THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXX OCTOBER 1951 NUMBER 4 PART II

THE BELL SYSTEM TECHNICAL JOURNAL

PUBLISHED QUARTERLY BY THE

AMERICAN TELEPHONE AND TELEGRAPH COMPANY

195 BROADWAY, NEW YORK 7, N.Y.

CLEO F. CRAIG, *President*

CARROLL O. BICKELHAUPT, *Secretary*

DONALD R. BELCHER, *Treasurer*

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each.

The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U.S.A.

The TD-2 Microwave Radio Relay System

By A. A. ROETKEN, K. D. SMITH and R. W. FRIIS

(*Manuscript Received July 5, 1951*)

The TD-2 microwave radio relay system is a recent addition to the telephone plant facilities for long distance communication. It is designed to supplement the coaxial system and to provide greatly expanded facilities for nationwide transmission of broad-band signals such as television pictures or large groups of message circuits. The system makes use of many microwave repeaters located 25 to 30 miles apart in line-of-sight steps. The great variety and number of components which make up such a system require the engineering of all components to close tolerances. This paper describes the system in some detail from the standpoints of overall objectives, component designs to meet such objectives and facilities for the maintenance of overall performance.

I. Introduction

SUPER-HIGH or microwave frequencies began to attract the interest of communication research engineers during the late '30s. The practical application of microwaves to commercial communication circuits was delayed by the outbreak of World War II, but the microwave techniques which had already been developed were employed to advantage in the prosecution of the war. The concentrated development effort and mass production of microwave equipment for military applications greatly expanded the engineering knowledge and production skill in this relatively new communications field. After termination of the war, it was possible again to devote the necessary development effort toward application of microwave techniques to commercial purposes. In the Bell System this effort was applied to the development and construction of a long-haul radio relay system.

A broad-band multi-channel radio relay system now connecting some of the main communication centers of the United States, as shown in Fig. 1, represents the combined efforts of a Bell System team since 1945.[1] This chain of stations carrying hundreds of message circuits or a television picture on each broad-band channel, in giant 25 to 30-mile strides across the country, has opened up a new radio field. The first step was the development of an experimental system placed in service in November 1947 between New York and Boston.[2] Upon the successful completion of this project objectives were established for a system, which is called the TD-2 Radio System, capable of extension to at least 4000 miles with upwards of 125 repeaters.

The TD-2 Radio System provides no new types of service but will supplement existing facilities such as the coaxial system. Therefore, TD-2 must provide comparable reliability, economy and quality of service. It is

Fig. 1—TD microwave radio relay routes.

contemplated that by the end of 1951 there will be over 20,000 broad-band channel miles of radio relay in operation in the Bell System. Of this, about two-thirds will be used for television service and one-third to provide over 600,000 circuit miles of telephone circuits.

II. TD-2 System—General

A radio relay system designed for long distances involves many problems new to radio but not new to long distance wire circuits. These problems are chiefly those of systems engineering to close transmission tolerances be-

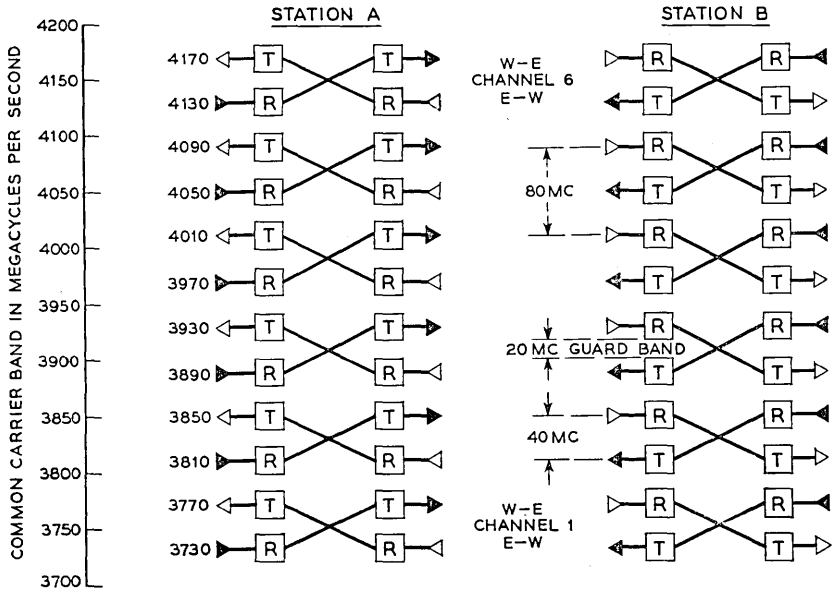


Fig. 2—TD-2 radio frequency plan.

cause of the many repeaters in tandem. To insure satisfactory systems operation, the transmission characteristics must remain stable over long periods of time to permit unattended operation. A reliable power plant and an alarm system are essential parts of the radio system.

A. *Description*

The TD-2 Radio System utilizes frequency modulation and provides twelve broad-band channels, six in each direction, spaced 40 megacycles apart in the 3700–4200 megacycle common carrier band. A frequency assignment chart is shown in Fig. 2 and a systems block diagram in Fig. 3. Two broad-band channels in opposite directions provide a two-way message

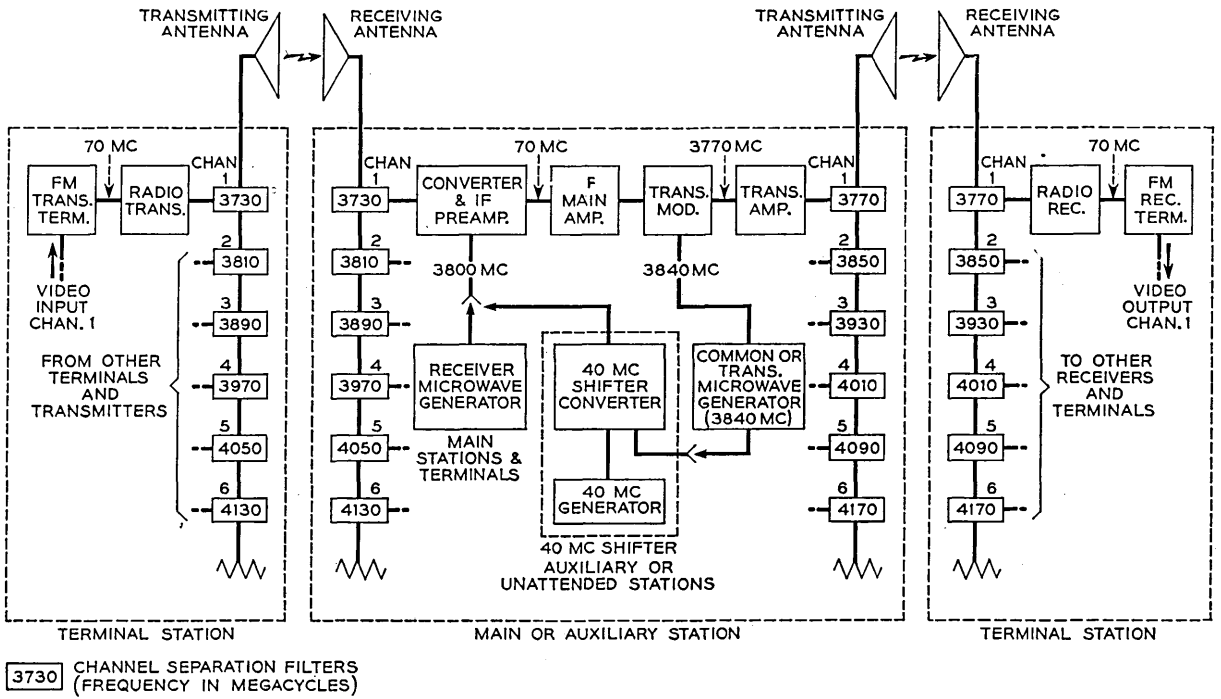


Fig 3—Radio system block diagram.

system having a capacity of hundreds of 4 kc message circuits. Alternately each of these broad-band channels can be used to provide a 4-megacycle video circuit of the kind required for present day black and white television or they may be used to provide broader band television circuits if the need for such circuits develops. The video or message input to a channel is frequency modulated on a 70-megacycle carrier, translated up to the microwave band, amplified and combined with the microwave output of other channels in the same direction. The combined output is carried through a single waveguide to a directive transmitting antenna[3] beamed toward the next station. At a repeater point the six-channel signal is received on a single antenna, separated by means of channel separation networks, and each channel converted to a 70-megacycle IF band for amplification. At a through repeater point this 70-megacycle IF signal again modulates the 4000-mega-

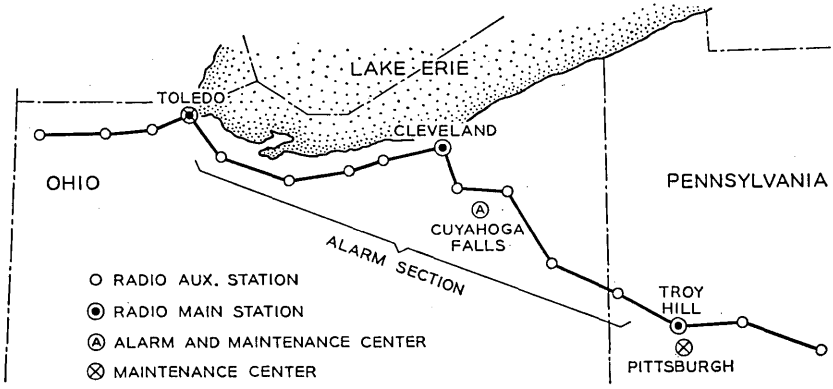


Fig. 4—Typical TD-2 route section.

cycle carrier, is amplified, added to the other channels in the same direction and delivered to the transmitting antenna. These are the simplified functions performed by the one-way radio repeater of the TD-2 System.

In order to feed a standard video or a multi-channel carrier signal into a TD-2 Radio System, an intermediate transmitter is required to frequency modulate these signals around a center frequency of 70 megacycles. This unit is known as the FM terminal transmitter. Likewise, at a receiving point an intermediate receiver is required to convert the 70-megacycle signal back to a video or carrier system signal. This unit is known as the FM terminal receiver.

A perspective of the system may be obtained from a typical route section as shown in Fig. 4. A long system is broken into sections by means of main repeater stations every few hundred miles. Auxiliary stations interconnect

the main stations. From an operating standpoint, main stations differ from auxiliary stations primarily in that each channel is terminated in IF switching circuits. This permits the removal of a channel for maintenance, or the replacement of a section which has failed by a spare circuit, by patching or remote control of the IF switching circuits. An alarm center may also be identified in Fig. 4 as an attended office to which a maximum of twelve repeater stations are connected by wire or radio for the purpose of reporting abnormal conditions that exist. Maintenance personnel are dispatched to unattended stations from this point. Not all TD-2 units are repaired and maintained at the radio stations. Maintenance centers are established along the route to service these units which require more elaborate test facilities than are provided at the stations. This requires the furnishing of certain spare units at the individual repeater stations.

B. *Route Selection and Towers*

The interconnecting of two or more communication centers by a radio relay system presents many new problems in plant engineering. The selection of hundreds of mountain top sites to obtain line-of-sight transmission between stations, sites which are accessible to roads and power lines, sites which permit reasonable tower heights and which are an economic balance of these and other factors was a new challenge to the plant engineering force.[4] In brief, these were accomplished first by a detailed study of topographical and road maps, inspection of sites selected and finally the measurement of the transmission loss of the path.

The construction of towers several hundred feet high also involved new thinking by the building engineers.[5] The type of structure used on the New York-Chicago section of the TD-2 System was somewhat influenced by the availability of materials during 1948 and 1949. Concrete structures were used for this section of the system as shown in Fig. 5 with steel towers appearing on the Omaha to San Francisco section. Where steel towers are used, conventional type single-story buildings house the radio and associated equipment as shown in Fig. 6. Double antenna decks are provided on towers where branching radio routes are required.

III. TD-2 Radio Equipment

A. *Repeater—General*

The design of the TD-2 microwave repeater follows in principle that of its predecessor for the NewYork-Boston system.[2] Rapid advancement in the development of microwave vacuum tubes and other repeater components during the period from 1945 to 1947 led to a general improvement of repeater components for the TD-2 System. The realization in late 1947 of a

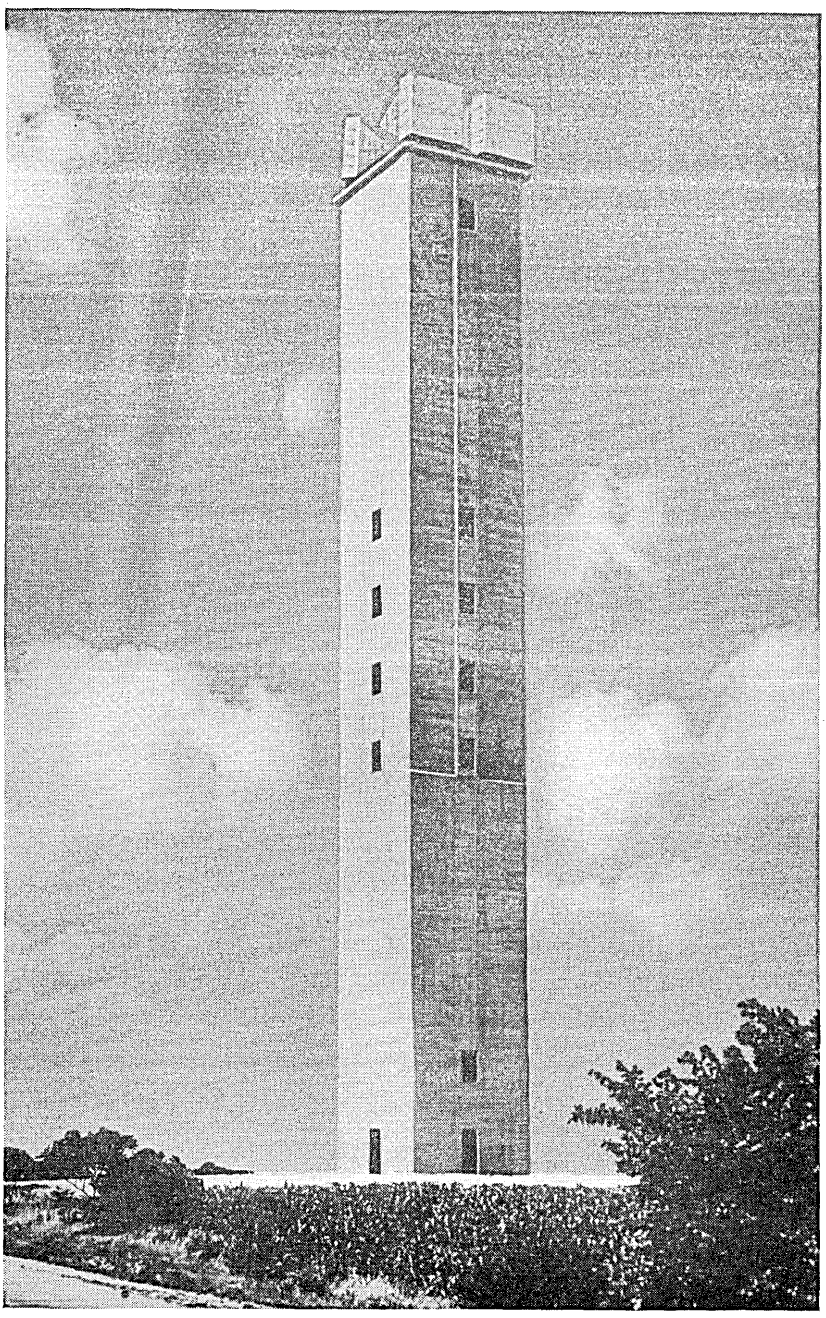


Fig. 5—190 foot concrete tower.

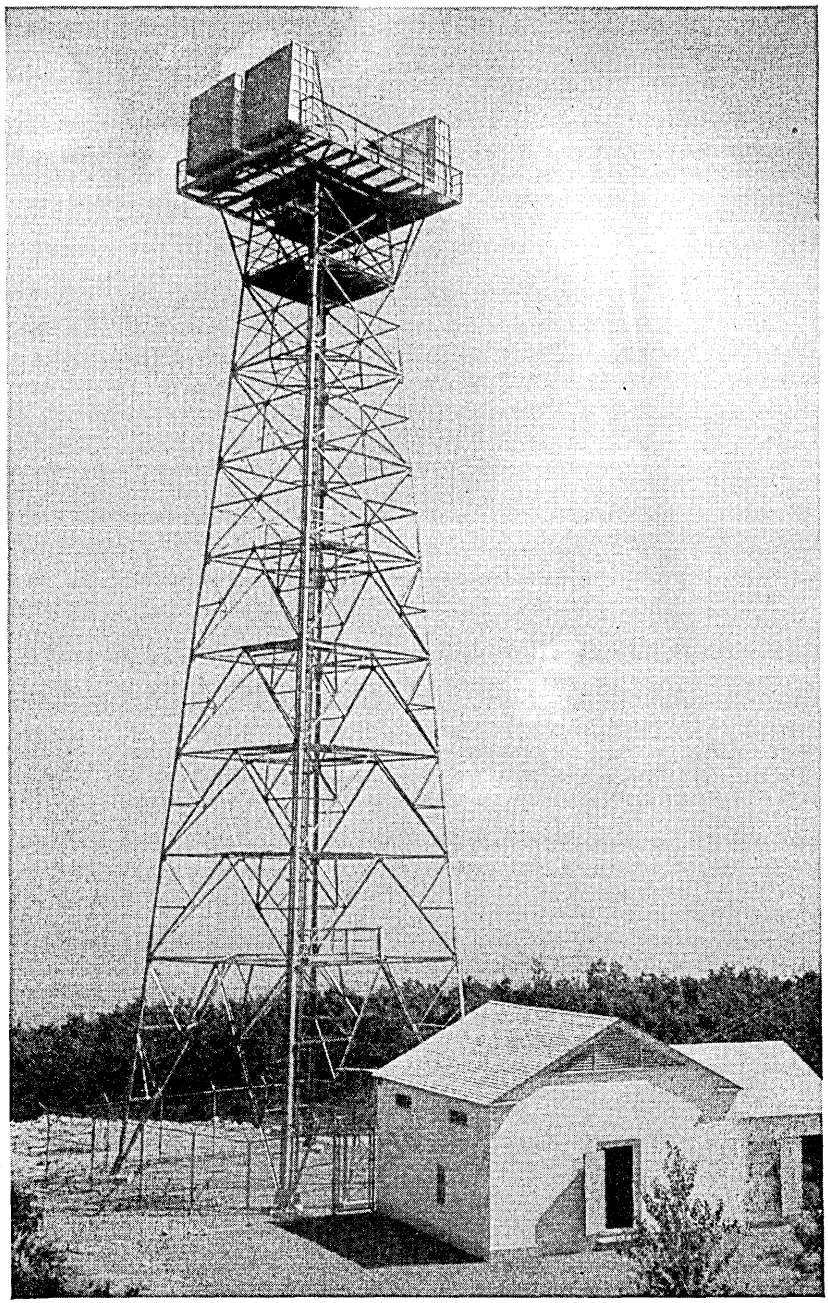


Fig. 6—125 foot steel tower.

practical triode amplifier for microwave application[6,7] was instrumental in determining a pattern for redesign, for it suggested the possibility of greatly simplifying the repeater while at the same time providing wider transmission bandwidths at greatly improved efficiency. By replacing the high voltage klystron (velocity variation type) amplifiers of the early repeaters with the new low voltage triodes, it became practical to design the system for battery operation—an important step toward increasing the system's reliability as it removed regulated rectifier tubes from the vulnerable portion of the system and eliminated the problem of hits during switchover from commerical to standby primary power. In the TD-2 System, vacuum tube heaters are generally operated from a 12-volt battery through dropping resistors,

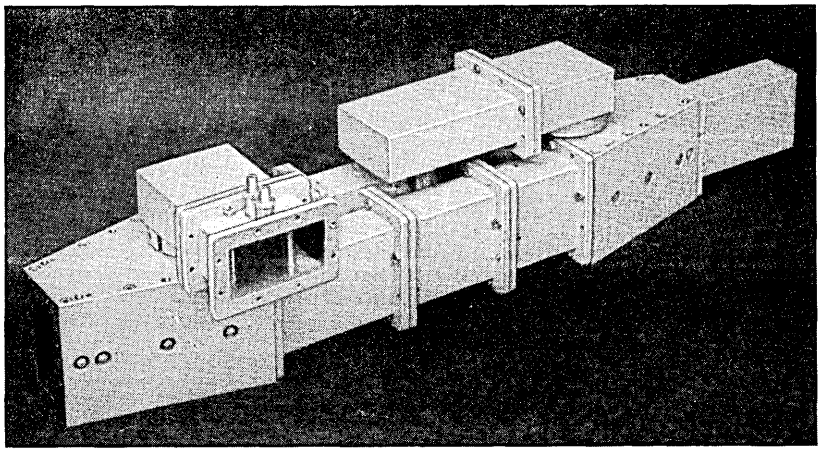


Fig. 7—Channel separation network.

thereby approximating constant heater power operation to increase tube life and reliability.

B. *Repeater Description*

A block diagram of a TD-2 repeater is shown in Fig. 3. An incoming microwave signal from a receiving antenna is selected by a channel separation network shown in Fig. 7. The signal is then combined in the receiver converter with energy from a beating oscillator source to provide an intermediate frequency signal band centered at 70 megacycles. Amplification, delay equalization and automatic gain control take place at the intermediate frequency of the radio receiver. In the transmitter this signal is combined in the transmitter modulator with a microwave source to provide a band offset 40 megacycles from the received frequency band. This signal is amplified and combined through transmitter channel separation networks with signals

from other channels for transmission through a common antenna. The 40-megacycle shift from receiving to transmitting frequencies is introduced in order to reduce the effect of crosstalk between transmitting and receiving antennas.

Main and auxiliary station repeaters differ in the following respects: An auxiliary repeater simply receives a particular channel signal, amplifies and transmits it to the next station. Here a common beating oscillator source for the transmitter and receiver, together with a stable 40-megacycle shifter, results in a systems frequency stability, for auxiliary stations alone, which is dependent only upon the stability of the 40-megacycle oscillator.[2] This feature cannot be used in the repeaters of a main station since each radio section between main stations must be independent of other sections for switching, branching, maintenance and terminating purposes. Here it is necessary to provide an independent oscillator source for each modulation process. In such an arrangement, errors in frequency add throughout the system and, therefore, the individual stability requirements for the oscillators are severe. In the TD-2 System this frequency stability is obtained by the use of a crystal controlled oscillator and harmonic generators. Two such microwave generators with temperature controlled crystals are used in each repeater bay at main stations, while one microwave generator and a 40-megacycle oscillator and shifter unit are used in each auxiliary station repeater.

A repeater bay using a 9-foot cable duct type framework is shown in Fig. 8. The top half of the bay contains the components which comprise the signal path through the repeater. These are the channel separation filters, image suppression filter,

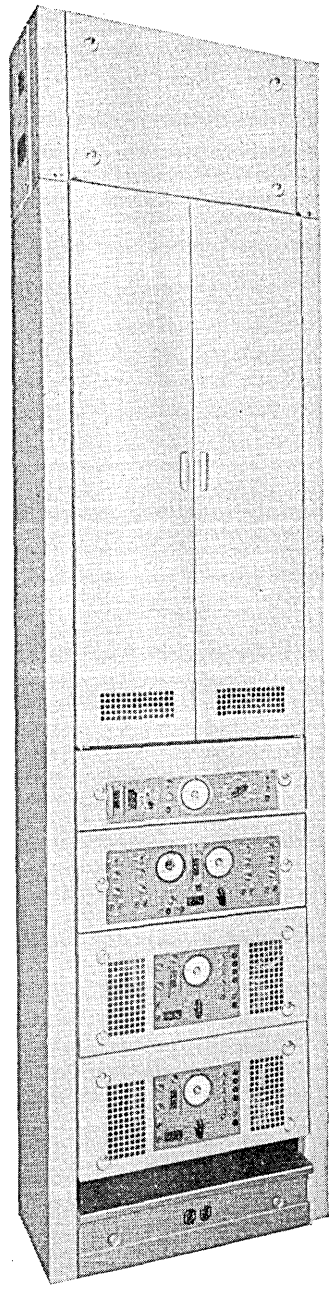


Fig. 8a

Fig. 8—Repeater bay.

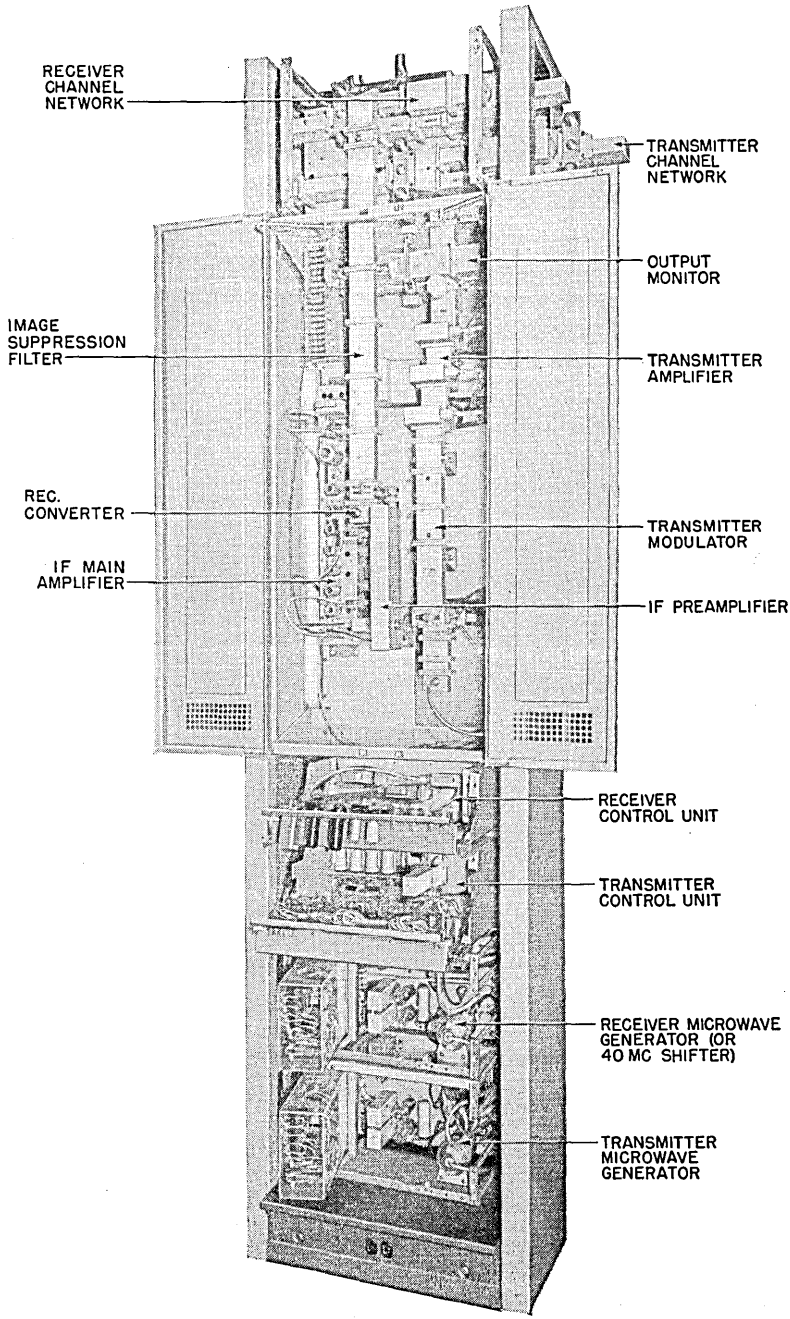

RECEIVER CHANNEL NETWORK
TRANSMITTER CHANNEL NETWORK
OUTPUT MONITOR
IMAGE SUPPRESSION FILTER
TRANSMITTER AMPLIFIER
REC. CONVERTER
TRANSMITTER MODULATOR
IF MAIN AMPLIFIER
IF PREAMPLIFIER
RECEIVER CONTROL UNIT
TRANSMITTER CONTROL UNIT
RECEIVER MICROWAVE GENERATOR (OR 40 MC SHIFTER)
TRANSMITTER MICROWAVE GENERATOR

Fig. 8b

receiving converter, IF preamplifier, delay equalizer, IF main amplifier, transmitting modulator and transmitting amplifier. The lower half of the bay contains four 19-inch wide oscillator and control units. These units, in the case of a main station repeater, are two microwave generators, a receiver control unit and a transmitter control unit. In the case of an auxiliary repeater, one of the microwave generators is replaced by a panel containing a 40-megacycle oscillator and shifter unit. All connections to the units of the bay are made by means of plugs and jacks for easy servicing.

A repeater receives a frequency modulated microwave signal at a normal level of about −38 dbm and transmits it at +27 dbm. Upward fades of 5 db and downward fades of 25 db are compensated to within about 1 db by automatic volume control action within the repeater. The amplitude characteristic is maintained flat to within 0.2 db over a 20-megacycle band.

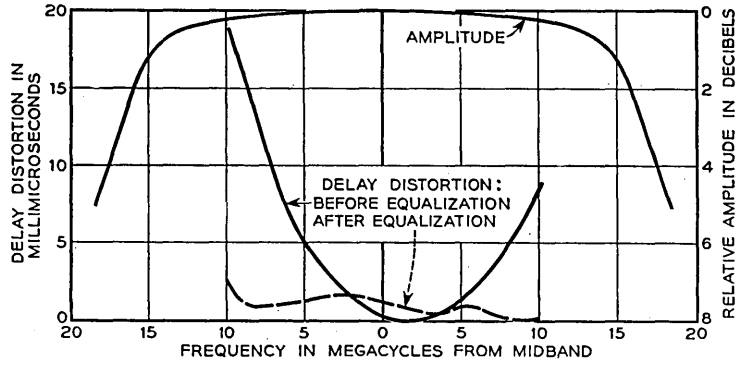


Fig. 9—Delay distortion & amplitude characteristics.

The amplitude and delay distortion characteristics of a repeater bay with and without delay equalization are shown in Fig. 9.

C. *Radio Receiver*

A channel separation network, as shown in Fig. 7, is required for each receiving channel. The network separates a particular channel from the six incoming 20-megacycle bandwidth channels for individual amplification and equalization in the repeater. It consists of two hybrid junctions and two band reflection filters which are tuned to the frequency band to be separated. An incoming signal is split into two parallel paths in passing through the first hybrid. A reflection filter in each of these paths returns the energy of the channel to be dropped to the first hybrid. By making the electrical path lengths from hybrid to filters differ by $\frac{1}{4}$ wavelength at the frequency of the channel to be dropped, the reflected signals are in phase op-

position at the hybrid and this results in transmission of the total signal energy through the fourth arm of the hybrid and into the receiving converter. The reflection filters are transparent to frequencies outside the desired 20-megacycle band. These frequencies are recombined in the second hybrid for transmission to the following channel separation networks.

An image suppression filter is located in the waveguide just ahead of the receiving converter. Its purpose is twofold: first, to provide discrimination against interfering signals which are the intermediate frequency image of the desired signal; and, second, to provide a critically spaced reflection of a beating oscillator component from the converter for control of the converter intermediate frequency output impedance.

The receiving converter is of the balanced crystal type in which the two crystals are mounted in a hybrid junction assembly. The signal input connection is by waveguide and the oscillator input connection is by coaxial line. An unbalanced output at intermediate frequency without the use of a balanced to unbalanced transformer is obtained by reversal of the polarity of one crystal relative to the other, permitting a parallel output connection. The 405A varistor unit, having symmetrical terminal design, was developed for this application. The preamplifier utilizes two 417A grounded grid triodes and has a gain of approximately 12 db. Its transmission band is centered at 70 megacycles and is flat to within 0.1 db over a 22-megacycle bandwidth. The converter-preamplifier has a net gain of approximately 6 db. The output of the preamplifier is coaxially connected through a delay equalizer to the input of the main IF amplifier.

The main IF amplifier shown in Fig. 10 has input and output impedances of 75 ohms and approximately 65 db gain. It consists of eight stages of amplification, the first being a 417A grounded grid triode followed by six stages of 404A pentodes and a 418A tetrode output stage. The input, output and interstage networks are all of the double-tuned impedance-matched type except the network between the sixth 404A pentode and the 418A output tetrode which is triple tuned and mismatched. Tuning of the triple-tuned network provides for adjustment of the over-all transmission band shape. Automatic gain control operating upon the grids of the first five pentode stages maintains the output power to within 1 db of a selected value between +4 dbm and +10 dbm for a 30 db range in input power. A low level bridging tap is available at the output of the main IF amplifier.

D. *Radio Transmitter*

The transmitter modulator consists of a 416A microwave triode mounted in a structure which provides a resonant cavity between cathode and grid and another between plate and grid. This cavity structure is used in both

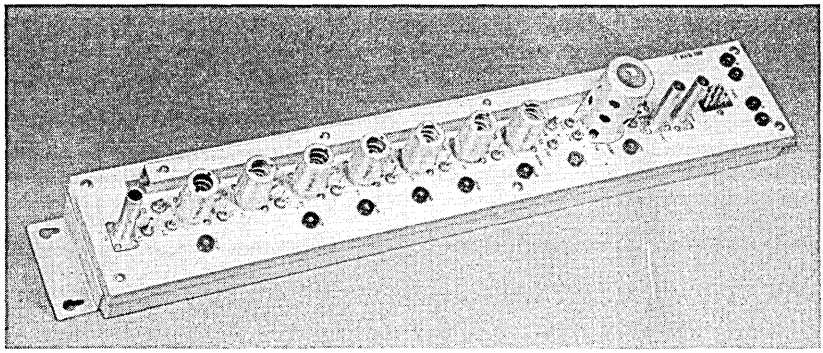


Fig. 10—Main IF amplifier.

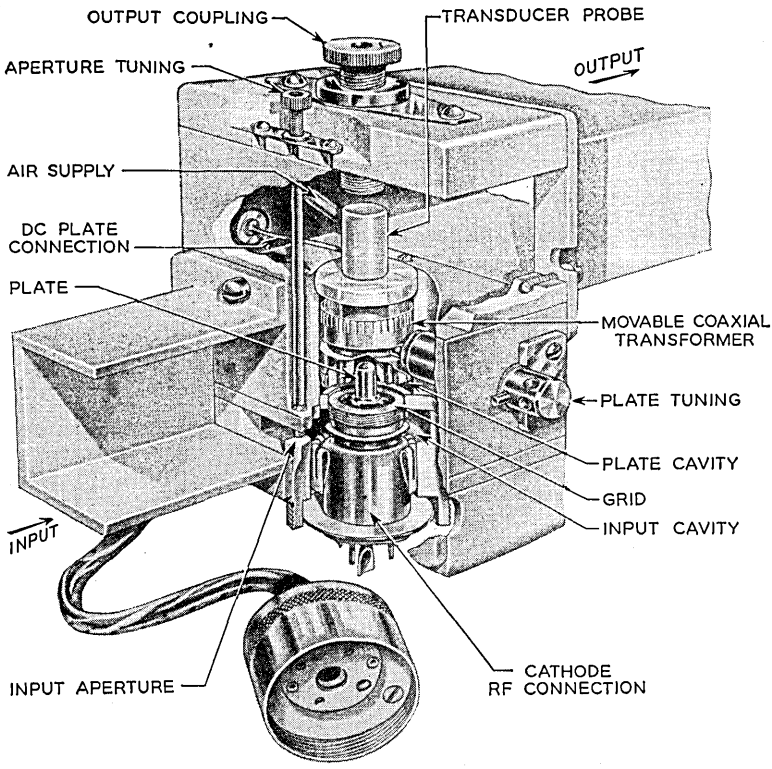


Fig. 11—416A tube cavity.

the modulator and the transmitter amplifiers and can best be described with the aid of a sectional view as shown in Fig. 11. The tube screws into the

cavity with the grid grounded directly to the body of the structure. The cathode of the tube is connected through an internal by-pass condenser to another part of the structure such that a cavity is formed around the tube between grid and cathode. An iris or aperture which is capacity tuned by a screw provides a means for coupling to the input waveguide. A coaxial cavity is formed around the plate which is tuned to the desired frequency by the movable coaxial transformer. The transformer couples the plate cavity to the transducer probe where the signal energy is transferred to the output waveguide.

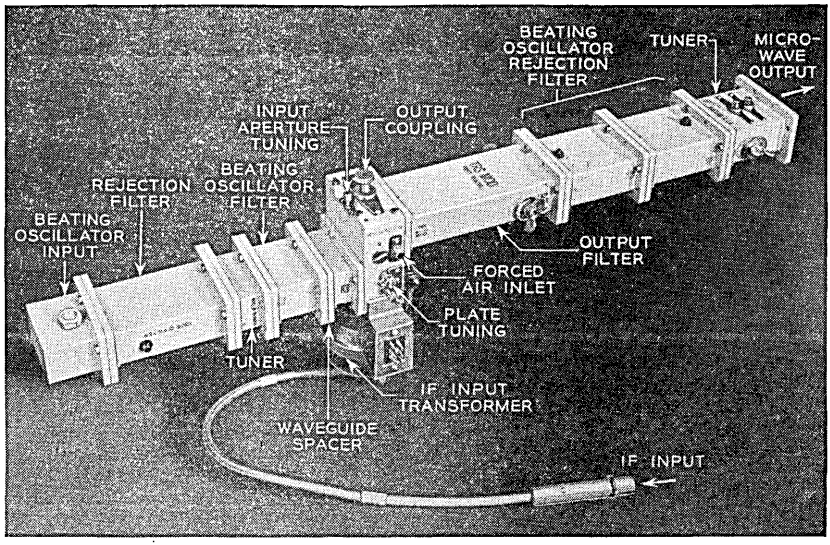


Fig. 12—Transmitter modulator.

The transmitter modulator is shown in Fig. 12. The oscillator power is applied to the cathode grid cavity through a tuner, a bandpass filter and a waveguide spacer. The IF power is applied between cathode and grid through a network which is mounted within a cylindrical compartment around the tube socket. The desired output sideband of the modulator is selected by a bandpass filter. Following this filter is a tuner unit which provides a means for adjusting the output impedance for a match with the following amplifier. A conversion gain of 9 db is realized in the process of shifting the IF frequency to the microwave band.

The modulator assembly is directly connected to the input of the transmitter amplifier, as may be seen in Fig. 8. An amplifier shown in Fig. 13 consists of three stages of 416A triodes mounted in cavity structures as

described above. The three stages are connected together in cascade through waveguide spacers and reactance tuners of such dimensions that the joining of each output cavity with the following input cavity (or filter section in the case of the output stage) forms a double-tuned critically coupled transformer. A flat over-all transmission characteristic is thereby obtained which is about 20 megacycles wide between points 0.1 db down. While capable of greater gain, the amplifier is adjusted to a gain of 18 db at an output power level of 0.5 watt. A double directional coupler in the waveguide between the transmitting amplifier and the transmitter channel separation filter provides monitoring and output alarm signals.

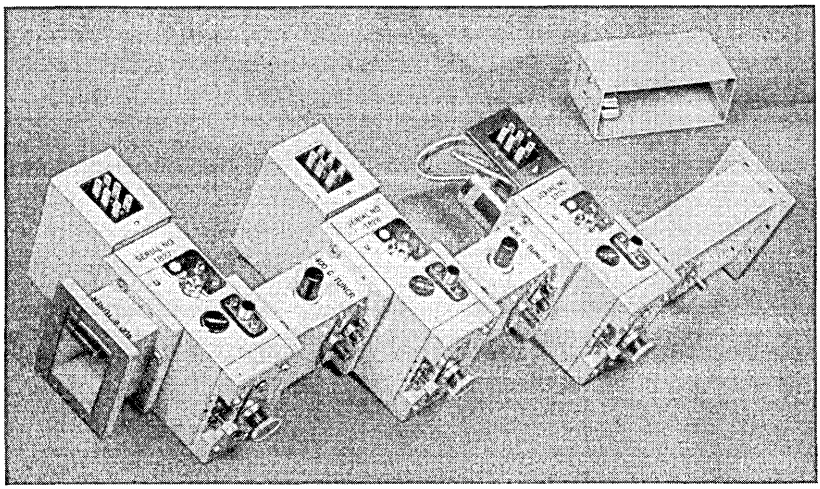


Fig. 13—Transmitter amplifier.

E. *Microwave Sources and Control Circuits*

The receiver control unit may be seen in Fig. 8. It contains a stabilized d-c amplifier for IF automatic gain control and level adjustments, and testing facilities for checking the performance of the receiver. The control unit also contains power controls and protection devices for the plate and filament circuits. The transmitter control unit contains controls for the application of power and bias to the transmitter and a means for metering various circuits.

The microwave generator, which furnishes about 200 milliwatts of beating oscillator power for the transmitter and receiver, is a stable microwave frequency source developed by harmonic generation from a quartz crystal in the vicinity of 18 megacycles. The multiplication takes place in six har-

monic generator stages, three doublers and three triplers. Only a few milliwatts of output power is required where the generator is used for the receiver beating oscillator alone, as in main stations or terminals. Here the final multiplier is operated as a sextupler, thereby permitting the elimination of the penultimate stage. At an auxiliary repeater, the receiving beating oscillator source is obtained from a 40-megacycle shifter converter, one input of which is from a part of the microwave generator output, and the other is from a crystal controlled 40-megacycle generator.

F. *Transmitter-Receiver Interconnections*

At auxiliary stations the IF output of the receiver is connected by a short coaxial line and 5 db resistance pad directly to the transmitting modulator in the same bay. This resistance pad is used as an impedance matching aid.

At main repeater stations the IF receiver output and the transmitter input are carried in coaxial lines to IF patching and switching equipment. With 30 to 60 feet of coaxial line between the receiver and transmitter, impedance match requirements are more severe than for short coaxial line connections. Here, a 6 db resistance pad is connected in the output line of the receiver and a 3 db resistance pad and buffer amplifier are connected in the input line of the transmitter modulator. The buffer amplifier consists of a single stage using a 418A tetrode and its gain may be set manually to provide -1 dbm to $+5$ dbm of signal power into the transmitter modulator as required. The bandwidth of the amplifier is approximately 20 megacycles and is sloped in such a manner as to approximately compensate for the small variation of loss over the band in the patching coaxial lines.

G. *IF Switching**

IF switching circuits are provided at terminals and main repeater points to facilitate maintenance operations as well as to provide flexibility for the changing requirements of network distribution. These switching and distributing operations are obtained by the use of unity gain amplifiers which are designated IF switching amplifiers and IF distributing amplifiers.

An IF switching amplifier functions as a single-pole double-throw switch for connection between intermediate frequency circuits of 75-ohm impedance. It has two input networks, each connected to a grid of a 404A pentode. The plates of the two tubes are connected in parallel to the output. Transmission through one or the other of the tubes is prevented by the application of a high negative grid bias to that tube. Switching the bias from one tube to the other thus permits the selection of either input signal. In most

* Prepared by T. R. D. Collins.

applications of the switching amplifier, signaling facilities are provided so that the switching operation can be controlled remotely.

An IF distributing amplifier provides three outputs from a single input, all at 75-ohm impedance. It consists of four 404A pentodes, the plate of one tube being connected to the grids of the other three tubes through an interstage network. Individual networks from the three output stages provide the desired distributing branches which are well isolated from each other electrically.

Switching and distributing amplifiers and a mounting framework are shown in Fig. 14. The two amplifiers have the same physical size and as many as five such units may be mounted in a frame on a plug-in basis. A number of such mounting frames are grouped and mounted on duct type bays to meet the needs of each switching and distributing location. Jack

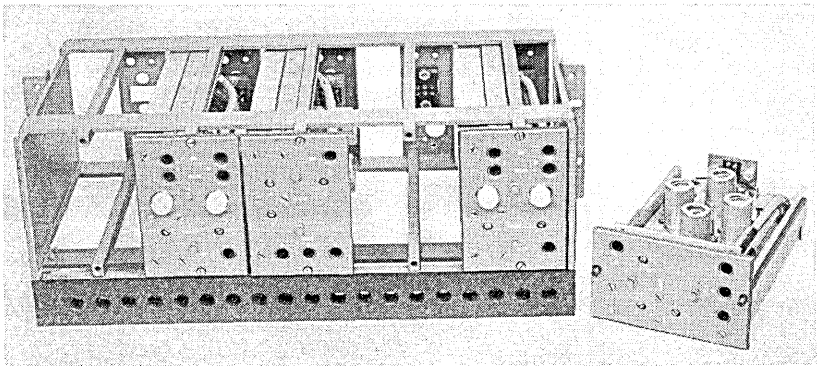


Fig. 14—Switching and distribution amplifiers.

fields associated with the mounting frames terminate the interbay coaxial trunks through which the switching and distributing connections are made.

Various combinations of switching and distributing amplifiers perform a large variety of interconnection functions within the system. Figure 15 indicates how these amplifiers may be used to replace a circuit which has failed by a spare circuit. At a transmitting terminal, the regular and spare channels may be paralleled. If a transmission failure occurs in channel 1 at one of the auxiliary repeater stations east of the main station, the failure of this channel is noted at the end of the system and service is switched to the spare channel 2. Since channel 1 is good except for the break east of the main station, the remote control for the switching amplifier in channel 1 is operated to switch output A of the channel 2 distributing amplifier to channel 1 radio transmitter. Thus channel 1 is connected in parallel with

channel 2 at this station and both a regular and a spare circuit are now available at succeeding stations.

H. *Automatic IF Switching**

At present IF switching is handled on a manual basis by attendants at the main stations or on a remote control basis over the order wire facilities. This type of switching is satisfactory for maintenance purposes but obviously is not fast enough to avoid a circuit interruption in replacing a circuit which has failed. The reliability of wire circuits will be difficult to meet in a long radio relay system without standby facilities because of vacuum tube failures and fading. Work is now under way to develop automatic IF switching facilities which will detect instantaneously a circuit

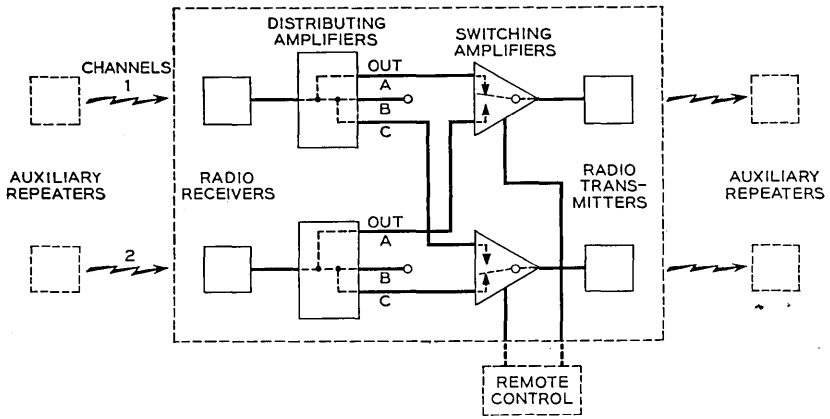


Fig. 15—IF switching and distributing amplifier. Interconnection diagram.

failure or an increase in noise on a radio channel and switch in a spare circuit for the poor section without circuit interruption. Fading data indicate that most deep fades which go beyond the range of the AGC circuit are of the selective type. Thus switching to a spare channel will provide frequency diversity advantages. With automatic switching it is believed the TD-2 System circuit outage time will not exceed that of wire circuits.

