maser research. Member American Physical Society, I.R.E., Elektrotechnischer Verein Österreichs (Austria).

HENRY J. LANDAU, A.B., 1953 and Ph.D., 1957, Harvard University; teaching fellow, Harvard, 1955–1957; Bell Telephone Laboratories, 1957—. He has been engaged in mathematical research in function theory and harmonic analysis. In 1959–60 he was on leave of absence from Bell Laboratories for study at the Institute for Advanced Study. Member American Mathematical Society, Phi Beta Kappa, Sigma Xi.

NORWOOD G. LONG, B.S. in E.E., 1956, Duke University; Naval Research Laboratory, 1954–1955; Bell Telephone Laboratories, 1956—. Since coming to the Laboratories he has been engaged chiefly in systems engineering of TASI and also systems studies on new devices such as Vocoder and nonsynchronous multiplexing systems. At present he is engaged in system engineering work on new submarine cable systems. Member I.R.E., Tau Beta Pi, Eta Kappa Nu, Phi Beta Kappa.

HENRY E. MEADOWS, JR., B.E.E., 1952, M.S. in E.E., 1953, Ph.D., 1959, Georgia Institute of Technology; Bell Telephone Laboratories, 1959—. He has been engaged in exploratory development in circuit theory, and, more recently, analysis of multimode waveguide transmission problems. Member Sigma Xi.

HOTZE MIEDEMA, Electrotechnis Ingenieur, 1947, Technische Hoogeschool, Delft, The Netherlands; Philips Telecommunications Industries, 1946–51; Canadian General Electric and Canadian Westinghouse, 1951–57; Bell Telephone Laboratories, 1957—. He has been engaged in work on transmission systems problems related to the application of TASI to the transatlantic submarine cable. During this period he made a special study of the effect of speech detector operation on speech transmission quality and TASI efficiency. He is now engaged in signal processing

studies and is working on military projects. Senior member I.R.E.; member Dutch Radio Society, Association of Professional Engineers of Ontario.

ROBERT MORRIS, A.B., 1957 and A.M., 1958, Harvard University; Operations Research Office, 1958–59; Bell Telephone Laboratories, 1960—. His work has been related to solving problems in queuing theory which have combinatorial complications, and evaluating methods for controlling errors in digital data systems. He has recently been engaged in research in computer programming. Member American Mathematical Society.

JOSEPH OTTERMAN, Diploma Ingenieur, 1947, Hebrew Institute of Technology (Technion), Haifa, Israel, M.S.E., 1952, University of Michigan; Ph.D., 1955, University of Michigan; Israeli Government Scientific Institute, 1950–1951; University of Michigan Research Institute, 1955–1959; ITT Federal Laboratories and Resident Visitor, Bell Telephone Laboratories, 1959–1961; Consultant, General Electric Company, 1961. Mr. Otterman has worked on ballistic measurements, network synthesis, rocket investigation of the upper atmosphere, shockwave propagation through the atmosphere, analog and digital simulation, topology and monitoring of communication systems, traffic congestion theory for direct and store-and-forward traffic, and satellite systems. Senior member I.R.E.; member American Physical Society, Association for Computing Machinery, Franklin Institute, Sigma Xi.

HENRY O. POLLAK, B.A., 1947, Yale University; Ph.D., 1951, Harvard University; Bell Telephone Laboratories, 1951—. He has been engaged in mathematical analysis of gunnery and missile systems and in mathematical research in communications. He has written technical papers on analysis, function theory and probability theory. At present he is acting Director, Mathematics and Mechanics Research Center. Member of the Mathematical Association of America, Governor of its New Jersey Section, and on its Committee on the Undergraduate Program. He also belongs to the Advisory Board of the School Mathematics Study Group.

MARSHALL G. SCHACHTMAN, B.S. and M.S. in E.E., 1958, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1955—. After early work on a new low-current telephone subscriber's set, he was engaged in designing circuits for use in test equipment for Radio Relay Systems. He later was engaged in systems engineering work on TASI

and on system studies of the use of light masers in communications. He is currently continuing his systems engineering work on new submarine cable systems. Associate member A.I.E.E.; member I.R.E., Eta Kappa Nu, Tau Beta Pi.

ARNOLD H. SCHEINMAN, B.E.E., 1948, City College of New York; M.E.E., 1959, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1951—. Since completing the Communications Development Training Program, he has been engaged in the development of circuits for the Crossbar Tandem System. He presently serves as a lecturer at the Polytechnic Institute of Brooklyn. Associate member A.I.E.E.; member Sigma Xi.

HENRY E. D. SCOVIL, B.A., 1948 and M.A., 1949, University of British Columbia; D. Phil., 1951, Oxford University; Nuffield Research Fellow, Oxford, 1951–52; faculty, University of British Columbia, 1952–55; Bell Telephone Laboratories, 1955—. He has been engaged in development of solid state devices at microwave frequencies.

S. H. TSIANG, B.S., 1947, University of Nanking; M.S., 1949, Carnegie Institute of Technology; Union Switch and Signal, 1949–56; Bell Telephone Laboratories, 1956—. Mr. Tsiang worked on maintenance, and administration circuits and requirements for an experimental electronic telephone central office. He is currently concerned with the systems planning of an electronic telephone switching system for production.

WERNER ULRICH, B.S., 1952, M.S., 1953, and Eng. Sc.D., 1957, Columbia University School of Engineering; Bell Telephone Laboratories, 1953—. Mr. Ulrich was in charge of automatic testing and maintenance facilities, and programs for an experimental electronic telephone central office. He is currently working on the system design of an electronic switching system. Member I.R.E., Tau Beta Pi.