I. *Television Monitoring*

Visual monitoring facilities are provided at terminal and main repeater stations for observing circuit performance. Auxiliary repeater stations may also be so equipped in special cases. At transmitting or receiving terminals, monitoring connections are bridged to the video cables which run to operat-

* Prepared by T. R. D. Collins.

ing centers. The equipment units which make up the monitoring facilities are an auxiliary IF amplifier, an FM receiver, a video amplifier and a video monitor. A combination of these units is assembled in a bay to fit the needs of each monitoring location.

IV. FM Terminal Equipment

A. *General*

The TD-2 System will transmit a standard RMA black and white television signal or a band of message channels built up on a frequency division basis as provided by the coaxial cable message terminals. The FM terminal transmitter converts either of these signals to a frequency-modulated signal centered at 70 megacycles for application to the radio transmitter. The FM terminal receiver recovers the television or carrier signal from a frequency-modulated 70-megacycle signal. Thus the FM terminal equipment provides the connecting links between the TD-2 radio equipment and other facilities.

In a long system it may be necessary to bring the radio signal down to voice and back up to radio frequency many times in order to add and drop message groups. Each such process will require FM receiving and transmitting terminal equipment which consequently establishes severe linearity requirements for this equipment. An objective in the development of the terminals was to meet long haul systems performance requirements with sixteen pairs of terminals in tandem.

B. *FM Transmitter*

A functional diagram of the FM terminal transmitter is shown in Fig. 16. It accepts a signal from an unbalanced 75-ohm line and delivers an FM signal centered at 70 megacycles to the radio transmitter. The input level may be adjusted from 0.2 volt to 2.5 volts peak-to-peak with an output level of 13 dbm at an impedance of 75 ohms. For television transmission with a ± 4 megacycle swing the tips of the synchronizing pulses are at 74 megacycles and the picture white at 66 megacycles. For message service the nominal deviation is centered about 70 megacycles. For television transmission the output is automatically clamped to a predetermined frequency during each synchronizing pulse. These differences in operation are described in more detail below.

1. *Description*

The input signal to the FM transmitter is applied through an adjustable attenuator to a video amplifier consisting of two similar three-stage feedback amplifiers in tandem which have a combined gain of 42 db. The video

amplifier output is applied to the repeller of a deviation oscillator described below.

A microwave heterodyne method of generating a 70-megacycle FM signal was selected because it was found possible to design a highly linear deviator in the microwave region. It also allows separate tests to be made of the transmitter and receiver linearity and thus facilitates maintenance.

A reflex klystron oscillator may be frequency-modulated by superimposing a modulating signal on the repeller d-c voltage. The rate of frequency change with change of repeller voltage passes through a minimum near the

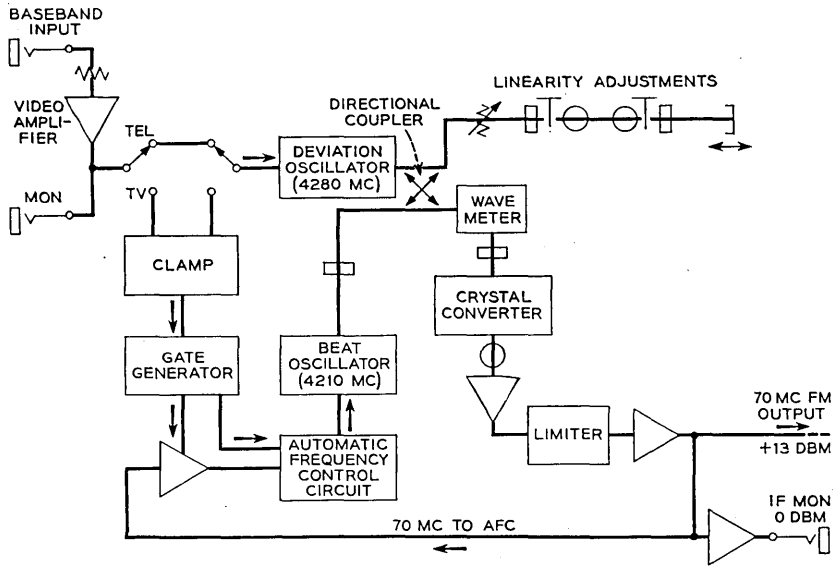


Fig. 16—Block diagram of FM terminal transmitter.

point of maximum power output. At a deviation of ±4 megacycles, the difference in FM sensitivity over the 8-megacycle swing would normally be sufficient to produce intolerable distortion. However, the operating frequency of a reflex oscillator is subject to modification by the load impedance seen by the oscillator. This effect is commonly called "pulling." In the deviation oscillator, this effect is made use of to provide deviation linearity over a range of more than 10 megacycles. The load circuit for the 4280-megacycle deviation oscillator consists of a variable attentuator, a short length of line, and a variable position short circuit. Adjustments of these two variables allows complete control of the reactance seen at the output

of the deviation oscillator. The length of circuit to the movable short is so chosen (about 35 inches) to provide the optimum rate of change of reactance with frequency. At optimum adjustment, the reactive component of the load pulls the frequency of the generator by just the amount necessary to straighten out the deviation curve. The deviation sensitivity is at the same time increased about 25%, which reduces the required video driving voltage.

A portion of the output signal of the deviation oscillator is fed through a directional coupler to a crystal microwave converter where it is mixed with a 4210-megacycle signal from another klystron to produce a 70-megacycle FM signal. The microwave output from the deviation oscillator is about 50 milliwatts, and after losses in the directional coupler and converter about one milliwatt of 70-megacycle FM output is available. This signal is amplified in a broad-band limiter-amplifier for application directly to the radio transmitter or indirectly through appropriate switching circuits.

2. *Clamper and AFC Circuit*

For television transmission the voltage supplied to the repeller of the deviation oscillator is clamped to a predetermined negative value during each synchronizing pulse in a conventional manner. This clamping action enables the transmission of video signal components down to direct current. For message telephone transmission the clamping circuit is disabled.

The automatic frequency control circuit used to control the frequency of the beat oscillator provides a high gain and stable AFC without a d-c amplifier. As shown in Fig. 16, a portion of the 70-megacycle output signal is diverted and after passing through a gated amplifier is applied to a discriminator. The discriminator network is of conventional design and the detector elements are germanium diodes. The direct-current output voltages are applied to the grids of two triodes acting as a pulse modulator. The anodes of these triodes are supplied with a high level positive pulse used for gating from a blocking oscillator associated with the clamper circuit. This oscillator is free running for message signals but is triggered by the synchronizing pulses when video signals are being transmitted. The unbalance voltage on the triode grids controls the amplitude and polarity of the pulse produced by this modulator. After two stages of a-c. amplification this error signal is combined with a second high level pulse from the same blocking oscillator source in a phase detecting circuit, and, after integration, the d-c. output of this detector is used for AFC. With television operation the gated amplifier operates only during synchronizing pulses, and the discriminator is adjusted for an output frequency of 74 megacycles. With multi-channel message operation, the gated amplifier is operated as a straight-through amplifier, and the discriminator is adjusted to hold an average output frequency of 70 megacycles.

C. *FM Receiver*

The FM receiver contains an IF amplifier, limiter, discriminator, and video amplifier, as indicated in Fig. 17. The input amplifier consists of two stages, each using a 404A pentode, with broad-band interstage networks. The two-stage instantaneous amplitude limiter has biased silicon varistors shunting the single-tuned plate loads of each of the 418A tubes. The bias voltages are so adjusted that the load impedance is high for signal voltages less than about one volt, and very low for any larger signal.

1. *Discriminator*

The discriminator circuit follows early conventional practice, in that two separately driven antiresonant circuits are used. The signal at the limiter output is fed to two 404A amplifier stages, one tuned above the signal band,

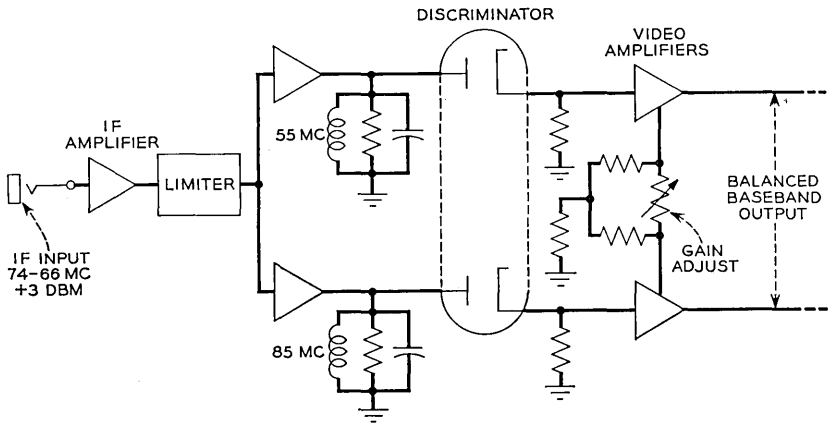


Fig. 17—Block diagram of FM terminal receiver.

the other below. The frequency-modulated signals produce amplitude variations of the voltage across these tuned circuits which are detected by diode rectifiers and applied to the video amplifier. A potentiometer in the cathode interconnection of the amplifier tubes provides a balance adjustment for the discriminator.

2. *Video Amplifier*

The video amplifier is a three-stage resistance-capacity coupled unit having negative feedback in each symmetrical half, and negative feedback to longitudinal voltages through a common cathode resistor. The gain is adjustable over a range of several db by means of a dual potentiometer which varies the common cathode resistance in each half of the amplifier. Whenever such an adjustment is made, a constant loop gain is maintained in the feedback system by varying simultaneously the local cathode de-

generation in the middle stages of the amplifier. A peak-to-peak voltmeter connected across one side of the balanced output is used to monitor the transmission level of television signals.

V. System Maintenance and Test Equipment

A. *General*

Most TD-2 stations are operated on an unattended basis. Test equipment is provided at each terminal, auxiliary and main station to perform the necessary maintenance functions. This consists of a radio test bay as shown in Fig. 18 for each auxiliary, main and terminal station, and an FM test console as shown in Fig. 19 at terminals and main stations where FM terminal equipment is provided. The philosophy is to provide sufficient test equipment at each station to isolate the trouble. When the unit in trouble requires extensive tests or repair, a station spare is substituted and the faulty unit is returned to a maintenance center. Maintenance centers are usually located in existing telephone offices along the route.

In maintaining the radio equipment each repeater bay is adjusted to provide a transmission band 20 megacycles wide, flat to within two-tenths of a db and centered about the assigned channel frequency. Trimming adjustments are provided on the receiver and transmitter to obtain this characteristic. This test involves the use of a swept signal source which is divided into a reference path and a path through the equipment under test, each of which is terminated in an identical detector. The outputs of these detectors are alternately applied to the vertical deflection amplifier of an oscilloscope at a 30-cycle rate, while a voltage proportional to the frequency excursion is applied to the horizontal amplifier. Generally, the vertical gain of the oscilloscope is adjusted so that a separation of one inch between the test and reference traces corresponds to a level difference of 1 db and the horizontal gain is adjusted so that one inch corresponds to a 10-megacycle frequency excursion. The reference trace is then matched to the test trace by adjustments of the equipment under test. The waveguide attenuators and directional couplers shown in Fig. 18 provide for testing over a wide range of levels.

B. *Radio Test Bay*

The radio test bay contains a microwave swept frequency oscillator, a combined microwave and IF power meter, a cathode ray oscilloscope, RF and IF wave meters, detectors and attenuators and associated power supplies.

The microwave sweep oscillator is adjustable in sweep range up to 70 megacycles over the 3700 to 4200 megacycle band. The frequency is swept

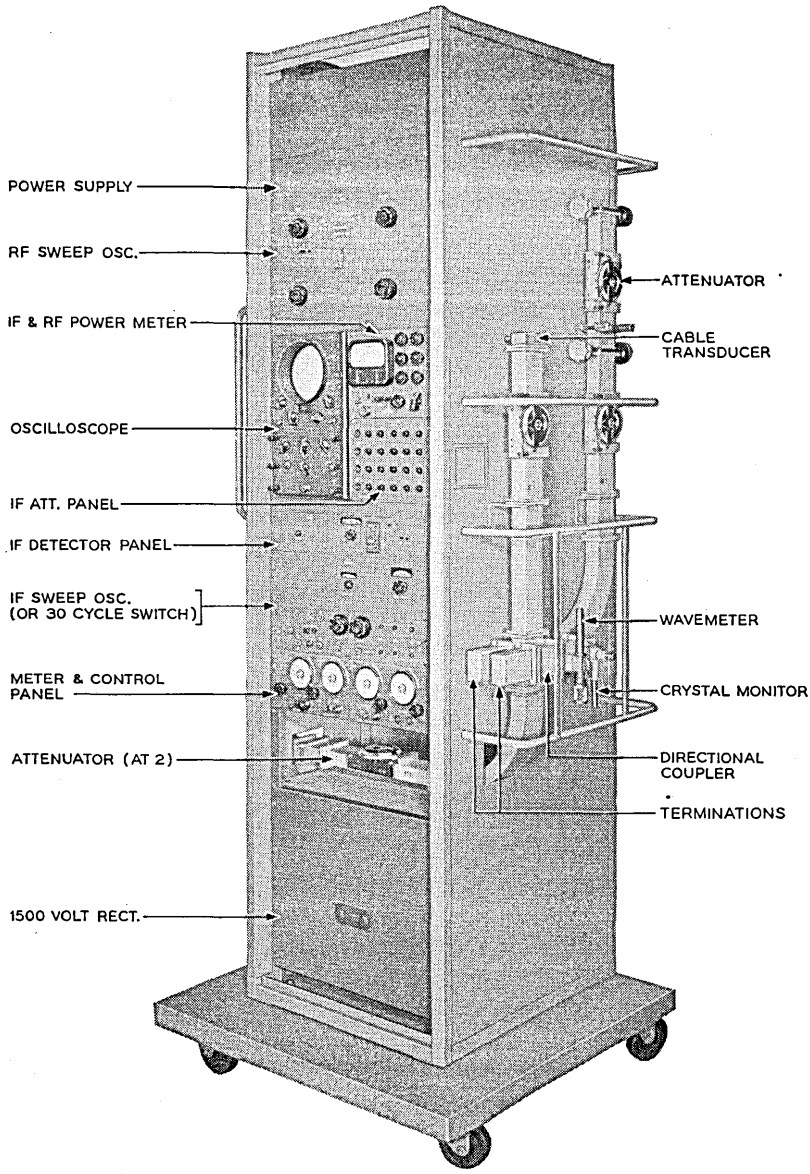


Fig. 18—Radio test bay.

by a motor driven reactive element in one of two cavities associated with the 402A velocity variation oscillator tube.

The RF and IF power meter consists of a temperature compensated thermistor bridge unit. It has separate input arrangements for the 3700 to 4200-megacycle and 50 to 90-megacycle bands. Accurate measurements of power may be made in the range from −10 dbm to +6 dbm.

The test bay used at maintenance centers has, in addition to the above equipment, a 50- to 90-megacycle swept frequency oscillator and associated detectors for the testing of intermediate frequency components. The opera-

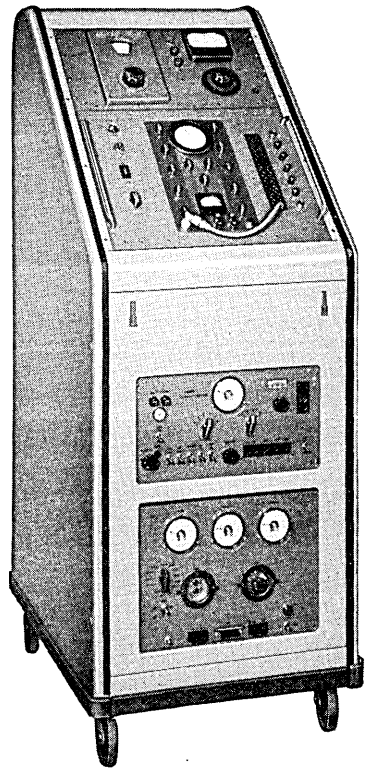


Fig. 19—FM terminal test console.

tions carried out at the maintenance center include the repair and realignment of defective equipment returned from the radio stations. The maintenance center test equipment includes facilities for accurate impedance match measurement in the microwave and IF range, for varistor matching tests, vacuum tube transconductance tests and general component tests which cannot be made at the radio station. Usually the maintenance centers are operated by the same staff that maintains the radio stations in the section.

C. *FM Terminal Test Console*

The terminal test console shown in Fig. 19 is used to measure FM deviation, linearity of the FM transmitter and receiver and for routine monitoring of wave forms at video frequencies. The equipment includes a conventional CW signal generator covering the range of 50 to 90 megacycles, a video "A" scope, an electronic switch and patching and terminating facilities. A rather unique linearity test set described below and an FM terminal receiver are also included.

1. *Deviation Measurement*

For deviation measurements, the IF signal being monitored is patched into one input of the IF electronic switch which switches between inputs at a 1200-cycle rate, and the CW signal generator into the other input. After detection by the FM receiver, the signals are applied to the oscilloscope and a straight line corresponding to the CW generator frequency is displayed superimposed on the video signal. By adjustment of the CW reference frequency, the instantaneous frequency of any signal component may be determined.

2. *FM Receiver Linearity*

For a measurement of linearity of the receiver discriminator, the linearity test set is connected to an FM transmitter which is patched to the receiver under test. The linearity test set supplies a low level 100 kc modulating voltage to the deviation oscillator of the transmitter and a high level 60-cycle voltage to the transmitter beat oscillator. For this test the transmitter AFC circuit is disabled. Under these conditions the signal applied to the receiver discriminator swings over approximately 10 megacycles at a 60-cycle rate and over a small range of less than one megacycle at a 100 kc rate. The 100 kc video component in the receiver output is then proportional to the slope of the discriminator response curve. The envelope of this 100 kc amplitude is recovered in the linearity test set and the a-c. component is applied to the oscilloscope vertical amplifier. The horizontal deflection is synchronized with the 60-cycle deviation. A 30-cycle switch changes the amplitude of the 100 kc signal by a calibrated amount to provide two separated traces on the screen and make the device self-calibrating.

3. *FM Transmitter Linearity*

For a measurement of transmitter linearity, the same setup used in the receiver test is made use of except that both the 100 kc small signal and 60-cycle large signal are applied to the deviation oscillator of the transmitter under test. The beat oscillator AFC circuit is allowed to operate with a time constant sufficiently rapid to follow the 60-cycle fluctuation of the deviation oscillator, but not the 100 kc component. Thus the 100 kc modulation component is applied over a 10-megacycle range of the deviation oscil-

lator characteristic, but is applied to the receiver at a fixed (70-megacycle) frequency, so that the receiver discriminator does not enter the measurement except as a fixed gain detector. While the transmitter is being tested as above, the magnitude and phase adjustments of the deviation oscillator load impedances are made as required to meet the desired linearity of deviation which is normally 1% over the 10-megacycle range.

VI. C1 Alarm and Control System*

The operation and maintenance of unattended repeater stations require a flexible and reliable alarm system whose performance is commensurate with the importance of the toll and television program services handled by the TD-2 System. The Cl alarm and control system has been developed for this purpose and, as its name implies, it serves two functions. The first is that of transmitting detailed alarm information from unattended repeater stations to the responsible alarm centers. The second function is that of transmitting orders, or remote control signals, from alarm centers to unattended stations.

The salient features of the C1 system may be summarized as follows:

1. It is a voice-frequency system, thus permitting its use with equal facility on cable pairs, open wire lines, or radio channels (or combinations thereof) capable of transmitting a 3000-cycle voice band.
2. It transmits a maximum of 42 separate alarms or indications from each unattended station to its associated alarm center.
3. It transmits a maximum of ten remote control orders in the opposite direction, that is, from an alarm center to each unattended station for whose operation it is responsible.
4. A maximum of twelve unattended stations may be associated with one alarm center.

A typical section of the TD-2 Radio Relay System is shown in Fig. 4. The alarm center for the section indicated is at Cuyahoga Falls, which in this case is also a maintenance center. Alarm centers and maintenance centers may be located at any attended central office or repeater station on existing cable and open wire routes.

Alarm signals are transmitted to the alarm center from the unattended station over a one-way, two-wire circuit as shown in Fig. 20. A four-wire local order circuit is used for voice communication between adjacent main radio stations and the intermediate unattended auxiliary repeater stations. The alarm centers and maintenance centers in that alarm section are also bridged on it. Remote control order signals from the alarm center are of such short duration that they can be transmitted without objectionable interference over one side of this four-wire local order circuit. An express

* Prepared by C. E. Clutts and G. A. Pullis.

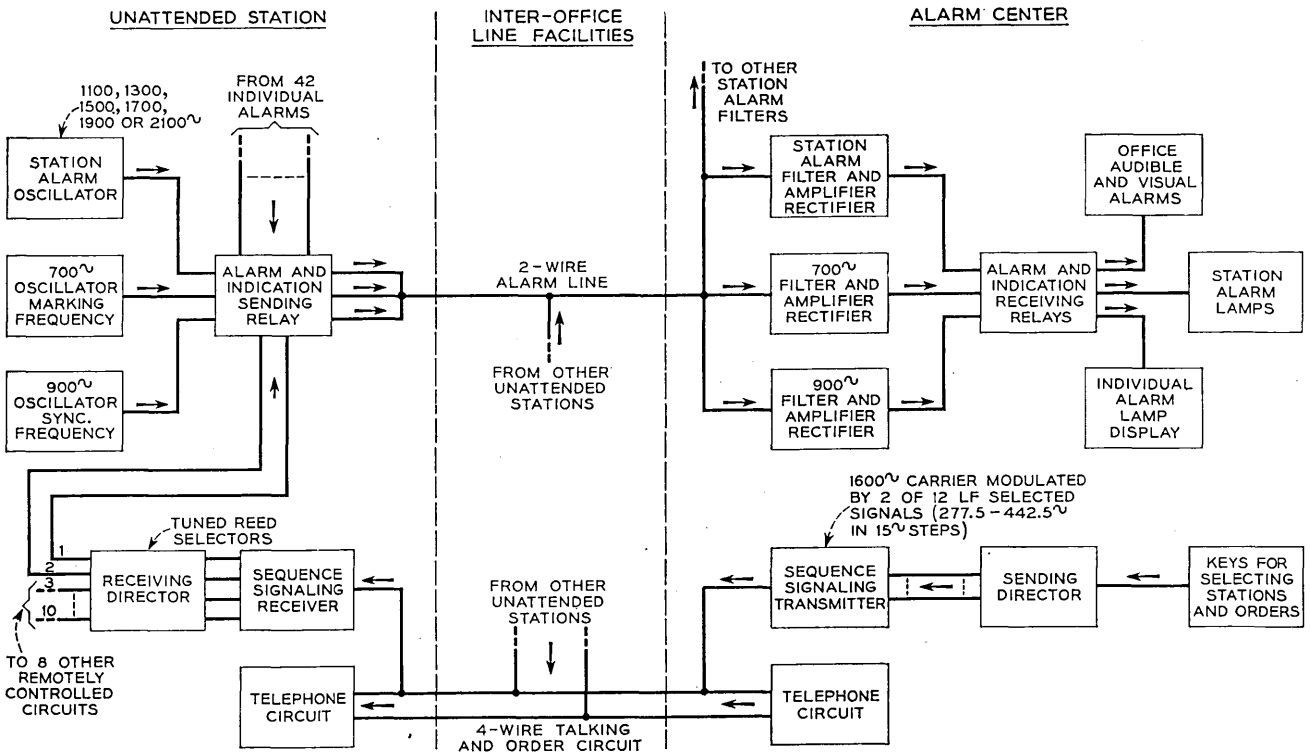


Fig. 20—Block diagram of C1 alarm and control system.

order circuit is used to link the terminal stations of a system with the main stations, alarm centers and maintenance centers for system-wise radio maintenance and traffic control.

A. *Station Alarms*

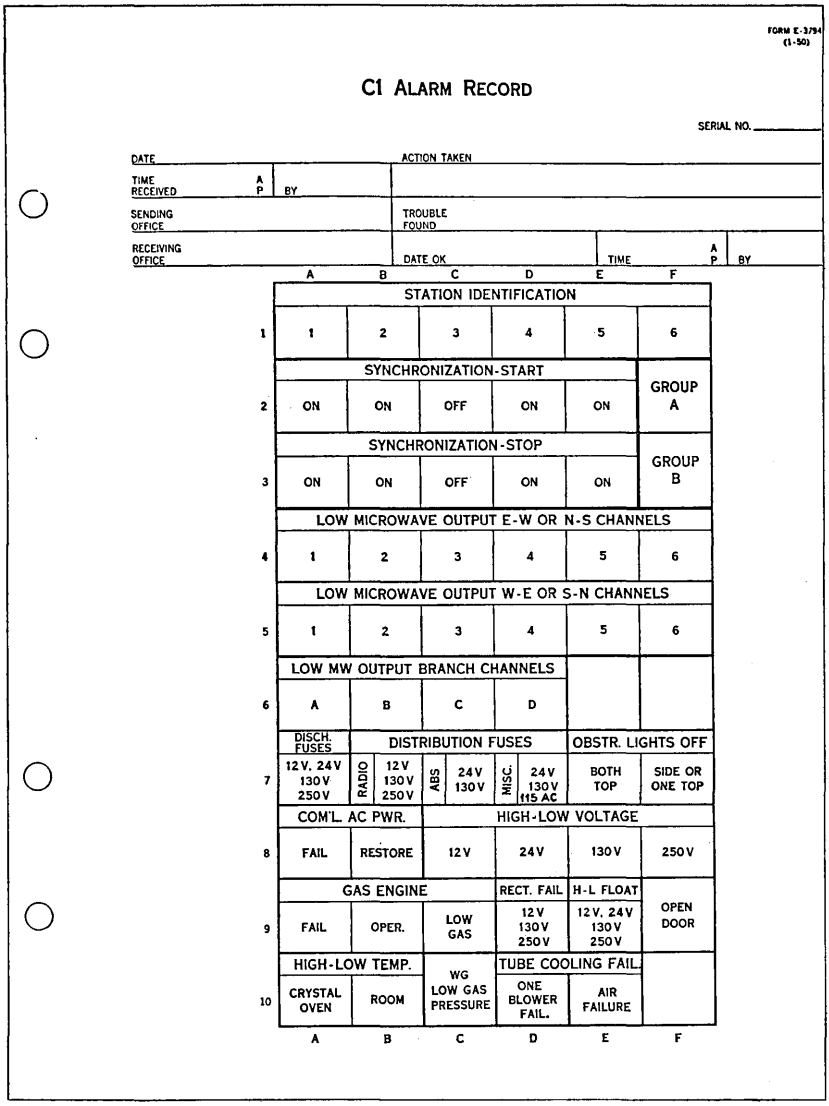
Each unattended station transmits a continuous and distinctive tone to its associated alarm center. Interruption of this tone for approximately ten seconds registers an audible and visual alarm at the alarm center, and, because each station is assigned a different frequency, the station whose tone is interrupted is easily identified. As many as six stations can report to an alarm center over a one-way, two-wire alarm line. Six more alarm sending stations can report to the same alarm center by bridging them on a second alarm line (usually in the opposite direction from the first), and providing a second set of receiving filters with associated amplifiers and detectors at the alarm center. In this manner an alarm center can identify the alarms from a maximum of twelve unattended repeater stations. The six station frequencies that can be used on.one alarm line are 1100, 1300, 1500, 1700, 1900, and 2100 cycles.

Each station alarm tone is selected at the alarm center by its associated receiving filter and individual amplifier-rectifier circuit. Automatic gain control action in the amplifier circuit permits a tone from the alarm line to vary ± 6 db from its normal value without interfering with the proper operation of the system.

B. *Individual Alarm Indications*

The station alarm reports that a particular unattended point is in trouble, but it does not tell what the specific trouble is. Supplementing the station alarm circuit is an individual alarm indication circuit that reports which, if any, of 42 possible alarm conditions exist at an unattended station. The alarm indication sending circuit does not start automatically but only in response to an order sent out from the alarm center. Thus, after receiving a station alarm, an attendant at the alarm center sends an order over the control system described later to that particular station directing it to scan the individual alarms and report those that have operated.

The report is transmitted over the alarm pair and received on a miniature lamp bank located in the key shelf of the alarm receiving bay at the alarm center. Of a total of 60 lamps in the key shelf, 42 are used for alarm indications, 8 for identifying the six east or west reporting stations, and 10 for checking synchronization of the indication sending and receiving circuits. Figure 21 is a copy of the form which is placed over the lamp display to



FORM E-3794 (1-50)

C1 ALARM RECORD

SERIAL NO. _____

DATE | ACTION TAKEN

TIME RECEIVED A P | BY

SENDING OFFICE | TROUBLE FOUND

RECEIVING OFFICE | DATE OK | TIME A P | BY

	A	B	C	D	E	F
	STATION IDENTIFICATION					
1	1	2	3	4	5	6
	SYNCHRONIZATION-START					GROUP A
2	ON	ON	OFF	ON	ON	
	SYNCHRONIZATION-STOP					GROUP B
3	ON	ON	OFF	ON	ON	
	LOW MICROWAVE OUTPUT E-W OR N-S CHANNELS					
4	1	2	3	4	5	6
	LOW MICROWAVE OUTPUT W-E OR S-N CHANNELS					
5	1	2	3	4	5	6
	LOW MW OUTPUT BRANCH CHANNELS					
6	A	B	C	D		
	DISCH. FUSES	DISTRIBUTION FUSES			OBSTR. LIGHTS OFF	
7	12V, 24V 130V 250V	RADIO 12V 130V 250V	ABS 24V 130V	MISC. 24V 130V 115 AC	BOTH TOP	SIDE OR ONE TOP
	COM'L. AC PWR.		HIGH-LOW VOLTAGE			
8	FAIL	RESTORE	12V	24V	130V	250V
	GAS ENGINE			RECT. FAIL	H-L FLOAT	OPEN DOOR
9	FAIL	OPER.	LOW GAS	12V 130V 250V	12V, 24V 130V 250V	
	HIGH-LOW TEMP.		WG LOW GAS PRESSURE	TUBE COOLING FAIL		
10	CRYSTAL OVEN	ROOM		ONE BLOWER FAIL.	AIR FAILURE	
	A	B	C	D	E	F

Fig. 21—C1 alarm record.

designate the lamps and provide a record of a specific alarm condition at an unattended station.

The alarm indication sending and receiving circuits utilize relay counting chains which scan over the 60 possible indications at a 5-cycle rate and

cause a 900-cycle pulse to be sent back to the alarm center for each indication scanned. Whenever an alarm condition or other indication is encountered, a 700-cycle pulse is transmitted simultaneously with the 900-cycle pulse. At the alarm center the pulses are selected by 700- and 900-cycle filters, and amplified and detected in the same manner as the station tones. The resultant d-c. pulses operate relays which in turn light particular lamps in the key shelf lamp display panel in the alarm center receiving bay whenever the two pulses are received simultaneously.

C. *Sequence Signaling Remote Controls*

As mentioned earlier the C1 alarm and control system is capable of transmitting as many as ten orders from the alarm center to a particular station in trouble. Typical orders to a repeater station may be an order to scan all alarm indications or an order to start the gas engine alternator. Sequence signaling transmitters and receivers are employed for the transmission of orders to the unattended repeater stations. Sequence signaling is an arrangement in which two separate signals sent in a predetermined sequence are translated by the receiver into an order. One hundred and thirty-two different orders can be transmitted from an alarm center through sequence combinations of two out of twelve modulating frequencies available in 15-cycle steps from 277.5 to 442.5 cycles. The C1 system makes use of 120 of these orders at those alarm centers which remotely control as many as twelve unattended repeater stations.

An attendant initiates an order by operating the key of the station to be called, the proper order key and a start key. This operation selects the proper two low frequencies which modulate a 1600-cycle carrier oscillator and the sequence in which they are sent. The incoming signal to the sequence signaling receiver at an unattended station is amplified and demodulated. The two low-frequency tones recovered from the 1600-cycle carrier are applied sequentially to the receiving director. The director identifies the tones by means of four accurately tuned reed selectors, recognizes their sequence and translates them into one of ten orders for that particular repeater station.

VIII. Power Equipment*

The TD-2 System is supplied by battery voltages of -12, $+130$ and $+250$ volts maintained by charging rectifiers which float the batteries within limits of $\pm 1\%$. A 24 volt battery to supply power to the C1 alarm and order wire circuits is also included in the power plan where necessary. The block diagram illustrated in Fig. 22 shows the inherent simplicity of the plant. During power failures the batteries carry the load until an automatic gas

* Prepared by J. M. Duguid.

engine alternator or diesel alternator is warmed up and assumes the load, at which time floated operation is resumed. An important characteristic is the absence of any direct switching in the load leads during power failures. The control equipment in all three battery plants is arranged for full automatic operation and additional charging rectifiers are switched in and out as required. After a power failure the rectifiers operate at full capacity until the battery is recharged, after which normal floating operation is resumed. Sufficient capacity is normally installed to give at least an eight-hour reserve

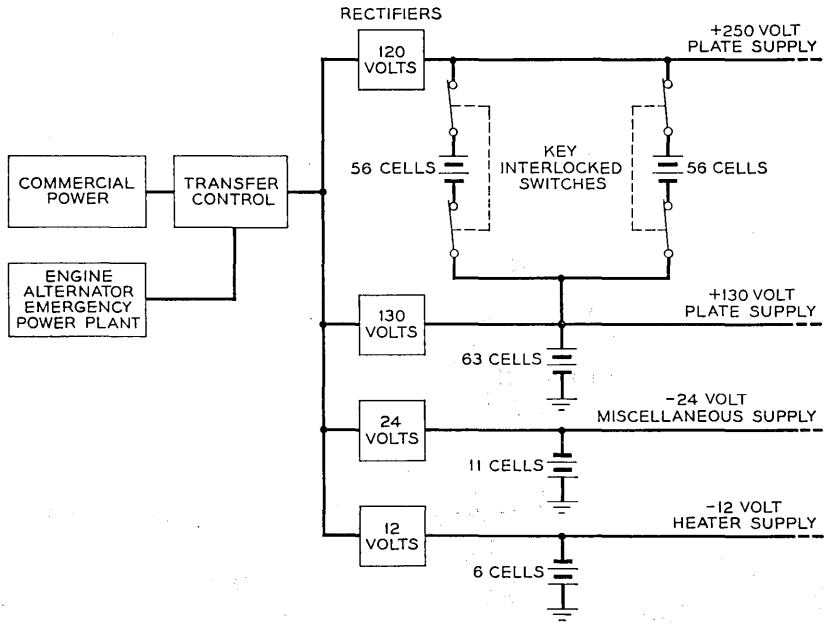


Fig. 22—Power plant block diagram.

to allow an attendant to get to the station in the event that the engine alternator fails to start.

A. *—12 Volt Supply*

The —12 volt heater supply consists of six battery cells floated by two or more parallel-connected 200-ampere full wave selenium rectifiers. The output voltage of the rectifier is controlled by a saturable reactor and regulating autotransformer in series with the primary of the stepdown power transformer which supplies the selenium bridge rectifier. The output voltage is automatically adjusted by the amount of d-c. current supplied to the saturable reactor by the electronic feedback control circuit in the rectifier. The

battery is floated at 13 volts and a discharge resistor in each fused discharge lead is adjusted during installation to drop the voltage to the normal limits of 11.0 ± 0.1 volts at the radio bays. Under 60-cycle a-c. power failure conditions before the gas engine or diesel alternator accepts the a-c. load, the radio bays may operate between their emergency limits of 9.9 to 11.5 volts.

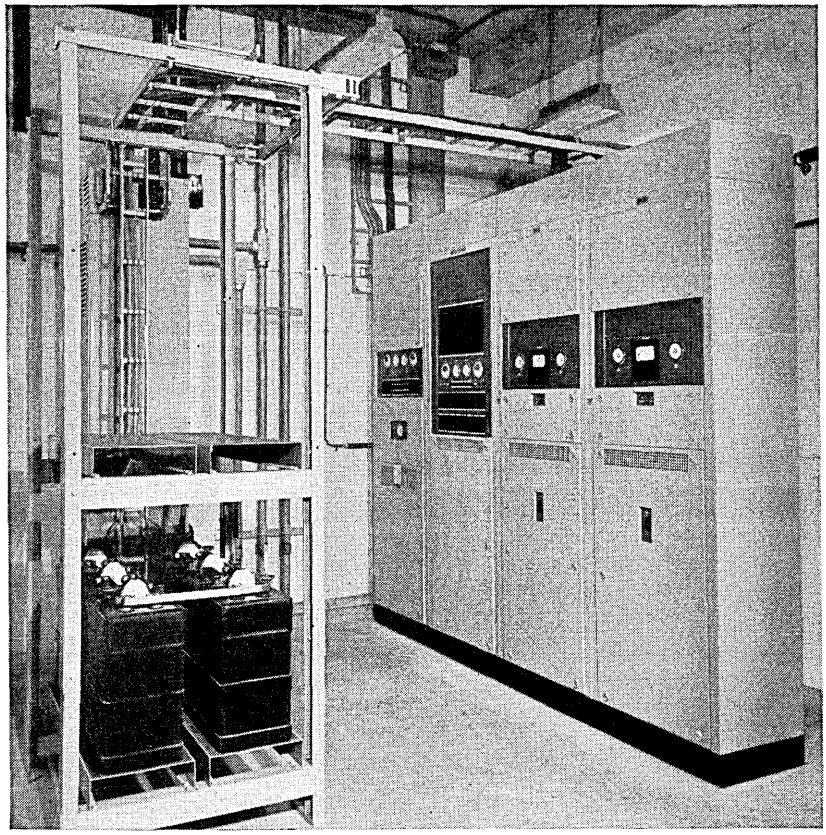


Fig. 23—12V and 24V power plant.

Figure 23 illustrates the installation in a typical tower of the −12 volt supply required for a main route of six radio channels in each direction.

B. *+130 Volt Supply*

The 130-volt plate supply consists of a 63-cell storage battery which is charged and floated by two to eight 8-ampere regulated tube rectifiers. As

shown on Fig. 22, this battery serves as the lower section of the 250-volt plate supply. Its capacity of 20 amperes is sufficient to supply the combined 130 and 250 volt loads. The regulated rectifiers normally float the plate battery at a voltage of 136 ± 1%. Under a-c. power failure conditions emergency limits of 116 to 140 volts are permissible. Due to the relatively high voltage involved and in order to insure maximum service and personnel protection, the rectifiers and their associated control and distribution equipment are mounted in sheet metal enclosures as shown in Fig. 24.

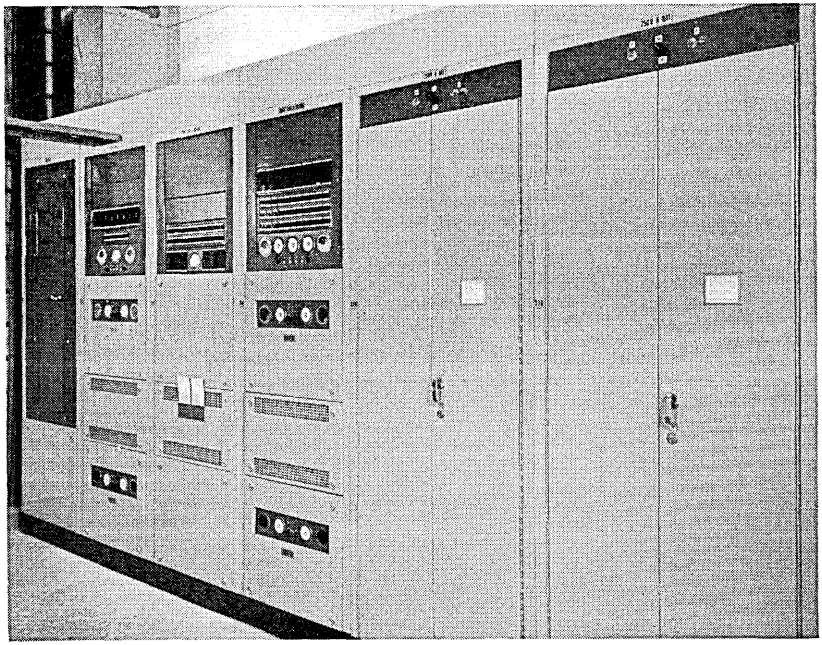


Fig. 24—130V power plant.

C. *+250 Volt Supply*

The 250-volt supply consists of duplicate 56 cell batteries in parallel which are in turn connected in series with the 63-cell 130-volt battery. Regulated thyratron rectifiers similar to those used in the 130-volt plate are connected across the 56 cells to float the load. The normal limits are 254 to 259 volts and the emergency limits are 224 to 266 volts. Each section of the 250-volt battery is housed in its own cabinet and key interlocked to protect maintenance personnel.

D. *−24 Volt Supply*

A −24 volt battery plant utilizing a regulated selenium charging rectifier capable of 6-ampere constant load supplies power for the alarm and order wire circuits. Voltage regulation is obtained by saturable reactors in a magnetic type of regulating circuit. This plant is shown as the extreme left bay in Fig. 23.

E. *Engine Alternator Reserve Plants*

The main route of the TD-2 system normally requires reserve engine alternators of 20 or 30 kw capacity. The initial sets used were of the automatic gasoline engine alternator type available in 20 to 60 kw capacity. The engines are fully automatic in operation. They accept the load after a predetermined period of commercial a-c. service failure and restore the load to the commercial service when it returns to normal. They are capable of long hours of operation under emergency conditions. Numerous alarms are available in the engine plant to indicate its status under all conditions. Recent development has made plants available similar to those mentioned above which are powered by automatic diesel engine driven alternators. It is expected that this latter type of engine will be used in the future where capacities of 20 kw or more are required.

VIII. Conclusion

The New York-Chicago section of the TD-2 transcontinental radio relay system was opened for service with the transmission of television network programs on September 1, 1950. The system was extended to Omaha on September 30, 1950. Similar systems were put into service during September between New York and Washington and between Los Angeles and San Francisco.

By the fall of 1951 a transcontinental microwave radio relay system will be in service between New York and San Francisco carrying television programs and hundreds of telephone messages. This system will augment present intercity toll facilities and, in conjunction with coaxial cable, will provide a nationwide network of broad-band channels capable of handling television transmission or large groups of telephone circuits.

The growth of broad-band channels during the next few years can be handled by the addition of channels to partially loaded TD-2 Systems and by new routes. Further expansion of radio relay systems into higher frequencies, 6,000 and 10,000 megacycle bands now set aside by FCC for common carrier use, appear to offer room for further expansion of systems comparable to TD-2.

References

1. H. T. Friis, "Microwave Repeater Research," *B. S. T. J.*, Vol. 27, No. 2, April 1948.
2. G. N. Thayer, A. A. Roetken, R. W. Friis and A. L. Durkee, "The New York-Boston Microwave Radio Relay System," *Proc. I.R.E.*, Vol. 37, pp. 183–188, February 1949.
3. W. E. Kock, "Metallic Delay Lenses," *B.S.T.J.*, Vol. 27, No. 1, January 1948.
4. C. E. Schooley and R. D. Campbell, "Spanning the Continent by Radio Relay," *Bell Telephone Magazine*, Vol. 29, No. 4, Winter 1950–51.
5. W. M. Marsters, "Radio Relay and Other Special Buildings," *Bell Telephone Magazine*, Vol. 29, No. 1, Spring 1950.
6. J. A. Morton and R. M. Ryder, "Design Factors of the B.T.L. 1553 Triode," *B.S.T.J.*, Vol. 29, No. 4, October 1950. (W.E.416A is the production version of the B.T.L. 1553 triode.)
7. A. E. Bowen and W. W. Mumford, "A New Microwave Triode: Its Performance as an Amplifier," *B.S.T.J.*, Vol. 29, No. 4, October 1950.

Deterioration of Organic Polymers

By B. S. BIGGS

(*Manuscript Received July 9, 1951*)

This paper is a general review of deterioration processes in polymers. It is pointed out that changes in properties with aging are usually the result of chemical reaction with components of the atmosphere. The mechanisms of these reactions and some methods of preventing or retarding them are discussed.

ORGANIC compounds which have enough inherent strength to be used as structural materials—e.g. rubbers, plastics, textiles, and surface coatings—belong to a class called polymers. The deterioration of these materials in service is a serious problem, probably equal in dollar value to corrosion of metals, and one or another aspect of it has been under study in the Laboratories for years.[1] Everyone is familiar with the tendering of cotton cloth and with the loss of strength of rubber with time, but except among people who work with them there is not a wide recognition of the fact that plastics also suffer extensive damage from the weather. This is probably because organic corrosion is usually not visible in its early stages even though deep-seated changes may be taking place throughout the body of the material. In its advanced state, however, such deterioration is easily observable, manifesting itself in loss of strength, erosion, warpage, development of cracks, loss of transparency, or in other ways depending on the material and the application. These changes are of obvious importance in most engineering uses, particularly in the Bell System where apparatus frequently is expected to last thirty or forty years, and it is therefore desirable that they be understood. It is the purpose of this article to review in a rather general way the causes and mechanisms of deterioration.

Even casual consideration reveals that both chemical and physical changes may occur. The loss of plasticizer from a plastic, for example, can induce warping and embrittlement without a change having occurred in the chemical nature of any of the component molecules. Alternate periods of high and low humidity can cause swelling and shrinking in such hydrophilic materials as nylon and cellulose acetate and if stresses are present this can result in permanent distortion[2] (Fig. 1). The swelling of rubber in contact with oils is another example of physical change (Fig. 2). These phenomena are generally well understood and are taken into account in careful engineering. The effect of chemical changes can be even more striking as illustrated in Figs. 3, 4 and 5, but their mechanisms are more obscure and require more detailed discussion.

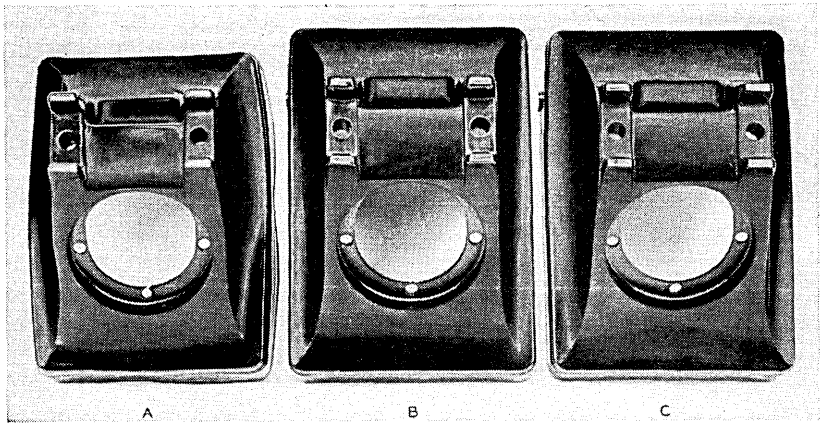


Fig. 1—Cellulose ester telephone housings.
A—Acetate after 7 cycles of high and low humidity.
B—Original.
C—Butyrate after 7 cycles of high and low humidity.

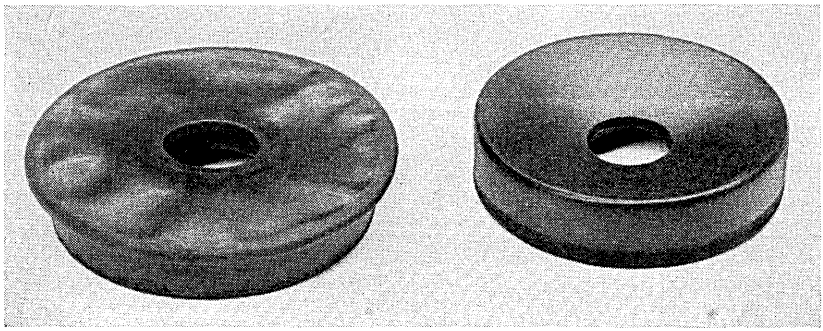


Fig. 2—Neoprene ear pad after one year's use, at left, and original at right.

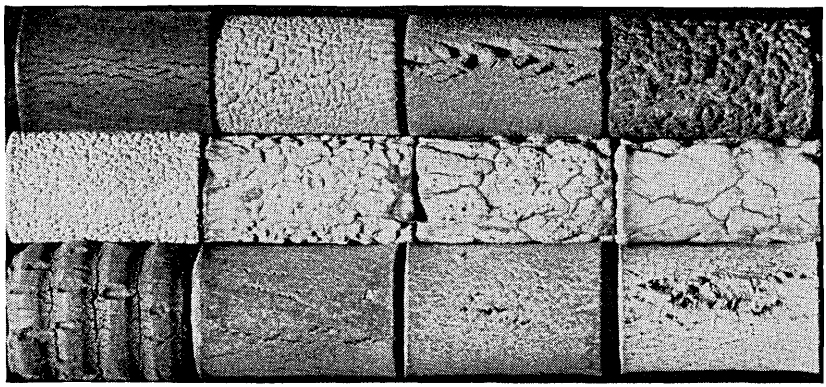


Fig. 3—Samples of rubber in various stages of weathering.

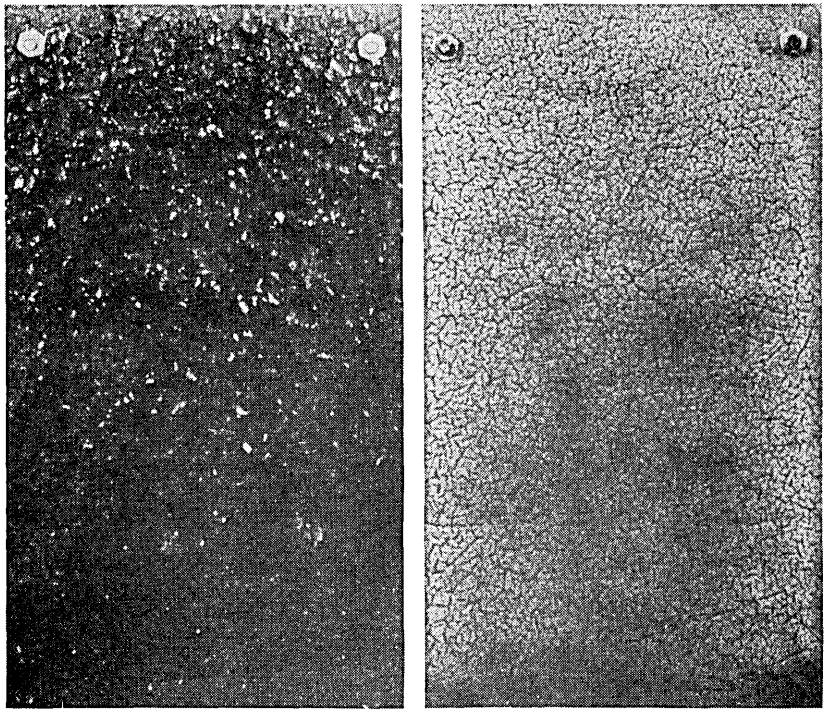


Fig. 4—Cellulose acetate panels exposed in Florida for six months.

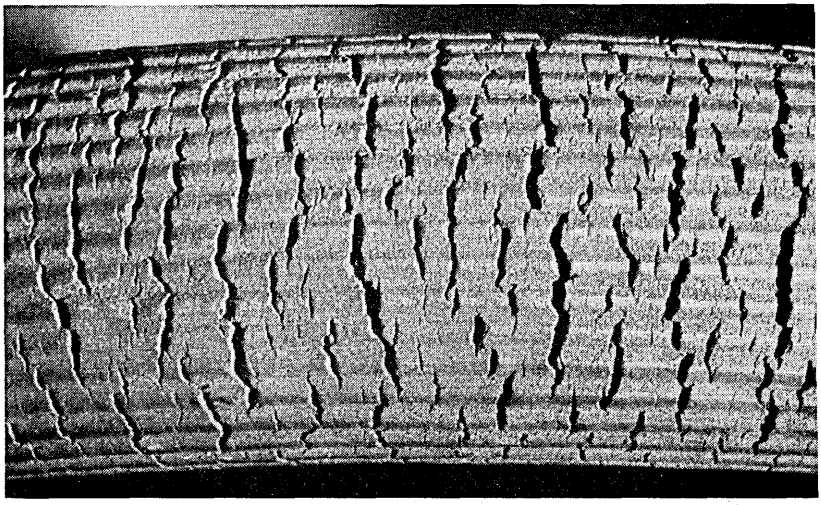
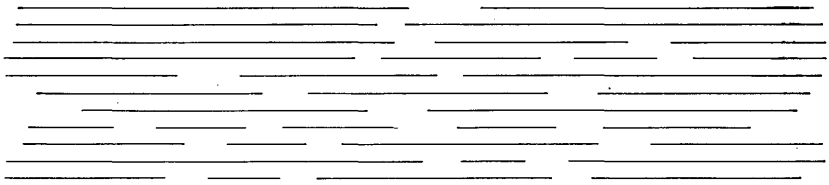


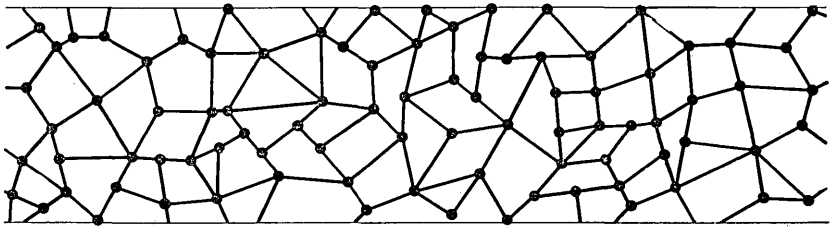
Fig. 5—Samples of rubber garden hose cracked by ozone.

The value of polymers as structural materials is derived entirely from the fact that they are composed of very large molecules. They are generally classified in two broad groups, the essentially linear or chain-like polymers comprising the thermoplastics and rubbers, and the very highly branched three-dimensional networks which are called thermoset materials. The fundamental difference between these groups is shown diagrammatically in Fig. 6 in which, for convenience, the linear polymers are shown as straight lines instead of in their usual randomly kinked shape.

The linear polymers are made up of molecules of finite average size, from a hundred to a thousand or more times as long as they are wide[3] and the



SCHEMATIC REPRESENTATION OF A LINEAR POLYMER



SCHEMATIC REPRESENTATION OF A THERMOSET POLYMER

Fig. 6—Schematic representation of a linear polymer, above, and of a thermoset polymer, below.

strength of the material is dependent on the size of these molecules much as the strength of a cotton thread is dependent on the length of the individual fibers of which it is composed. The forces holding the aggregate together are the cumulative interchain forces. In thermoplastics these forces may be quite strong. In rubbers they are weak until the rubber is vulcanized. Vulcanization connects the chain-like molecules into a loose three-dimensional network, but the number of cross-links is very low compared to typical thermoset polymers being only about one or two for every hundred chain atoms.[4] Vulcanized rubbers are therefore still largely linear polymers and their deterioration follows the pattern of the thermoplastics. The

thermoset materials, of which the phenolic resins are typical examples, are so highly interconnected that the molecular weight can be considered to be infinite. Each molding, for example, may consist of a single molecule. Because of their extensive internal cross-bracing their deterioration is usually a surface phenomenon.[5] It will be discussed later in this memorandum. The paragraphs which follow immediately will refer to linear polymers.

Any material chosen for an engineering application obviously must possess desirable characteristics and "corrosion" or deterioration changes these characteristics in some undesirable way. There are three ways in which a system of chain-like molecules can change: 1) the chains may be cut into smaller pieces, 2) the chains may be tied together by cross-links, and 3) the nature of any side groups along the chain may be modified. All of these changes have been found to occur during normal weathering of polymers and the properties of the product are determined by the extent of each change.[6]

The first type, chain scission, is usually the most serious because it cuts at the very essence of polymeric nature which is high molecular weight. As molecular weight is lowered, strength is lowered and ultimately is lost completely. To continue the analogy to a cotton thread, the individual fibers become so short that they cease to overlap each other adequately. Tough horny polyethylene, for example, deteriorates to something akin to paraffin wax. If chain scission occurs extensively in rubbers, portions of chains are cut loose from the relatively few cross-links and the product will appear to have become unvulcanized. This phenomenon is well known with natural rubber and is called "reversion".[7] (Fig. 7)

The second type of change caused by aging, the introduction of ties or cross-links, is not usually of great importance in plastics unless carried to an extreme when the rigidity and brittleness of thermoset polymers might result. As a matter of fact, the introduction of a few cross-links in a thermoplastic, without accompanying chain scission, probably serves to toughen the material. In rubbers, however, where high elongation is a desired property and is derived from the uncoiling of the molecules under stress, introduction of cross-links beyond those necessary for vulcanization tends to "shorten" the material and can eventually stiffen it to the point that it loses serviceability. The introduction of cross-links increases the density, and frequently when the surface of a plastic or rubber has been cross-linked extensively it develops an "alligator" or "mud crack" pattern resulting from excessive shrinkage.

The third type of change, the modification of side groups, normally has little effect on the strength of a polymer, but may have a pronounced effect on the dielectric properties, solubility, moisture absorption, etc., depending

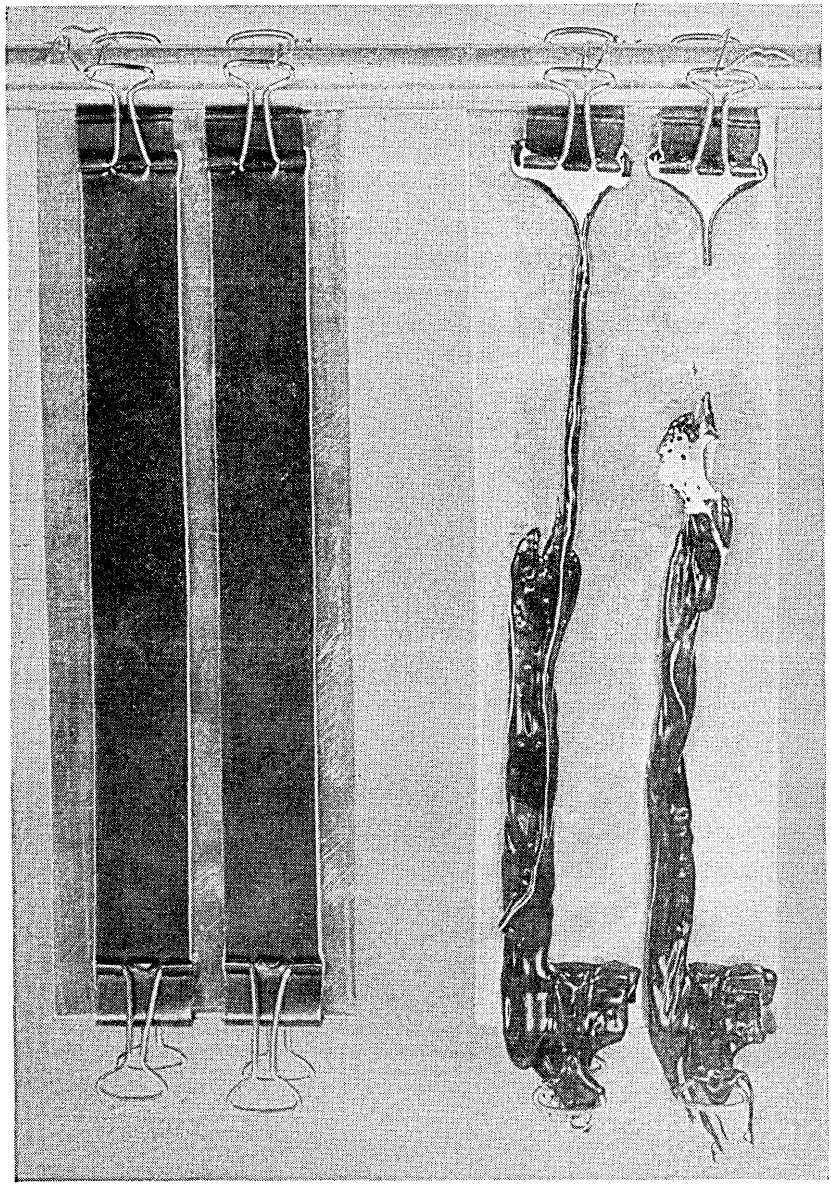


Fig. 7—Natural rubber tapes before and after oxygen bomb treatment.

on the nature of the groups introduced or modified. As indicated above, during normal deterioration all of these types of change are proceeding

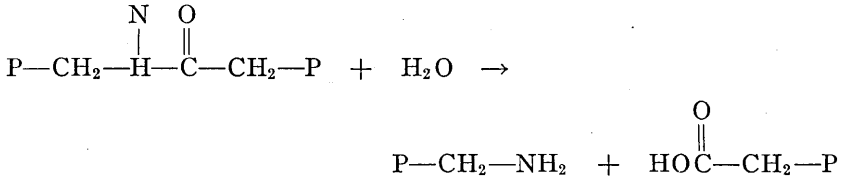
simultaneously to greater or less extent. The rates of the reactions vary from one material to another, and the same conditions which degrade natural rubber to a soft gum may cause neoprene or GR-S to become harder and stiffer.

Returning now to the thermoset polymers, one sees that neither occasional chain scission nor occasional cross-linking can have an important effect on the mechanical properties of a thermoset polymer since every part of the structure is tied to the rest of it by many bonds. For this reason the most conspicuous changes of thermoset materials on exposure to weather are on the surface and are of the third type discussed above.

From the viewpoint of physical structure all of the elements of deterioration are covered in the above paragraphs. However, nothing has been said about the agencies which cause the chemical changes or the mechanisms by which they are brought about. These agencies and mechanisms become the most important objects of study. One type of change—the cross-linking of molecules—in certain cases can occur by self-reaction under the influence of heat or light in complete absence of other chemicals. Self-reaction is not, however, an important effect in materials which are in engineering use. The changes which lead to the loss of utility of polymers during aging are caused by *chemical reaction with the environment.* Usually this environment is the atmosphere. There are normally three substances in the atmosphere which under various circumstances may be considered reactive toward organic compounds, namely water vapor, ozone, and oxygen. The next section will discuss the ways in which these chemicals bring about the destruction or organic polymers.

Water

The chemical reaction of water with organic compounds is limited to materials which contain hydrolyzable groups either as part of their original composition or as a result of oxidation. Examples of such groups are esters, amides, nitriles, acetals, and certain types of ketones. The reaction is illustrated with an amide linkage, the unaffected portions of the molecule being represented by the letter P:



When these vulnerable groups are present as substituents on a polymer chain composed exclusively of carbon-to-carbon bonds their hydrolysis

may affect certain properties of the material (dielectric constant, power factor, insulation resistance, water absorption) but in general the molecular weight of the polymer is unaffected. When the vulnerable group is a link in the skeletal chain, however, the result of hydrolysis is much more serious because it constitutes scission of the primary chain and hence a lowering of molecular weight. Polymers which are subject to this kind of scission are polyesters, polyamides, cellulose and cellulose derivatives (ethers and esters). Hydrolysis is accelerated by high temperature and is catalyzed by acids and alkalies, and hence many polymers of the classes listed are stable only when kept neutral. Polyesters in particular are usually easily hydrolyzed and it is this fact which has been the main barrier to their greater commercial utilization. Hydrolysis as such is a well known reaction and is taken into account in current engineering with materials which are subject to it. For example, nylon molding powder is shipped dry in sealed containers to keep the moisture content low until after the molding operation which requires that the nylon be heated to a high temperature,[8] and cellulose esters undergo repeated careful neutralizations and washes after esterification to reduce acidity.[9] The extent to which water plays a role in the deterioration of hydrocarbon materials which are first attacked by oxidation is not yet known, but it is certainly secondary to the oxidation itself. An important effect of rain in outdoor weathering is the washing away of water soluble oxidation products with consequent exposure of new surface. Another effect is the removal of water soluble compounding ingredients. This may be distinctly beneficial as in the case of polyester rubbers vulcanized by acid-producing catalysts,[10] or harmful as in certain polyvinyl chloride formulations which contain water soluble protective agents.

Ozone

Ozone is an extremely reactive chemical which is present in the air in extremely small amounts, ranging from 0 to 10 parts per hundred million. In this low concentration it has not been shown to have any effect on chemically saturated materials, but it is a very serious hazard for unsaturated compounds. Natural rubber and several synthetic rubbers fall in this class (Fig. 8). Ozone is a specific reagent for carbon-to-carbon double bonds, forming an ozonide which undergoes rearrangement resulting in chain scission.[12] When rubber is not being stretched the attack of ozone appears to be negligible, but when it is under stress the attack has very serious consequences resulting in transverse cuts which may sever the piece of rubber.[11, 13] Apparently the initial attack, starting in regions of highest local stress, cuts enough chains to cause a crack to open, and this exposes new surface and concentrates the stress so that the crack grows.

The practical significance of the reaction of ozone on rubber is very great since almost all rubber articles which undergo any appreciable stretching in service are in some degree subject to attack. Exceptions are articles composed of certain specialty rubbers such as silicones, Hypalon*, and some Thiokols. These are saturated materials and hence are not attacked. Neoprene and Butyl rubber are more resistant than natural rubber or GR-S, Butyl because it is only slightly unsaturated, and neoprene because its double bond is considerably deactivated by the adjacent chlorine atom.[14]

$$\begin{matrix} Cl & H \\ | & | \\ -C=C- \end{matrix}$$

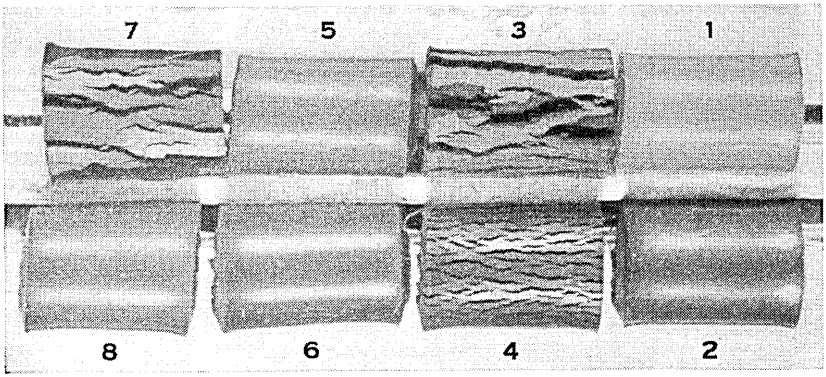


Fig. 8—Samples of various rubber compounds after exposure to ozone. (1) Silicone; (2) Hypalon; (3) Buna-N; (4) natural rubber; (5) & (6) Neoprene; (7) GR-S; (8) Butyl.

Large additions of pigments or plasticizers lower the ozone resistance of neoprene. The measure which has been found most effective for protecting rubber compounds from ozone is the inclusion of several percent of wax. The amount required varies with the type of wax, the polymer, and the other compounding ingredients, the absorptive power of any pigments present being an important factor. By proper compounding neoprene can be made extremely resistant to the attack of ozone, and the other unsaturated rubbers can be greatly improved. The chief effect of temperature changes on the cracking of rubber by ozone is in changing the solubility of wax in the rubber. At elevated temperature the wax film may redissolve and leave the rubber unprotected. This is illustrated in Fig. 9 which shows a tape wrapping which has been attacked on the sunny side, not by the light, but by ozone enabled to reach the rubber because the sun's heat had redissolved the wax in it.

* A chlorinated, sulphonated polyethylene manufactured by the Du Pont Company.

Oxygen

The degradative agent of most general attack and of greatest economic importance is oxygen, which is capable sooner or later of bringing about change in almost any organic material. Even disregarding the oxidation of dead organic matter in nature, which is aided by bacteria and fungi, one finds many examples of oxidation familiar to the layman. The development of rancidity in foods is a common one. The production of sludge-forming acids in engine oils, and the spontaneous combustion of rags soaked with linseed oil are others. The loss of strength of cotton cloth after a few years of service is very largely due to oxidation although mildew or other fungus attack may have played a part depending on circumstances.[15, 16, 17] That changes

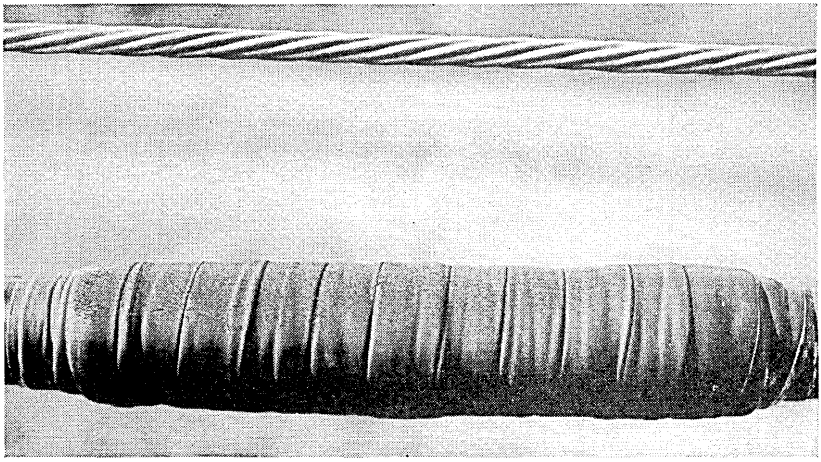


Fig. 9—A tape-wrapped splice after 6-weeks of exposure outdoors. An example of the acceleration of the ozone reaction by heat.

in polymers are indeed the result of oxidation is easily demonstrated in the laboratory by exposing samples to heat or to ultraviolet light in the presence and in the absence of oxygen. The results of such an experiment are shown in Table I, in which solution viscosity is used as a measure of molecular weight. It is seen that in nitrogen neither heat nor light brought about any serious loss of molecular weight.

Similar work has been reported with natural rubber with the conclusion that in an inert atmosphere rubber would retain its original properties "for at least thirty years".[18]

Gross Effects of Oxidation of Polymers

Severe oxidation of organic polymers results in the drastic changes mentioned in the introduction and is easily detected. Photo-oxidation of poly-

ethylene,[19] nylon and cellulose esters,[20] for example, causes crazing, cracking, embrittlement, and in extreme cases granulation of the sample. (Fig. 10) In polyvinyl chloride it leads to hardening and discoloration.[21] In natural rubber, GR-S, and neoprene it causes the development of "mud-crack" patterns or "alligatoring" of the surface and loss of elongation. Thermal oxidation leads to embrittlement of thermoplastics, to "shortening" or loss of elongation in neoprene,[22, 23] nitrile rubbers, and GR-S, and to reversion or the development of tackiness in Butyl rubber and sometimes in natural rubber. As pointed out earlier, these varying effects result from the relative rates of cross-linking and chain-scission reactions. The mechanisms by which oxygen can attack polymers are discussed in the next paragraphs.

TABLE I
SOLUTION VISCOSITY OF CELLULOSE ACETATE BUTYRATE

Original	1.77
After 4 Weeks Exposure to UV Light at Room Temperature	
In Nitrogen	1.60
In Oxygen	.15
After 150 hrs. at 150°C	
In Nitrogen	1.78
In Oxygen	.52

Mechanism of Oxidation Leading to Chain Scission

The reaction of organic compounds with atmospheric oxygen, frequently called "auto-oxidation" or "autoxidation", has been of interest to chemists for a long time and a voluminous literature on the subject has accumulated.[24, 25, 26] While most of the work done has been on small molecules rather than on polymers it is becoming apparent that much of the mechanism of oxidation is the same and what has been learned on small molecules can be applied to large.[27, 28, 29] This is fortunate since polymers do not lend themselves readily to normal chemical manipulations. While it might be expected that different compounds would be attacked by oxygen in different ways a general mechanism has emerged which appears to be characteristic for aliphatic hydrocarbon structures and is probably applicable to many of the polymeric materials in current engineering use. It can be described as an autocatalytic free radical chain reaction.[30, 31, 32]

The sequence of events is believed to be as follows: Free radicals are produced in the substrate from the energy of heat or of light. They may arise from the decomposition of unstable groupings such as the —O—O—

bond in peroxides or by the dissociation of a relatively more stable bond such as —C—C— or —C—H. Needless to say, the ease with which such cracking occurs is influenced by chemical structure. These free radicals, which may be produced in very minute amount, react with oxygen to form

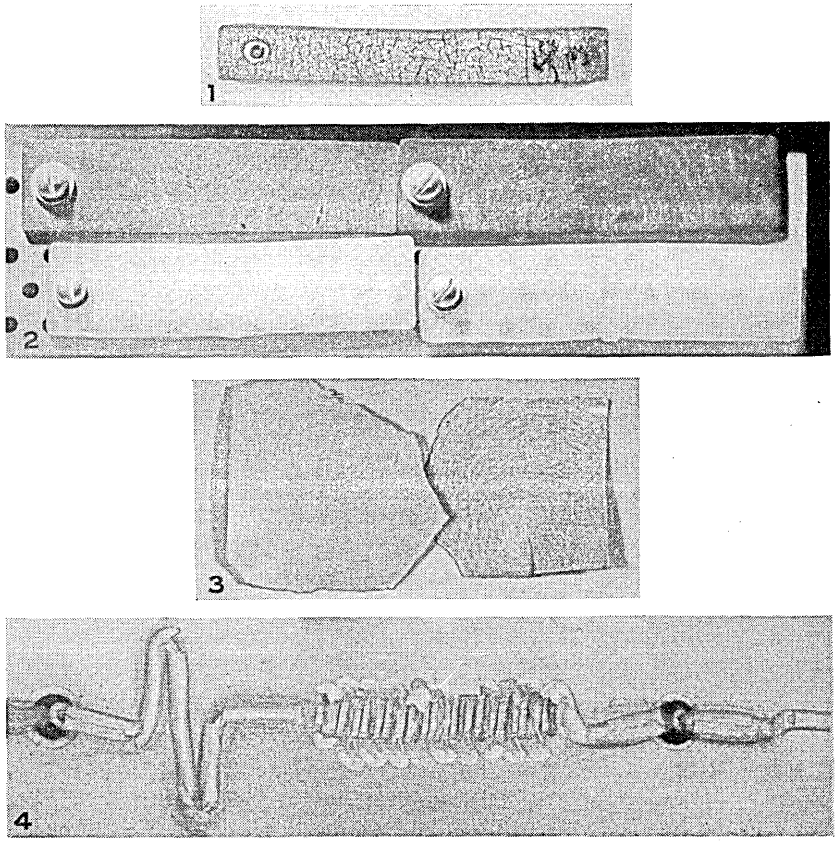


Fig. 10—(1) Cellulose acetate exposed six months at Murray Hill, N. J.
(2) Nylon test panels exposed 5 months at Yuma, Arizona.
(3) Clear Polyethylene sheet exposed 3 years at Murray Hill, N. J.
(4) Clear polyethylene coated wire exposed 3 years at Murray Hill, N. J.

peroxidic radicals. This is illustrated by the following chemical equation in which the radical is represented by the letter R and the fact that it is "free" or reactive is indicated by the dot.

$$R\cdot + O_2 \rightarrow R—O—O\cdot$$

The peroxidic radicals are also reactive entities, but their affinity is for hydrogen atoms and they tend to abstract hydrogen from some other molecule of substrate, thus:

$$ROO\cdot + RH \rightarrow ROOH + R\cdot$$

(These equations were written by Bäckström for the oxidation of benzaldehyde[30] and have been adopted by many others.)[31, 6, 33] The latter reaction results from molecular collision with formation of intermediate additive complexes which decompose into the indicated products and in general many ineffective collisions will occur before reaction takes place. The molecule of substrate which loses hydrogen in this way is now a free radical and it repeats the process, reacting first with oxygen and then with another molecule of substrate. This *linear* chain reaction continues until two radicals unite by collision with each other, thus terminating two chains. The word linear is italicized in the previous sentence to emphasize that this part of the reaction is not in itself autocatalytic. The autocatalytic nature of the oxidation stems from the fact that the product of the reaction as outlined is a hydroperoxide, ROOH. Such compounds are relatively unstable and slowly decompose into free radicals which initiate new chains. This might go as follows:

$$ROOH \rightarrow RO\cdot + \cdot OH$$

$$RO\cdot + RH \rightarrow ROH + R\cdot \text{ and } \cdot OH + RH \rightarrow HOH + R\cdot$$

Thus, though the original rate of generation of free radicals from cracking might have been very low, the combined rate increases quite rapidly since each molecule of peroxide produced in the chain reaction becomes a potential source of new radicals. Eventually the rate reaches what appears to be a steady state and finally levels off. A typical oxygen absorption curve for a liquid hydrocarbon is shown in Fig. 11. The region of fast reaction has received attention from those interested in the oxidation of small molecules but it is unimportant to people interested in polymers because it has been shown by various workers that only slight oxidation is required to destroy the useful properties of a polymer.[34] By the time oxidation has proceeded far enough to be getting into a rapid rate it has already resulted in enough chain scissions to have lowered the molecular weight below useful levels. (A simple calculation will illustrate this point. Suppose a polymer molecule whose molecular weight is 32,000 reacts with one molecule of oxygen (mol. wt. 32) and a chain scission results. The molecular weight of the polymer molecule will have been halved by reaction with .1% of its weight of oxygen. Not every reaction with oxygen results in chain scission of course;[27] but, even so, the amount of oxygen required to ruin the polymer is very small.)

The principal effect of the reactions described above is to introduce the hydroperoxide group into the polymer at various points. It is in the decomposition of these peroxides that chain scission occurs. Studies of the decomposition of the tertiary peroxides produced by oxidation of various dialkyl

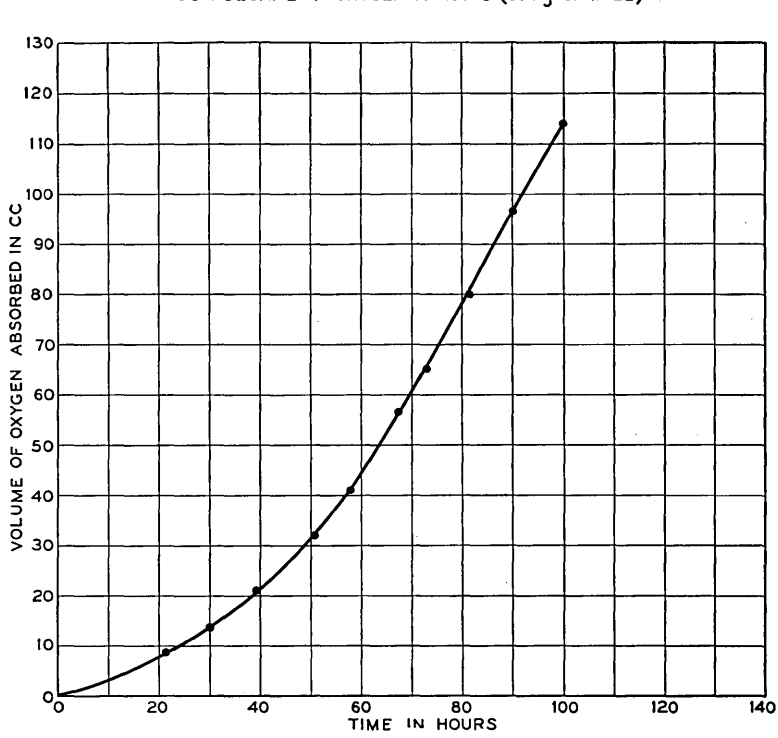
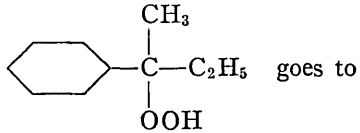
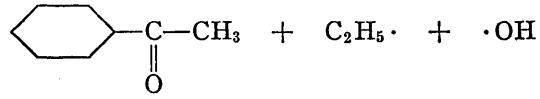


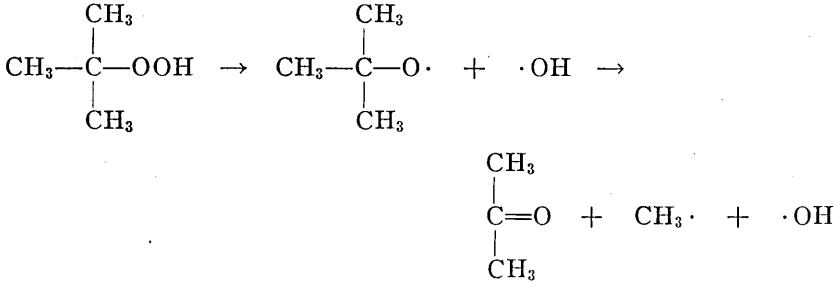
Fig. 11—Octadecane in oxygen at 105° C. (2.4 g sample)

phenyl methanes have shown that the product is invariably an alkyl phenyl ketone in which the alkyl group is the shorter of the two originally present.[35]

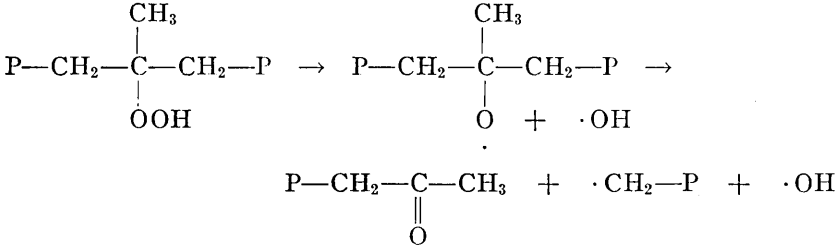
Thus:  goes to



the C—C bond attaching the ethyl group having been broken. The decomposition of tertiary butyl hydroperoxide has been shown similarly to proceed by the following reactions:[36, 37]

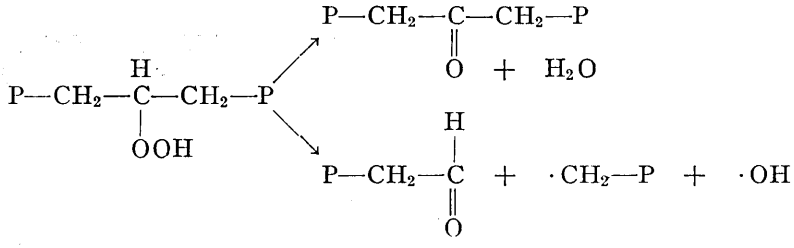


Here again a carbon-to-carbon bond has been broken. The implication of this work for a substituted polyethylene, for example, is quite clear. If oxidation occurred at a tertiary hydrogen as it most probably would,[38, 39, 40] the decomposition of the resulting tertiary peroxide would be as follows (the unaffected portions of the polyethylene chain being represented by the letter P):



Thus the polyethylene chain would be cut.

The course of events is less clear for secondary peroxides. Here at least two paths are possible:



If the decomposition proceeds according to the top arrow, no chain scission results, whereas if it takes the lower course the polymer is divided. The aldehyde is subsequently oxidized to acid. The fact that short fatty acids are produced in the oxidation of straight chain hydrocarbons such as octadecane

is proof that secondary peroxides or peroxidic radicals can decompose by the chain splitting process. That they do not decompose exclusively by that mechanism is shown by the high yield of tetralone obtained from the decomposition of tetralin hydroperoxide.

The mechanisms outlined, while certainly not complete, are adequate to account for the chain scission type of oxidative deterioration of many plastics and rubbers. The degradation of chlorine bearing plastics such as polyvinyl chloride and polyvinylidene chloride, while also being caused by oxygen and being energized by light and heat, is not believed to follow the patterns out-

TABLE II
FIELD RESULTS ON SAMPLES OF NATURALLY AGED NEOPRENE JACKETING
(FROM DROP WIRE)*

	Original Months Exposure at	Tensile Strength psi 2218	Elongation, % 330
Chester, N. J.	15	2635	215
	31	2655	225
	57	2510	205
Stone Harbor, N. J.	21	1990	190
	64	2485	185
	78	2615	175
Miami, Fla.	14	2540	195
	48	2215	140
	60	2260	125
	74	2410	150
	87	2450	130
	109	2520	120
San Antonio, Tex.	11	2395	160
	22	2300	145
	34	2585	180
	45	2165	135
Brawley, Cal.	15	1980	165
	58	2405	165

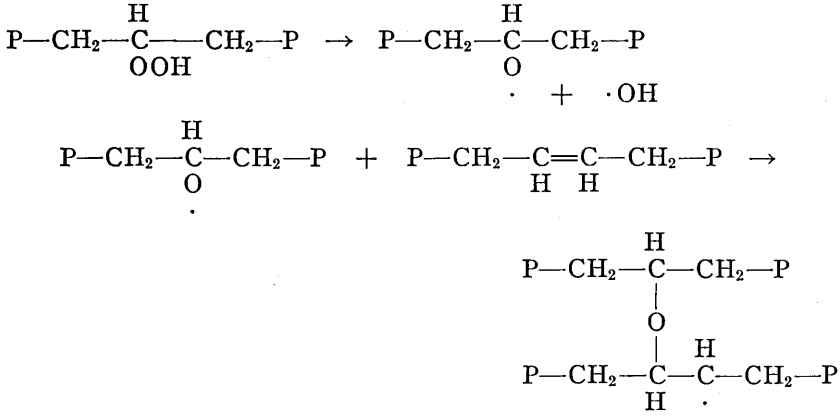
* From a paper by G. N. Vacca, R. H. Erickson and C. V. Lundberg[22]

lined above. The first step here is reported[21,41,42] to be the elimination of hydrogen chloride with introduction of a double bond, which makes the loss of more HCl easier and also increases the oxidizability.

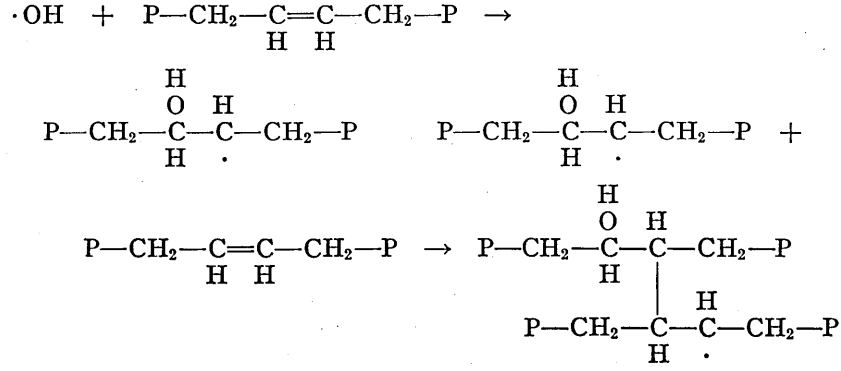
Cross-Linking Resulting from Oxidation

The second important effect of oxidation of polymers is cross-linking. This is of great consequence only with unsaturated compounds and these are principally the rubbers. If cross-linking is the dominant reaction (as it usually is on neoprene, GR-S, and the nitrile rubbers) the result is a decrease in elongation and an increase in hardness without a loss in tensile strength.

The strength may actually increase as shown in Table II. The rubber tends to "shorten" and ultimately ceases to be rubbery. Carried to an extreme condition, oxidized rubbers can resemble hard rubber or the phenolic resins. The detailed mechanisms of cross-linking are not worked out but certain deductions can be made about the reaction. That it is not just a polymerization of the double bonds can be shown by the fact that its rate in the absence of oxygen is extremely slow. That it is probably induced by free radicals can be inferred from the work on "vulcanization" of unsaturated polyesters with peroxides in which it was evident that a free radical attacking a double bond initiated the cross-linking reaction.[10, 43] The sequence might be as follows:[28]



Thus a link has been introduced. The new radical can react with another radical, it can react with oxygen to form another peroxide group, it can react with a double bond in another chain to form an additional cross-link, or when antioxidant is present it can react with antioxidant. The hydroxy radical resulting in the original decomposition of the peroxide could initiate a similar series of reactions resulting in one or more cross-links as follows:



The factor that determines whether or not cross-linking will be dominant in the aging of an unsaturated material must be the chemical structure of the polymer (and its peroxide.) The mode of decomposition of the peroxide which, of course, is a function of structure probably has the most important effect. While cross-linking can occur in saturated materials, as shown by the vulcanization of saturated polyester rubbers with peroxides, its rate is never high enough to result in a condition that could be called deterioration. Both polyethylene and cellulose acetate butyrate can undergo enough gelation on outdoor exposure to become insoluble, but if this were the only change occurring their toughness would be improved rather than degraded by it. Their deterioration in strength is due entirely to chain scission.

Modification of Side Groups by Oxidation

All the oxidation reactions discussed result in the introduction of oxygen into the polymer composition. If the polymer is one which already contains a high percentage of oxygen such as cellulose or even nylon, this may have little effect. If the polymer is a hydrocarbon, however, its power factor will be raised markedly. As a matter of fact the measurement of power factor is a very sensitive way of detecting the addition of oxygen to polyethylene. Except where the polymer is being used for its low power factor, however, the change in side groups will be secondary to the change resulting from chain scission and cross-linking.

Acceleration of Oxidation

The foregoing description of the mechanisms of auto-oxidation makes apparent several ways in which oxidation may be accelerated beyond what might be called the natural rate for a pure material. Since oxidation is a free radical process an obvious way to accelerate it is to add free radicals or materials which produce free radicals. Addition of peroxides to organic compounds generally accelerates the rate of oxidation.[33] Similarly the oxidation of a relatively stable material is accelerated if there is left in it a small amount of a chemical which itself is easily oxidized to peroxides. For example, an addition of turpentine greatly accelerates the air-oxidation of paraffin wax.[44] The addition to polyethylene of an unsaturated polymer such as natural rubber would probably have a similar effect.

It is apparent that the amount by which the rate of oxidation of a substrate is accelerated by peroxides, whether the latter are added as such or are self-generated, is dependent on the rate of decomposition of the peroxide. The latter rate can be accelerated by the presence of certain metallic ions and hence they act as catalysts for oxidation reactions. Copper is particularly active in this regard in natural rubber, and the rubber industry long ago learned to avoid it[45] (Fig. 12). Other metals which have been found to

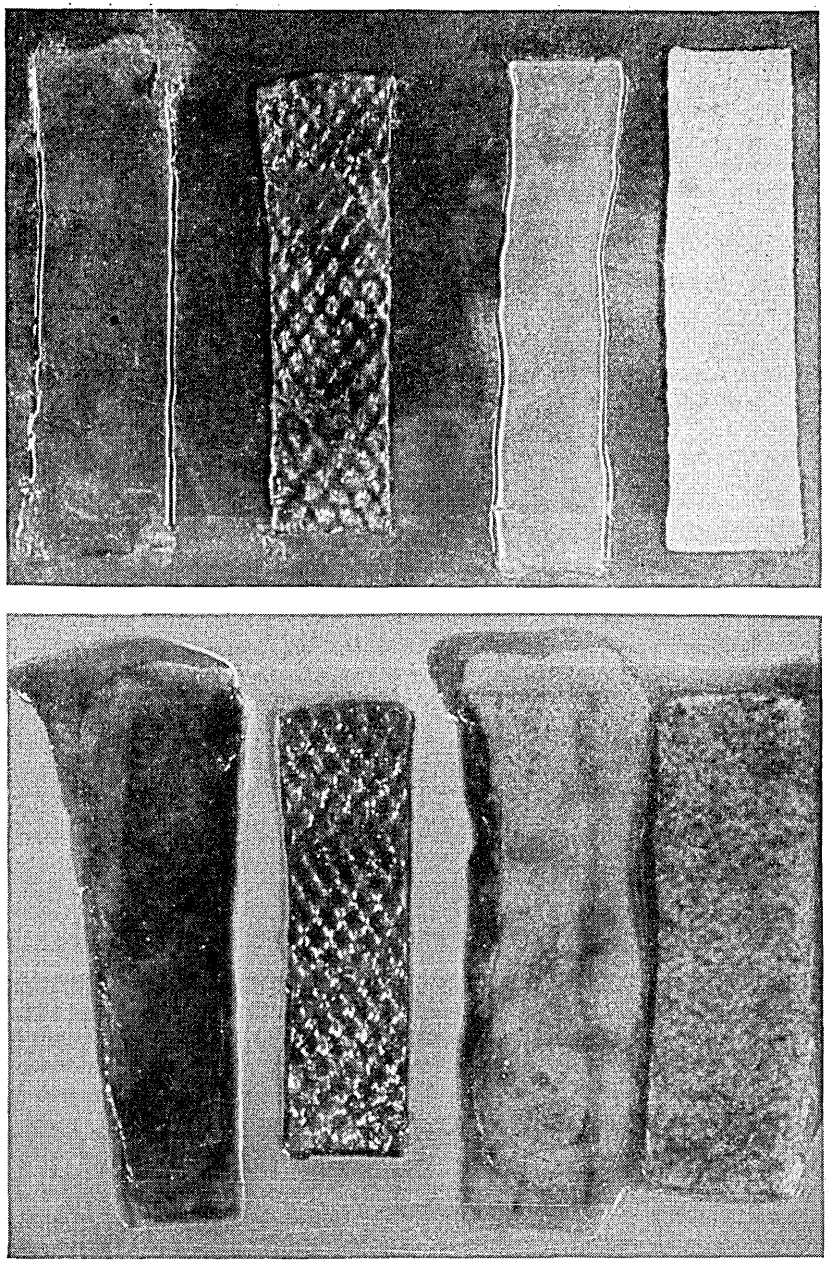


Fig. 12—Various samples of natural rubber oven aged on tin, above; and on copper, below.

cause poor aging at various times are cobalt, manganese, and iron.[46] Since the "drying" of paint is an oxidative reaction, and since rapid drying is a desirable feature, the paint industry has found it advantageous, to use certain metallic salts as "paint dryers".[47]

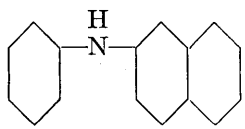
Retardation of Oxidation

Antioxidants

It was discovered by Moureu and Dufraisse about 1918[48] that the oxidation of many organic compounds could be very greatly retarded by the addition of small quantities of certain other chemicals, which they called "antioxygens." Although the mechanisms which they postulated for the action of these materials were later found to be incorrect, their discovery led to the wide use of such protective agents in industry, particularly in rubber which needs this protection badly. The action of what are now called "antioxidants" becomes clear when one understands the free radical chain mechanism of oxidation outlined above. Antioxidants are chain stoppers.[4] By interposing themselves in the chain reaction they terminate it by giving rise to relatively inert free radicals[50, 51] (stabilized by resonance). For example, the antioxidant, designated HA, could act in the following way:

$$ROO\cdot + HA \rightarrow ROOH + \cdot A$$

In this case the antioxidant satisfies the peroxidic radical by giving it the hydrogen atom it needs, but the residual radical ·A is not sufficiently reactive toward oxygen to continue the chain. A typical antioxidant is β-phenyl naphthylamine,



It was pointed out earlier that many ineffective collisions of the radical ROO· with substrate molecules occur before reaction takes place. If the reactivity of ROO· toward HA is sufficient that few ineffective collisions take place, then small concentrations of HA in the substrate will be adequate to stop each chain at a very early stage. This not only saves all those substrate molecules which would otherwise have become links in these chains but, by so doing, it limits the number of molecules of peroxide produced and thus keeps the rate of initiation of new chains at a low level. The degree of protection by antioxidants varies with the length of the "natural" chain reaction (which is a function of the ratio of effective to

ineffective collisions in the absence of an inhibitor and depends on chemical structure) and on the efficiency of the antioxidant but, in some cases, particularly with liquids, very remarkable protection is obtainable as shown in Fig. 13. (The oxidation of the control sample is so fast at this high temperature that the autocatalytic period is not evident.) The effect is less in solids but is still of great value. Antioxidants are of the greatest benefit where the rate of initiation is low, a condition usually true of thermal oxidation. The

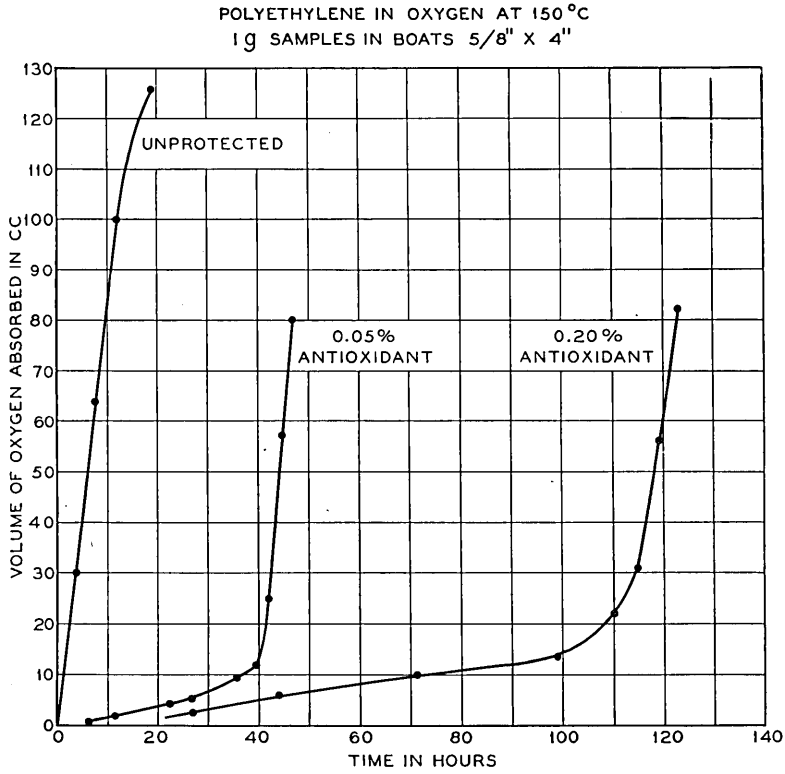


Fig. 13—Polethylene in oxygen at 150° C, 1 g samples in boats $\frac{5}{8}'' \times 4''$.

reason is that since antioxidant is *consumed*[52, 27] in doing its job the rather limited amounts which can be added from a practical point of view (usually not over 1 or 2%) do not last long if the rate of chain formation is very high. This explains the oft stated fact that an antioxidant is far more effective if added before oxidation starts, than if added after oxidation has proceeded for a while.[53] In the latter case enough peroxide will have been produced to overwhelm the antioxidant relatively quickly. This is also the explanation of

the fact that antioxidants are limited in their effectiveness against photooxidation. The rate of initiation of chains in a material exposed to sunlight is so great that any antioxidant present is used up relatively fast. Furthermore, the oxidation of the antioxidant itself is rapid in sunlight and hence if it were not removed in the one way it would be in the other.

There are many chemicals in current use as antioxidants and more are being created all the time. Most of them are either phenols or aromatic amines. It is frequently asked why one antioxidant is more effective than another, if indeed such differences do exist. The answer is not altogether clear but certain statements can be made about it. In the first place, for any given substrate there are usually several antioxidants which are equally good. However, gradations of effectiveness of many commercial antioxidants can be demonstrated. Many factors can exert an influence on this. Some are solubility in the substrate, volatility, inertness toward the substrate. Beyond these are the reactivity of the antioxidant toward free radicals, both hydrocarbon and peroxidic, and the relative stability of the free radical left when the antioxidant reacts. Undoubtedly, some intermediate level of reactivity is desirable in an antioxidant[54, 55] and this desired level probably varies from one substrate to another.

Light Screens

It was mentioned above that antioxidants are of little effect against relatively strong photooxidation because of the overwhelming rate of generation of chains. The most serious problems of deterioration in the Bell System are, of course, in outdoor applications, and it is quite clear that this is because of exposure to short wave light. The extensive commerical use of unprotected material outdoors came about because of a lack of appreciation of this fact. Once this vulnerability or organic materials to light is appreciated the remedy is obvious, at least in principle, and that is to shut off the light. For this purpose there are many pigments available as well as many light-absorbing organic compounds. A great deal of work has been done with various substrates in testing the effects of the absorbers, and this can be summarized as follows:

In the class of light colored pigments, none offers complete protection. Most of them have a slight effect; a few are fairly helpful; and a few are actually harmful, acting as photosensitizers. Of the darker pigments several are quite effective but the outstanding ones are lead chromates, iron oxides and carbon black, the last being the best. A study of the effect of various types of carbon black in various concentrations in polyethylene has been reported[19] wherein it is shown that under the most favorable conditions the useful life of polyethylene, as judged by accelerated tests, can be extended

at least 30 fold. It was shown in this work that for best results the carbon black should be finely divided and well dispersed. The use of polyethylene as a sheath material on outdoor cable would not have been practical without the protective effect of carbon black. The efficacy of carbon black as a light screen is apparently quite general although detailed studies have been made only with polyethylene, rubber,[56] cellulose esters,[57] and polyvinyl chloride,[58] in all of which it is effective.

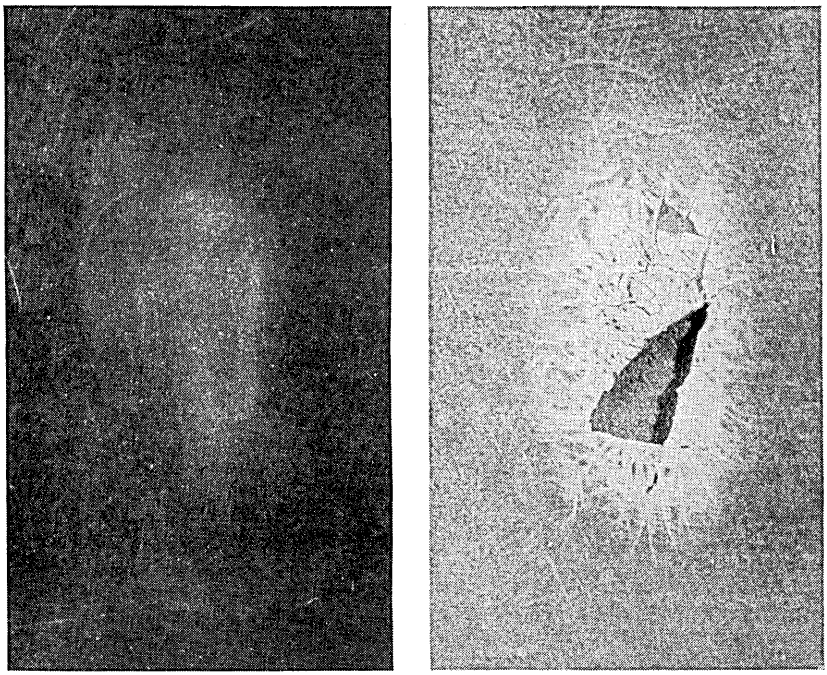


Fig. 14—Cellulose Acetate Butyrate panels exposed to concentrated beams of UV light filtered through pyrex bottles filled with water. Sample at left contains 1% carbon black. Sample at right contains 1% Salol.

In many applications of plastics and rubbers, colors are desirable and for these the use of carbon black is, of course, precluded. Some success has resulted from the use of organic materials which are transparent to visible light but which absorb in the ultraviolet. For example, phenyl salicylate, known as Salol, at a concentration of 1% is a fairly effective protective agent for transparent cellulose esters.[20] Such effects appear to be quite specific, since Salol is not nearly so effective in most other polymers (it is reported to be effective in Saran[59]), and many other compounds which are even better

absorbers of ultraviolet light are much less effective in cellulose esters. Some of them actually are sensitizers. Even in cellulose esters Salol at a concentration of 1% is a poor second to carbon black, giving in accelerated tests less than half the life imparted by 1% of a well dispersed, finely divided carbon black.[57] (Fig. 14)

REFERENCES

1. G. T. Kohman, *J. Phys. Chem. 33,* 226 (1929).
2. R. Burns, *A. S. T. M. Bulletin* #134, pg. 27 (May 1950).
3. T. Alfrey, "Mechanical Behavior of High Polymers," pp. 464–465, Interscience Publishers, (1948).
4. P. J. Flory, *Chem. Reviews, 35,* 51 (1944).
5. L. H. Campbell, A. H. Falk, R. Burns, *Proc. A. S. T. M. 46,* 1465 (1946).
6. R. B. Mesrobian and A. V. Tobolsky, *Jl. Polymer Science, 2,* 463 (1947).
7. C. C. Davis and J. T. Blake, "Chemistry & Technology of Rubber," pp. 538, Reinhold Publishing Co., N. Y. (1937).
8. Du Pont *Technical Service Bulletin* No. 8B, March 1950.
9. C. J. Malm and C. L. Crane, *U. S. Patent* #2,346,498.
10. B. S. Biggs, R. H. Erickson and C. S. Fuller, *Ind. and Eng. Chem., 39,* 1096 (1947).
11. J. Crabtree and A. R. Kemp, *Ind. and Eng. Chem. 38,* 278 (1946).
12. A. Rieche, R. Meister, H. Santhoff, H. Pfeiffer, *Liebig's Ann. Chem., 553,* 187 (1942).
13. R. G. Newton, *Jl. of Rubber Research, 14,* 27 (1945).
14. C. R. Noller, J. F. Carson, H. Martin, K. S. Hawkins, *Jl. Am. Chem. Soc. 58,* 24 (1936).
15. J. D. Dean et al, *Am. Dyestuff Reporter 36,* 705 (1947).
16. J. D. Dean and R. K. Worner, *Am. Dyestuff Reporter 36,* 405 (1947).
17. G. S. Egerton, *Am. Dyestuff Reporter 36,* 561 (1947).
18. Admiralty Engineering Lab., *Journal of Rubber Research 15,* 737 (1946).
19. V. T. Wallder, W. J. Clarke, J. B. DeCoste and J. B. Howard, *Ind. and Eng. Chem. 42,* 2320 (1950).
20. L. W. A. Meyer and W. M. Gearhart, *Ind. and Eng. Chem. 37,* 232 (1945).
21. V. W. Fox, J. G. Hendricks, H. F. Ratti, *Ind. and Eng. Chem. 41,* 1774 (1949).
22. G. N. Vacca, R. H. Erickson, and C. V. Lundberg, *Ind. and Eng. Chem. 43,* 443 (1951).
23. D. C. Thompson and N. L. Cotton, *Ind. and Eng. Chem. 42,* 892 (1950).
24. Symposium on Oxidation, *Trans. Faraday Soc. 42* (1946).
25. K. C. Bailey, Retardation of Chemical Reactions, Longmans, N. Y. (1937).
26. H. H. Zuidema, *Chem. Reviews 38,* 197 (1946).
27. L. Bateman, *Trans. of Inst. of Rubber Ind. 26,* 246 (1950).
28. A. V. Tobolsky, *India Rubber World 118,* 363 (1948).
29. J. L. Bolland and P. TenHave, *Trans Faraday Soc. 45,* 93 (1949).
30. H. L. J. Bäckström, *Zeit. für Physichalische Chem. B25,* 99 (1934).
31. L. Bateman and G. Gee, *Proc. Royal Soc. 195,* 376 (1949).
32. E. H. Farmer, G. F. Bloomfield, A. Sundralingham, and D. A. Sutton, *Trans. Faraday Soc. 38,* 348 (1942).
33. J. L. Boland, *Proc. Royal Soc. 186,* 218 (1946).
34. R. Houwink, *Kautschuk 17,* 67 (1941).
35. H. N. Stephens, *Jl. Am. Chem. Soc. 50,* 2523 (1928); *57,* 2380 (1935).
36. J. H. Raley, F. F. Rust and W. E. Vaughn, *Jl. Am. Chem. Soc. 70,* 1336 (1948).
37. N. A. Milas and D. M. Surgenor, *Jl. Am. Chem. Soc. 68,* 205 (1946).
38. H. S. Taylor and J. O. Smith, *Jl. Chem. Physics 8,* 543 (1940).
39. A. D. Walsh, *Trans. Faraday Soc. 42,* 269 (1946).
40. P. George and A. D. Walsh, *Trans. Faraday Soc. 42,* 272 (1946).
41. R. F. Boyer, *Jl. Phys. and Colloid Chem. 51,* 80 (1947).
42. P. I. Pavlovich, *Legkaya Prom.* (1945). 23 C.A. *40,* 7699 (1946).
43. W. O. Baker, *Jl. Am. Chem. Soc. 69,* 1125 (1947).
44. F. E. Francis, *Jl. Chem. Soc. 121,* 502 (1922).
45. C. O. Weber, "The Chemistry of India Rubber," pp. 220 and 299, Charles Griffin and Co., London, 1902.

46. A. Van Rosse and P. Dekker, *Ind. and Eng. Chem. 18*, 1152 (1926).
47. *Proc. of the Scientific Sec. Nat. Paint, Varnish and Lacquer Assoc.*, Circ. 546, pp. 307 (1938).
48. C. Moureu and C. Dufraisse, *Chem. Reviews 3*, 113 and ref. cited therein (1926).
49. J. A. Christianson, *Jl. Phys. Chem. 28*, 145 (1924).
50. J. L. Bolland and P. TenHave, *Trans. Faraday Soc. 43*, 201 (1947).
51. H. S. Taylor, *A. S. T. M. Proc. 32 Part II*, 9 (1932).
52. H. N. Alyea and H. L. J. Bäckström, *Jl. Am. Chem. Soc. 51*, 90 (1929).
53. A. M. Wagner and J. C. Brier, *Ind. and Eng. Chem. 23*, 46 (1931).
54. L. F. Fieser, *Jl. Am. Chem. Soc. 52*, 5204 (1930).
55. C. D. Lowry, C. G. Dryer, G. Egloff, and J. C. Morrell, *Ind. and Eng. Chem. 24* 1375 (1932).
56. W. N. Lister, *Trans. Inst. of Rubber Ind. 8*, 241 (1932).
57. R. H. Erickson, unpublished work.
58. V. T. Wallder and J. B. DeCoste, unpublished work.
59. R. F. Boyer, *U. S. Patent* 2,429,155.

The Development of Electron Tubes for a New Coaxial Transmission System

By G. T. FORD and E. J. WALSH

(*Manuscript Received July 27, 1951*)

1. Introduction

AS THE demand for long distance telephone circuits has increased, new transmission systems capable of handling more channels per conductor have been developed. Also the advent of television has created a demand for broad band channels for network facilities. One of the latest developments now nearing completion is the L3 Coaxial System.

Three new tubes have been developed specifically to meet the exacting requirements of this system: two tetrodes, the W.E. 435A and W.E. 436A, and a triode, the W.E. 437A. All three types are used in the line and office amplifiers. The new tubes make possible a substantially higher level of broad band amplifier performance compared to their predecessors. They represent the result of improvements made by applying well known basic principles through new tube-making techniques. These techniques have been developed largely within the framework of existing conventional telephone tube manufacturing methods.

The development of special, small, low power vacuum tubes for high frequency application in the Bell System began in 1934. The tube program was instituted originally as part of a research project in the field of radio communications. When the development of the L1 Coaxial System began it was recognized that similar tubes would be needed. Part of the tube development effort was therefore directed toward the coaxial requirements. Work on the W.E. 384A and W.E. 386A tubes used in the L1 system was completed in 1939 as an outgrowth of this program.

The demand for amplification over wider frequency bands resulted in further development work along the same lines. During World War II this effort was applied to the development of the 6AK5 tube which became available early in 1943 and was used widely in IF amplifiers in radar equipment. Shortly after the war the W.E. 408A tube was developed for telephone repeater uses. This is a long life version of the 6AK5 tube having the same electrical characteristics except for the heater voltage and current. The W.E. 404A tube appeared in telephone circuits in 1949. This tube, having a higher figure of merit than the W.E. 408A, provided improved performance in the IF amplifiers used in the New York to Boston radio

relay system and in the TD2 radio relay system. The W.E. 435A, W.E. 436A and W.E. 437A tubes are the latest types to come out of this long range program.

It will be seen in what follows that the key to continued development along these lines has been improvements in the techniques of grid making to meet the basic objective of providing a grid which can be spaced very close to the cathode and which, in effect, acts as a uniform potential plane controlling the current drawn from the cathode without offering any physical obstruction. This objective is approached by using many turns of very small diameter wire for the grid winding. The reason for the close grid-cathode spacing is that the transconductance or sensitivity depends on this factor. Although the increase in input capacitance which results is a disadvantage because of its effects on the interstage circuits, this disadvantage is more than compensated for by the higher transconductance obtained.

2. Principles of Design

2.1 *Requirements*

The overall requirements for the L3 system, and the manner in which they are related to the tube parameters, are very complex. However, in its simplest terms, the objective for the L3 system is to provide on one coaxial pipe a facility suitable for the simultaneous transmission over a 4000-mile circuit of a television signal and 600 one-way telephone channels or, alternatively, 1800 one-way telephone channels when no television channel is required. The transmission band being provided is from approximately 0.3 MC to approximately 8 MC. The amplifier needed to compensate for the cable attenuation must meet very exacting requirements with respect to gain-frequency characteristics, stability, noise, and linearity.

The design features necessary to provide suitable electron tubes for use in the L3 amplifiers are closely related to the requirements mentioned above for the amplifiers. In general terms, the tube design objectives are: (1) high transconductance-capacitance ratio (figure of merit), (2) minimum excess phase shift or phase delay, (3) low noise, (4) well controlled modulation, (5) long life, (6) interchangeability, and (7) lowest cost consistent with the first six objectives. In the material which follows, each of these objectives will be discussed in detail and its relationship to the system objectives brought out.

2.2 *Figure of Merit*

Figure of merit is of particular importance. It is a direct measure of the bandwidth over which the required amplification can be obtained. In gen-

eral, a given factor of improvement in the figure of merit can be translated directly into a wider transmission band providing more communication channels.

For a two-terminal type of interstage such as that used in the L3 amplifier, the figure of merit is

$$F = GB = \frac{KG_m}{2\pi(C_1 + C_2)} \tag{1}$$

where F is the figure of merit, G is the voltage amplification, B is the bandwidth between the frequencies where the gain is 3 db below that at the center frequency, K is a constant whose value depends on the particular interstage design, G_m is the transconductance of the tube, C_1 is the input

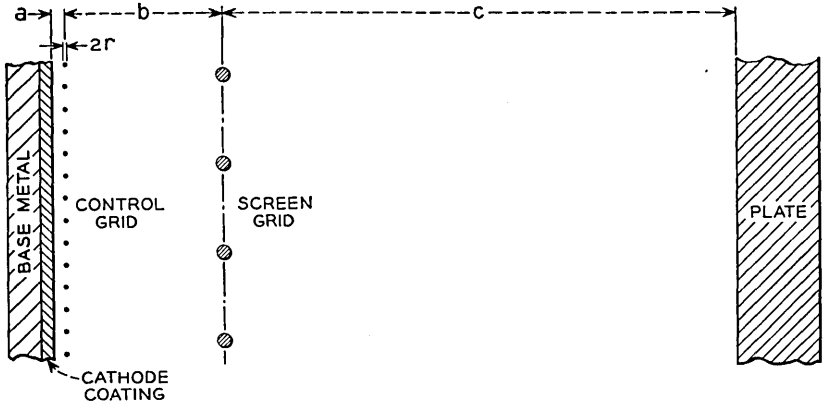


Fig. 1—Geometry of the 436A tube.

capacitance, and C_2 is the output capacitance. This figure of merit is directly applicable to a tetrode operated as a small-signal voltage amplifier and is a well known relationship.[1] It will be used to show how the tube design factors influence the figure of merit of the W.E. 435A and W.E. 436A tubes.

Using equation (1) and applying certain simplifying assumptions which can be made without materially affecting the results, expressions are derived in the appendix showing the relationship between the figure of merit and the tube parameters. Equations (2) and (4) in the appendix show how the figure of merit is affected by the grid-cathode spacing "a", the grid-screen spacing "b", the screen-plate spacing "c", and the grid wire radius "r". See Fig. 1. Equation 3 gives the required screen voltage for the as-

[1] "Characteristics of Vacuum Tubes for Radar Intermediate Frequency Amplifiers," G. T. Ford, *B.S.T.J.*, Vol. XXV, p. 389, July, 1946.

sumed current density and geometry. Since these expressions are rather involved, the manner in which the various factors influence the figure of merit can be brought out best by a series of curves. Figures 2, 3, and 4 show how F is affected by changes in "a", "b", and "c". Figures 6 and 7 show how the screen voltage required to get the assumed current density with a given bias E_{c1} varies with "a" and "b" (equation 3). The screen voltage is essentially independent of "c".

These relationships are also applicable to the W.E. 437A tube with minor modifications.

2.21 *Design Considerations*

How the various factors in equations (2), (3), and (4) affect the figure of merit will be discussed in detail. They are listed in Table I.

The factor M is the ratio of the plate current to the cathode current. The figure of merit is directly proportional to this factor. M can be increased

TABLE I

Factor	Design Values	Practical Design Considerations
M	0.75	Mechanical, plate-grid capacitance
I_0	50 MA/cm^2	Stability of emission, life
a	0.00635 cm	Mechanical
b	0.0444 cm	Screen voltage, mechanical
c	0.150 cm	Formation of potential min.
r	0.00038 cm	Mechanical
E_{c1}	-1.5 volts	Grid current
E_{c2}	150 volts	Dissipation, life

by using smaller wire in the screen grid, the minimum practical wire size being determined by the mechanical rigidity and heat dissipation capability required. M can also be increased by reducing the number of turns on the screen, but this is limited by the necessity for sufficient shielding effect to meet the requirement that the plate-grid capacitance be less than a specified value.

Since the figure of merit is directly proportional to the cube root of the cathode current density I_0, the improvement with increasing I_0 is not very rapid. The problems of obtaining uniform initial performance and long life are aggravated by increasing the current density, for several reasons. There is no direct evidence to show that high current density per se causes accelerated loss of available emission. In fact there is some evidence to the contrary.[2] However, there is ample evidence that phenomena

[2] "Influence of Density of Emission on the Life of Oxide Cathodes," S. Wagener, *Nature*, p. 357, Aug. 27, 1949.

usually associated with high current density tend to shorten the life. Higher electrode temperatures, higher potentials, and the production of more ions are the major items in this category. It is presumed that the shorter life found under these conditions is due to the greater rate of contamination of the cathode by material from the other parts of the tube. Great efforts have been made to find and to use processing techniques which will minimize this kind of limitation and to introduce constituents into the cathode which will counteract such deterioration. The situation at the time the L3

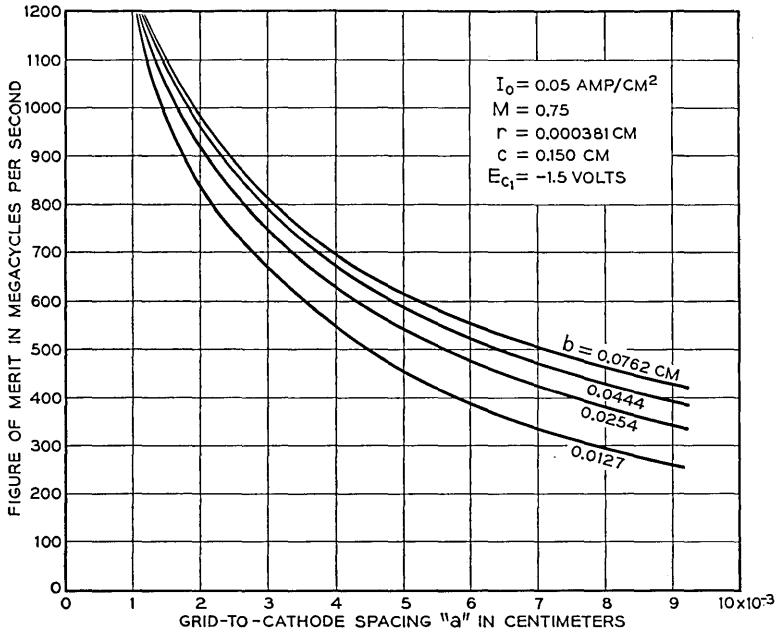


Fig. 2—Figure-of-merit vs. grid-to-cathode spacing.

tubes were being developed was that 50 MA/cm² was as high a current density as seemed to be consistent with the long life required.

It is apparent from the curves in Fig. 2 that the figure of merit increases rapidly as the grid-cathode spacing "*a*" is reduced. The limitation here is mechanical and manifests itself in two ways. One is the practical difficulty of spacing the parts so closely with sufficient accuracy. The other is the problems associated with fabricating grids wound with wire of small enough diameter to make effective use of the close grid-cathode spacing. This part of the subject will be discussed in detail later. It is one of the most important aspects of the design of the L3 tubes.

It would appear from Fig. 3 that the grid-screen spacing "*b*" should be as large as possible. However, the required screen voltage increases as "*b*" increases, and it is desirable to keep the screen voltage low. Therefore, "*b*" is made as low as possible without causing too much penalty on figure of merit. A good compromise value for "*b*" depends on the range of grid-cathode spacing "*a*" being considered, but it will usually be from 0.005 cm to 0.020 cm for close spaced tubes.

Figure 4 shows that the figure of merit increases as "*c*" increases, but there is very little advantage in making it more than 0.040–0.050 cm. Making it much larger also increases the outside dimensions of the struc-

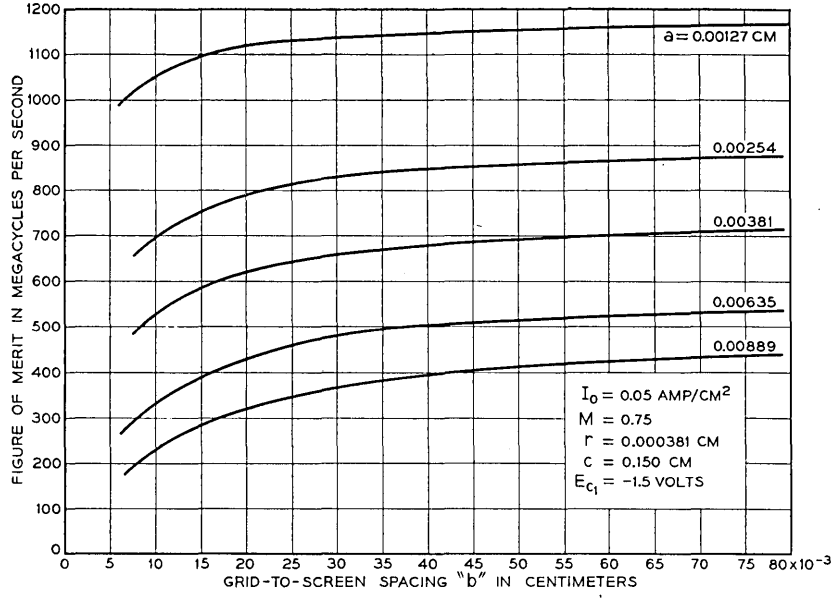


Fig. 3—Figure-of-merit vs. grid-to-screen spacing.

ture unnecessarily, and eventually leads to a spacing which will cause irregularities in the plate current-plate voltage characteristic due to space charge effects in the screen-plate space.

Although it is not apparent from the curves or from what has been said above, it is desirable to have "*r*" as small as possible. It is obvious that "*r*" must be less than $\frac{1}{2n}$, otherwise the grid is completely closed. Under the assumption that $na = 1$, this means that "*r*" must be less than $0.5a$ if there is to be open space between the grid wires. Actually, it is desirable to have not more than 30% of the projected area of the grid closed, which

means that "r" should be less than $0.15a$. In addition to this consideration, it is desirable to have "r" considerably less than $0.15a$ so that the required screen voltage will be as low as possible. This comes about because the amplification factor μ increases as "r" is increased, other quantities held constant, and equation (3) shows that E_{c2} increases as μ increases. The diagram shown in Fig. 5 illustrates the trend in grid-cathode spacings and grid wire sizes. The W.E. 416A tube (formerly BTL 1553) represents the

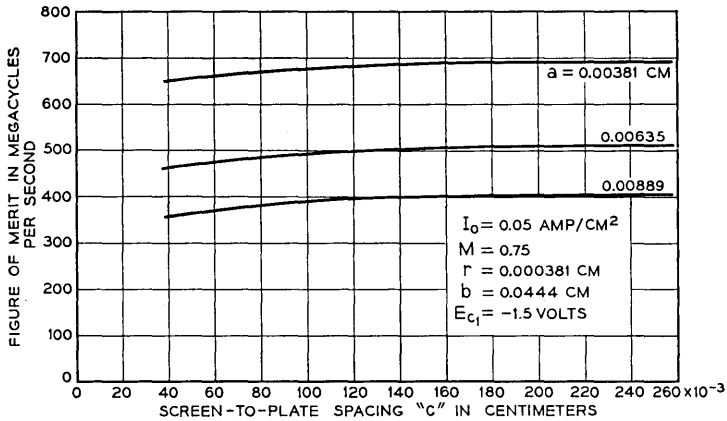


Fig. 4—Figure-of-merit vs. screen-to-plate spacing.

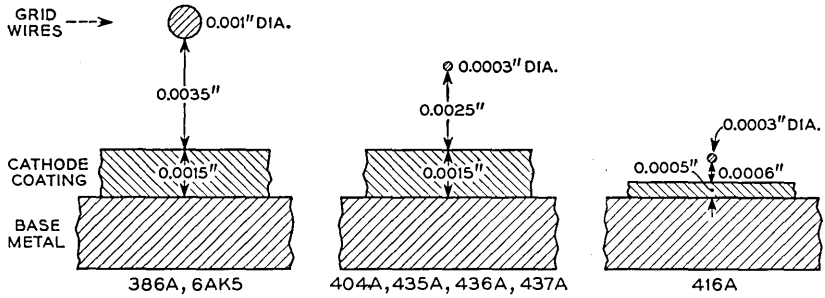


Fig. 5—Trend in spacing and grid wire size.

greatest extension of this trend reported as far as grid-cathode spacing is concerned.[3]

The figure of merit increases as the absolute value of the bias E_{c1} is reduced, since I_0 increases. However, a minimum bias value of about -1.5 volts is usually necessary in order to avoid undesirable effects due to the

[3] "Design Factors of the Bell Telephone Laboratories 1553 Triode," J. A. Morton and R. M. Ryder, *B.S.T.J.*, Oct. 1950.

collection of electrons of high initial velocity by the control grid. Such grid currents contribute to the noise, cause input loading, and may also cause excessive signal distortion.

Since I_0 increases as the screen voltage E_{c2} is increased, the figure of merit likewise increases. It is desirable to keep E_{c2} as low as possible for at least three reasons. The most important is that high screen voltage will generally have an adverse effect on tube life. The second is that low power consumption is desirable for economic reasons and the third is that it helps from the standpoint of maintaining low temperatures of the components in an

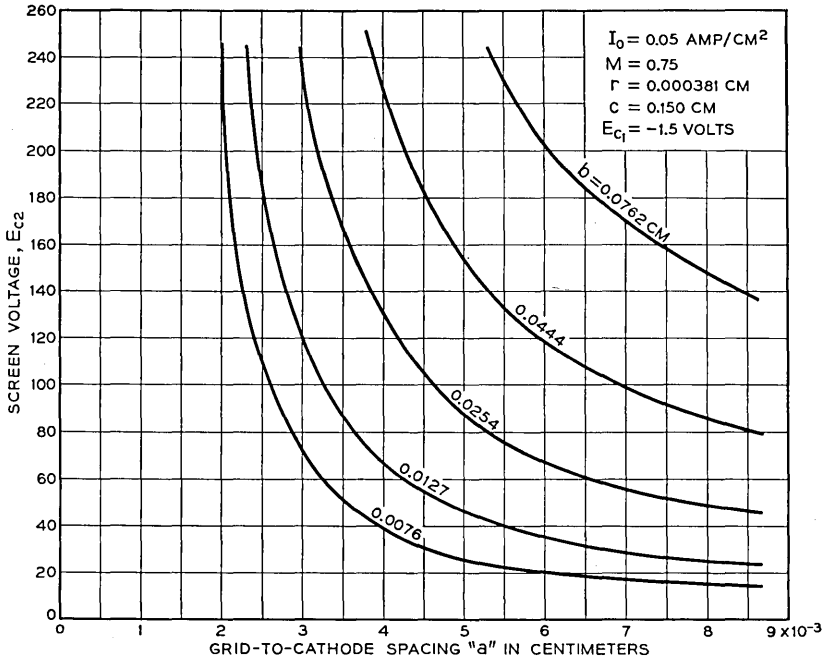


Fig. 6—Screen voltage vs. grid-to-cathode spacing.

amplifier. Figures 6 and 7 show how E_{c2} depends on "a" and "b". The requirement that E_{c2} be kept low means that the range of "a" and "b" which can be used is restricted.

2.3 *Phase Shift*

The effects of electron transit time and lead inductance in the tubes must be taken into account in order to meet the amplifier requirements with respect to phase margin. In order to maintain stable operation over

the desired transmission band, the gain and phase characteristics must be controlled up to about 200 MC. The amount of phase shift at the frequency where the gain becomes unity ("cross-over point") is of particular interest. In the L3 amplifier this frequency is about 40 MC. Phase shift introduced by electron transit time and by lead inductance is referred to as "excess phase." Ideally, of course, the excess phase would be zero.

The time required for an electron to travel from the cathode to the plate is of the order of 10^{-10} seconds. This corresponds to about 5° of excess phase at 40 MC. Close spacings and high electrode potentials tend to reduce the

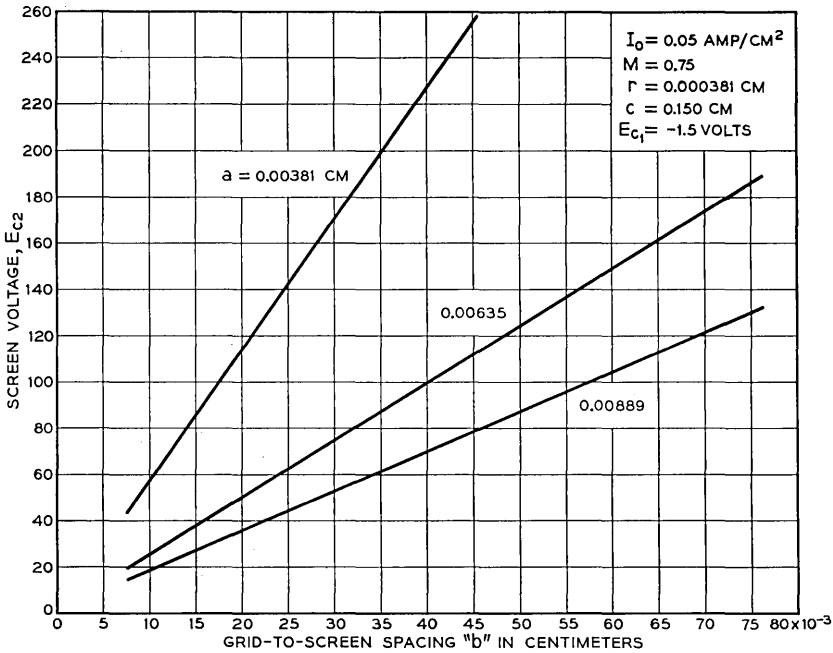


Fig. 7—Screen voltage vs. grid-to-screen spacing.

transit time. However, the considerations discussed in Section 2.2 have been the major factors in setting the spacings and potentials because the transit time, though important, is far less so than the figure of merit.

By using relatively heavy lead wires and mounting the tube structure in such a way as to make the lead wires as short as possible, the additional excess phase due to the lead wires has been minimized so that it amounts to about 5° at 40 MC.

In order to insure adequate margin against a singing condition, the amplifier has been designed to have about 20°–30° less phase shift at 40 MC

with these tubes than that which will cause singing. With this situation, it can be seen that any substantial factor of increase in the excess phase introduced by the tubes, or any other components, could begin to reduce the phase margin seriously.

2.4 *Noise*

Fluctuation noise is an important factor in the W.E. 435A used in the first stage of the input amplifier and in the W.E. 436A used in the first stage of the output amplifier. There is adequate margin against the effect of low frequency noise components such as microphonics, power frequency hum, and "sputter noise" if reasonable precautions in tube and circuit design are taken. From a design standpoint, the fluctuation noise is min'mized by adopting a combination of cathode temperature and current density drawn such that, with a normally active cathode, the space current is substantially space charge limited, with ample margin for some loss of cathode activity in service before the temperature limited condition is approached. When the temperature limited region is reached, the noise is substantially higher than for the space charge limited condition. The temperature and the cathode current density ratings for these tubes have been set at values which take these considerations into account.

2.5 *Modulation*

Since a major purpose of using feedback is to reduce the modulation products arising in the amplifiers, the more nearly an ideal linear transfer characteristic can be approached in the tube design the better, because less feedback is required to obtain a given grade of system performance. Unfortunately, however, a conventional triode or tetrode type of vacuum tube operating under normal space charge limited conditions necessarily has a transfer characteristic which is non-linear. Several possible special structures which might give less modulation were explored, but none were found which would provide the required figure of merit and be sufficiently stable and reproducible.

Considerable emphasis was placed on the problem of controlling the variation in modulation from tube to tube. The most important factors are grid-cathode spacing, uniformity of grid pitch, and cathode activity. Although these factors must be well controlled for other reasons also, the special requirements on modulation necessitated a thorough investigation.

The effect of the grid-cathode spacing can be expressed in terms of the d-c. plate current and the signal level. For a triode having an idealized three-halves-power transfer characteristic with $\frac{d\mu}{dE_{c1}} = 0$ as in Section 2.2,

and for small signals, the ratio of the fundamental signal current to the second harmonic component for the case of a very small load impedance is

$$\frac{I_p}{I_{2p}} = 12\,\frac{I_b}{I_p} \quad \text{(see appendix for derivation)} \tag{11}$$

This means that, for a given signal current amplitude I_p in the output, a tube having the assumed characteristics will give a ratio which depends only on the d-c. plate current, which in turn is very sensitive to changes in grid-cathode spacing.

A study of the variations in grid pitch and their effects on modulation in a particular experiment showed that reducing the standard deviation of the pitch distance from 16% to about 7% reduced the second-order modulation by 4 db. Since the second-order modulation must be reduced by feedback which is at a premium in the L3 system, this experiment showed that control of the grid pitch was important, and that periodic checks on this factor would be desirable in manufacture.

The effect of cathode activity on modulation was studied in diodes so as to eliminate the effects of grid variations. The variations in modulation from tube to tube were found to be about the same as when grids were present. The geometry of the diodes was so closely controlled that dimensional variations could not account for the differences in the modulation levels. This part of the investigation led to a recognition of the importance of obtaining the best possible uniformity of cathode activity. It also became apparent that the surface condition of the anode was a factor, and that it is therefore desirable to maintain a high degree of cleanliness of the electrodes to which positive potentials are applied.

2.6 *Life*

Long tube life is a very important requirement in the L3 system. The most important consideration is the effect of the life on the reliability of the system. There is also the obvious effect of the life on maintenance costs.

Short life tends to reduce the reliability of a system which contains a great number of tubes because the potential failures cannot be predicted so accurately as when the life is long, without a prohibitively costly amount of testing. Even with the most frequent and accurate testing procedure which might be considered, it would be amazing if more than 90% of the potential failures were replaced before causing transmission trouble. To illustrate the effect of short life, consider a 100-mile section of L3 line. There will be five tubes in each of 24 amplifiers. If the performance of any one of the total of 120 tubes becomes poor enough to make the circuit uncommercial, that section must be taken out of service until the defective

tube (or amplifier) is replaced. Now, for a going system, with 120 tubes, and assuming an abnormally short life of say 1200 hours, a tube will fail every ten hours on the average unless preventative testing is used. Even if very frequent testing be done in order to replace 90% of the potential failures before they occur, one circuit interruption every 100 hours may be expected.

It is evident that a life many times greater than that assumed in this illustration is imperative if reliable service is to be obtained and costly maintenance avoided. Laboratory life tests predict that a tube life of at least 15,000 hours may be expected in the L3 system. The actual results will depend on the extent to which the operating conditions are closely controlled, the severity of the field rejection limits, and the ability of the tube factory to control the processing.

2.7 *Interchangeability*

The objective is to make the characteristics of the tubes sufficiently uniform so that tubes may be replaced at will without circuit adjustments being needed. In the L3 amplifiers, the circuits have been designed so that a relatively wide range of characteristics can be accepted for individual tubes. However, it is essential that the average characteristics be held in close control from one manufacturing lot to another. This has been provided for by setting up distribution requirements which will be discussed further in a later section.

2.8 *Cost*

As will be seen from the description which follows, it has been possible to meet the L3 requirements with tube designs which do not require too great departures from the manufacturing methods employed for conventional telephone tubes. With a reasonable demand, it is accordingly expected that the tube costs should be moderate.

3. Design Description and Characteristics

3.1 *Mechanical Description and Mechanical Problems*

Figure 8 shows the three L3 tubes along with some of the earlier high figure of merit tubes. The W.E. 386A (left hand side) was designed to be soldered directly into the circuits and had its input lead at the stem end while the output lead came out through the top of the bulb. The flexible leads used for soldering purposes, and the double-ended lead construction,

add materially to the cost of tube construction and testing. Early in the L3 tube development the question of factory cost compared with circuit performance was weighed and it was decided that the advantage of lower tube cost plus the very large advantage of simple plug-in tubes would outweigh

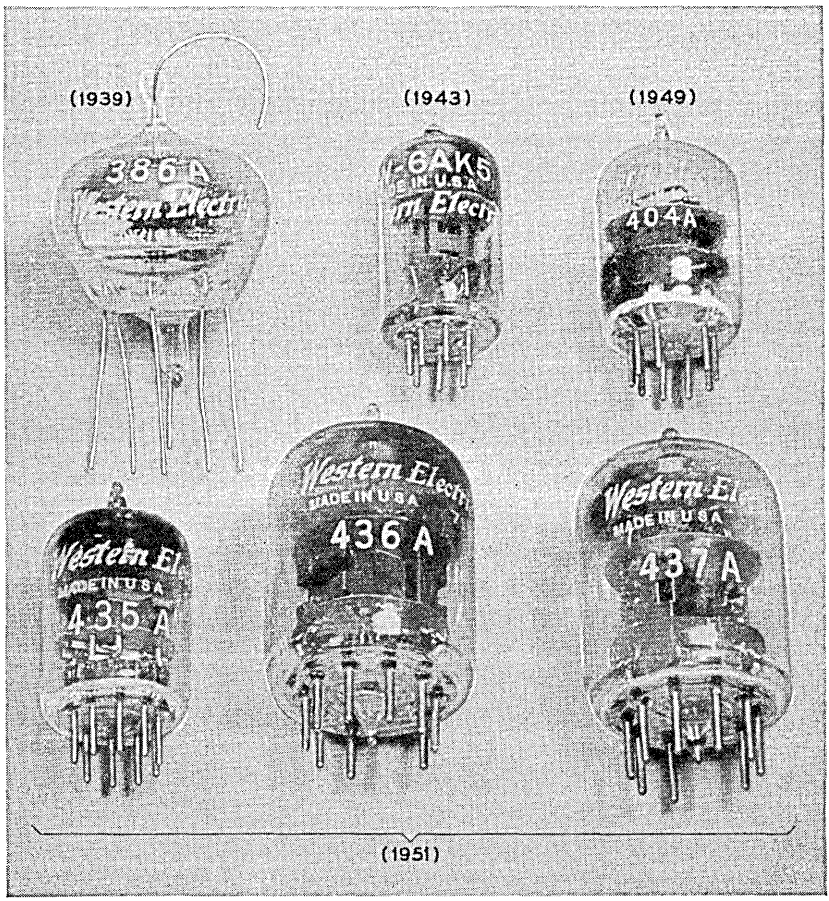


Fig. 8—The 386A, 408A, 404A, 435A, 436A and the 437A tubes approximately actual size.

the cost in performance. Accordingly, all three L3 tubes are of the stiff pin, plug-in type designed to fit existing sockets. The price paid for obtaining the advantages of lowered cost and interchangeability has been a loss in feedback of approximately 2 db per amplifier.

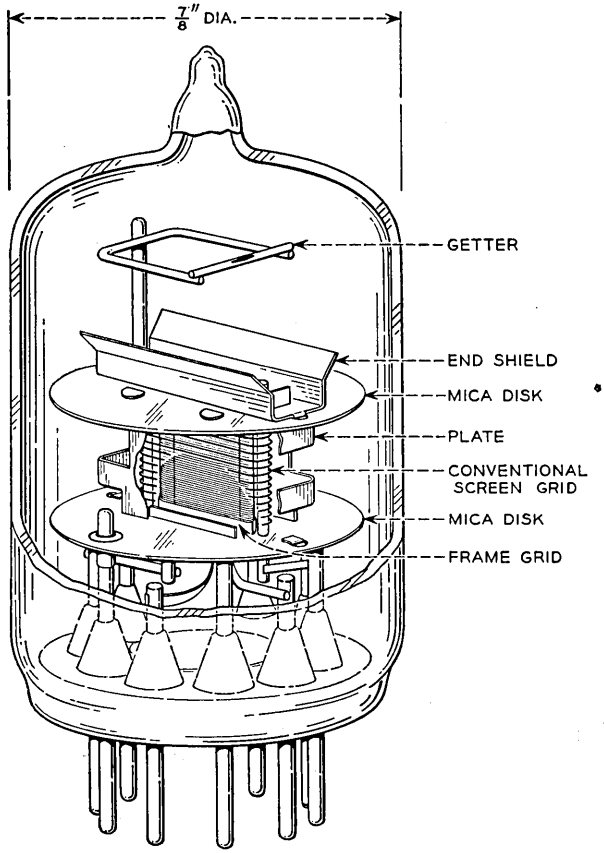


Fig. 9—Cutaway view of the 435A.

Figures 9, 10, and 11 are the cutaway views of the 435A, 436A, and 437A tubes. The overall dimensions of the L3 tubes are:

	435A	436A	437A
Max. seated height	$1\frac{1}{2}''$	$1\frac{5}{8}''$	$1\frac{5}{8}''$
Max. diameter	$\frac{7}{8}''$	$1\frac{3}{16}''$	$1\frac{3}{16}''$
Number of pins	9	9	9
Pin circle diameter	$\frac{15}{32}''$	$\frac{11}{16}''$	$\frac{11}{16}''$

Conventional construction for small repeater tubes may be thought of as two mica wafers between which are assembled the active elements of the tube. The mica wafers serve to support and space the tube elements. This

mica and element assembly is then mounted upon a glass stem or platform which is next sealed into a glass bulb after which the exhaust and activation procedures complete the tube. These cutaway views show that these L3 system tubes are very similar to conventional tubes. The reason for wanting to continue with the conventional type structures is simply that of tube

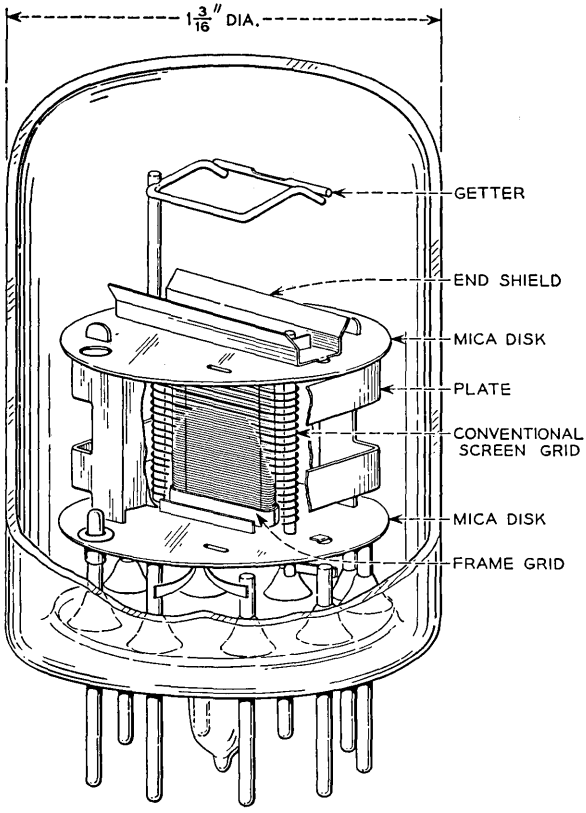


Fig. 10—Cutaway view of the 436A.

cost. A production line such as that for the W.E. 408A or 6AK5 tubes could very readily be changed over to any one of these tubes. The only change needed would be in the control grid supply and in the dimensional control procedures.

The principal distinctive design feature in these tubes, compared to earlier repeater tubes, is the "frame" type of control grid which was first introduced in a somewhat different form in the W.E. 404A[4] and W.E.

[4] "The 404A- A Broadband Amplifier Tube," G. T. Ford, *Bell Laboratories Record*, Vol. XXVII Feb. 1949.

418A tube. The L3 frame grids are illustrated in Fig. 12, together with a conventional control grid from the 6AK5 tube and the control grid from the 404A vacuum tube. The conventional grid consists of two large side rods, usually of nickel, around which is spirally wound the grid lateral wire. The lateral wire is joined to the side rod at each intersecting point by first knifing a groove into the side rod, laying the lateral wire into the groove,

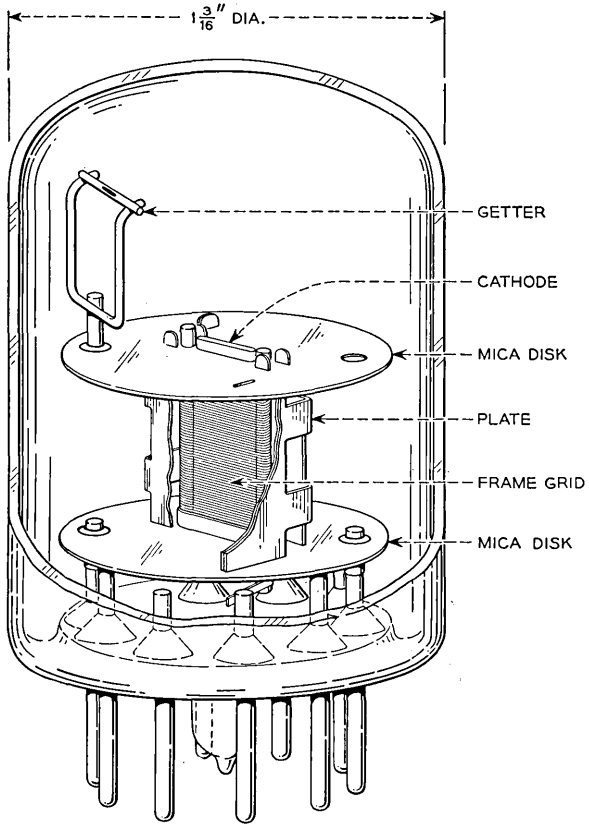


Fig. 11—Cutaway view of the 437A.

and then swaging the groove closed. Since, in these conventional grids, the lateral wire is usually larger than 0.0008″ diameter, the grid is self supporting and needs no strengthening members. For the high figure of merit tubes, control grid lateral wires of the order of 0.0003″ diameter are needed. Wire of this diameter is not self supporting in the necessary lengths and for that reason the two large side rods are first joined together by the cross straps

which are located at the ends of the grid proper. These then form a rigid frame around which the very fine lateral wire can be spiraled without any danger of having laterals out of place. It can be seen that this technique produces the extremely flat grid plane which is necessary for the desired

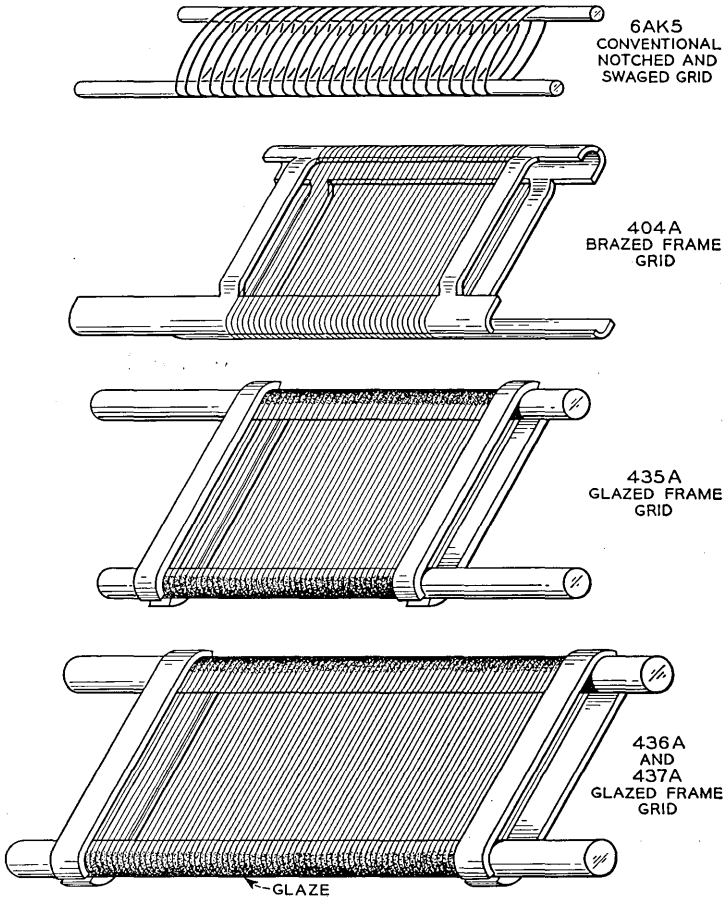


Fig. 12—Control grids for the 408A or 6AK5, the 404A and the L3 carrier tubes.

tube performance, and which is the real difference between these high figure of merit tubes and the more conventional tubes.

The fabrication of the 404A frame grid has been discussed in a previous article.[5] That article also mentioned the 418A grid which is a side rod type

[5] "Fine-Wire Type Vacuum Tube Grid," E. J. Walsh, *Bell Laboratories Record*, Vol. XXVIII April 1950.

frame grid. The L3 grids are a further development. The major difference between the earlier frame grids and these is in the method of bonding the 0.0003″ lateral wire to the side rods. In the earlier grids a gold braze was used to bond the laterals to the side rods. This necessitated heating the unit to approximately 1070°C to flow the gold. The newer grids have the lateral wires bonded by a glass glaze which allows the process to be carried out at approximately 700°C. There is a differential expansion between molybdenum and tungsten of about five to four. The net result of the reduction of temperature in this process is that the tungsten wires are stretched less at the lower temperature and thus when returned to room temperature have higher residual tensions. This is important because the higher the residual tension the higher the resonant frequency of the lateral wires. This in turn means that the noise level of the tube due to vibration or shock will be reduced since loose grid wires will give rise to microphonic noises. Tighter wires also decrease the possibility of grid to cathode shorts. Tests have shown that an increase of about 25% in the resonant frequencies of the lateral wires can be expected as a result of using the glass glazing technique as compared to the gold brazing technique.

It is interesting to note that the residual stress in the lateral wires of these grids is of the order of 200,000 pounds per square inch. This figure is roughly ten times as great as the allowable working stress for steel beams such as are used in the construction industry.

When the glazing technique is used, the grid is gold plated after the glazing operation has been completed. The gold is used to inhibit thermionic emission from the grid wires. This is a necessity for tubes of this type when used in the circuits for which they were designed. The need for the plating exists because of the proximity of the grid wires to the hot cathode and their unfortunately favorable position for receiving a deposition of active material from the cathode during its processing and operation. The desired amount of gold on the grid wires is that which will cause a diameter increase of about 0.00002″. This is an extremely difficult increase to measure because the measurement must be non-destructive, since it is made on the finished grid and is used as a production control. The method used to date has been an optical measurement at a magnification of about 500×.

A very high degree of precision, compared to that previously available, has been obtained for some of the parts whose dimensions are critical. The cathode sleeve is now obtainable with minor axis limits of ±0.0003″. The mica discs are now made with the critical holes to that same tolerance. The frame grid side rod is made to ±0.0001″. These are the basic elements of the tube and, after inspection has shown them to be acceptable, their assembly becomes close to that of a conventional tube. The inspection of

these parts is difficult when production numbers are considered. The micas in particular presented a serious problem. Mica sheet is composed of a large number of laminations many of which are of the order of 0.0001″ in thickness. When the mica discs are punched out, these laminations leave, not smooth edge holes as do metal stampings, but rather a large number of minute jagged edges. The method used to check these was an optical one in which the mica was projected at about 40 times size onto a glass screen on which engraved lines acted as go-no-go gages. This reduced tool and human error considerably. The cooperation of several industrial concerns which supply some of the critical parts and the measuring instruments was very helpful in obtaining the desired tolerances.

It was evident from the start of the development of the L3 tubes that the performance requirements for high gain conventional structure tubes would be pushing to the limit the available process controls and measuring techniques. A statistical quality control program was put into effect on the tubes after the final laboratory design had been crystallized. The statistical study covered the tube dimensions and the data collected from those tubes after they had been processed. The net result of the study was to indicate that better measuring methods and process controls are needed.

With the amount of d-c. feedback employed in the working circuits, the space current does not vary too rapidly with tube geometry. In the case of the most critical spacing, that between the grid and the cathode, a 10% change in the spacing would be expected to cause only about 2.5% change in space current. However, the transconductance is more sensitive to the grid-cathode spacing, with a 14% change in transconductance to be expected for a 10% change in the spacing. This comes about because the transconductance is a function of the spacing, even at a fixed space current.

Since a 10% change in spacing is only 0.00025 inch, the importance of close tolerances on the parts dimensions controlling it is evident. The test specification limits on transconductance permit a variation of about ±25%, so that the 0.00025 inch change in spacing would use up over half of the allowed deviation. Preproduction runs at the Laboratories have shown that the tubes are practical and that their performance in the amplifier circuits has justified their design.

3.2 *Electrical Characteristics*

The nominal electrical characteristics are shown in Table II. The corresponding characteristics for the earlier types 386A, 6AK5 and 404A are also given for comparison. The last row in the table shows figure of merit values which are a measure of the circuit performance. The tabulated values for the figure of merit were calculated, taking into account the effect of

space charge on the input capacitance, and include a total allowance of 3 mmf. for socket and wiring capacitance (input plus output). The L3 tubes are somewhat better than the 404A and substantially better than the 6AK5. The figures of merit tabulated will not check with the values shown in Figs. 2–4 since the curves were calculated for cold tube capacitances and zero socket and circuit capacitances.

TABLE II

Classification	386A Pentode	6AK5 408A Pentodes	404A Pentode	435A Tetrode	436A Tetrode	437A Triode
Heater voltage	6.3	6.3 20.0	6.3	6.3	6.3	6.3 volts
Heater current	0.150	0.175 .05	0.30	0.30	0.45	0.45 amps
Plate current	7.5	7.5	13	13	25	40 ma
Screen current	2.5	2.5	4.5	3.5	8	— ma
Transconductance	4000	5000	12500	15000	28000	45000 μmhos
Input capacitance*	3.6	3.9	7.0	7.8	15.2	11.5 mmf
Output capacitance*	2.6	2.85	2.5	2.5	3.3	0.9 mmf
Plate-grid capacitance*	0.025	0.01	0.03	0.025	0.05	3.5 mmf
Figure of merit**	61	72	123	146	165	— mc

* Cold capacitances.

** This is the frequency at which unity voltage amplification would occur with a simple parallel tuned circuit interstage. Allowances have been made for stray capacitances and for the increase in input capacitance when a tube is energized. No figure is given for the 437A tube because the relations derived for the earlier stages in the amplifier do not apply to the output stage.

3.3 *Performance in Repeater*

The positions of the tubes in an auxiliary repeater are indicated in the diagram of Fig. 13. The overall insertion gain is a little over 30 db at 4 mc. While the noise contributions of the first 435A tube and the 436A tube are important, they have been reduced to 48 db below one volt and 62 db below one volt, respectively, for 1200 repeaters. The second 435A tube and the "lower" 437A tube appearing in Fig. 13 are the major contributors to the modulation. The expected modulation levels from these tubes are those associated with single tone ratios of about 34 db for the fundamental to second harmonic ratio and 70 db for the fundamental to third harmonic ratio, with a grid signal level of 0.1 volt r.m.s.

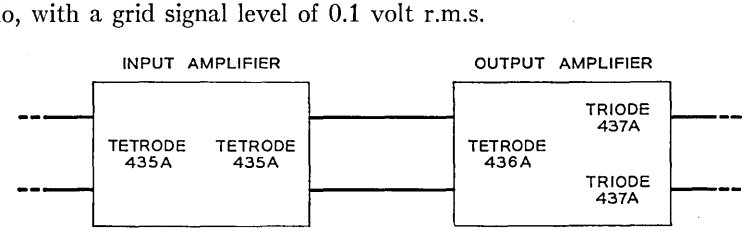


Fig. 13—Position of tubes in the input and output amplifiers for the L3 carrier system.

The gain-band performance in this repeater, or in any other circuit, will be less than that shown in the curves in Figs. 2–4 since the total shunt capacitances in a working circuit are always larger than the cold tube capacitances used in calculating the inherent figure of merit.

3.4 *Test Specifications*

The test specifications for the L3 tubes were written with the L3 system requirements as the prime consideration. In addition to the usual tests made on small tubes, a modulation test was included for the 435A and 437A tubes because of the importance of this characteristic in terms of system performance. In order to avoid penalties which can result from unwanted systematic deviations which pile up in a long system, requirements have been set up which will control the distribution of transconductance, modulation, and some of the most critical interelectrode capacitances. By the application of suitable quality control methods, it is expected that these requirements can be met economically and that such measures will prevent the manufacture of large numbers of tubes having average characteristics far off the design values. When simple go, no-go limits are used, it is not economical to set close enough limits to attain the desired control of the average characteristics.

4. Conclusions and Future Possibilities

The fundamental problem in the development of repeater tubes for broad band coaxial systems has been to devise means for utilizing closer and closer grid-cathode spacings without sacrificing life performance.

The closer spacings have been made possible by devising rigid control grid supporting structures which can be wound with very small diameter wire. The wire is held under tension by the supporting frame so that a flat winding is produced which can be spaced very close to a flat cathode.

The possibilities for the development of tubes which will provide still better broad band amplification depend to a great extent upon the kind of system design to be considered. If higher figure of merit, as defined in this article, can be utilized, considerable improvement can be realized with space charge controlled tubes such as the 435A, 436A and 437A by using mounting arrangements which provide more precise means of establishing and maintaining the critical dimensions.

Acknowledgements

Several members of the technical staff and their assistants have contributed materially in solving the numerous technical problems which arose during the development of these tubes. In addition, those who fabricated

the grids and assembled the experimental tube models made important contributions in terms of skill and painstaking effort. It is not practical to name all of the persons involved.

The development of broad band tubes over the last fifteen years was under the direction of the late Dr. H. A. Pidgeon until 1943, and Mr. J. O. McNally from 1943 to date. The writers wish to acknowledge the importance of their helpful guidance and encouragement.

APPENDIX

Meanings of Symbols

F = Figure of merit
G = Voltage amplification
B = Bandwidth between 3 db points
K = Interstage circuit constant
G_m = Grid-plate transconductance
C_1 = Input capacitance
C_2 = Output capacitance
n = Turns per unit distance on grid
a = Grid-cathode spacing
b = Grid-screen spacing
c = Screen-plate spacing
A = Area of active structure
I_b = Plate current, d-c
M = Ratio of plate current to cathode current
E_{c1} = Grid-cathode voltage
E_{c2} = Screen-cathode voltage
μ = Grid-screen amplification factor
I_0 = Cathode current density, d-c
r = Grid wire radius
I_p = Amplitude of fundamental component of a-c plate current
I_{2p} = Amplitude of second harmonic component of a-c plate current
i_p = Fundamental a-c plate current component
i_{2p} = Second harmonic plate current component
G_0 = Perveance factor
E_g = Amplitude of grid signal voltage
p = Frequency $\times 2\pi$
t = Time
i = a-c plate current

Units: length in cms; practical electrical units; time in seconds.

Assumptions

I $$na = 1$$

The ratio of the pitch distance to the grid-cathode spacing is held constant. This is done so that the effect of the variation in field along the cathode surface resulting from the finite grid pitch distance will be small and the same throughout the discussion.

II $$C_1 = 0.0885 \times 10^{-12} A \left(\frac{1}{a} + \frac{1}{b}\right)$$

$$C_2 = \frac{0.0885 \times 10^{-12} A}{c}$$

The input and output spaces are treated as if they can be represented by ideal condensers. This amounts to assuming that the grids are perfect planes and that there are no end effects. The effects of space charge on the capacitances are neglected, and the socket and wiring capacitances are also neglected. This means that the resulting calculated figure of merit represents the limiting value inherent in the tube structure.

III $$I_B = \frac{2.33 \times 10^{-6} MA \left(\frac{E_{c2}}{\mu} + E_{c1}\right)^{3/2}}{a^2 \left(1 + \frac{1}{\mu} \frac{a+b}{a}\right)^{3/2}}$$

The expression for plate current assumed is an idealized one, but holds fairly well for these tubes.

IV $$\frac{d\mu}{dE_{c1}} = 0$$

It is assumed that the triode amplification factor is independent of the control-grid voltage. This holds fairly well under the conditions of assumption I.

V When the interstage consists of a single parallel tuned circuit, $K = 1$. This case is assumed.

Derivations

Beginning with the above assumptions, and substituting in equation (1), equations (2), (3) and (4) can be derived. The procedure will be outlined below:

$$F = \frac{KG_m}{2\pi(C_1 + C_2)} \qquad (1)$$

$$F = \frac{4.74 \times 10^8 M I_0^{1/3}}{a^{4/3}\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right)\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{2}$$

$$E_{c2} = \mu\left[5.68 \times 10^3 a^{4/3} I_0^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right) - E_{c1}\right] \tag{3}$$

$$\mu = \frac{2.73\frac{b}{a} - \log_{10}\cosh\left(2\pi\frac{r}{a}\right)}{\log_{10}\coth\left(2\pi\frac{r}{a}\right)} \tag{4}$$

Substitutions for C_1, C_2, and K in (1) are made from assumptions II and V. G_m is found by differentiating the plate current expression (assumption III) with respect to E_{c1}, remembering that μ is independent of E_{c1} according to assumption IV.

$$G_m = \frac{3}{2}\frac{2.33 \times 10^{-6} MA\left(\frac{E_{c2}}{\mu}+E_{c1}\right)^{1/2}}{a^2\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}} \tag{5}$$

From assumption III we can write

$$\left(\frac{E_{c2}}{\mu}+E_{c1}\right)^{1/2} = \frac{I_B^{1/3} a^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{1/2}}{(2.33 \times 10^{-6})^{1/3} M^{1/3} A^{1/3}} \tag{6}$$

Substituting in (5)

$$G_m = \frac{3}{2}\frac{(2.33 \times 10^{-6}) MA I_B^{1/3} a^{2/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{1/2}}{a^2\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}(2.33 \times 10^{-6})^{1/3} M^{1/3} A^{1/3}}$$

$$G_m = \frac{1.5(2.33 \times 10^{-6})^{2/3} M^{2/3} A^{2/3} I_B^{1/3}}{a^{4/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{7}$$

Since $I_B = MI_0A$ by definition,

$$G_m = \frac{2.64 \times 10^{-4} MA I_0^{1/3}}{a^{4/3}\left(1+\frac{1}{\mu}\frac{a+b}{a}\right)} \tag{8}$$

Substituting in (1),

$$F = \frac{2.64 \times 10^{-4} M A I_0^{1/3}}{a^{4/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right) 2\pi \left[.0885 \times 10^{-12} A \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c}\right)\right]} \tag{9}$$

Collecting the constants and cancelling the A's,

$$F = \frac{4.74 \times 10^{8} M I_0^{1/3}}{a^{4/3}\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c}\right)\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)} \tag{2}$$

The expression for E_{c2} can be found by substituting $I_B = MI_0A$ in assumption III and solving for E_{c2}.

$$MI_0A = \frac{2.33 \times 10^{-6} MA \left(\frac{E_{c2}}{\mu} + E_{c1}\right)^{3/2}}{a^2\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)^{3/2}}$$

$$\frac{E_{c2}}{\mu} + E_{c1} = \frac{a^{4/3} I_0^{2/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right)}{1.76 \times 10^{-4}}$$

$$E_{c2} = \mu\left[5.68 \times 10^{3} a^{4/3} I_0^{2/3}\left(1 + \frac{1}{\mu}\frac{a+b}{a}\right) - E_{c1}\right] \tag{3}$$

The expression for μ can be found by applying assumption I and substituting $n = \frac{1}{a}$ in the Vogdes-Elder formula* for a plane structure.

$$\mu = \frac{2\pi nb}{2.303 \log_{10} \coth (2\pi nr)} - \frac{\log_{10} \cosh (2\pi nr)}{\log_{10} \coth (2\pi nr)} \tag{10}$$

Substituting $n = \frac{1}{a}$

$$\mu = \frac{2.73 \frac{b}{a}}{\log_{10} \coth\left(2\pi \frac{r}{a}\right)} - \frac{\log_{10} \cosh\left(2\pi \frac{r}{a}\right)}{\log_{10} \coth\left(2\pi \frac{r}{a}\right)} \tag{4}$$

Equation (11) can be derived by considering a particular structure and introducing a small sinusoidal signal $E_g \cos pt$ added to the d-c. voltage in the plate current expression, Assumption III. This can be written as

$$I_B + i = G_0\left(\frac{E_{c2}}{\mu} + E_{c1} + Eg \cos pt\right)^{3/2} \tag{12}$$

* F. B. Vogdes and Frank R. Elder, *Phys. Rev.*, 24, pp. 683–689, Dec., 1924.

For zero signal this becomes

$$I_B = G_0 \left(\frac{E_{c2}}{\mu} + E_{c1} \right)^{3/2} \tag{13}$$

For a pure resistance load which is small compared to the plate resistance, the fundamental component, neglecting contributions from third and higher order terms, is

$$i_p = G_m E_g \cos pt \tag{14}$$

Neglecting the contributions of fourth and higher order terms, the second harmonic component is

$$i_{2p} = I_B \frac{3}{16} \frac{E_g^2}{\left(\frac{E_{c2}}{\mu} + E_{c1} \right)^2} \cos 2pt \tag{15}$$

This is found by expanding (12) into the binomial series. From Assumption III and equation (5),

$$\left(\frac{E_{c2}}{\mu} + E_{c1} \right)^2 = \frac{9 I_B^2}{4 G_m^2} \tag{16}$$

The amplitude of the second harmonic component, from (15) is

$$I_{2p} = \frac{3}{16} \frac{I_B E_g^2}{\left(\frac{E_{c2}}{\mu} + E_{c1} \right)^2} \tag{17}$$

Substituting for $\left(\frac{E_{c2}}{\mu} + E_{c1} \right)^2$ from (16) in (17).

$$I_{2p} = \frac{G_m^2 E_g^2}{12 I_B} \tag{18}$$

From (14),

$$G_m^2 E_g^2 = I_p^2 \tag{19}$$

Substituting from (19) in (18),

$$I_{2p} = \frac{I_p^2}{12 I_B} \tag{20}$$

This can be written

$$\frac{I_p}{I_{2p}} = 12 \frac{I_B}{I_p} \tag{11}$$

Telephone Traffic Time Averages

By JOHN RIORDAN

(*Manuscript Received April 25, 1951*)

This paper describes the determination of the first four semi-invariants of the distribution of the average, over an arbitrary time interval, of traffic carried by a telephone system with an infinite number of trunks, during a period of statistical equilibrium. Both finite and infinite numbers of independent call sources are considered, and the distribution function of call holding times is left general.

1. Introduction

FOR mathematical studies of telephone traffic, like those of call loss or delay which are used in trunking engineering, the traffic is considered as a flow of probability in time. In the period of most importance, the busy hour, this flow is usually regarded as stationary; that is to say, the probability of a given number of busy trunks, or the probability of delay of an incoming call (or any other probability of the system which comes in question) is taken as independent of the particular moment in the busy hour at which the system is examined. The system is said to be in statistical equilibrium.

For such theoretical studies, the statistical quantities which determine these probabilities, like the rate at which calls appear, are of course taken as given, but in the application they must be determined by observations, such as those being taken in the current extensive program of traffic measurements. Here a difficulty appears. To abridge the extensive amount of observational material, either measurements are made of traffic averages over periods small compared to the busy hour (but not small enough to be neglected) or the measurements of continuous recorders are averaged by hand. It may be noticed here that for application of the results given below the traffic averages obtained by measurements must be those of a continuous device which records all traffic changes and not, as in some measuring devices, those obtained from a number of "looks" at points within the averaging interval. But to use these measurements in determining the traffic parameters by standard sampling theory, a corresponding theoretical study of the averages is necessary.

Such a study, within limits to be described presently, is given here. No attempt is made to describe the sampling studies possible from the results reached. These seem to be of many kinds, not necessary to describe, but for

concreteness it may be mentioned that the most important, at the moment, seems to be that of setting confidence limits for the average traffic.

The most important of the limits to this study are those implied by the assumptions of statistical equilibrium with fixed average, and an infinite number of trunks. The former limits application to periods in which, roughly speaking, average traffic is neither rising nor falling; the latter is justified only by the extreme mathematical difficulties produced by assuming otherwise. The traffic variable is the number of busy trunks in a period of statistical equilibrium. For pure chance call input, the call holding time characteristic is left arbitrary throughout the development, but main interest lies in the two extreme cases of constant holding time and exponential holding time, which are examined in detail.* For calls from a limited number of sources, results are obtained only for exponential holding time.

More precisely, if $N(t)$ is the random variable for the number of busy trunks at time t, the variable studied, the average number of calls in an interval of length T, is

$$M(T) = \frac{1}{T}\int_0^T N(t)\,dt \tag{1}$$

The question is: What are the statistical properties of $M(T)$?

The results given are the first four cumulants (semi-invariants) of $M(T)$, which seem to have the simplest expressions. For the convenience of the reader it may be noticed that the first cumulant is the mean, the second the second moment about the mean which is the variance, the third the third moment about the mean, and the fourth the fourth moment about the mean less three times the square of the variance.

In all cases the mean of $M(T)$ is the mean of $N(t)$ and for pure chance call input is called b, the average number of calls in unit average holding time, h.

The other cumulants for pure chance call input, k_n, have the general expression

$$k_n = b\,\frac{n(n-1)}{T^n}\int_0^T dx g(x)(T-x)x^{n-2}; \qquad n = 2,3,4$$

with

$$g(x) = \frac{1}{h}\int_x^\infty f(t)\,dt$$

* F. W. Rabe [6] has reported results for these two cases for relatively long averaging intervals, which are verified below. I owe my interest in this problem to a report on Rabe's work made by Messrs. Gibson, Hayward and Seckler in a probability colloquium at Bell Telephone Laboratories initiated and directed by Roger Wilkinson.

and $f(t)$ the probability that a call lasts at least t, that is, the distribution function of holding times. The specializations of this, for constant holding time and exponential holding time, appear in section 4. The results for finite source input have a similar character.

The procedure in obtaining these is as follows. The cumulants are determined from the ordinary moments (about the origin) and the latter are determined by the integration of expectations. Thus the first moment, the mean is determined from

$$E[M(T)] = \frac{1}{T}\int_0^T E[N(t)]\,dt = E[N(t)] \tag{2}$$

where $E(x)$ is written for the expectation or mean of x.

Similarly the second moment is given by

$$E[M^2(T)] = \frac{1}{T^2}\int_0^T \int_0^T E[N(t)N(u)]\,dt\,du \tag{3}$$

and so on for higher moments. Correlation effects appear in (3) in $E[N(t)N(u)]$ and are included in the development by formulation of transition probabilities, that is, those probabilities determining the traffic flow in time. The transition probability $P_{jk}(t)$ is defined as the probability of transition in t from j calls in progress (busy trunks) to k calls in progress, and fixes the inter-relatedness of call probabilities at different time epochs. Only for large values of t are these probabilities independent.

Hence, the first task is to determine these simple transition probabilities, then those of double and triple transitions, then the expected values of pairs, triples and quadruples of numbers of busy trunks, and finally the moments.

2. Transition Probabilities

For exponential holding time, and infinite sources, infinite trunks, these probabilities have already been determined by Conny Palm [5]. Palm's work has been summarized both by Feller [1] and by Jensen [3], and describes the whole process, not merely the equilibrium condition. For the equilibrium condition, a different procedure,* similar to that used by Newland [4] for another purpose, allows the assumption of a more general holding time characteristic.

* Thanks are due S. O. Rice for suggesting this, as well as for many corrections and improvements. I also have had the advantage of a careful reading of the mss. by E. L. Kaplan.

For infinite sources, and calls arriving individually and collectively at random with average density a, the well-known formula for the probability that exactly k calls arrive in time interval t is the Poisson

$$\pi_k(t) = e^{-at}(at)^k/k! \tag{4}$$

Then, if $P_{ij}(t; k)$ is the conditional probability of transition from i to j when k calls arrive in time t,

$$P_{ij}(t) = \sum_{k=0}^{\infty} P_{ij}(t; k)\pi_k(t) \tag{5}$$

Consider $P_{ij}(t; 0)$, that is the (conditional) transition probabilities when no calls arrive. Let the probability that a call lasts at least t be $f(t)$, so that the average holding time h is given by

$$h = \int_0^{\infty} u[-f'(u)]\, du = \int_0^{\infty} f(u)\, du \tag{6}$$

The i calls initially in process are independent of each other. Select one of them and suppose the time from its arrival (its age) is t_1. Then the probability that it will also exist t units later is the conditional probability $f(t + t_1)/f(t_1)$. Since in equilibrium conditions all moments of arrival have equal probability, the corresponding probability for an arbitrary call is

$$g(t) = \int_0^{\infty} f(t + t_1)\, dt_1 \div \int_0^{\infty} f(t_1)\, dt = \frac{1}{h}\int_t^{\infty} f(u)\, du \tag{7}$$

Hence the transitional probability $P_{ij}(t; 0)$ is the binomial expression

$$P_{ij}(t; 0) = \binom{i}{j} g^j(1 - g)^{i-j} \tag{8}$$

and its generating function is

$$P_i(t, x; 0) = \sum P_{ij}(t; 0)x^j = [1 + (x - 1)g]^i \tag{9}$$

In (8) and (9), for brevity, the argument of g is omitted.

Now, suppose one call arrives in interval t. The moment of arrival is uniformly distributed in t; that is, if u_1 is the moment of arrival,

$$Pr(u < u_1 < u + du) = du/t$$

and the probability that a call arriving at an arbitrary moment will be in existence at time t is, say,

$$Q(t) = \int_0^t f(t - u)\frac{du}{t} = \frac{1}{t}\int_0^t f(u)\, du = \frac{h}{t}(1 - g(t)) \tag{10}$$

The corresponding generating function is

$$1 - Q(t) + xQ(t) = 1 + (x - 1)Q(t)$$

and, since calls arriving are independent, the generating function for k calls arriving is

$$[1 + (x - 1)Q]^k$$

and

$$P_i(t, x; k) = [1 + (x - 1)g]^i[1 + (x - 1)Q]^k \tag{11}$$

Hence, finally by (5),

$$\begin{aligned} P_i(t; x) &= \sum P_{ij}(t)x^j \\ &= [1 + (x - 1)g]^i \sum_{k=0}^{\infty} [1 + (x - 1)Q]^k \frac{e^{-at}(at)^k}{k!} \\ &= [1 + (x - 1)g]^i \exp (x - 1)\, at\, Q \\ &= [1 + (x - 1)g]^i \exp (x - 1)\, ah\, (1 - g) \end{aligned} \tag{12}$$

The last step uses (10).

This is the generating function for the simplest transition probabilities, and is quite like Palm's result; indeed, for exponential holding time $g = f = e^{-t/h}$. The probabilities themselves are obtained by expansion of the generating function in powers of x, or by substituting g for $e^{-t/h}$ in Palm's result. But they are not needed here; the generating function is most apt for determining the averages of interest, as will appear.

Before going on to the other transition probabilities, it is interesting to notice certain checks of equation (12). In statistical equilibrium the traffic has Poisson density (Palm l.c.) that is, in the present notation

$$Pr(N(t) = k) = e^{-b}b^k/k!$$

where $b = ah$. This of course is independent of time. Then, if $N(0)$ has this density, so should $N(t)$ as determined from $N(0)$ and the transition probabilities implicit in (12). This is verified by

$$\begin{aligned} \sum P_i(t,x)e^{-b}b^i/i! &= \exp (x - 1)b(1 - g) \sum [1 + (x - 1)g]^i \frac{e^{-b}b^i}{i!} \\ &= \exp [(x - 1)b(1 - g) - b + b + (x - 1)bg] \\ &= \exp (x - 1)b. \end{aligned} \tag{13}$$

Also, $g(0) = 1$ and $g(\infty) = 0$ so that

$$P_i(0, x) = [1 + (x - 1)]^i = x^i \tag{14}$$

$$P_i(\infty, x) = \exp (x - 1)b \tag{15}$$

showing that in zero time no transit to another state is possible, and in infinite time the equilibrium probabilities are reached no matter what the initial state has been.

Finally, in a Markov process (cf. Feller [2], Chap. 15) the simple transition probabilities alone are needed since

$$P_{ijk}(t, u) = P_{ij}(t)P_{jk}(u)$$

A test for this is the Chapman-Kolomogorov equation, namely

$$P_{ik}(t + u) = \sum_j P_{ij}(t)P_{jk}(u)$$

Using (12), the corresponding relation of generating functions is

$$[1 + (x - 1)g(t + u)]^i \exp b(x - 1)[1 - g(t + u)]$$
$$= [1 + (x - 1)g(t)g(u)]^i \exp b(x - 1)[1 - g(t)g(u)];$$

so the process is Markovian only if

$$g(t + u) = g(t)g(u)$$

which is true only for exponential holding time.

For the next transition probability $P_{ijk}(t, u)$, consider first the condition in which no call arrives in the whole interval $t + u$. As before

$$P_{ij}(t) = \binom{i}{j} g_t^j (1 - g_t)^{i-j}$$

where for convenience g_t is written for $g(t)$. For the next transit, however, there is a difference, namely

$$P_{jk}(u) = \binom{j}{k} \left(\frac{g_{t+u}}{g_t}\right)^k \left(1 - \frac{g_{t+u}}{g_t}\right)^{j-k}$$

since g_{t+u}/g_t is the conditional probability that a call which has lasted t will last u more; $P_{jk}(u)$ is the conditional probability of a transit from j to k in u, given the transit i to j in t.

The generating function for the double transition probabilities in this case is, then,

$$\sum_j \sum_k P_{ijk}(t, u; 0)x^j y^k = [1 + (x - 1)g_t + x(y - 1)g_{t+u}]^i \tag{16}$$

Now suppose a single call arrives at random in interval t. As before, the probability that it will occupy a trunk at time t is $Q(t) = ht^{-1}(1 - g(t))$

and the conditional probability that it will also occupy a trunk at time $t + u$ is

$$\frac{1}{t}\int_0^t f(t + u - x)\,dx \div Q(t)$$

or

$$\frac{g(u) - g(t + u)}{1 - g(t)} = R(t, u), \quad \text{say.}$$

The corresponding generating function, with x and y the indicators of calls at t and $t + u$, resp. is

$$1 - Q(t) + Q(t)[1 - R(t, u)]x + Q(t)R(t, u)xy$$

or

$$1 + (x - 1)(1 - g(t))h/t + x(y - 1)[g(u) - g(t + u)]h/t$$

The generating function for c calls in this interval is this expression raised to the c'th power, since calls arrive independently; and since c calls arrive with probability $e^{-at}(at)^c/c!$, the generating function for calls arriving in this interval is

$$\begin{aligned}\sum [1 + (x - 1)Q + x(y - 1)QR]^c e^{-at}(at)^c/c! \\ = \exp b[(x - 1)(1 - g(t)) + x(y - 1)(g(u) - g(t + u))]\end{aligned} \tag{17}$$

For brevity Q and R have been written for $Q(t)$ and $R(t, u)$.

Finally the generating function for calls arriving in t, $t + u$, is

$$\exp b(y - 1)(1 - g(u)) \tag{18}$$

Hence

$$\begin{aligned}\sum_j \sum_k P_{ijk}(t, u)x^j y^k = [1 + (x - 1)g(t) + x(y - 1)g(t + u)]^i \\ \cdot \exp b[(x - 1)(1 - g(t)) + (y - 1)(1 - g(u)) \\ + x(y - 1)(g(u) - g(t + u))]\end{aligned} \tag{19}$$

By similar argument, the generating function for triple transition probabilities is

$$\begin{aligned}\sum_j \sum_k \sum_t P_{ijkt}(t, u, v)x^i y^k z^t \\ = [1 + (x - 1)g_t + x(y - 1)g_{t+u} + xy(z - 1)g_{t+u+v}]^i \\ \cdot \exp b((x - 1)(1 - g_t) + (y - 1)(1 - g_u) + \\ (z - 1)(1 - g_v) + x(y - 1)(g_u - g_{t+u}) + \\ y(z - 1)(g_v - g_{u+v}) + xy(z - 1)(g_{u+v} - g_{t+u+v}))\end{aligned} \tag{20}$$

3. Expected Correlations

Correlation expectations, like $E[N(t)N(u)]$ in equation (3), are needed for evaluation of the moments of $M(T)$. They may be determined from the transition probability generating functions, if it is agreed, as a matter only of convenience, that the time epochs t, u, v, etc. are in that order ($t \leq u \leq v \leq \cdots$). Since, on the assumption of statistical equilibrium, the call probabilities at the first epoch t, are independent of its value, as already noticed, this value may be taken as zero without loss of generality.

Thus for the second moment it is sufficient to determine

$$\varphi(u) = E[N(0)N(u)] = \sum ip_i \sum jP_{ij}(u) \tag{21}$$

with $p_i = Pr[N(0) = i] = e^{-b}b^i/i!$

Write

$$G_u(x, y) = \sum p_i x^i \sum P_{ij}(u)y^j$$

By (12), this is the same as

$$G_u(x, y) = \exp b[x - 1 + y - 1 + (x - 1)(y - 1)g(u)]$$

or

$$H_u(x, y) = G_u(x + 1, y + 1) \equiv \exp b(x + y + xyg(u))$$

and

$$\begin{aligned} \varphi(u) &= \frac{\partial^2 H}{\partial x \partial y}\bigg|_{x,y=0} \\ &= b^2 + bg(u) \end{aligned} \tag{22}$$

In the same way the second order correlation expectation, that is

$$\varphi(u, v) = E[N(0)N(u)N(u + v)],$$

is obtained from

$$G_{u,v}(x, y, z) = \sum p_i x^i \sum \sum P_{ijk}(u, v)y^j z^k$$

and

$$\begin{aligned} H_{u,v}(x, y, z) &= G_{u,v}(x + 1, y + 1, z + 1) \\ &= \exp b(x + y + z + xyg(u) + yzg(v) + x(y + 1)zg(u + v)) \end{aligned}$$

Hence

$$\varphi(u, v) = b^3 + b^2[g(u) + g(v) + g(u + v)] + bg(u + v) \tag{23}$$

Finally, the third order correlation turns out to be

$$\begin{aligned}\varphi(u, v, w) &= E[N(0)N(u)N(u+v)N(u+v+w)] \\ &= b^4 + b^3[g(u) + g(v) + g(w) \\ &\quad + g(u+v) + g(v+w) + g(u+v+w)] \\ &\quad + b^2[g(u+v) + g(v+w) + 2g(u+v+w)] \\ &\quad + b^2[g(u)g(w) + g(u+v)g(v+w) \\ &\quad + g(v)g(u+v+w)] + bg(u+v+w)\end{aligned} \tag{24}$$

As will appear, the arrangement of terms in (22), (23) and (24) corresponds to the expansion of ordinary moments in terms of cumulants (semi-invariants); e.g. (24) corresponds to $m_4 = b^4 + 6b^2k_2 + 4bk_3 + 3k_2^2 + k_4$ with k_i the i'th cumulant (for the Poisson of mean b, $k_i = b$).

4. Moments

Moments are obtained from these results by integrations. As already noted, equation (2), the first moment is b for any holding time distribution.

Since there are two ways of ordering the epochs t, u, the second moment is

$$\begin{aligned}E[M^2(T)] &= \frac{2}{T^2}\int_0^T dt \int_0^t du\, \varphi(t-u) \\ &= b^2 + \frac{2b}{T^2}\int_0^T dt \int_0^t du\, g(t-u) \\ &= b^2 + \frac{2b}{T^2}\int_0^T dx\, g(x)(T-x)\end{aligned} \tag{25}$$

The last step is by the formula for reversing the order of integration indicated by

$$\int_0^T dt \int_0^t du = \int_0^T du \int_u^T dt$$

The variance or second central moment, which is also the second cumulant k_2, is then

$$\begin{aligned}\text{Var}\,[M(T)] &= E[(M(T) - b)^2] \\ &= E[M^2(T)] - b^2 \\ &= \frac{2b}{T^2}\int_0^T dx\, g(x)(T-x)\end{aligned} \tag{26}$$

Since there are $3! = 6$ ways of ordering 3 epochs, the third moment may be written

$$E[M^3(T)] = \frac{6}{T^3}\int_0^T dt \int_0^t du \int_0^u dv\, \varphi(t-u, u-v)$$

$$= b^3 + \frac{6b^2}{T^3}\int_0^T dt \int_0^t du \int_0^u dv[g(t-w) + g(u-v) + g(t-v)]$$

$$+ \frac{6b}{T^3}\int_0^T dt \int_0^t du \int_0^u dv\, g(t-v)$$

Here the first triple integral is immediately evaluated by use of the identity

$$2\int_0^T dt \int_0^t du \int_0^u dv[g(t-u) + g(u-v) + g(t-v)]$$

$$= \int_0^T \int_0^T \int_0^T dt\, du\, dv\, g(|\, t-v\,|)$$

$$= 2T\int_0^T dx\, g(x)(T-x)$$

$$= T^3 k_2/b$$

The last triple integral, by successive inversions of integration order, turns out to be

$$\frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x$$

Hence finally

$$E[M^3(T)] = b^3 + 3bk_2 + \frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x \tag{27}$$

and

$$\begin{aligned} k_3 &= E[(M(T) - b)^3] \\ &= E[M^3(T)] - 3b\, E[M^2(T)] + 2b^3 \\ &= E[M^3(T)] - 3b\, k_2 - b^3 \\ &= \frac{6b}{T^3}\int_0^T dx\, g(x)(T-x)x \end{aligned} \tag{28}$$

The fourth moment is given by

$$
\begin{aligned}
E[M^4(T)] &= \frac{24}{T^4}\int_0^T dt \int_0^t du \int_0^u dv \int_0^v dw\, \varphi(t-u, u-v, v-w) \\
&= b^4 \\
&+ \frac{24}{T^4}\Bigg\{ b^3 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw [g(t-u) + g(t-v) \\
&\qquad + g(t-w) + g(u-v) + g(u-w) + g(v-w)] \\
&+ b^2 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw\, [g(t-v) + g(u-w) + 2g(t-w)] \\
&+ b^2 \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw [g(t-u)g(v-w) + \\
&\qquad g(t-v)g(u-w) + g(t-w)g(u-v)] \\
&+ b \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw[g(t-w)]\Bigg\}
\end{aligned}
$$

Employing the identities

$$
\begin{aligned}
4\int_0^T \int_0^T \int_0^T \int_0^T & dt\, du\, dv\, dw\, [g(t-u) + g(t-v) + g(t-w) \\
&+ g(u-v) + g(u-w) + g(v-w)] \\
&= \int_0^T \int_0^t \int_0^u \int_0^v dt\, du\, dv\, dw\; g(|\, t-u\,|) \\
&= 2T^2 \int_0^T dx\; g(x)(T-x) = T^4 k_2/b,
\end{aligned}
$$

$$
\begin{aligned}
8\int_0^T \int_0^t \int_0^u \int_0^v & dt\, du\, dv\, dw\, [g(t-u)\; g(v-w) + g(t-v)\; g(u-w) \\
&+ g(t-w)\; g(u-v)] \\
&= \int_0^T \int_0^T \int_0^T \int_0^T dt\, dw\, dv\, dw\; g(|\, t-u\,|) g(|\, v-w\,|) \\
&= 4\left[\int_0^T dx\; g(x)(T-x)\right]^2 = T^4 k_2^2/b^2,
\end{aligned}
$$

and successive inversion of order of integration, the final result turns out to be

$$E[M^4(T)] = b^4 + 6b^2k_2 + 4bk_3 + 3k_2^2 + \frac{12b}{T^4}\int_0^T dx\, g(x)(T-x)x^2 \tag{29}$$

and

$$\begin{aligned} k_4 &= E[(M(T)-b)^4] - 3E[(M(T)-b)^2] \\ &= \frac{12b}{T^4}\int_0^T dx\, g(x)(T-x)x^2 \end{aligned} \tag{30}$$

It is a tempting surmise that

$$k_n = b\,\frac{n(n-1)}{T^n}\int_0^T dx\, g(x)(T-x)\,x^{n-2}$$

but this has not been proved. Note that for $g(x) = 1$, $k_n = b$, the cumulant of the Poisson, as it should.

For the two cases of chief interest, constant and exponential holding times, the function $g(x)$, *in average holding time units* (that is, $x = t/h$) is given by

c.h.t. $g(x) = 1 - x \quad x < 1$; $= 0 \quad x > 1$

e.h.t. $g(x) = e^{-x}$

and the results are as follows:

Cumulant	Constant Holding Time T < 1	T > 1
k_2	$b(1 - T/3)$	$bT^{-1}(1 - 1/3T)$
k_3	$b(1 - T/2)$	$bT^{-2}(1 - 1/2T)$
k_4	$b(1 - 3T/5)$	$bT^{-3}(1 - 3/5T)$
	Exponential Holding Time	
k_2	$2bT^{-2}[T - 1 + e^{-T}]$	
k_3	$6bT^{-3}[T - 2 + (T+2)e^{-T}]$	
k_4	$12bT^{-4}[2T - 6 + (T^2 + 4T + 6)e^{-T}]$	

It may be worth noting that, if the surmise is correct, for constant holding time

$$k_n = b\left[1 - \frac{n-1}{n+1}T\right] \qquad T < 1$$

$$= \frac{b}{T^{n-1}}\left[1 - \frac{n-1}{n+1}\frac{1}{T}\right] \qquad T > 1$$

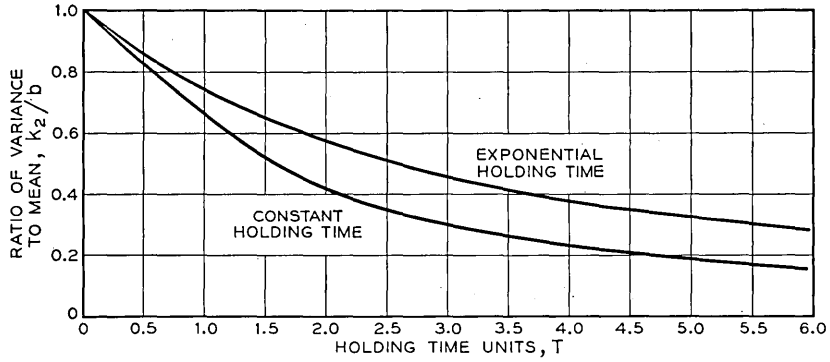


FIG. 1.—Comparison of variances of average traffic for constant and exponential holding times.

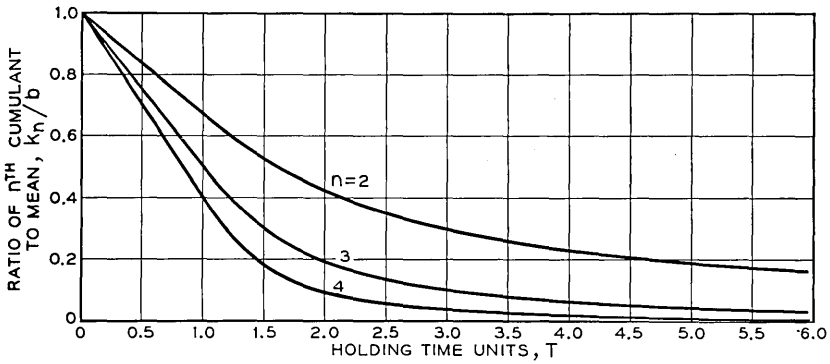


FIG. 2.—Cumulants k_2, k_3, and k_4 for constant holding time.

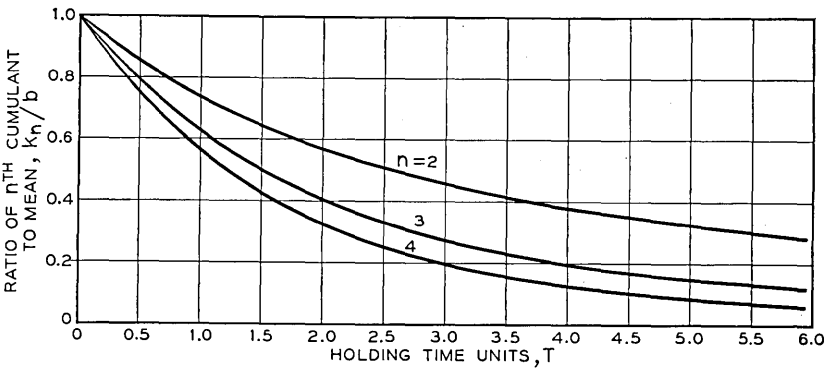


FIG. 3.—Cumulants k_2, k_3 and k_4 for exponential holding time.

and for exponential holding time

$$k_n = b\,\frac{n(n-1)}{T^n}\,[(n-2)!\,T - (n-1)! + e^{-T}\,(T+\alpha)^{n-2}]$$

where in the last term $(T+\alpha)^{n-2}$ is a symbolic expression or shorthand for

$$(T+\alpha)^{n-2} = \sum_0^{n-2} \binom{n-2}{m} T^{n-2-m}\,\alpha_m$$

and $\alpha_m = (m+1)!$; e.g.

$$(T+\alpha)^3 = T^3 + 6T^2 + 18T + 24$$

For small values of T, the two cases coalesce ($e^{-x} \approx 1 - x$) and at $T = 0$ approach b as they should. For large values of T, and constant holding time,

$$k_n \sim b/T^{n-1}, \qquad (n = 2, 3, 4);$$

for exponential holding time

$$k_n \sim n!b/T^{n-1}, \qquad (n = 2, 3, 4).$$

For $n = 2$, these results agree with Rabe [6].

As T increases, for either holding time, the cumulants are progressively smaller, and the approximation of the distribution of $M(T)$ by a normal curve (which has all cumulants, except the first and second, zero) improves. This is what follows from the central limit theorem if the subdivision of T into a large number of intervals results in mutually independent random variables (cf. Rice [7] 3.9).

Figure 1 shows a comparison of the variances (k_2) for the two holding time cases. Figure 2 shows a comparison of the cumulants k_2, k_3 and k_4 for constant holding time, and Fig. 3 shows the same thing for exponential holding time.

5. Finite Sources—Exponential Holding Time

The generating function for transitional probabilities for N subscribers, each originating calls independently with probability λ, and for exponential holding time, as given by Jensen (l.c.) is as follows:

$$P_i(t, x) = [1 + q_1(x-1)]^i[1 + q_0(x-1)]^{N-i} \qquad (31)$$

with

$$q_0 = p - pe^{-(\lambda+\gamma)t}$$
$$q_1 = p + q \quad \text{“}$$
$$p = 1 - q = \lambda/(\lambda+\gamma)$$
$$\gamma = 1/h$$

It should be noticed that for $t = \infty$, $q_0 = q_1 = p$ and

$$P_i(\infty, x) = [1 + p(x - 1)]^N \tag{32}$$

The right hand side is the binomial generating function and, as independent of i, is the generating function for the statistical equilibrium probabilities; that is

$$Pr\,[N(t) = k] = \binom{N}{k} p^k\, q^{N-k}$$

Also the process is Markovian since

$$\begin{aligned}\sum_k x^k \sum_j P_{ij}(t)P_{jk}(u) &= \sum_j P_{ij}(t)[1 + q_{1u}(x - 1)]^j\,[1 + q_{0u}(x - 1)]^{N-j} \\ &= [1 + (q_{0u} + q_{1t}\,q_{1u} - q_{1t}\,q_{0u})(x - 1)]^i \\ &\quad [1 + (q_{0u} + q_{0t}\,q_{1u} - q_{0t}\,q_{0u})(x - 1)]^{N-i}\end{aligned}$$

and

$$q_{0u} + q_{1t}\,q_{1u} - q_{1t}\,q_{0u} = q_{1,\,t+u}$$

$$q_{0u} + q_{0t}\,q_{1u} - q_{0t}\,q_{0u} = q_{0,\,t+u}$$

Here it has been convenient to indicate by the double subscript the dependence of q_0 and q_1 on a time variable.

Moments are obtained by the process given in detail for the infinite source case. For brevity it is convenient to use the binomial cumulants which are as follows

$$\kappa_2 = Npq$$

$$\kappa_3 = Npq(q - p)$$

$$\kappa_4 = Npq(1 - 6pq)$$

and the modified time variable $T_1 = (\lambda + \gamma)T$. Then the results are

$$\begin{aligned}k_2 &= 2T_1^{-2}\kappa_2[T_1 - 1 + e^{-T_1}] \\ k_3 &= 6T_1^{-3}\kappa_3[T_1 - 2 + (T_1 + 2)e^{-T_1}] \\ k_4 &= 12T_1^{-4}((\kappa_4 + \kappa_2^2 N^{-1})[2T_1 - 6 + (T_1^2 + 4T_1 + 6)e^{-T_1}] \\ &\quad - \kappa_2^2 N^{-1}[1 - (T_1^2 + 2)e^{-T_1} + e^{-2T_1}])\end{aligned}$$

These of course bear a strong resemblance to the infinite source case (exponential holding time), to which they converge.

BIBLIOGRAPHY

1. W. Feller, "On the theory of stochastic processes with particular reference to applications," *Proc. Berkeley Symposium on Math. Statistics and Probability*, Univ. of California Press, 1949.
2. W. Feller, "An introduction to probability theory and its applications," New York, 1950.
3. A. Jensen, An elucidation of Erlang's statistical works through the theory of stochastic processes, "The Life and Works of A. K. Erlang," Copenhagen, 1948.
4. W. F. Newland, A method of approach and solution to some fundamental traffic problems, *P.O.E.E. Jl.*, *25*, 119–131 (1932–3).
5. Conny Palm, "Intensitätsschwangungen in fernsprechverkehr," *Ericsson Technics*, *44*, 1–189 (1943).
6. F. W. Rabe, "Variations of telephone traffic," *Elec. Comm. 26*, 243–248 (1949).
7. S. O. Rice, "Mathematical analysis of random noise," *Bell System Technical Journal*, *23*, 282–332 (1944); *24*, 46–156 (1945).

The Reproduction of Magnetically Recorded Signals

R. L. WALLACE, JR.

(*Manuscript Received July 9, 1951*)

For certain speech studies at the Bell Telephone Laboratories, it has been necessary to design some rather specialized magnetic recording equipment. In connection with this work, it has been found experimentally and theoretically that introducing a spacing of d inches between the reproducing head and the recording medium decreases the reproduced voltage by $54.6\,(d/\lambda)$ decibels when the recorded wavelength is λ inches. For short wavelengths this loss is many decibels even when the effective spacing is only a few ten-thousandths of an inch. On this basis it is argued that imperfect magnetic contact between reproducing head and recording medium may account for much of the high-frequency loss which is experimentally observed.

Introduction

WITHIN the last few years there has been increasing use of magnetic recording in various telephone research applications (examples are various versions of the sound spectrograph used in studies of speech and noise). Some of these uses[1] have required a reproducing head spaced slightly out of contact with the recording medium. Experimental studies were made to determine the effect of such spacing and the results were found to be expressible in an unexpectedly simple form. The general equation derived is believed to be fundamental to the recording problem and to account for much of the high-frequency loss that is found in both in- and out-of-contact systems.

This paper discusses results of the experimental study and presents for comparison some theoretical calculations based on an idealized model.

Measurements of Spacing Loss

In order to measure the effect of spacing between the reproducing head and the medium, an experiment was set up as indicated in Fig. 1. The recording medium used was a 0.0003 inch plating of cobalt-nickel alloy[2] on the flat surface of a brass disc approximately 13 inches in diameter by $\frac{1}{4}$ inch thick.

This disc was made with considerable care to insure that the recording surface was as nearly plane and smooth as possible and that it would turn reasonably true in its bearings. Speeds of 25 and 78 rpm were provided.

[1] R. C. Mathes, A. C. Norwine, and K. H. Davis, "Cathode-Ray Sound Spectroscope," *Jl. Acous. Soc. Am.*, *21*, 527 (1949).

[2] Plating was done by the Brush Development Company.

The ring-type record-reproduce head shown in Fig. 1 was lapped slightly to obtain a reasonably good fit with the surface of the disc.

A single-frequency recording was made with the head in contact with the disc using a-c. bias in the usual way. Then the open circuit reproduced signal level was measured, first with the head in contact, and then after introducing paper shims of various thickness between the reproducing head and the medium. Thus the effect of spacing was measured at a particular frequency and recording speed. The signal was then erased and the process was repeated for other recorded frequencies and for several record-reproduce speeds. Measurements were also made for cases in which the recording and reproducing speeds were different. Considerable care was required to keep the disc and head sufficiently clean so that reproducible results could be obtained.

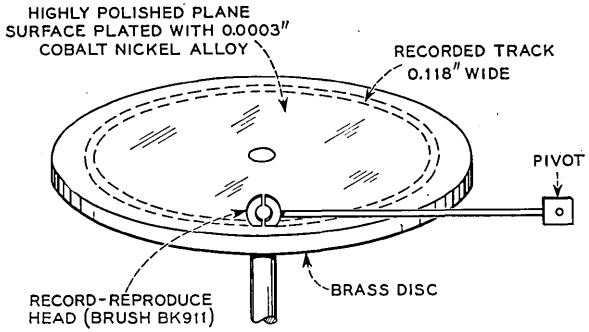


Fig. 1—Mechanical arrangement of recording set up. The one head served for recording, playback, and erase.

Figure 2 shows typical response curves measured at 21 in./sec. with the reproducing head in contact and with 0.004 inch spacing. The difference between these two curves will be called the spacing loss corresponding to this spacing and speed. From these data and more of the same sort it is found that, within experimental error, spacing loss can be very simply related to spacing and the recorded wavelength, λ, by the empirical equation,

$$\text{Spacing loss} = 55(d/\lambda) \text{ decibels} \qquad (1)$$

where spacing loss is the number of decibels by which the reproduced level is decreased when a spacing of d inches is introduced between the reproducing head and a magnetic medium on which a signal of wavelength λ inches has been recorded.

The fact that this expression fits the experimental data reasonably well is indicated in Fig. 3 where spacing loss data measured at a number of different speeds, frequencies, and spacings are plotted against d/λ.

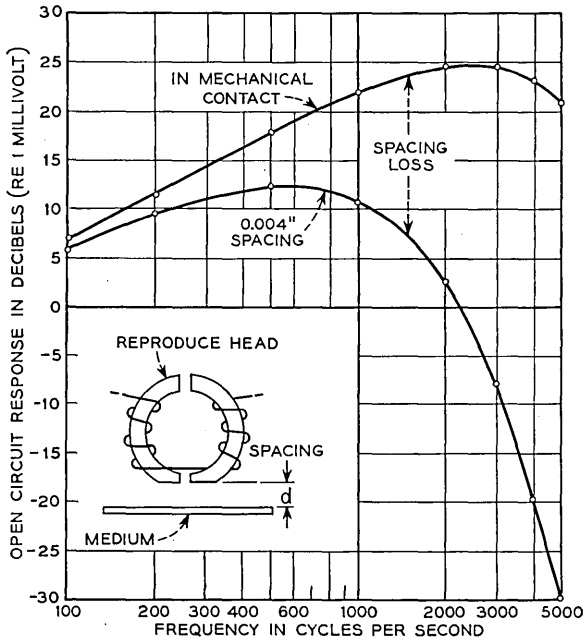


Fig. 2—Response curves taken at 21 in./sec. Recordings were made with head in contact and were played back first with head in contact and then with a spacing of 4 mils between head and disc.

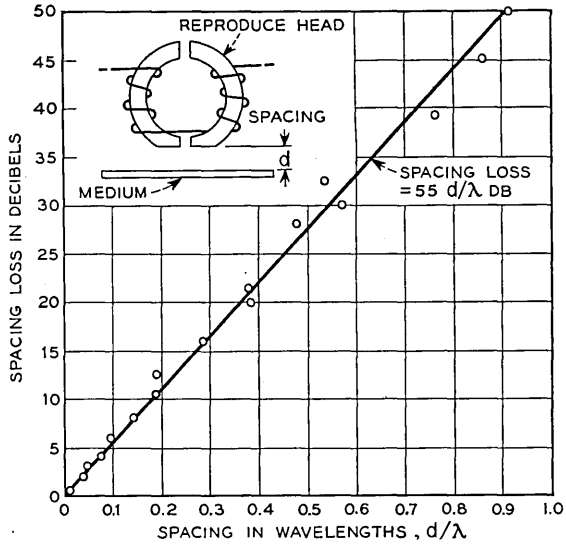


Fig. 3—Data obtained as in Fig. 2 show spacing loss approximately equal to $55(d/\lambda)$ decibels.

Implications of the Experiment

In this section it will be assumed that equation (1) holds true in all cases where the spacing d is sufficiently small and the recorded track is sufficiently wide so that end effects are negligible. If this is true, as it seems experimentally to be, then it is indeed surprising how great can be the effect of even a very small spacing when the recorded wavelength is small. For example, take the case of a 7500 cps signal recorded at 7.5 in./sec. in which case the wavelength is 0.001 inch. A particle of dust which separated the tape from the reproducing head by one-thousandth of an inch would decrease the reproduced level by 55 db. A spacing of 0.0001 inch would produce a quite noticeable 5.5 db effect and even at 0.00001 inch spacing the 0.55 db loss would be measurable in a carefully controlled experiment.

In view of the magnitudes involved, it seems probable that this spacing loss may play a significant role even in cases where the reproducing head is supposed to be in contact with the medium. For example, it has been known for some time that chattering of the tape on the reproducing head or changes in the degree of contact due to imperfect smoothness of the tape can result in amplitude modulation of the reproduced signal and thereby give rise to "modulation noise" or "noise behind the signal."

With the aid of equation (1) it is possible to estimate the magnitude of the noise provided some assumption is made about the wave form of the modulation. To take a simple case, suppose that the roughness of the tape were such as to sinusoidally modulate the spacing by a very small amount and at a low frequency. The reproduced signal would then be modulated and would contain a sideband on each side of the center frequency. The energy in these two sidebands constitutes the modulation noise in this case. If it is required that this noise be 40 db down on the signal, then one can calculate the maximum permissible excursion of the tape away from the reproducing head. This turns out to be $1.1(10)^{-5}$ cm. or about one-sixth of the wavelength of the red cadmium line! Of course, the one mil wavelength assumed in this example is about as short as is often used and the effect becomes less severe as the wavelength is increased. This is one of the reasons that speeds greater than 7.5 in./sec. are used for highest quality reproduction.

One can also make some rough qualitative inferences about the effect of the thickness of the recording medium on the shape of the response curve. As can be seen from equation (1) or from Fig. 2, low frequencies can be reproduced with very little loss in amplitude in spite of considerable spacing between the reproducing head and the medium while high frequencies (i.e. short wavelengths) may be appreciably attenuated by even 0.0001 inch

spacing between the head and the medium. With this in mind it is easy to see that at high frequencies only a thin layer of the medium nearest the reproducing head will contribute to the reproduced signal. In this case (short λ) increasing the thickness of the medium beyond a certain amount can have no effect on the reproduced level simply because the added part of the tape is too far from the head to make its effect felt. Consider the effect of increasing the thickness of the medium from 0.3 mil to 0.6 mil when the wavelength is one mil. Since the spacing loss for 0.3 mil spacing at $\lambda = 1$ mil is 16.5 db, the signal contributed by the lower half of a medium 0.6 mils thick cannot be less than 16.5 db lower than that contributed by the upper half and hence the increase in thickness can do no more than to raise the reproduced level by 1.2 db.

At a lower frequency for which $\lambda = 100$ mil, however, the corresponding spacing loss is only 0.165 db and in this case the two halves of the tape can contribute almost equally with the result that doubling the thickness of the medium can almost double the reproduced signal voltage.

Qualitatively, then, one might expect that increasing the thickness of the recording medium, other things being equal, would increase the response to low frequencies and leave the high frequency response relatively unaltered. This is in agreement with data published by Kornei.[3]

The estimates of magnitudes just given rest on assumptions which cannot be proved except by further experiments. It has been implicitly assumed, for example, that the medium is uniformly magnetized throughout its thickness and this may not be the case. It does seem perfectly safe, however, to conclude that at a wavelength of one mil that part of the medium which lies deeper than about 0.3 mil from the surface cannot contribute appreciably to the reproduced signal. Furthermore, as the wavelength is decreased beyond this point the thickness of the effective part of the tape decreases in inverse proportion to λ with the result that the available flux also decreases. For this reason the "ideal" response curve cannot continue indefinitely to rise at 6 db per octave as it does at low frequencies. In fact, when the effective part of the tape becomes thin enough, the available flux will decrease at 6 db per octave and just cancel the usual 6 db per octave rise, giving an "ideal" response curve which rises 6 db per octave at low frequencies but which eventually becomes flat, neither rising nor falling with further increase in frequency.

Spacing loss may contribute in still another way to the frequency response characteristic of a magnetic recording system in which the reproducing head makes contact with the medium. It is well known to those who work

[3] Otto Kornei, "Frequency Response of Magnetic Recording," *Electronics*, p. 124, August, 1947.

with magnetic structures such as are used in transformers and the like that intimate mechanical contact between two parts of a magnetic circuit does not imply intimate magnetic contact. In fact, even when great care is taken in fitting such parts together, measurements invariably show an effective air gap between them and the effective width of this gap usually amounts to appreciably more than one mil. One reason for this is that the permeability of soft materials such as are used in the cores of transformers and reproducing heads is very sensitive to strain. Even the light cold working which a surface receives in being ground flat is sufficient to impair very seriously the permeability of a thin surface layer.

In view of this it is to be expected that the magnetic contact between reproducing head and medium is less than perfect. If cold working during the fabrication of the head or due to abrasion by the recording medium should result in an effective air space between head and medium amounting to as much as one mil, the effect on frequency response would be pronounced indeed. At a recording speed of 7.5 in./sec. this amount of spacing would cause a loss of 7.3 db at 1000 cps, 14.6 db at 2000 cps, 21.9 db at 3000 cps, 29.2 db at 4000 cps, etc.

It seems certain that in a practical recording system some loss of this sort must occur. The problem of determining the magnitude of the loss or in other words the amount of the effective spacing in a practical case is, however, a difficult one. So far, no direct experimental method for its determination has been found.

Theoretical Calculations for an Idealized Case

In the preceding section an experimentally determined spacing loss function has been discussed. It was shown that as the reproducing head is moved away from the recording medium the reproduced signal level decreases. This means that the magnetic flux through the head decreases. If the distribution of magnetization in the recording medium were known, it should be possible to compute the flux through the head and thereby to derive the spacing loss function on a theoretical basis. Unfortunately it seems almost impossible to do this calculation in an exact way because very little is known about the magnetization pattern in the medium and because the geometry of the usual ring type head makes the boundary value problem an exceedingly difficult one to solve.

It is possible, however, to obtain a solution for an idealized case which bears at least some resemblance to the practical situation and this solution will be presented. The results must, of course, be viewed with due skepticism until they can be proved experimentally or else recalculated on the basis of better initial assumptions. It is hoped, however, that in some

measure they may serve as a guide to a better understanding of the magnetic reproducing process.

The Idealized Recording Medium

The problem will be reduced to two dimensions by assuming an infinitely wide and infinitely long tape of finite thickness δ. A rectangular coordinate system will be chosen in such a way that the central plane midway between the upper and lower surfaces of the recording medium lies in the x–y plane. It will be assumed that the medium is sinusoidally magnetized in such a way that in the medium the intensity of magnetization is given by

$$I_x = I_m \sin (2\pi x/\lambda)$$
$$I_y = I_z = 0. \tag{2}$$

Equations (2) say that the recording is purely longitudinal. In a practical case, of course, the recorded signal is neither purely longitudinal nor purely perpendicular but rather contains components of both sorts. In Appendix I it is shown that the frequency response does not depend on the relative amounts of these two components and hence that the computed results are equally valid whether the recorded signal is purely longitudinal, purely perpendicular, or a mixture of the two.

Appendix II contains calculations for the case of a round wire sinusoidally magnetized along its axis, and for a plated wire. These results, though much different in mathematical form, are shown to be very similar to the results for a flat medium.

The Idealized Reproducing Head

Figure 4 shows a semi-practical version of the sort of idealized reproducing head which will be treated.

It consists of a bar of core material with a single turn of exceedingly fine wire around it. This head is imagined to be spaced d inches above the surface of the recording medium. If the dimensions of the bar are made large enough, the amount of flux through it will obviously be as great as could be made to pass through any sort of head which makes contact with only one side of the tape and so the open circuit reproduced voltage per turn is as high as can be obtained with any practical head.

Suppose a very narrow gap is introduced in this head where the single turn coil was and that the magnetic circuit is completed by a ring of core material as shown in Fig. 5.

If the permeability of the head is very high and the gap very small then the flux which passed through the single turn coil of Fig. 4 will now pass

through the ring of Fig. 5 and can be made to thread through a coil of many turns wound on the ring. In so far as this is true, calculations based on this bar type head are applicable to ring type heads.

If the bar of Fig. 4 is now allowed to become infinite in length, width, and thickness, the flux density in it can be computed and the flux per unit width can be evaluated. This calculation is outlined in Appendix I. If the tape moves past the head with a velocity v in the x direction, the repro-

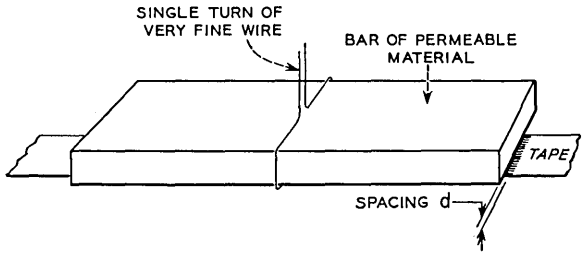


Fig. 4—Idealized bar-type reproducing head.

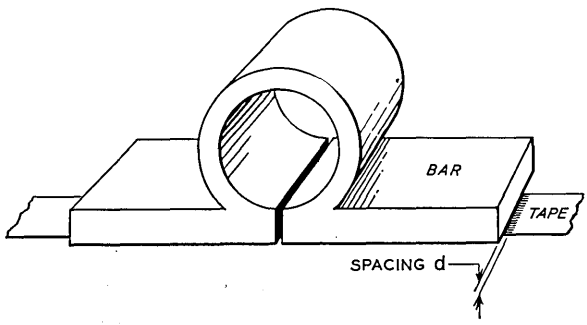


Fig. 5—Idealized ring-type reproducing head.

duced voltage should be proportional to the rate of change of flux. In the appendix this is shown to be

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi W v I_m (1 - e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{3}$$

where $\frac{d\phi_x}{dt}$ is the rate of change of flux in W cm. width of the reproducing head measured in Maxwells per sec.,

μ is the permeability of the reproducing head,

W is the width in cm. of the reproducing head (and of the recorded track in a practical case),

v is the velocity in cm./sec. with which the recording medium passes the *reproducing* head.

I_m is the peak value of the sinusoidal intensity of magnetization in the recording medium measured in gauss,

δ is the thickness of the recording medium measured in the same units as λ,

λ is the recorded wavelength measured in any convenient units,

d is the effective spacing between the reproducing head and the surface of the recording medium measured in the same units as λ, and

ω is 2π times the reproduced frequency in cycles per sec.

Note that equation (3) applies to a ring type head with no back gap. If the head has a back gap then not all the available flux will thread through the ring. Some of it will return to the medium through the scanning gap and hence will not contribute to the reproduced voltage. This does not affect the shape of the frequency response curve but does contribute a constant multiplying factor (less than unity) to the right hand side of equation (3). The value of this factor depends on the reluctances of the gaps and of the magnetic parts of the reproducing head. If the reluctance of the magnetic parts is negligible and the reluctance of the back gap is equal to the reluctance of the front gap then the available flux will divide equally in the two gaps and the factor will be one-half. This factor will not be considered further in this paper because it does not contribute to the shape of the response curve but only to the absolute magnitude of the reproduced voltage. It could be interpreted as reducing the effective number of turns on the reproducing head to a value somewhat lower than the actual number of turns.

Spacing Loss

The term $e^{-2\pi d/\lambda}$ tells how the reproduced voltage depends on spacing. In order to compare this computed effect with the experimentally observed one it is necessary to put it in decibel form by computing twenty times the $\log_{10}$ of $e^{-2\pi d/\lambda}$. This gives

$$\text{Spacing Loss} = 54.6\ (d/\lambda)\ \text{decibels}.$$

This agrees very well indeed with the experimentally determined equation (1) in which the constant is 55 instead of the computed 54.6. The computed spacing loss function is plotted in Fig. 6.

Thickness Loss

The effect of the thickness of the recording medium shows up in the term $(1 - e^{-2\pi\delta/\lambda})$. At low frequencies for which the wavelength is much greater than the thickness of the medium this reduces to $2\pi\delta/\lambda$. In this case the reproduced voltage is proportional to the thickness of the medium and to frequency. This is the familiar six db per octave characteristic.

At high frequencies, however, when $\lambda \ll \delta$ the term reduces to unity

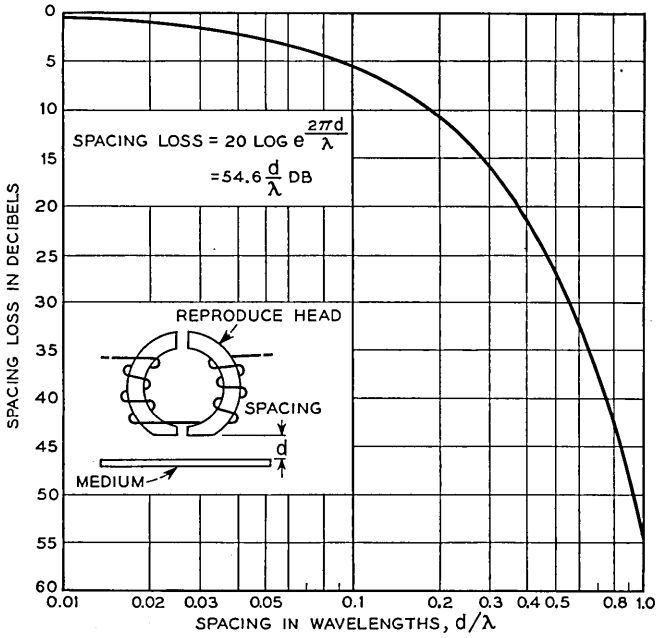


Fig. 6—Computed spacing loss as a function of d/λ.

and the computed "ideal" response is flat with frequency and independent of the thickness of the medium.

If the term $(1 - e^{-2\pi\delta/\lambda})$ is rewritten as

$$(2\pi\delta/\lambda)\left[\frac{1 - e^{-2\pi\delta/\lambda}}{2\pi\delta/\lambda}\right]$$

then the part in parenthesis accounts for a 6 db per octave characteristic and the part in brackets accounts for a loss *with respect to this 6 db per octave characteristic.* This loss, which will be called Thickness Loss[4], is given by

[4] It seems somewhat awkward to speak of "Thickness Loss" when nothing is actually lost by making the medium thick. The only excuse for this way of splitting the terms is that it makes for ease in comparing measured and computed curves.

$$\text{Thickness Loss} = 20 \log_{10} \frac{2\pi\delta/\lambda}{1 - e^{-2\pi\delta/\lambda}} \text{ db} \tag{5}$$

where λ is the recorded wavelength and δ is the thickness of the recording medium. This function is plotted in Fig. 7.

Comparison with Experiment

The most elementary consideration of the magnetic recording process indicates that when the recording signal current is held constant the open circuit reproduced voltage should be a function of frequency, increasing by 6 db for each octave increase in frequency. Experimental response curves tend to show this 6 db per octave characteristic when the recorded wave-

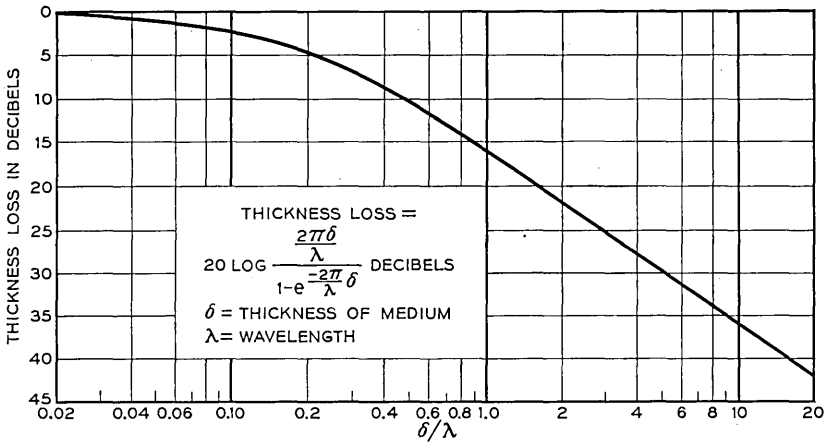


Fig. 7—Computed thickness loss as a function of δ/λ.

length is moderately long and the frequency moderately low. This makes it possible to draw a 6 db per octave line on the measured response characteristic in such a way as to coincide with the low-frequency part of the measured response characteristic. As the frequency is increased the measured curve tends to fall more and more below the 6 db per octave line. This is because several kinds of loss come into play as the wavelength decreases or as the frequency increases. Among these losses are:

1. Self demagnetization,
2. Eddy current and other losses in the recording and reproducing heads, and
3. Gap loss due to the finite scanning slit in the reproducing head.

The work presented in the first sections of this paper indicates that the

following two kinds of loss should be added to this list:

4. Spacing loss due to imperfect magnetic contact between the reproducing head and the recording medium, and
5. Thickness loss.

Of these five losses three can be evaluated quantitatively either by direct measurement or by calculation from theory. The remaining two are self-demagnetization and spacing loss.

In this section the known losses will be evaluated for a particular recording system. This leads to a response curve which can be compared with the measured curve. The difference between the two curves should be due to self-demagnetization and to spacing loss provided the above list of losses is complete.

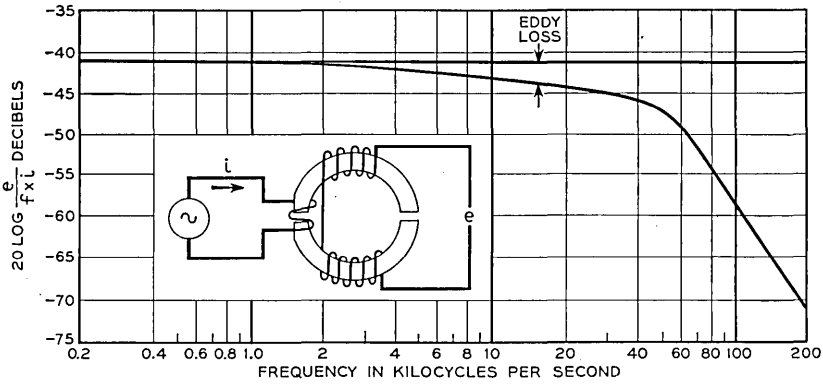


Fig. 8—Measured eddy current loss as a function of frequency.

The recording system used is the one shown in Fig. 1 with the speed set at 15.5 in./sec. for both recording and reproducing. A constant signal current of 0.1 ma was used for recording with the 55 kc bias adjusted to give maximum open circuit reproduced voltage.

Eddy current losses were measured as indicated in Fig. 8 by sending a measured constant current i through a small auxiliary winding around the pole tip and measuring the open circuit voltage developed across the normal winding of the head. Any departure of this measured voltage from a 6 db per octave increase with increasing frequency is due to losses in the head which will be loosely called eddy current losses. Other kinds of loss may enter into this measurement (as, for example, loss due to the self-capacitance of the winding) but in the frequency range of interest, eddy losses predominate.

By a completely different sort of measurement,[5] J. R. Anderson has arrived at a similar value for eddy current loss in this type of head and has shown that approximately the same loss occurs in both the recording and the reproducing process. For this reason it seems proper to assume that eddy currents account for just twice the loss measured by the method of Fig. 8.

The loss due to the finite gap in the reproducing head is computed from the well known relation.[6]

$$\text{Gap loss} = 20 \log_{10} \frac{\pi g/\lambda}{\sin (\pi g/\lambda)}$$

where g is the effective gap width in inches and λ is the recorded wavelength in inches.

Thickness loss is computed from equation (5). It must be remembered that this loss was derived on the assumption of uniform magnetization throughout the thickness of the recording medium. This may be a fairly good approximation to the actual state of affairs for a thin medium such as the one being considered, but obviously if the thickness of the medium is large compared with the width of the recording gap then the recording field will not penetrate uniformly through the medium and the derived thickness loss function will not apply.

The derived equation (3) indicates that at low frequencies the reproduced voltage should be proportional to the thickness of the medium. If the thickness of the medium is increased beyond the limit to which the recording field can penetrate, this will no longer be the case and further increase in thickness will have no effect on the response.

Data presented by Kornei[3] on the cobalt-nickel plating being considered here shows that the low-frequency response is approximately proportional to the thickness of the medium for values of thickness between 0.075 mil and 0.5 mil. This may be taken as an indication of approximately uniform penetration through these thicknesses and hence tends to indicate that the derived thickness loss function should be applicable in the case of the 0.3 mil plating being considered here.

The effects of these losses are shown in Fig. 9 along with measured frequency response data. Consider first the experimentally measured response

[5] In unpublished work, J. R. Anderson of the Bell Telephone Laboratories has made use of the fact that eddy losses depend on frequency while all other magnetic recording losses depend on wavelength. By recording a single frequency and playing back at various speeds he determined the loss on playback. By recording various frequencies with recording speed adjusted to give constant recorded wavelength and using a single playback speed he evaluates the eddy loss in the recording process.

[6] S. J. Begun, "Magnetic Recording," p. 84, Murray Hill Books, Inc., New York.

data shown as circles falling near the lowest curve. Some of the measured points have been omitted to avoid crowding but enough remain to show the trend. At low frequencies these points fall along a line of approximately 6 db per octave.

A straight 6 db per octave line labeled 1 has been drawn through these points and extended as shown in the figure. This line is the base from which the various losses must be subtracted. Curve 2 shows the effect of subtracting the computed thickness loss. When eddy losses and gap loss are

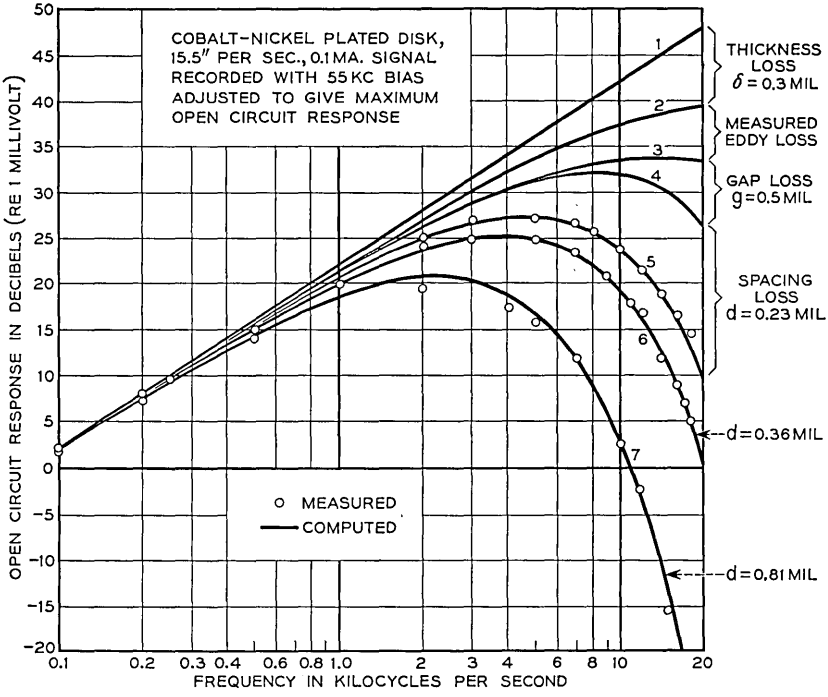


Fig. 9—Computed response curves and measured response points.

also taken into account, curve 4 is obtained. The difference between this curve and the lowest measured response points is presumably due either to self-demagnetization, to spacing loss, or perhaps to both.

There is one clue which may be of help in deciding how much of this loss should be attributed to self-demagnetization and how much to spacing loss. This clue comes from the fact that the form of the spacing loss function is known. Any part of the loss which is due to spacing must follow the equation

$$\text{Spacing Loss} = 54.6\,(d/\lambda)\ \text{db}$$

whereas there is reason to believe that the effects of self-demagnetization cannot possibly account for more than something like ten or fifteen db loss and hence could not follow the equation given above.

In view of this it seems reasonable to try as a first guess the assumption that all the unexplained loss is due to spacing.

If this assumption properly accounts for the shape of the measured response curve there will be at least some reason to suppose it may be correct; particularly so if the required amount of effective spacing seems reasonable.

The lowest solid curve, No. 7 of Fig. 9, has been computed on this basis. That is, a spacing loss corresponding to 0.81 mil effective spacing has been subtracted from curve 4. It is seen that this computed response curve fits reasonably well with the measured points. Furthermore, 0.81 mil effective spacing corresponds to quite reasonably good magnetic contact.

If this interpretation of the measured data is correct then it is obvious that the high-frequency response could be improved a great deal if more intimate magnetic contact between the reproducing head and the recording medium could be achieved. To this end an attempt was made to lap the surface of the head in such a way as to remove material very gently and slowly. After lapping, the response was appreciably improved as indicated by the set of measured points around curve 6. This curve was computed assuming an effective spacing of 0.36 mil. Note that the computed curve now fits the measured points very well indeed.

After still more lapping,[7] the measured response points around curve 5 were obtained. In this case it is necessary to assume only 0.23 mil effective spacing in order to account for the measured curve. Further lapping failed to give further improvement in response but a defect in the head which may account for this has since been found and it is believed that with great care one might actually measure something very close to curve 4.

To summarize, this is what seems to have been found. It is possible to compute a response curve taking into account gap loss, eddy current losses, and thickness loss. If this curve is compared with the final measured response curve it is found that the measured curve gives less high-frequency response than was computed. The difference between the two curves is just the right sort of function of frequency and of just the right magnitude to be accounted for by an effective spacing of 0.00023 inch between the reproducing head and the recording medium. It seems probable that the effective spacing could not have been much smaller than this value and therefore it may be correct to assume that practically all the unexplained

[7] After each lapping it was found that smaller values of bias current sufficed to give maximum reproduced voltage. This is presumably because the improved magnetic contact made the bias current more effective.

high-frequency loss is due to spacing. This would imply that, under the conditions of this experiment, self-demagnetization has a negligible effect on frequency response.

In any case it seems clear that the intimacy of magnetic contact between the reproducing head and the medium can have a very pronounced effect on high-frequency response. The condition of the surface of the reproducing head (and of the tape) may have more effect on high-frequency response than any other single factor.

In a very fine piece of pioneering work Lübeck[8] found empirically that a term of the form $e^{-\lambda_1/\lambda}$ was needed to account for the shape of measured response curves. Guckenburg[9] has recently written more on this subject. Both authors have assumed that this term has to do with self-demagnetization and that λ_1 is determined by the magnetic properties of the recording medium. The experiment just discussed and the theory presented in this paper suggest, on the other hand, that λ_1 is not a function of the magnetic properties of the recording medium but rather is determined by the intimacy of magnetic contact between reproducing head and medium. If this is the case then Lubeck's λ_1 is related to the d of this paper through the equation $\lambda_1 = 2\pi d$.

Guckenburg reports $\lambda_1 = 100\mu$ for the best available medium. This corresponds to $d = 0.625$ mil and yields a response curve a little better than the poorest measured curve of Fig. 9. The best measured curve of Fig. 9. corresponds to $\lambda_1 = 37\mu$.

ACKNOWLEDGEMENTS

The author wishes to express his appreciation for the encouragement and guidance of Mr. R. K. Potter, Mr. J. C. Steinberg, and Mr. W. E. Kock.

APPENDIX I

THE FIELD DUE TO A FLAT SINUSOIDALLY MAGNETIZED MEDIUM

The Field in Free Space

It is convenient to begin by evaluating the field inside and outside the recording medium when the medium is in free space. By making use of the

[8] H. Lübeck, "Magnetische Schallaufzeichnung mit Filmen und Ringkopfen," *Akustiche Zeit.*, 2, 273 (1937).

[9] W. Guckenburg, "Die Wechselbeziehungen zwichen Magnettonband und Ringkipf bie der Wiedergabe," *Funk und Ton*, 4, 24 (1950).

method if images, this solution can be used to find the fields which exist when the medium is under an idealized reproducing head of permeability μ. Also, the free space solution may be of use in evaluating the effect of self demagnetization since the demagnetizing field is computed.

Let the recording medium be an infinite plane sheet of thickness δ and choose rectangular coordinates so that the central plane of the medium lies in the x–y plane as shown in Fig. 10.

Let the permeability of the recording medium be unity and let the intensity of magnetization inside it be given by

$$\begin{aligned} I_x &= I_m \sin (2\pi x/\lambda) \\ I_y &= I_z = 0 \end{aligned} \tag{6}$$

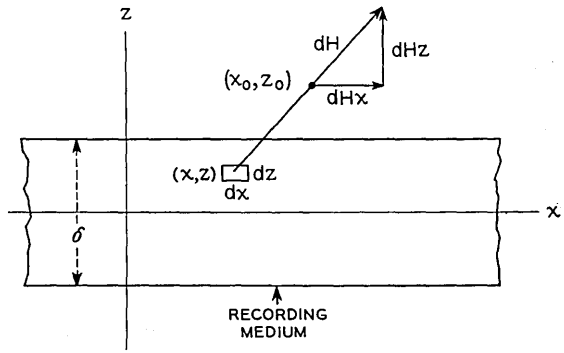


Fig. 10—Coordinate system for flat tape calculations.

This is equivalent to a volume density of "magnetic charge" given by

$$\begin{aligned} \rho &= -\operatorname{div} I \\ &= -\frac{dI_x}{dx} \\ &= -(2\pi I_m/\lambda) \cos (2\pi x/\lambda) \end{aligned} \tag{7}$$

The problem then is to compute the field at a point (x_0, z_0) due to this charge. Consider the field at (x_0, z_0) due to the element $dx\ dz$ at (x, z). This element amounts to an infinitely long line of uniform charge density. The field due to such a distribution is directed perpendicular to the line and has a magnitude equal to twice the linear charge density divided by the distance from the point to the line.

In the present case this leads to

$$\begin{aligned} dH_x &= -(4\pi I_m/\lambda)\,\frac{(x_0 - x)}{(x_0 - x)^2 + (z_0 - z)^2}\cos(2\pi x/\lambda)\,dx\,dz \\ dH_z &= -(4\pi I_m/\lambda)\,\frac{(z_0 - z)}{(x_0 - x)^2 + (z_0 - z)^2}\cos(2\pi x/\lambda)\,dx\,dz \end{aligned} \tag{8}$$

The total field at (x_0, z_0) is obtained by integrating with respect to x over the range $-\infty$ to $+\infty$ and with respect to z over the range $-\delta/2$ to $+\delta/2$. In carrying out the integration over x it is convenient to make the substitution

$$\begin{aligned} (x_0 - x)/(z_0 - z) &= p \\ dx &= -(z_0 - z)\,dp \end{aligned} \tag{9}$$

Neglecting terms which obviously integrate to zero, this gives

$$\begin{aligned} H_x &= (4\pi I_m/\lambda)\sin(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2}\left[\int_{\infty}^{-\infty}\frac{p\sin[2\pi(z_0 - z)p/\lambda]}{1 + p^2}\,dp\right]dz \\ H_z &= (4\pi I_m/\lambda)\cos(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2}\left[\int_{\infty}^{-\infty}\frac{\cos[2\pi(z_0 - z)p/\lambda]}{1 + p^2}\,dp\right]dz \\ & z_0 \geq z \end{aligned} \tag{10}$$

The integrals in brackets can be found in tables.[10] Carrying out the integration gives

$$\begin{aligned} H_x &= -(4\pi^2 I_m/\lambda)\sin(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2} e^{-2\pi(z_0 - z)/\lambda}\,dz \\ H_z &= -(4\pi^2 I_m/\lambda)\cos(2\pi x_0/\lambda)\int_{-\delta/2}^{\delta/2} e^{-2\pi(z_0 - z)/\lambda}\,dz \\ & z_0 \geq z \end{aligned} \tag{11}$$

which integrate to

$$\begin{aligned} H_x &= -2\pi I_m \sin(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}[e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ H_z &= -2\pi I_m \cos(2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}[e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ & z_0 \geq \delta/2 \end{aligned} \tag{12}$$

[10] D. Bierens de Haan, "Nouvelles Tables D'Integrales Définies," p. 223, Leide, Engels, 1867.

Below the recording medium, that is for $z_0 \leq -\delta/2$,

$$\begin{aligned} H_x &= -2\pi I_m \sin(2\pi x_0/\lambda) e^{+2\pi z_0/\lambda} [e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ H_z &= 2\pi I_m \cos(2\pi x_0/\lambda) e^{+2\pi z_0/\lambda} [e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}] \\ & z_0 \leq -\delta/2 \end{aligned} \tag{13}$$

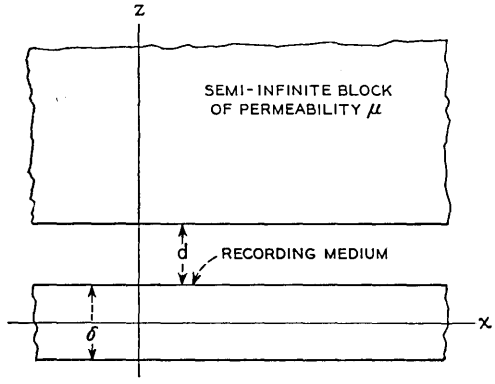


Fig. 11—Flat tape under idealized reproducing head.

Inside the recording medium,

$$\begin{aligned} H_x &= -(4\pi^2 I_m/\lambda) \sin(2\pi x_0/\lambda) \\ &\quad \cdot \left[\int_{-\delta/2}^{z_0} e^{-2\pi(z_0-z)/\lambda}\, dz + \int_{z_0}^{\delta/2} e^{+2\pi(z_0-z)/\lambda}\, dz \right] \\ H_z &= -(4\pi^2 I_m/\lambda) \cos(2\pi x_0/\lambda) \\ &\quad \cdot \left[\int_{-\delta/2}^{z_0} e^{-2\pi(z_0-z)/\lambda}\, dz - \int_{z_0}^{\delta/2} e^{+2\pi(z_0-z)/\lambda}\, dz \right] \end{aligned} \tag{14}$$

which integrate to

$$\begin{aligned} H_x &= -2\pi I_m \sin(2\pi x_0/\lambda)[2 - e^{-\pi\delta/\lambda}(e^{-2\pi z_0/\lambda} + e^{2\pi z_0/\lambda})] \\ H_z &= 2\pi I_m \cos(2\pi x_0/\lambda) e^{-\pi\delta/\lambda}(e^{-2\pi z_0/\lambda} - e^{2\pi z_0/\lambda}) \\ & \delta/2 \geq z_0 \geq -\delta/2 \end{aligned} \tag{15}$$

The Fields in and Under the Reproducing Head

The idealized reproducing head amounts simply to a semi-infinite block of high permeability material with a flat face spaced a distance d above the surface of the recording medium as shown in Fig. 11.

The problem of most interest is that of finding the x component of magnetic induction, B_x, at any point (x_0, z_0) in the idealized head and integrating this with respect to z_0 to determine the total flux passing through unit width (in the y direction) of a plane $x = x_0$. This plane will then be allowed to move with a velocity v by putting $x_0 = vt$ and the time rate of change of flux will be computed. Except for the effects of eddy currents, self demagnetization, gap loss, etc. (which are treated separately) this rate of change of flux should be proportional to the open circuit reproduced voltage. This is the only result of which direct use will be made but for the sake of completeness all the field components will be evaluated not only in the idealized head but also at all other points.

This problem is completely analogous to the problem of a point charge in front of a semi-infinite dielectric treated by Abraham and Becker[11] and can be solved by use of the method of images.

The Field Inside the High Permeability Head

By analogy with the treatment of Abraham and Becker, the value of B in the high permeability head is computed as though this head filled all space and as though the recording medium were polarized to a value $2\mu/(\mu + 1)$ times the actual value of polarization present. This gives directly from equations (12),

$$\begin{aligned} B_x &= -[2\mu/(\mu + 1)]2\pi I_m \sin (2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}(e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ B_z &= -[2\mu/(\mu + 1)]2\pi I_m \cos (2\pi x_0/\lambda)e^{-2\pi z_0/\lambda}(e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ & z_0 \geq d + \delta/2 \end{aligned} \tag{16}$$

The Field Below the Reproducing Head

Again by analogy with the treatment of Abraham and Becker, the field outside the idealized head is computed as though no head were present. The field is that due to the actual magnetized medium plus the field due to an image of the medium (centered about $z = 2d + \delta$). The intensity of magnetization of the image medium is $-(\mu - 1)/(\mu + 1)$ times the intensity of magnetization of the actual medium.

The field due to the image medium is computed from equations (13) after suitable modification. The required modifications are:

1. Multiply the right hand sides by $-(\mu - 1)/(\mu + 1)$ to take account of the magnitude and sign of the image magnetization as just discussed, and
2. Replace z_0 by $z_0 - (2d + \delta)$ to take account of the position of the image.

[11] M. Abraham and R. Becker, *The Classical Theory of Electricity and Magnetism.* p. 77, Blackie and Son Limited, London, 1937.

This gives the field due to the image plane as

$$\begin{aligned} H_{xi} &= 2\pi I_m \frac{\mu - 1}{\mu + 1} \sin (2\pi x_0/\lambda) e^{2\pi(z_0 - 2d - \delta)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ H_{zi} &= -2\pi I_m \frac{\mu - 1}{\mu + 1} \cos (2\pi x_0/\lambda) e^{2\pi(z_0 - 2d - \delta)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ & \qquad z_0 \leq d + \delta/2 \end{aligned} \tag{17}$$

To this must be added the field due to the real medium which is given by equations (12) when $\delta/2 \leq z_0 \leq d + \delta/2$, by equations (15) when $-\delta/2 \leq z_0 \leq \delta/2$, and by equations (13) when $z_0 \leq -\delta/2$.

Performing this addition gives the following results:

Between the head and the recording medium,

$$\begin{aligned} H_x &= -2\pi I_m \sin (2\pi x_0/\lambda) e^{-2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ &\qquad \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - 2z_0)/\lambda}\right] \\ H_z &= -2\pi I_m \cos (2\pi x_0/\lambda) e^{-2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ &\qquad \cdot \left[1 + \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - 2z_0)/\lambda}\right] \\ &\qquad d + \delta/2 \geq z_0 \geq \delta/2 \end{aligned} \tag{18}$$

Inside the recording medium,

$$\begin{aligned} H_x &= -2\pi I_m \sin (2\pi x_0/\lambda) \\ &\cdot \left[2 - e^{-\pi\delta/\lambda}(e^{2\pi z_0/\lambda} + e^{-2\pi z_0/\lambda}) - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - z_0)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda})\right] \\ H_z &= -2\pi I_m \cos (2\pi x_0/\lambda) \\ &\cdot \left[e^{-\pi\delta/\lambda}(e^{2\pi z_0/\lambda} - e^{-2\pi z_0/\lambda}) + \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta - z_0)/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda})\right] \\ &\qquad \delta/2 \geq z_0 \geq \delta/2 \end{aligned} \tag{19}$$

Below the recording medium,

$$\begin{aligned} H_x &= -2\pi I_m \sin (2\pi x_0/\lambda) e^{2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ &\qquad \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta)/\lambda}\right] \\ H_z &= 2\pi I_m \cos (2\pi x_0/\lambda) e^{2\pi z_0/\lambda} (e^{\pi\delta/\lambda} - e^{-\pi\delta/\lambda}) \\ &\qquad \cdot \left[1 - \frac{\mu - 1}{\mu + 1} e^{-2\pi(2d + \delta)/\lambda}\right] \\ &\qquad z_0 \leq -\delta/2 \end{aligned} \tag{20}$$

The Flux per Unit Width in the Idealized Reproducing Head

The desired flux per unit width is computed from

$$\phi_x = \int_{d+\delta/2}^{\infty} B_x \, dz \tag{21}$$

where B_x is given by equation (16). Performing the indicated integration gives

$$\phi_x = -\frac{2\mu}{\mu+1} 2\pi\delta I_m \sin(2\pi x_0/\lambda) \left[\frac{1-e^{-2\pi\delta/\lambda}}{2\pi\delta/\lambda}\right] e^{-2\pi d/\lambda} \tag{22}$$

If the reproducing head moves past the recording medium with a velocity v so that $x_0 = vt$,

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu+1} 4\pi v I_m (1-e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{23}$$

where ω is 2π times the reproduced frequency. This is the result for unit width of the reproducing head. For a width of W cm.,

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu+1} 4\pi W v I_m (1-e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{24}$$

The Case of Perpendicular Magnetization

Equation (23) was derived for the case of pure longitudinal magnetization as defined by equations (6). It will now be shown that this same result is obtained for $d\phi_x/dt$ if the magnetization is purely perpendicular, that is if

$$\begin{aligned} I_z &= -I_m \cos(2\pi x/\lambda) \\ I_x &= I_y = 0 \end{aligned} \tag{25}$$

In this case the divergence of I is zero except at the surface of the tape and this magnetization is equivalent to a surface distribution of magnetic charge on the top and bottom surfaces of the tape. The magnitude of this charge density is just equal to I_z so that on the top surface of the tape there is a surface density of charge given by

$$\sigma = -I_m \cos(2\pi x/\lambda) \qquad \text{at } z = \delta/2 \tag{26}$$

and on the bottom surface of the tape there is a surface density of charge given by

$$\sigma = I_m \cos(2\pi x/\lambda) \qquad \text{at } z = -\delta/2 \tag{27}$$

Since the permeability of the recording medium is assumed to be unity, this problem reduces to that of finding $d\phi_x/dt$ due to two infinitely thin

tapes of the sort to which equation (23) applies. One of these tapes is at $z = \delta/2$ and the other at $z = -\delta/2$.

The problem then is to rewrite equation (23) for a very thin tape and in terms of surface density of charge. As δ approaches zero, equation (23) reduces to

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m (2\pi\delta/\lambda) e^{-2\pi d/\lambda} \cos(\omega t) \tag{28}$$

From equation (7), the volume density of charge in this tape is

$$\rho = -(2\pi I_m/\lambda) \cos(2\pi x/\lambda)$$

But as δ approaches zero, the longitudinally magnetized tape to which equation (28) applies becomes equivalent to a surface distribution of magnetic charge of surface density equal to $\delta\rho$. This amounts, for the thin longitudinally magnetized tape, to a surface charge density of

$$\sigma_1 = -(2\pi\delta/\lambda) \cos(2\pi x/\lambda) \tag{29}$$

But the charge density on the top side of the perpendicularly magnetized tape is given by equation (26). Comparing these two values shows that the surface charge density in the thin longitudinally magnetized tape is just $2\pi\delta/\lambda$ times as great as the surface charge density on top of the perpendicularly magnetized medium. This means that $d\phi_x/dt$ due to the top side of the perpendicularly magnetized tape can be obtained by dividing the right hand side of equation (28) by $2\pi\delta/\lambda$. This gives

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m e^{-2\pi d/\lambda} \cos(\omega t) \tag{30}$$

due to the top side of the tape.

The contribution from the bottom side is obtained from equation (30) by replacing d by $d + \delta$ (since the bottom side is spaced $d + \delta$ from the reproducing head) and changing the sign. Adding these two contributions gives for the total

$$\frac{d\phi_x}{dt} = -\frac{\mu}{\mu + 1} 4\pi v I_m (1 - e^{-2\pi\delta/\lambda}) e^{-2\pi d/\lambda} \cos(\omega t) \tag{31}$$

This is the same as equation (23) and so the desired result has been established.

Note from equations (6) and (24) that in order to get the same result for the perpendicular and longitudinal cases it was necessary to assume a 90-degree phase difference between I_x and I_z. The usual type of recording head lays down a pattern of magnetization which is neither purely per-

pendicular nor purely longitudinal but the two components are always in phase. This means that the two contributions to $d\phi_x/dt$ add as vectors at 90 degrees. If the intensity of magnetization in the recording medium is held constant while the relative values of perpendicular and longitudinal components are changed, the only effect on the reproduced signal is a change of phase.

APPENDIX II

THE FIELD DUE TO A ROUND WIRE

In Appendix I the field due to a sinusoidally magnetized flat medium such as a tape has been calculated and the rate of change of flux in an

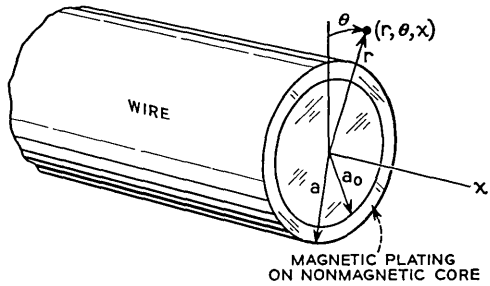


Fig. 12—Coordinate system for round wire calculations.

idealized reproducing head has been evaluated. The analogous calculations for a round wire have also been carried through and it is the purpose of this section to present some of the results. The derivation of these results seems too tedious and long to be presented here.

The Recording Medium

Let the recording medium be a wire, the axis of which lies along the x axis as shown in Fig. 12. Let the radius of the wire be a. To take account of plated wires as well as solid magnetic ones, let the wire have a nonmagnetic core of radius a_0. Let the cylindrical shell between a_0 and a be magnetized sinusoidally in the x direction so that

$$\begin{aligned} I_x &= I_m \operatorname{Sin}(2\pi x/\lambda) \\ I_r &= I_\theta = 0 \end{aligned} \tag{32}$$

By putting $a_0 = 0$ in the expressions which follow it will be possible to obtain the result for a solid magnetic wire.

The Field in Free Space

If no reproducing head is present to disturb the field distribution, the computed field components at a point (x_0, r_0) are

$$\begin{aligned} H_x &= -4\pi I_m \operatorname{Sin}(2\pi x_0/\lambda) K_0(2\pi r_0/\lambda)[(2\pi a/\lambda) I_1(2\pi a/\lambda) \\ &\qquad - (2\pi a_0/\lambda) I_1(2\pi a_0/\lambda)] \\ H_r &= -4\pi I_m \operatorname{Cos}(2\pi x_0/\lambda) K_1(2\pi r_0/\lambda)[(2\pi a/\lambda) I_1(2\pi a/\lambda) \\ &\qquad - (2\pi a_0/\lambda) I_1(2\pi a_0/\lambda)] \\ r_0 &\geq a \end{aligned} \tag{33}$$

A discussion and tabulation of the I and K functions can be found in Watson's "Theory of Bessel Functions."[12]

The field due to a solid magnetic wire is obtained by setting $a_0 = 0$ in equations (33). This gives

$$\begin{aligned} H_x &= -4\pi I_m \operatorname{Sin}(2\pi x_0/\lambda)(2\pi a/\lambda) K_0(2\pi r_0/\lambda) I_1(2\pi a/\lambda) \\ H_r &= -4\pi I_m \operatorname{Cos}(2\pi x_0/\lambda)(2\pi a/\lambda) K_1(2\pi r_0/\lambda) I_1(2\pi a/\lambda) \\ r_0 &\geq a \end{aligned} \tag{34}$$

The Rate of Change of Flux in an Idealized Head

It has not been possible to carry out the calculations for an idealized head which is a satisfactory approximation to the grooved ring-type head often used in wire recording. The results presented below will apply only to reproducing heads which completely surround the wire. In this case the idealized head is an infinitely large block of core material of permeability μ pierced by a cylindrical hole of radius R in which the wire is centered as shown in Fig. 13. At any point (x_0, r_0) in the permeable medium the components of flux density can be shown to be

$$\begin{aligned} B_x &= \alpha H_x \\ B_r &= \alpha H_r \\ r_0 &\geq R \end{aligned} \tag{35}$$

where

$$\alpha = \frac{\mu}{(\mu - 1)(2\pi R/\lambda) I_0(2\pi R/\lambda) K_1(2\pi R/\lambda) + 1} \tag{36}$$

and H_x and H_r are given by equation (33).

[12] G. N. Watson, "A Treatise on the Theory of Bessel Functions," p. 79, 361, 698, Cambridge Univ. Press, 1922.

The total flux through a plane $x = x_0$ in the permeable medium is obtained by integrating

$$\phi_x = \int_R^\infty B_x(2\pi r)\, dr \tag{37}$$

This gives

$$\phi_x = -2\lambda^2\alpha I_m \sin(2\pi x_0/\lambda)(2\pi R/\lambda)K_1(2\pi R/\lambda)[(2\pi a/\lambda)I_1(2\pi a/\lambda) - (2\pi a_0/\lambda)I_1(2\pi a_0/\lambda)] \tag{38}$$

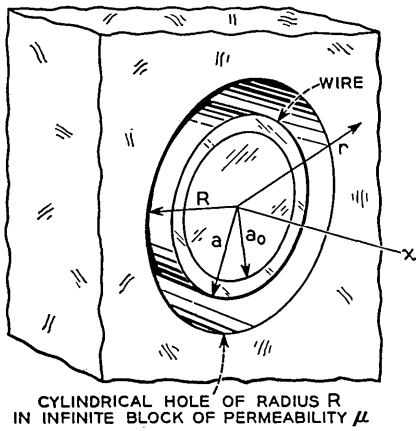


Fig. 13—Round wire surrounded by idealized reproducing head consisting of an infinite block of core material of permeability μ.

If the plane $x = x_0$ moves with a velocity v with respect to the wire so that $x_0 = vt$, then

$$\frac{d\phi_x}{dt} = -4\pi\lambda\alpha v I_m \cos(\omega t)(2\pi R/\lambda)K_1(2\pi R/\lambda)[(2\pi a/\lambda)I_1(2\pi a/\lambda) - (2\pi a_0/\lambda)I_1(2\pi a_0/\lambda)] \tag{39}$$

where $\omega = 2\pi f$ and f is the reproduced frequency.

Special Cases

Equation (39) can be used to compute the response of a simple reproducing head consisting of a single turn of very fine[13] wire as shown in Fig. 14.

In this case $\mu = 1$ and equation (36) shows that $\alpha = 1$. Furthermore if the wire is solid so that $a_0 = 0$, equation (39) reduces to

[13] Unless the diameter of the wire is small compared to the recorded wavelength there will be additional loss not accounted for by 39.

$$\frac{d\phi_x}{dt} = -4\pi\lambda v I_m \cos(\omega t)(2\pi R/\lambda)(2\pi a/\lambda)K_1(2\pi R/\lambda)I_1(2\pi a/\lambda) \quad (40)$$

As λ approaches infinity, $K_1(2\pi R/\lambda)$ approaches $\lambda/2\pi R$ and $I_1(2\pi a/\lambda)$ approaches $\pi a/\lambda$ so that, for very long wavelengths, equation (40) reduces to

$$\frac{d\phi_x}{dt} = -4\pi I_m v(2\pi/\lambda)(\pi a^2)\cos(\omega t) \quad (41)$$

This relation (which could have been derived in a much simpler manner) should be useful for the experimental determination of the intensity of magnetization, I_m.

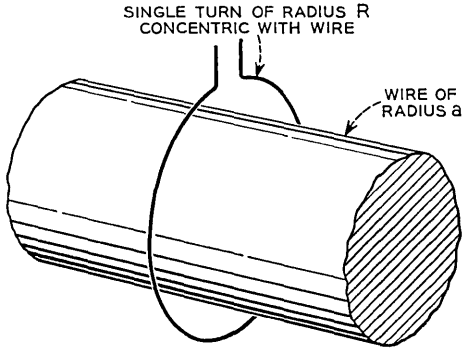


Fig. 14—Elementary reproducing head consisting of a single turn of wire.

Another case of some interest corresponds to a high permeability reproducing head which surrounds the wire. In this case μ is great enough so that equation (36) reduces to

$$\alpha = \frac{1}{(2\pi R/\lambda)I_0(2\pi R/\lambda)K_1(2\pi R/\lambda)} \quad (42)$$

If it is assumed, in addition, that the wire is solid so that $a_0 = 0$, then equations (42) and (39) give

$$\frac{d\phi_x}{dt} = -4\pi\lambda v I_m \cos(\omega t)(2\pi a/\lambda)I_1(2\pi a/\lambda)/I_0(2\pi R/\lambda) \quad (43)$$

Comparison between Round Wire and Flat Medium Response

It is interesting to compare equation (43) with equation (24) to see how the response characteristic of a round wire compares with that of a tape. Assuming $\mu \gg 1$, the appropriate equation for the flat medium is

$$\frac{d\phi_x}{dt} = -4\pi W v I_m \cos(\omega t)(1 - e^{-(2\pi\delta/\lambda)})e^{-2\pi d/\lambda} \quad (44)$$

To compare equations (43) and (44), consider first the limiting cases of very long and very short wavelength. As λ approaches infinity they reduce to

$$\frac{d\phi_x}{dt} = -\pi a^2(8\pi^2 v/\lambda)I_m \cos(\omega t) \tag{45}$$

for the wire and

$$\frac{d\phi_x}{dt} = -\delta W(8\pi^2 v/\lambda)I_m \cos(\omega t) \tag{46}$$

for the tape.

These two expressions are identical provided the cross section area of the wire, (πa^2), is the same as that of the recorded track on the tape, (δW).

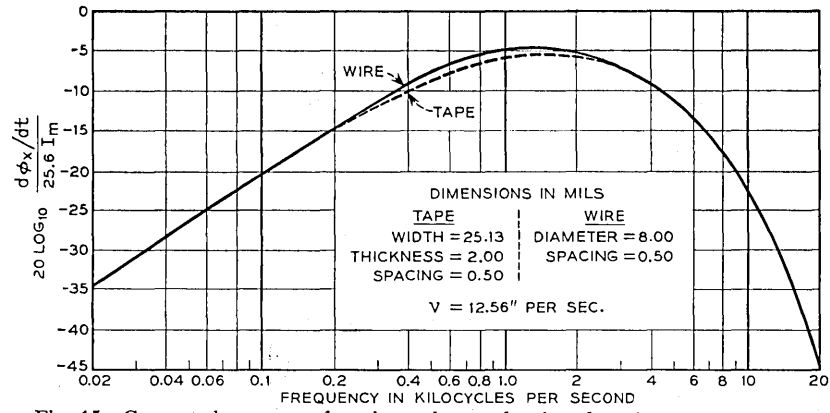


Fig. 15—Computed responses for wire and tape showing that the responses are very similar provided the dimensions of the wire and tape are suitably related.

As λ approaches zero, the two expressions reduce to

$$\frac{d\phi_x}{dt} = -4\pi v(2\pi a)\sqrt{R/a}\,e^{-2\pi(R-a)/\lambda}I_m \cos(\omega t) \tag{47}$$

for the wire, and

$$\frac{d\phi_x}{dt} = -4\pi v(W)e^{-2\pi d/\lambda}I_m \cos(\omega t) \tag{48}$$

for the tape.

Suppose that the reproducing head makes reasonably good contact with the wire so that $\sqrt{R/a} \doteq 1$. In this case equations (47) and (48) are identical provided the circumference of the wire, $(2\pi a)$, is the same as the width of the recorded track on the tape and provided also the effective spacing between reproducing head and medium is the same in the two cases, $(d = R - a)$. In both cases only a thin surface layer of the recording medium is effective in producing high frequency response. For this reason the

high-frequency response is independent of the "thickness" of the medium and is directly proportional to the "width" of the track provided $2\pi a$ is interpreted as the width of track on a wire.

The comparisons which have just been made indicate that if the dimensions of a wire and of a tape are suitably related, the two media should give identical response at very high and very low frequencies provided they are equally magnetized. The dimensional requirements are

$$\begin{aligned} \pi a^2 &= \delta W, \\ 2\pi a &= W, \text{ and} \\ R - a &= d \end{aligned} \tag{49}$$

In order to show how the computed responses compare at intermediate frequencies, numerical calculations have been made for a special case in which equations (49) are satisfied. The case chosen is that of a wire 8 mils in diameter moving at a velocity of 12.56 in./sec. past a reproducing head which is effectively one half mil out of contact with the wire ($R - a = 0.5(10)^{-3}$ in.). By equations (49) the corresponding flat medium is a tape which is 2 mils thick and 25.13 mils wide. The tape is assumed to be moving with a velocity of 12.56 in./sec. past a reproducing head which is also effectively one half mil out of contact ($d = 0.5(10)^{-3}$ in.). In this case the numerical constants in equations (43) and (44) are equal. That is,

$$8\pi^2 av = 4\pi W v = 25.6 \text{ cm.}^2/\text{sec.}$$

and the quantity to be computed and compared for the two cases is

$$20 \log_{10} \frac{d\phi_x/dt}{25.6 I_m}$$

The computed curves are shown in Fig. 15 from which it can be seen that they coincide at low and high frequencies as planned and that furthermore they differ by no more than 1.5 db in the middle range of frequencies.

As has been pointed out, equation (43) applies only to the unusual case in which the head completely surrounds the wire. The similarity of the two curves of Fig. 15, however, suggests a way of computing approximately the response to be expected when the wire head makes contact with only a part of the circumference of the wire. It suggests that the computation be carried out as though the wire were a flat medium of suitably chosen dimensions. In order to make the high frequency end come out right one would expect that W in equation (44) should be given a value equal to the length of the arc of contact between the wire and the head. To make the low frequency end come out right, δ must be given a value which makes the cross section area of the tape equal to that of the wire, i.e. such that $\delta W = \pi a^2$.

Some Results Concerning the Partial Differential Equations Describing the Flow of Holes and Electrons in Semiconductors

By R. C. Prim, III

(*Manuscript Received June 22, 1951*)

The subject equations are investigated with the aim of establishing some general properties of the flow fields which they describe, including the existence or non-existence of classes of exact solutions having certain formal properties. The results include a number of geometric characteristics of the vector fields involved, a suggestive reformulation of the partial differential equations restricting carrier concentration and electrostatic potential, and several classes of exact solutions involving arbitrary constants and/or functions. Of particular interest is a family of solutions in closed form for the steady-state, no-recombination case involving an arbitrary harmonic function in three dimensions.

TABLE OF CONTENTS

A. INTRODUCTION

THIS paper is concerned with the system of relations describing the flow of holes and electrons in the interior of a homogeneous semiconductor subject to the assumption of constant temperature, electrical neutrality, and constant difference in concentrations of ionized donor and acceptor centers. These relations are:

$$\operatorname{div} \overset{\circ}{\mid\mid}_p = -e\left[\mathcal{R} + \frac{\partial p}{\partial t}\right] \tag{1}$$

$$\operatorname{div} \overset{\circ}{\mid\mid}_n = e\left[\mathcal{R} + \frac{\partial n}{\partial t}\right] \tag{2}$$

$$\overset{\circ}{||}_p = -\mu_p e \left[p \operatorname{grad} \mathfrak{V} + \frac{kT}{e} \operatorname{grad} p \right] \tag{3}$$

$$\overset{\circ}{||}_n = -\mu_n e \left[n \operatorname{grad} \mathfrak{V} - \frac{kT}{e} \operatorname{grad} n \right] \tag{4}$$

$$n - p = n_0 - p_0 \equiv N \text{ (a constant)} \tag{5}$$

$$n, p \geq 0 \tag{6}$$

$$\overset{\circ}{||} = \overset{\circ}{||}_p + \overset{\circ}{||}_n \tag{7}$$

wherein

n: concentration of negative carriers (electrons)
p: concentration of positive carriers (holes)
n_0: thermal equilibrium value of n
p_0: thermal equilibrium value of p
$\overset{\circ}{||}_p$: hole current density vector
$\overset{\circ}{||}_n$: electron current density vector
$\overset{\circ}{||}$: total current density vector
t: time variable
e: magnitude of electronic charge
k: Boltzmann's constant
μ_p: hole mobility constant
μ_n: electron mobility constant
T: absolute temperature (assumed constant with time and uniform)
$\mathfrak{V}$: potential of electrical intensity field
$\mathfrak{R}$: electron-hole recombination rate function (will usually be regarded as depending on $p - p_0$ and $n - n_0$ or equivalent variables).

These relations have fundamental application to transistor electronics, photoelectric effects, and related phenomena. Detailed discussions of their physical bases will be found in References 1 and 3. In brief, (1) and (2) are conservation conditions for the positive and negative carriers; (3) and (4) express the dependence of the local current densities on the electrostatic potential gradient and on the carrier concentration gradients (i.e., on conduction and diffusion); (5) expresses the condition of electrical neutrality under the assumption of a constant difference in concentrations of ionized donor and acceptor centers; and (6) and (7) are self evident.

The present study is directed toward the discovery of (1) general properties of the flow fields inside semiconductors and (2) families of exact solutions to the flow equations. The approach to the latter objective is through

the "inverse method" which has proved very useful in the study of various non-linear partial differential equation systems in mechanics. In the inverse method, one proceeds by formal devices suggested by the equations under study to try to find families of solutions to the equations which involve arbitrary constants or, preferably, arbitrary functions. This is done without reference to any preconceived boundary value problems. After a pool of such families of solutions is available, it can be examined from the point of view of finding boundary value problems of interest consistent with any of the solutions in hand. The likelihood of finding solutions of interest in this way is of course greatly enhanced when the solutions involve arbitrary functions. Aside from providing solutions of some useful boundary value problems, the solutions found by the inverse method constitute a reference bank of non-trivial exact solutions against which to check numerical methods and approximation schemes (based, for example, on the assumption that a particular term can be neglected) for solving problems of more immediate practical interest.

J. Bardeen has demonstrated (in Reference 2) how the steady-state behavior of contact-semiconductor combinations can be explained on the basis of the characteristics of (1) the flow field inside the semiconductor and (2) those of the barrier layer at the contact. The present study is concerned in this connection only with the first of these influences. It provides, for example, a complete solution for the spherically symmetric flow field without recombination for arbitrary currents—a generalization of the zero-total current solution given by Bardeen. In the absence of surface recombination this spherically symmetric solution provides the hemispherically symmetric flow field in the neighborhood of a point contact on a plane surface and remote from other electrodes or surfaces. This spherically symmetric solution is contained as a particular case in a family of solutions involving an arbitrary harmonic function in three dimensions. Other choices of the harmonic function can be made to yield flow fields associated with numerous electrode configurations of immediate practical interest, for example that of the type-A transistor.

The objective of the present paper is to find (or establish the non-existence of) broad classes of solutions, and not to undertake detailed studies of any particular solutions. Such detailed studies of particular cases from the family of solutions mentioned above (and from other families found in this study) will form the subject matter of papers dealing with specific flow field configurations. However, in order to illustrate the interpretation of mathematical arbitrary constants in terms of basic physical parameters, the analysis of the spherically symmetric solution mentioned above is car-

ried up to the point of actual substitution of numerical values in the formulae.

Note: In the following, functions and constants described as "arbitrary" are to be considered as being subject nevertheless to the restrictions implied by (6). In any particular case it is an elementary matter to determine these restrictions and we shall not usually carry out this detail. Also, "arbitrary" functions are subject to appropriate differentiability conditions readily evident in any particular case.

B. Some Properties of the Current Density Vector Fields

Several interesting properties of the current density vector fields $\overset{\circ}{\mathfrak{U}}_p$, $\overset{\circ}{\mathfrak{U}}_n$, and $\overset{\circ}{\mathfrak{U}}$ are easily found from (3)–(5).

It is evident that (3) and (4) can be rewritten as

$$\overset{\circ}{\mathfrak{U}}_p = -e\mu_p\, p \operatorname{grad}\left(\mathcal{V} + \frac{kT}{e} \ln p\right) \tag{8}$$

and

$$\overset{\circ}{\mathfrak{U}}_n = -e\mu_n\, n \operatorname{grad}\left(\mathcal{V} - \frac{kT}{e} \ln n\right). \tag{9}$$

From (3), (4), and (7) we have

$$\overset{\circ}{\mathfrak{U}} = -e(\mu_n n + \mu_p p) \operatorname{grad} \mathcal{V} + kT \operatorname{grad} (\mu_n n - \mu_p p) \tag{10}$$

which because of (5) can be rewritten as

$$\overset{\circ}{\mathfrak{U}} = -e\,(\mu_n\, n + \mu_p\, p) \operatorname{grad}\left[\mathcal{V} - \frac{kT}{e} \frac{\mu_\mu - \mu_p}{\mu_n + \mu_p} \ln\,(\mu_n\, n + \mu_p\, p)\right]. \tag{11}$$

Now (8), (9) and (11) are all of the form

$$\mathfrak{u} = \phi \operatorname{grad} \psi$$

and hence obviously satisfy the condition

$$\mathfrak{u} \cdot \operatorname{curl} \mathfrak{u} = 0.$$

Therefore we have

Theorem 1: $\overset{\circ}{\mathfrak{U}}_p$, $\overset{\circ}{\mathfrak{U}}_n$, and $\overset{\circ}{\mathfrak{U}}$ are surface-normal vector fields.

From (8)–(10) we find, using (5)

$$\operatorname{curl} \overset{\circ}{\mathfrak{U}}_p = -e\mu_p \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{12}$$

$$\operatorname{curl} \overset{\circ}{||}_n = -e\mu_n \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{13}$$

and

$$\operatorname{curl} \overset{\circ}{||} = -e(\mu_n + \mu_p) \operatorname{grad} p \times \operatorname{grad} \mathcal{V}, \tag{14}$$

whence

Theorem 2:

$$\frac{\operatorname{curl} \overset{\circ}{||}_p}{\mu_p} = \frac{\operatorname{curl} \overset{\circ}{||}_n}{\mu_n} = \frac{\operatorname{curl} \overset{\circ}{||}}{\mu_n + \mu_p}.$$

That is, curl $\overset{\circ}{||}_p$, curl $\overset{\circ}{||}_n$, and curl $\overset{\circ}{||}$ are constant multiples of one another.

and

Theorem 3: $\overset{\circ}{||}_p$, $\overset{\circ}{||}_n$, and $\overset{\circ}{||}$ are irrotational if and only if

$$\operatorname{grad} p = 0 \qquad (p = p(t))$$

or

$$\operatorname{grad} \mathcal{V} = 0 \qquad (\mathcal{V} = \mathcal{V}(t))$$

or

$$\mathcal{V} = \mathcal{V}(p, t).$$

The following interesting relations can be obtained from (8) and (9) (they are really consequences of Theorem 1):

$$\operatorname{curl} \overset{\circ}{||}_p = \operatorname{grad} \ln p \times \overset{\circ}{||}_p \tag{15}$$

and

$$\operatorname{curl} \overset{\circ}{||}_n = \operatorname{grad} \ln n \times \overset{\circ}{||}_n. \tag{16}$$

Now from (3) – (5) we find

$$\overset{\circ}{||}_p \times \overset{\circ}{||}_n = e\mu_n\mu_p kT(n + p) \operatorname{grad} p \times \operatorname{grad} \mathcal{V} \tag{17a}$$

$$= \tfrac{1}{2}\mu_n\mu_p kT(n + p) \operatorname{grad} (n + p) \times \operatorname{grad} \mathcal{V} \tag{17b}$$

$$= \tfrac{1}{4}e\mu_n\mu_p kT \operatorname{grad} (n + p)^2 \times \operatorname{grad} \mathcal{V} \tag{17c}$$

$$= \tfrac{1}{4}e\mu_n\mu_p kT \operatorname{curl} [(n + p)^2 \operatorname{grad} \mathcal{V}] \tag{17d}$$

and

$$\frac{\overset{\circ}{||}_p}{\mu_p e} - \frac{\overset{\circ}{||}_n}{\mu_n e} = \operatorname{grad} \left[N\mathcal{V} - \frac{kT}{e}(n + p) \right] \tag{18}$$

and

$$\frac{\overset{\circ}{||}_p}{\mu_p e} + \frac{\overset{\circ}{||}_n}{\mu_n e} = -(n + p) \operatorname{grad} \mathcal{V}. \tag{19}$$

[*Note:* As is suggested by (18) and (19), the total carrier concentration

$$\mathcal{P} \equiv n + p = N + 2p = 2n - N \quad (\mathcal{P} \geq |N|)$$

will frequently appear as the "natural" concentration variable in the relations with which we shall be working. Hence, expressions involving p, or p and n will often be replaced in the sequel by their equivalents in terms of the variable $\mathcal{P}$. It will be noted that

$$\text{grad } \mathcal{P} = 2 \text{ grad } p = 2 \text{ grad } n.]$$

Equations (17) and (19) yield at once the following theorems:

Theorem 4: The vector field

$$\overset{\circ}{\mathbb{I}}_p \times \overset{\circ}{\mathbb{I}}_n = \overset{\circ}{\mathbb{I}} \times \overset{\circ}{\mathbb{I}}_n = \overset{\circ}{\mathbb{I}}_p \times \overset{\circ}{\mathbb{I}}$$

is solenoidal.

Theorem 5: The vector field

$$\left(\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} - \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n}\right) \text{ is irrotational with a potential } (-eN\mathcal{U} + kT\mathcal{P}).$$

Theorem 6: The vector field

$$\left(\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} + \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n}\right) \text{ is surface-normal (to the surfaces of constant } \mathcal{U}).$$

Theorem 7: $\overset{\circ}{\mathbb{I}}_p$, $\overset{\circ}{\mathbb{I}}_n$, $\overset{\circ}{\mathbb{I}}$, grad $\mathcal{U}$, and grad p are coplanar vectors.

Theorem 8: The flow lines of any two of the fields $\overset{\circ}{\mathbb{I}}_p$, $\overset{\circ}{\mathbb{I}}_n$, and $\overset{\circ}{\mathbb{I}}$ coincide if and only if

$$\text{grad } p = 0 \qquad (p = p(t))$$

or

$$\text{grad } \mathcal{U} = 0 \qquad (\mathcal{U} = \mathcal{U}(t))$$

or

$$\mathcal{U} = \mathcal{U}(p, t).$$

Also, from (17) and (19) we obtain the curious relations:

$$\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} \times \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n} = -\frac{kT}{2} \text{ grad } \mathcal{P} \times \left(\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} + \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n}\right) \tag{20a}$$

$$= -\frac{kT}{2} \mathcal{P} \text{ curl } \left(\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} + \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n}\right) \tag{20b}$$

$$= -\frac{kT}{2} \text{ curl } \left[\mathcal{P}\left(\frac{\overset{\circ}{\mathbb{I}}_p}{\mu_p} + \frac{\overset{\circ}{\mathbb{I}}_n}{\mu_n}\right)\right]. \tag{20c}$$

Finally, by taking the divergence of (7) and making use first of (1) and (2) and then of (5), we obtain:

[*Theorem 9:* The vector field $\overset{\circ}{||}$ is solenoidal.

C. Formulation of Partial Differential Equation System Restricting $\mathcal{P}$ and $\mathcal{U}$

A very convenient formulation of the partial differential equations restricting $\mathcal{P}$ and $\mathcal{U}$ is suggested by (18) and (19). Taking the divergence of these equations and substituting (1) and (2) into the results we obtain:

$$\text{div grad}\left(N\mathcal{U} - \frac{kT}{e}\mathcal{P}\right) = -\alpha\left(\mathcal{R} + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right) \tag{21}$$

and

$$\text{div}\ (\mathcal{P}\ \text{grad}\ \mathcal{U}) = \beta\left(\mathcal{R} + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right) \tag{22}$$

wherein for brevity we have set

$$\alpha \equiv \frac{1}{\mu_p} + \frac{1}{\mu_n}$$

and

$$\beta \equiv \frac{1}{\mu_p} - \frac{1}{\mu_n}$$

and shall henceforth assume $\beta \neq 0$, i.e., $\mu_p \neq \mu_n$. Equations (21) and (22) yield immediately a derived equation not containing explicitly the terms introduced by recombination and time variations:

$$\text{div grad}\left(N\mathcal{U} - \frac{kT}{e}\mathcal{P}\right) = -\frac{\alpha}{\beta}\,\text{div}\ (\mathcal{P}\ \text{grad}\ \mathcal{U}) \tag{23a}$$

or

$$\text{div}\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\text{grad}\ \mathcal{U} - \frac{kT}{e}\,\text{grad}\ \mathcal{P}\right] = 0 \tag{23b}$$

or

$$\text{div}\left(\left(\mathcal{P} + \frac{\beta N}{\alpha}\right)\text{grad}\left[\mathcal{U} - \frac{\beta kT}{\alpha e}\ln\left(\mathcal{P} + \frac{\beta N}{\alpha}\right)\right]\right) = 0. \tag{23c}$$

Either the set (21) and (22) or one of the forms of (23) together with either (21) or (22) constitutes a basic set of two partial differential equations determining $\mathcal{P}$ and $\mathcal{U}$. We are here considering $\mathcal{R}$ as $\mathcal{R}(\mathcal{P})$.

It will be observed that (23) is equivalent to the condition

$$\operatorname{div} \overset{\circ}{||} = 0 \tag{24}$$

established as Theorem 9.
(In terms of $\mathcal{P}$, (10) becomes

$$\overset{\circ}{||} = -\frac{e(\mu_n - \mu_p)}{2}\left[\left(\frac{\alpha}{\beta}\mathcal{P} + N\right)\operatorname{grad} \mathcal{V} - \frac{kT}{e}\operatorname{grad} \mathcal{P}\right]. \tag{25}$$

In most of the following sections we shall find it expedient to consider separately the cases $N \neq 0$ and $N = 0$ (associated respectively with semiconductors of the extrinsic and intrinsic conductivity types). *For the case* $N \neq 0$, use will be made frequently of new dependent variables $\mathfrak{U}$ and $\mathfrak{IC}$ defined by:

$$\mathfrak{U} \equiv \frac{kT}{eN}\mathcal{P} \tag{26}$$

$$\mathfrak{IC} \equiv \mathcal{V} - \frac{kT}{eN}\mathcal{P} = \mathcal{V} - \mathfrak{U}. \tag{27}$$

That is,

$$\mathcal{P} \equiv \frac{eN}{kT}\mathfrak{U} \tag{28}$$

$$\mathcal{V} \equiv \mathfrak{U} + \mathfrak{IC} \tag{29}$$

will be substituted into relations involving $\mathcal{P}$ and $\mathcal{V}$ to obtain the corresponding relations in terms of $\mathfrak{U}$ and $\mathfrak{IC}$. Incidentally, it will be noted that $\mathfrak{U}$ and $\mathfrak{IC}$ have the dimensions of voltage.

In terms of $\mathfrak{U}$ and $\mathfrak{IC}$ the basic equations (21)–(23) can be written:

$$\operatorname{div} \operatorname{grad} \mathfrak{IC} = -\frac{\alpha}{N}\left[\mathcal{R} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right] \tag{30}$$

$$\operatorname{div} [\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})] = \frac{\beta kT}{eN}\left[\mathcal{R} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right], \tag{31}$$

$$\operatorname{div}\left[\operatorname{grad} \mathfrak{IC} + \frac{\alpha e}{\beta kT}\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})\right] = 0 \tag{32}$$

wherein $\mathcal{R}$ will be considered as $\mathcal{R}(\mathfrak{U})$.

It will be observed that, in the absence of recombination and time variation, (30)–(32) reduce to

$$\operatorname{div} \operatorname{grad} \mathfrak{IC} = 0 \tag{33}$$

and $[N \neq 0]$

$$\operatorname{div} [\mathfrak{U} \operatorname{grad} (\mathfrak{U} + \mathfrak{IC})] = 0 \tag{34}$$

The elegant form of this set of equations furnished the original motivation for the introduction of the variables $\mathfrak{U}$ and $\mathfrak{IC}$. The comparable equations for $N = 0$ are

$$\text{div grad } \mathcal{P} = 0 \tag{35}$$

$$[N = 0]$$

$$\text{div } [\mathcal{P} \text{ grad } \mathcal{V}] = 0. \tag{36}$$

D. The Recombination Rate Function $\mathcal{R}$

In order to avoid undue confusion in the sequel we shall at this point make some clarifying remarks concerning the function $\mathcal{R}$. As was stated in the Introduction, we basically regard $\mathcal{R}$ as a function of $p - p_0$ and $n - n_0$. However, because of (5), any expression in $p - p_0$ and $n - n_0$ can be replaced by one in which (say) p is the only field variable quantity. It is then convenient to regard $\mathcal{R}$ as a function of p and write it $\mathcal{R}(p)$. When dealing with expressions in terms of $\mathcal{P}$ and of $\mathfrak{U}$, it is convenient to regard $\mathcal{R}$ as a function of one of these variables and to indicate this fact by writing $\mathcal{R}(\mathcal{P})$ or $\mathcal{R}(\mathfrak{U})$. When we do this we do not mean that $\mathcal{R}(\mathcal{P})$ (say) is the same algebraic function of $\mathcal{P}$ as $\mathcal{R}(p)$ is of p, but rather that $\mathcal{R}(p)$ is the function of p obtained when one substitutes $\mathcal{P} = N + 2p$ into $\mathcal{R}(\mathcal{P})$.

For example, *for constant mean lifetime recombination*

$$\mathcal{R}(p) \equiv \frac{1}{\tau_0}(p - p_0) \tag{37a}$$

$$\mathcal{R}(\mathcal{P}) \equiv \frac{1}{2\tau_0}(\mathcal{P} - \mathcal{P}_0) \tag{37b}$$

$$\mathcal{R}(\mathfrak{U}) \equiv \frac{eN}{2\tau_0 kT}(\mathfrak{U} - \mathfrak{U}_0) \tag{37c}$$

with τ_0 constant;
and *for mass-action recombination*

$$\mathcal{R}(p) \equiv \frac{1}{n_0\tau_0}[p(p + N) - p_0 n_0] \tag{38a}$$

$$\mathcal{R}(\mathcal{P}) \equiv \frac{1}{2\tau_0(\mathcal{P}_0 + N)}(\mathcal{P}^2 - \mathcal{P}_0^2) \tag{38b}$$

$$\mathcal{R}(\mathfrak{U}) \equiv \frac{e^2N^2}{2k^2T^2\tau_0\,(\mathcal{P}_0 + N)}(\mathfrak{U}^2 - \mathfrak{U}_0^2). \tag{38c}$$

E. Addition of Arbitrary Time Functions to $\mathcal{U}$ and $\mathcal{K}$

Since only the gradient of $\mathcal{U}$ appears in the basic equations (21) and (22), it is evident that if

$$\mathcal{U} = \mathcal{U}(x, y, z, t)$$

and

$$\mathcal{P} = \mathcal{P}(x, y, z, t)$$

are a pair of functions satisfying (21) and (22), then so also are

$$\tilde{\mathcal{U}} = \mathcal{U}(x, y, z, t) + \tilde{m}(t)$$

and

$$\tilde{\mathcal{P}} = \mathcal{P}(x, y, z, t)$$

where $\tilde{m}(t)$ is an arbitrary time function.
And since $\mathcal{U} = \mathfrak{U} + \mathcal{K}$, if

$$\mathcal{K} = \mathcal{K}(x, y, z, t)$$

and

$$\mathfrak{U} = \mathfrak{U}(x, y, z, t)$$

are a pair of functions satisfying (30)–(32), so also are

$$\tilde{\mathcal{K}} = \mathcal{K}(x, y, z, t) + \tilde{m}(t)$$

and

$$\tilde{\mathfrak{U}} = \mathfrak{U}(x, y, z, t).$$

These arbitrary additive functions with zero gradients are physically trivial in that they merely reflect the arbitrariness of the reference voltage level. They will, however, be retained for the sake of formal completeness whenever they appear in the subsequent analyses.

F. Summary of Solutions for No Recombination or Time Variation

The next ten sections of this paper (Sections G–Q) contain a sequence of detailed analyses in which is determined the existence or non-existence of solution fields having certain prescribed formal properties. In most of these studies time variability and recombination are admitted and the analysis includes the establishment of the class of recombination rate functions $\mathcal{R}$ consistent with the property under consideration. In those cases where solutions are found to exist, they are expressed in the simplest convenient

terms: in closed form, or as solutions of an ordinary differential equation, or as solutions of a single partial differential equation. The solutions found usually involve arbitrary constants and/or arbitrary functions of various kinds.

The present section is intended to provide a skimpy but compact sampling of the results obtained in the next ten sections. It will be confined to a simple listing of solutions found and furthermore will contain only the forms to which these solutions reduce when recombination and time variation are excluded. (Some solutions are lost under this reduction.) A heading will indicate the section(s) from which the solution comes as well as the formal property associated with each solution.

For the sake of conciseness and simplicity the symbols denoting arbitrary constants and functions in this section are independent of those employed in the later sections. They are to be interpreted as follows:

A, B: arbitrary constants
$h(x, y, z)$: any harmonic function
(or with subscript)

$(\tilde{\mathcal{V}}, \tilde{\mathcal{P}})$: any given solution field

[G. grad $\mathcal{V} = 0$]

$$\begin{cases} \mathcal{V} = A \\ \mathcal{P} = h(x, y, z) \end{cases}$$

[H, I. grad $\mathcal{P} = 0$]

$$\begin{cases} \mathcal{V} = h(x, y, z) \\ \mathcal{P} = A \end{cases}$$

[J. grad $\mathfrak{C} = 0$, $N \neq 0$]

$$\begin{cases} \mathcal{V} = A + \sqrt{h(x, y, z)} \\ \mathcal{P} = \dfrac{Ne}{kT} \sqrt{h(x, y, z)} \end{cases}$$

[K, L. $\mathcal{V} = \mathcal{V}(\mathcal{P})$, $N \neq 0$]

$$(A \neq 0) \begin{cases} \mathcal{V} = h(x, y, z) + A\Lambda\left[\dfrac{B - h(x, y, z)}{A}\right] \\ \mathcal{P} = \dfrac{Ne}{kT} A\Lambda\left[\dfrac{B - h(x, y, z)}{A}\right] \end{cases}$$

(For definition of function Λ see Equation (87) and Figs. 1 and 2.)

[K, M. $\mathcal{U} = \mathcal{U}(\mathcal{P})$, $N = 0$]

$$\begin{cases} \mathcal{U} = A \ln h(x, y, z) + B \\ \mathcal{P} = h(x, y, z) \end{cases}$$

[N, O. grad $\mathcal{P}$·grad $\mathcal{U} = 0$]

$$\begin{cases} \mathcal{U} = h_1(x, y, z) \\ \mathcal{P} = h_2(x, y, z) \\ \quad \text{provided} \\ \text{grad } h_1(x, y, z) \cdot \text{grad } h_2(x, y, z) = 0 \end{cases}$$

[N. grad $\mathfrak{U}$·grad $\mathfrak{K} = 0$, $N \neq 0$]

$$\begin{cases} \mathcal{U} = \sqrt{h_1(x, y, z)} + h_2(x, y, z) \\ \mathcal{P} = \dfrac{Ne}{kT} \sqrt{h_1(x, y, z)} \\ \quad \text{provided} \\ \text{grad } h_1(x, y, z) \cdot \text{grad } h_2(x, y, z) = 0 \end{cases}$$

[P. grad $\mathcal{P}$·grad $h = 0$]

$$\begin{cases} \mathcal{U} = \tilde{\mathcal{U}} + h(x, y, z) \\ \mathcal{P} = \tilde{\mathcal{P}} \\ \quad \text{provided} \\ \text{grad } \tilde{\mathcal{P}} \cdot \text{grad } h(x, y, z) = 0. \end{cases}$$

G. Solutions with $\mathcal{U} = \mathcal{U}(t)$

Our point of view in general is that $\mathcal{P}$ and $\mathcal{U}$ (or $\mathfrak{U}$ and $\mathfrak{K}$) are functions of three space coordinates and time, so that $\mathcal{U} = \mathcal{U}(t)$ implies for example that $\dfrac{\partial \mathcal{U}}{\partial x} = \dfrac{\partial \mathcal{U}}{\partial y} = \dfrac{\partial \mathcal{U}}{\partial z} = 0$. That is to say, we now seek solutions for which everywhere

$$\text{grad } \mathcal{U} = 0. \tag{39}$$

From (21) and (22) this condition gives us the following restrictions on

$\mathcal{P}$ (and none on $\mathcal{U}(t)$):

$$\text{div grad } \mathcal{P} = 0 \tag{40}$$

and

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t} = 0. \tag{41}$$

By operating with div grad on (41) we obtain

$$\mathcal{R}''(\mathcal{P}) = 0$$

(we consistently use primes to denote differentiation with respect to the argument of a function of a single variable—e.g.,

$$\mathcal{R}''(\mathcal{P}) \equiv \frac{d^2\mathcal{R}(\mathcal{P})}{d\mathcal{P}^2}\Big)$$

whence,

$$2\mathcal{R}(\mathcal{P}) = A\mathcal{P} + B \tag{42}$$

(A, B: arbitrary constants). Substituting (42) into (41) we obtain

$$\frac{\partial \mathcal{P}}{\partial t} + A\mathcal{P} = -B$$

whence

$$\mathcal{P} = c(x, y, z)\epsilon^{-At} - B/A \qquad (A \neq 0) \tag{43a}$$

or

$$\mathcal{P} = c(x, y, z) - Bt \qquad (A = 0). \tag{43b}$$

From (36) it follows that

$$\text{div grad } c(x, y, z) = 0, \tag{44}$$

that is, c must be harmonic.

In brief, if $\mathcal{R}(\mathcal{P})$ is of the form given in (42), any $\mathcal{U}(t)$ and (43) constitute solutions to the flow equations for any harmonic $c(x, y, z)$. Other forms of $\mathcal{R}(\mathcal{P})$ admit no solutions with $\mathcal{U} = \mathcal{U}(t)$.

It is evident that when recombination is absent time variation is also absent, and vice versa. The solutions reduce in this case to:

$$\mathcal{U} = C \qquad (C\text{: arbitrary constant}) \tag{45}$$

$$\mathcal{P} = c(x, y, z). \tag{46}$$

H. Solutions with $\mathcal{P} = \mathcal{P}(t)$, $N \neq 0$

The condition

$$\operatorname{grad} \mathcal{P} = 0 \tag{47}$$

yields from (21) and (22)

$$\left(N + \frac{\alpha}{\beta}\mathcal{P}\right) \operatorname{div} \operatorname{grad} \mathcal{U} = 0 \tag{48}$$

and

$$\mathcal{P} \operatorname{div} \operatorname{grad} \mathcal{U} = \beta\left[\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{d\mathcal{P}}{dt}\right]. \tag{49}$$

Two cases arise for $\mathcal{R} \not\equiv 0$:

Case 1:

$$\mathcal{P} = -\frac{\beta N}{\alpha} \tag{50}$$

and

$$\operatorname{div} \operatorname{grad} \mathcal{U} = -\frac{\alpha}{N}\mathcal{R}(\mathcal{P})\Big|_{\mathcal{P} = -\frac{\beta N}{\alpha}} \tag{51}$$

Case 2:

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{d\mathcal{P}}{dt} = 0$$

or

$$D - t = \int \frac{d\mathcal{P}}{2\mathcal{R}(\mathcal{P})} \qquad (D\text{: arbitrary constant}) \tag{52}$$

and

$$\operatorname{div} \operatorname{grad} \mathcal{U} = 0. \tag{53}$$

When recombination is absent, these cases reduce to:

$$\mathcal{P} = E \qquad (E\text{: arbitrary constant}) \tag{54}$$

and (53).

When time variation is absent, Case 2 again yields (53) and (54).

It should be recalled that $\mathcal{U}$ can depend on t as well as x, y, z; so that arbitrary functions of t play the role of arbitrary constants in (51) and (53), whenever time variation is allowed.

I. Solutions with $P = P(t)$, $N = 0$

For $N = 0$, only Case 2 of the previous section occurs, because the condition $\mathcal{P} = 0$ (implying no carriers!) is of no interest.

J. Solutions with $\mathfrak{C} = \mathfrak{C}(t)$, $N \neq 0$

For grad $\mathfrak{C} = 0$, (30) and (32) yield:

$$\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t} = 0 \tag{55}$$

and

$$\text{div grad } \mathfrak{U}^2 = 2 \text{ div } \mathfrak{U} \text{ grad } \mathfrak{U} = 0. \tag{56}$$

Taking the div grad of (55) multiplied by $\mathfrak{U}$ we obtain

$$\text{div grad } \mathfrak{U}\mathfrak{R}(\mathfrak{U}) = 0$$

whence, because of (56)

$$\frac{4kT}{eN}\,\mathfrak{U}\mathfrak{R}(\mathfrak{U}) = F\mathfrak{U}^2 + G \qquad (F, G\text{: arbitrary constants})$$

or

$$\frac{4kT}{eN}\,\mathfrak{R}(\mathfrak{U}) = F\mathfrak{U} + G\mathfrak{U}^{-1}. \tag{57}$$

Substituting this permitted form for the recombination rate function into (55) we obtain

$$\frac{\partial \mathfrak{U}^2}{\partial t} + F\mathfrak{U}^2 = -G \tag{58}$$

whence

$$\mathfrak{U} = \sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \qquad (F \neq 0) \tag{59a}$$

or

$$\mathfrak{U} = \sqrt{f(x, y, z) - Gt} \qquad (F = 0). \tag{59b}$$

From (56), $f(x, y, z)$ is subject to

$$\text{div grad } f(x, y, z) = 0. \tag{60}$$

In summary, if and only if $\mathfrak{R}(\mathfrak{U})$ has the form (57), there are solutions for which $\mathfrak{C} = \mathfrak{C}$ (t) (arbitrary). The $\mathfrak{U}$ is given by (59) in which f is an arbitrary harmonic function of x, y, z.

In terms of $\mathcal{P}$ and $\mathcal{U}$ these solutions are given by:

$$\mathcal{R}(\mathcal{P}) = \frac{F}{4}\mathcal{P} + \left(\frac{eN}{kT}\right)^2 \frac{G}{4}\mathcal{P}^{-1}, \tag{61}$$

$$\mathcal{P} = \frac{eN}{kT}\sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \qquad (F \neq 0) \tag{62a}$$

or

$$\mathcal{P} = \frac{eN}{kT}\sqrt{f(x, y, z) - Gt} \qquad (F = 0), \tag{62b}$$

and

$$\mathcal{U} = \mathcal{K}(t) + \sqrt{f(x, y, z)\epsilon^{-Ft} - G/F} \quad (F \neq 0) \tag{63a}$$

or

$$\mathcal{U} = \mathcal{K}(t) + \sqrt{f(x, y, z) - Gt} \qquad (F = 0). \tag{63b}$$

For no recombination ($\mathcal{R} \equiv 0$), these results specialize to (59b), (60), (62b), and (63b) with G set equal to zero. It should be noted (see (55)) that absence of time variation implies absence also of recombination.

K. Solutions with $\mathcal{U} = \mathcal{U}(\mathcal{P}, t)$, grad $\mathcal{P} \neq 0$

In Theorems 3 and 8 of Section B we have shown that some very interesting properties are implied by the condition

$$\text{grad } \mathcal{U} \times \text{grad } \mathcal{P} = 0. \tag{64}$$

In sections G–I we have treated the cases grad $\mathcal{U} = 0$ and grad $\mathcal{P} = 0$. We now turn to the remaining possibility leading to (64):

$$\mathcal{U} = \mathcal{U}(\mathcal{P}, t) \text{ with grad } \mathcal{P} \neq 0. \tag{65}$$

Substitution of (65) into (23b) leads to

$$\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e}\right] \text{div grad } \mathcal{P} + \frac{\partial}{\partial \mathcal{P}}\left[\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e}\right](\text{grad } \mathcal{P})^2 = 0. \tag{66}$$

Two cases arise.

Case 1:

$$\left(N + \frac{\alpha}{\beta}\mathcal{P}\right)\frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} = 0.$$

This condition clearly satisfies (66) and leads to

$$\mathcal{U}(\mathcal{P}, t) = g(t) + \frac{\beta kT}{\alpha e} \ln \left| \mathcal{P} + \frac{\beta N}{\alpha} \right| \quad (g(t): \text{arbitrary function}). \tag{67}$$

The restriction on $\mathcal{P}$ is then provided by the result of substituting (67) into (21):

$$\operatorname{div} \operatorname{grad} \left[\mathcal{P} - \frac{\beta N}{\alpha} \ln \left| \mathcal{P} + \frac{\beta N}{\alpha} \right| \right] = \alpha \left[\mathcal{R}(\mathcal{P}) + \frac{1}{2} \frac{\partial \mathcal{P}}{\partial t} \right]. \tag{68}$$

Any $\mathcal{P}$ satisfying (68) constitutes with (67) a solution having the property desired.

If (65) is substituted into (25) it will be found that the condition

$$\left(N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} = 0$$

is equivalent to $\overset{\circ}{||} = 0$, so that Case 1 is characterized by zero total current.

Case 2:

$$\left(N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} \neq 0.$$

In this case (66) can be written in the form

$$\frac{\operatorname{div} \operatorname{grad} \mathcal{P}}{(\operatorname{grad} \mathcal{P})^2} = - \frac{\partial}{\partial \mathcal{P}} \ln \left[\left(N + \frac{\alpha}{\beta} \mathcal{P} \right) \frac{\partial \mathcal{U}}{\partial \mathcal{P}} - \frac{kT}{e} \right] = \phi(\mathcal{P}, t). \tag{69}$$

From (69) it follows that $\mathcal{P}$ must be of the form $\mathcal{P}(h, t)$ with

$$\operatorname{div} \operatorname{grad} h(x, y, z, t) = 0. \tag{70}$$

In summary we have

Theorem 10: If $\mathcal{U} = \mathcal{U}(\mathcal{P}, t)$ with grad $\mathcal{P} \neq 0$, then either $\overset{\circ}{||} = 0$ or $\mathcal{U} = \mathcal{U}(h, t)$ and $\mathcal{P} = \mathcal{P}(h, t)$ with div grad $h(x, y, z, t) = 0$.

We shall investigate the restrictions on the functions $h(x, y, z, t)$, $\mathcal{U}(h, t)$, and $\mathcal{P}(h, t)$ in the next two sections.

Theorem 10 remains unchanged if recombination is absent. If time variation is absent, it simply drops t as a variable in the functions mentioned in the theorem. If both recombination and time variation are absent, the theorem can be strengthened to:

Theorem 11: If both recombination and time variation are absent and $\mathcal{U} = \mathcal{U}(\mathcal{P})$, then $\mathcal{U} = \mathcal{U}(h)$ and $\mathcal{P} = \mathcal{P}(h)$ with div grad $h(x, y, z) = 0$.

L. Solutions with $\mathfrak{V} = \mathfrak{V}(h, t)$, $\mathfrak{P} = \mathfrak{P}(h, t)$, grad $\mathfrak{P} \neq 0$, div grad $h = 0$, $N \neq 0$

For formal reasons we shall work, not with the conditions $\mathfrak{P} = \mathfrak{P}(h, t)$ and $\mathfrak{V} = \mathfrak{V}(h, t)$, but with the equivalent conditions

$$\mathfrak{U} = \mathfrak{U}(h, t) \text{ and } \mathfrak{JC} = \mathfrak{JC}(h, t). \tag{71}$$

The condition grad $\mathfrak{P} \neq 0$ now implies $\dfrac{\partial \mathfrak{U}}{\partial h} \neq 0$.

Substitution of (79) into (30) and (32) yields—after use of (70):

$$\frac{\partial^2 \mathfrak{JC}}{\partial h^2} (\text{grad } h)^2 = -\frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t} + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial h}\frac{\partial h}{\partial t}\right] \tag{72}$$

and

$$\frac{\partial}{\partial h}\left(\left[\frac{\beta kT}{\alpha e} + \mathfrak{U}\right]\frac{\partial \mathfrak{JC}}{\partial h} + \mathfrak{U}\frac{\partial \mathfrak{U}}{\partial h}\right) = 0. \tag{73}$$

From (73) we get

$$\frac{\partial \mathfrak{JC}}{\partial h} = \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial h}}{\dfrac{\beta kT}{\alpha e} + \mathfrak{U}} \tag{74}$$

($j(t)$: arbitrary function)

which yields upon substitution into (72):

$$\frac{\partial}{\partial h}\left[\frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial h}}{\dfrac{\beta kT}{\alpha e} + \mathfrak{U}}\right](\text{grad } h)^2 = -\frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\left(\frac{\partial \mathfrak{U}}{\partial h}\frac{\partial h}{\partial t} + \frac{\partial \mathfrak{U}}{\partial t}\right)\right] \tag{75}$$

in which $\mathfrak{U}$, $\dfrac{\partial \mathfrak{U}}{\partial h}$, and $\dfrac{\partial^2 \mathfrak{U}}{\partial h^2}$ are, of course, functions of h and of t.

In determining the combined implications of (75) and (70) three cases arise according to whether or not $\dfrac{\partial^2 \mathfrak{JC}}{\partial h^2} = 0$ or grad $(\text{grad } h)^2 = 0$.

Case 1:

$$\frac{\partial^2 \mathfrak{JC}}{\partial h^2} \neq 0, \qquad \text{grad } (\text{grad } h)^2 \neq 0.$$

In this case no satisfactory interpretation has been found when time variability is present.

When time variation is absent, we work with the conditions

$$\mathfrak{U} = \mathfrak{U}(h); \qquad \mathfrak{IC} = \mathfrak{IC}(h)$$

with

$$\text{div grad } h(x, y, z) = 0 \tag{76}$$

and arrive at counterparts of (74) and (75):

$$\mathfrak{IC}' = \frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}} \qquad \left(\gamma \equiv \frac{\beta k T}{\alpha e}\right) \qquad (H\text{: arbitrary constant}) \tag{77}$$

and

$$\mathfrak{IC}'' (\text{grad } h)^2 = \left(\frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' (\text{grad } h)^2 = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U}). \tag{78}$$

From (78) it is evident that $\mathfrak{R} \neq 0$ implies $\mathfrak{IC}'' \neq 0$ and grad $h \neq 0$. So we have

$$(\text{grad } h)^2 = \frac{-\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U})}{\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)'} \tag{79}$$

which is of the form

$$[\text{grad } h(x, y, z)]^2 = \phi(h). \tag{79a}$$

Now from (79a) follows

$$\text{grad } h \times \text{grad } (\text{grad } h)^2 = 0 \tag{80}$$

which implies that the vector lines of the field grad h are all straight. Since h is harmonic, this restricts the choice of h to the potential fields associated with a uniform parallel flow, a straight line source, or a point source. Hence, for suitably chosen rectangular coordinates (x, y, z), circular cylindrical coordinates (ρ, θ, z) or spherical polar coordinates (r, θ, ϕ), the only possibilities, are respectively

$$h = x \rightarrow (\text{grad } h)^2 = 1 \tag{81a}$$

or

$$h = \ln \frac{1}{\rho} \rightarrow (\text{grad } h)^2 = \frac{1}{\rho^2} = \epsilon^{2h} \tag{81b}$$

or

$$h = \frac{1}{r} \rightarrow (\text{grad } h)^2 = \frac{1}{r^4} = h^4. \tag{81c}$$

The possibility $h = x$ violates one defining condition for the present case (i.e., grad (grad $h)^2 \neq 0$) and hence will be left for consideration in Case 3. The remaining two possibilities lead respectively to the following forms of ordinary differential equation for the determination of $\mathfrak{U}(h)$:

$$\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' + \frac{\alpha}{N} \epsilon^{-2h} \mathfrak{R}(\mathfrak{U}) = 0 \tag{82b}$$

or

$$\left(\frac{S - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right)' + \frac{\alpha}{N} \frac{1}{h^4} \mathfrak{R}(\mathfrak{U}) = 0. \tag{82c}$$

Given any $\mathfrak{U}(h)$ satisfying one of these equations, the associated $\mathfrak{IC}(h)$ is obtained by integration from (77):

$$\mathfrak{IC}(h) = \int \left(\frac{H - \mathfrak{U}\mathfrak{U}'}{\gamma + \mathfrak{U}}\right) dh + J \qquad (J: \text{arbitrary constant}). \tag{83}$$

It is evident from (72) that Case 1 does not exist if both recombination and time variation are absent.

Case 2:

$$\frac{\partial^2 \mathfrak{IC}}{\partial h^2} = 0$$

In considering this case we shall exclude the condition $\frac{\partial \mathfrak{IC}}{\partial h} = 0$ because it has been included in Section J.

From the condition $\frac{\partial^2 \mathfrak{IC}}{\partial h^2} = 0$ we have

$$\mathfrak{IC} = k(t)h + \ell(t) \qquad (k(t), \ell(t): \text{arbitrary functions}) \tag{84}$$

with $k \neq 0$. This shows that $\mathfrak{IC}$ itself is a harmonic function and we can without loss of generality use it in place of h.

Equations (74) and (75) now yield the two conditions on $\mathfrak{U}(\mathfrak{IC}, t)$, $\mathfrak{R}(\mathfrak{U})$, and $\mathfrak{IC}(x, y, z, t)$:

$$\frac{j(t) - \mathfrak{U} \frac{\partial \mathfrak{U}}{\partial \mathfrak{IC}}}{\mathfrak{U} + \gamma} = 1 \tag{85}$$

and

$$\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\left(\frac{\partial \mathfrak{U}}{\partial \mathfrak{C}}\frac{\partial \mathfrak{C}}{\partial t} + \frac{\partial \mathfrak{U}}{\partial t}\right) = 0. \tag{86}$$

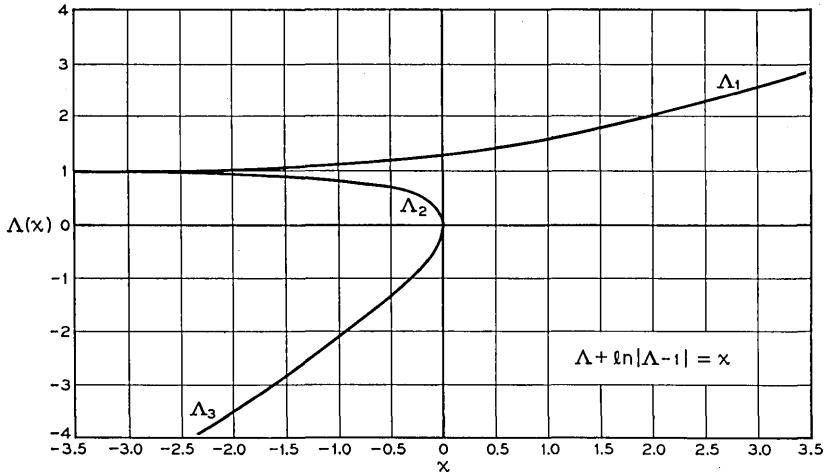


Fig. 1—The transcendental function $\Lambda(x)$.

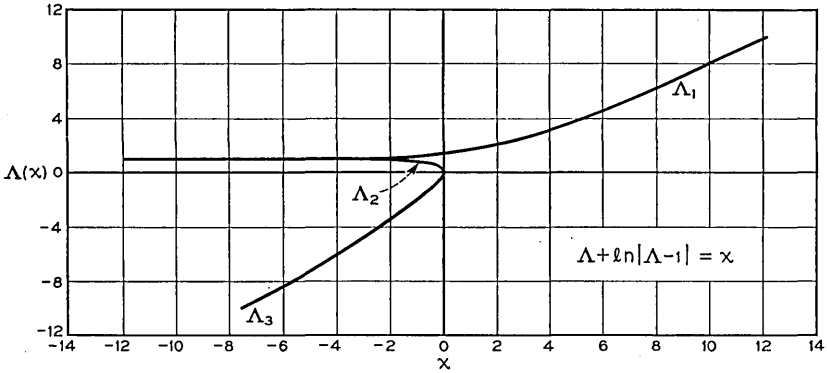


Fig. 2—The transcendental function $\Lambda(x)$.

For the integration of (85) we need the transcendental algebraic function of a single real variable defined by

$$\Lambda(x) + \ln|\Lambda(x) - 1| = x. \tag{87}$$

This function is plotted in Figs. 1 and 2. It will be observed that x is always a single-valued function of Λ; while Λ is a single-valued function of x for $x > 0$, a double-valued function for $x = 0$, and a triple-valued function for

$x < 0$. The single-valued monotone functions Λ_1, Λ_2, and Λ_3 are defined respectively by the restrictions $\Lambda > 1$, $1 > \Lambda \geq 0$, and $\Lambda \leq 0$. When Λ is used without subscript it is implied that either Λ_1, Λ_2, or Λ_3 can be used. It will be useful to remember that

$$\Lambda'(x) = \frac{\Lambda(x) - 1}{\Lambda(x)}. \tag{88}$$

In terms of the function Λ, (85) integrates to

$$\mathfrak{U}(\mathfrak{IC}, t) = [j(t) - \gamma]\Lambda\left[\frac{m(t) - \mathfrak{IC}}{j(t) - \gamma}\right] \qquad (j \neq \gamma) \tag{89a}$$

(m(t): arbitrary function)

or

$$\mathfrak{U}(\mathfrak{IC}, t) = m(t) - \mathfrak{IC} \qquad (j = \gamma). \tag{89b}$$

The latter case ($j = \gamma$) corresponds to $\mathfrak{V} = \mathfrak{V}(t)$ and hence was included in Section G. Therefore in the following we shall consider only $j \neq \gamma$.

Now by making use of (89a) and (88), (86) can be rewritten in the form:

$$\frac{2kT}{eN}\frac{\mathfrak{U}\mathfrak{R}(\mathfrak{U})}{\mathfrak{U} - j + \gamma} + \frac{j'\mathfrak{U}^2}{(j - \gamma)(\mathfrak{U} - j + \gamma)} = \frac{\partial \mathfrak{IC}}{\partial t} + \frac{j'(m - \mathfrak{IC})}{j - \gamma} - m' \tag{90}$$

$$\left(\text{primes denoting here } \frac{d}{dt}\right).$$

We now observe that the right side of (90) is harmonic, while the left side is a function only of $\mathfrak{IC}$ and t. From this it follows that the right side can be written in the form:

$$\frac{\partial \mathfrak{IC}}{\partial t} + \frac{j'(m - \mathfrak{IC})}{j - \gamma} - m' = q(t)\left[\frac{m - \mathfrak{IC}}{j - \gamma}\right] + r(t). \tag{91}$$

From (90), (91) and (89a) follows

$$\frac{2kT}{eN}\mathfrak{R}(\mathfrak{U}) = -\frac{j'}{j - \gamma}\mathfrak{U} + \frac{\mathfrak{U} - j + \gamma}{\mathfrak{U}}\left(r + \frac{q}{j - \gamma}\mathfrak{U} + q\ln\left|\frac{\mathfrak{U}}{j - \gamma} - 1\right|\right). \tag{92}$$

Since (92) is of the form

$$\mathfrak{R}(\mathfrak{U}) = \phi(\mathfrak{U}, t),$$

The result of taking $\left(\frac{\partial}{\partial t}\right)_{\mathfrak{U} = \text{const}}$ of the right side must be zero identi-

cally in $\mathfrak{U}$. Making use of the algebraic lemma that

$$A + Bx + \frac{C}{x} + \left(D + \frac{E}{x}\right) \ln \left| \frac{x}{F} - 1 \right| \equiv 0$$

implies $A=B=C=D=E=0$, we arrive at two possibilities:
Possibility 1:

$$(q(t) = r(t) = 0 \quad \text{and} \quad j - \gamma = K\epsilon^{-Lt})$$

(K, L: arbitrary constants)

This yields from (92), (89a) and (91):

$$\frac{2kT}{eN} \mathfrak{R}(\mathfrak{U}) = L\mathfrak{U} \tag{93}$$

$$\mathfrak{U}(\mathfrak{IC}, t) = K\epsilon^{-Lt} \Lambda \left[\frac{1}{K} \epsilon^{Lt} (m(t) - \mathfrak{IC}) \right] \tag{94}$$

and

$$\mathfrak{IC}(x, y, z, t) = \epsilon^{-Lt} s(x, y, z) + \epsilon^{-Lt} \int \epsilon^{Lt} \left[Lm(t) + m'(t) \right] dt \tag{95}$$

where $s(x, y, z)$ is any harmonic function.
Possibility, 2:

$$(j(t) = M, q(t) = Q, r(t) = R)$$

(M, Q, R: arbitrary constants)

This yields from (92), (89a) and (91):

$$\frac{2kT}{eN} \mathfrak{R}(\mathfrak{U}) = \frac{\mathfrak{U} - M + \gamma}{\mathfrak{U}} \cdot \left(R + \frac{Q}{M - \gamma} \mathfrak{U} + Q \ln \left| \frac{\mathfrak{U}}{M - \gamma} - 1 \right| \right) \tag{96}$$

$$\mathfrak{U}(\mathfrak{IC}, t) = (M - \gamma) \Lambda \left[\frac{m(t) - \mathfrak{IC}}{M - \gamma} \right] \tag{97}$$

and

$$\mathfrak{IC}(x, y, z, t) = \epsilon^{Qt/(M-\gamma)} u(x, y, z) + \epsilon^{Qt/(M-\gamma)} \int \epsilon^{-(Qt/(M-\gamma))} \left[R + m'(t) + \frac{Q}{M - \gamma} m(t) \right] dt \tag{98}$$

where $u(x, y, z)$ is any harmonic function.

In the absence of recombination, Possibilities 1 and 2 lead to the same result: Equation (97) and

$$\mathfrak{IC}(x, y, z, t) = u(x, y, z) + m(t). \tag{99}$$

In the absence of time variation, (86) shows that recombination is necessarily absent, too, so the results reduce to

$$\mathfrak{U}(\mathfrak{IC}) = \tilde{A}\Lambda\left(\frac{\tilde{B} - \mathfrak{IC}}{\tilde{A}}\right) \tag{100}$$

with $\mathfrak{IC}(x, y, z)$ any harmonic function and $\tilde{A}$ and $\tilde{B}$ arbitrary constants. This solution for the case grad $\mathfrak{IC} \neq 0$, together with that given by (59b) and (60) (with $G = 0$) for the case grad $\mathfrak{IC} = 0$, constitute a veritable gold mine of useful solutions because of the arbitrary harmonic function involved. An example involving a particular choice of $\mathfrak{IC}$ will be examined in Section R.

Case 3:

$$\frac{\partial^2 \mathfrak{IC}}{\partial h^2} \neq 0, \qquad \text{grad (grad h)}^2 = 0.$$

In this case $(\text{grad } h)^2$ is a function of t so that (75) can be written in the form

$$\frac{\partial h}{\partial t} = \phi(h, t)$$

From this it follows (because div grad $h = 0$) that

$$h(x, y, z, t) = \tilde{a}(t)\tilde{b}(x, y, z) + \tilde{c}(t) \tag{101}$$

with

$$\text{div grad } \tilde{b}(x, y, z) = 0. \tag{102}$$

The condition grad $(\text{grad } h)^2 = 0$ now requires further that

$$\text{grad (grad } \tilde{b})^2 = 0. \tag{103}$$

But any $\tilde{b}(x, y, z)$ satisfying both (100) and (101) can, by suitable choice of axes, be written

$$\tilde{b} = Sx \qquad (S\text{: constant}).$$

This leaves us with exactly the same totality of solutions as we could have obtained by setting $\mathfrak{U} = \mathfrak{U}(x, t), \mathfrak{IC} = \mathfrak{IC}(x, t)$ in the first place. So we replace h by x in (74) and (75) and obtain:

$$\frac{\partial \mathfrak{IC}}{\partial x} = \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}} \tag{104}$$

and

$$\frac{\partial}{\partial x}\left[\frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}}\right] + \frac{\alpha}{N}\left[\mathfrak{R}(\mathfrak{U}) + \frac{eN}{2kT}\frac{\partial \mathfrak{U}}{\partial t}\right] = 0. \tag{105}$$

Any $\mathfrak{U}(x, t)$ satisfying (105) can be substituted into (104) to obtain $\mathfrak{IC}(x, t)$ from

$$\mathfrak{IC}(x, t) = \tilde{f}(t) + \int^{(x)} \frac{j(t) - \mathfrak{U}\dfrac{\partial \mathfrak{U}}{\partial x}}{\gamma + \mathfrak{U}}\, dx \tag{106}$$

($\tilde{f}(t)$: arbitrary function).

If recombination is absent, $\mathfrak{R}(\mathfrak{U})$ disappears from (105). If time variation is absent, $\dfrac{\partial \mathfrak{U}}{\partial t}$ disappears from (103) and $j(t)$ and $\tilde{f}(t)$ are replaced by arbitrary constants. In the latter case, the standard change of variables

$$\mathfrak{W}(\mathfrak{U}) \quad \text{for} \quad \frac{d\mathfrak{U}}{dx} \qquad \mathfrak{W}(\mathfrak{U})\frac{d}{d\mathfrak{U}} \quad \text{for} \quad \frac{d}{dx} \tag{107}$$

reduces the solution of the second order equation (105) to the solution of a first-order equation followed by a quadrature. If both recombination and time variation are absent, the substitution (107) reduces the solution of (105) to two quadratures.

A set of equations equivalent to the steady-state $\left(\dfrac{\partial}{\partial t} \equiv 0\right)$ forms of (104) and (105) has been the subject of an extensive numerical investigation by W. van Roosbroeck (Reference 1) for the recombination rate functions given in (37) and (38).

M. Solutions with $\mathfrak{V} = \mathfrak{V}(h, t)$, $\mathfrak{P} = \mathfrak{P}(h, t)$, Grad $\mathfrak{P} \neq 0$, Div Grad $h = 0$, $N = 0$

For these conditions (21) and (23b) yield

$$\frac{\partial^2 \mathfrak{P}}{\partial h^2}(\text{grad } h)^2 = \frac{\alpha e}{kT}\left[\mathfrak{R}(\mathfrak{P}) + \frac{1}{2}\left(\frac{\partial \mathfrak{P}}{\partial h}\frac{\partial h}{\partial t} + \frac{\partial \mathfrak{P}}{\partial t}\right)\right] \tag{107}$$

and

$$\frac{\partial}{\partial h}\left[\frac{\partial \mathfrak{P}}{\partial h} - \frac{\alpha e}{\beta kT}\mathfrak{P}\frac{\partial \mathfrak{V}}{\partial h}\right](\text{grad } h)^2 = 0. \tag{108}$$

Since we do not here allow grad $h = 0$, (108) implies

$$\frac{\partial \mathcal{U}}{\partial h} = \frac{\gamma\left[\frac{\partial \mathcal{P}}{\partial h} - \tilde{g}(t)\right]}{\mathcal{P}} \qquad \gamma \equiv \frac{\beta k T}{\alpha e} \qquad (\tilde{g}(t)\text{: arbitrary function}) \tag{109}$$

Case 1:

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} \neq 0, \qquad \text{grad (grad } h)^2 \neq 0.$$

In this case, as in the associated case in Section L, the implications of (107) together with

$$\text{div grad } h(x, y, z, t) = 0$$

are not known when time variation is present.

When time variation is absent, we work with the conditions

$$\mathcal{P} = \mathcal{P}(h) \quad \text{and} \quad \mathcal{U} = \mathcal{U}(h)$$

with

$$\text{div grad } h(x, y, z) = 0$$

and arrive at counterparts of (107) and (108):

$$\mathcal{P}'' \cdot (\text{grad } h)^2 = \frac{\alpha e}{kT} \mathcal{R}(\mathcal{P}) \tag{110}$$

and

$$\left(\mathcal{P}' - \frac{1}{\gamma} \mathcal{P}\mathcal{U}'\right)' = 0. \tag{111}$$

Proceeding as in the analysis of Case 1 of Section L, we infer that h must be of the kind given by (81b) or (81c). The associated second-order differential equations restricting $\mathcal{P}(h)$ are then, respectively:

$$\mathcal{P}'' - \frac{\alpha e}{kT} \epsilon^{-2h} \mathcal{R}(\mathcal{P}) = 0 \tag{112a}$$

and

$$\mathcal{P}'' - \frac{\alpha e}{kT} \frac{1}{h^4} \mathcal{R}(\mathcal{P}) = 0. \tag{112b}$$

The $\mathcal{U}(h)$ associated with any solution of (112) can be obtained by integration from

$$\mathcal{U}(h) = \tilde{C} + \int \frac{\gamma \mathcal{P}' - \tilde{D}}{\mathcal{P}} \, dh \qquad (\tilde{C}, \tilde{D}\text{: arbitrary constants}). \tag{113}$$

It will be noted from (110) that simultaneous absence of recombination and time variation is inconsistent with the defining conditions of this case.

Case 2:

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} = 0.$$

We shall exclude the possibility of $\frac{\partial \mathcal{P}}{\partial h} = 0$ because it is included in Section I. Then, proceeding as in Case 2 of Section L, we conclude that $\mathcal{P}$ itself is a harmonic function and can be used in place of h. (107) and (109) then become

$$\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t} = 0 \tag{114}$$

and

$$\frac{\partial \mathcal{V}}{\partial \mathcal{P}} = \frac{\gamma[1 - \tilde{g}(t)]}{\mathcal{P}} \tag{115}$$

or

$$\mathcal{V}(\mathcal{P}, t) = \gamma[1 - \tilde{g}(t)] \ln \mathcal{P} + \tilde{j}(t) \qquad (\tilde{j}(t): \text{arbitrary function}). \tag{116}$$

Because $\frac{\partial \mathcal{P}}{\partial t}$ is harmonic and a function of $\mathcal{P}$, we have

$$2\mathcal{R}(\mathcal{P}) = -\frac{\partial \mathcal{P}}{\partial t} = E\mathcal{P} - \tilde{F}$$

($\tilde{E}$, $\tilde{F}$: arbitrary constants)

whence

$$2\mathcal{R}(\mathcal{P}) = \tilde{E}\mathcal{P} - \tilde{F} \tag{117}$$

and

$$\mathcal{P}(x, y, z, t) = \epsilon^{-\tilde{E}t}\tilde{m}(x, y, z) - \frac{\tilde{F}}{\tilde{E}} \qquad (E \neq 0) \tag{118a}$$

or

$$\mathcal{P}(x, y, z, t) = \tilde{m}(x, y, z) + \tilde{F}t \qquad (\tilde{E} = 0) \tag{118b}$$

where $\tilde{m}(x, y, z)$ is an arbitrary *harmonic* function.

If recombination is absent, these results specialize to (116) and (118b)

with $\tilde{F} = 0$. If time variation is absent it follows from (114) that recombination is absent, too, and the results specialize to

$$\mathcal{U}(\mathcal{P}) = \tilde{G} + \tilde{H} \ln \mathcal{P} \tag{119}$$

with $\mathcal{P}(x, y, z)$ any harmonic function and $\tilde{G}$ and $\tilde{H}$ arbitrary constants. These solutions play the same role for the intrinsic semiconductor ($N = 0$) that (100) does for the extrinsic ($N \neq 0$).

Case 3:

$$\frac{\partial^2 \mathcal{P}}{\partial h^2} \neq 0, \qquad \text{grad (grad } h)^2 = 0.$$

In this case it can be shown, just as in Case 3 of Section L, that no generality is lost by considering $\mathcal{P} = \mathcal{P}(x, t)$ and $\mathcal{U} = \mathcal{U}(x, t)$ in place of $\mathcal{P}(h, t)$ and $\mathcal{U}(h, t)$. Equations (107) and (109) then become

$$\frac{\partial^2 \mathcal{P}}{\partial x^2} = \frac{\alpha e}{kT}\left[\mathcal{R}(\mathcal{P}) + \frac{1}{2}\frac{\partial \mathcal{P}}{\partial t}\right] \tag{120}$$

and

$$\frac{\partial \mathcal{U}}{\partial x} = \frac{\gamma\left[\dfrac{\partial \mathcal{P}}{\partial x} - \tilde{g}(t)\right]}{\mathcal{P}}. \tag{121}$$

Any solution of (120) when substituted into (121) gives an associated $\mathcal{U}$ from

$$\mathcal{U}(x, t) = \tilde{q}(t) + \gamma \int^{(x)} \frac{\dfrac{\partial \mathcal{P}}{\partial x} - \tilde{g}(t)}{\mathcal{P}}\, dx. \tag{122}$$

If recombination is absent, $\mathcal{R}(\mathcal{P})$ merely vanishes from (120). If time variation is absent, the functions $\tilde{g}(t)$ and $\tilde{q}(t)$ are replaced by arbitrary constants and the standard change of variables

$$\begin{aligned} &\mathfrak{u}(\mathcal{P}) \quad \text{for} \quad \frac{d\mathcal{P}}{dx} \\ &\mathfrak{u}(\mathcal{P})\frac{d}{d\mathcal{P}} \quad \text{for} \quad \frac{d}{dx} \end{aligned} \tag{123}$$

leads to a solution of (120) in two quadratures. An equivalent solution is given by W. van Roosbroeck in Reference 1. From (120) it follows that recombination and time variation cannot simultaneously be absent for Case 3.

N. Construction of Solutions from Orthogonal Harmonic Fields, $N \neq 0$

There are many known examples of pairs of harmonic functions $h_1(x, y, z)$ and $h_2(x, y, z)$ that have orthogonal vector fields—that is, for which

$$\text{grad } h_1 \cdot \text{grad } h_2 = 0 \tag{124}$$

with grad $h_1 \not\equiv 0$ and grad $h_2 \not\equiv 0$. [E.g., the real and imaginary parts of any analytic function of a complex variable.] From any such pair of functions we can construct the following solutions of (33) and (34):

$$\mathfrak{U} = h_1; \qquad \mathfrak{IC} = h_2 - h_1 \tag{125}$$

and

$$\mathfrak{U} = \sqrt{h_1}; \qquad \mathfrak{IC} = h_2. \tag{126}$$

In terms of $\mathcal{P}$ and $\mathcal{V}$ these solutions are

$$\mathcal{P} = \frac{Ne}{kT} h_1; \qquad \mathcal{V} = h_2 \tag{127}$$

and

$$\mathcal{P} = \frac{Ne}{kT} \sqrt{h_1}; \qquad \mathcal{V} = \sqrt{h_1} + h_2. \tag{128}$$

The validity of the solution (125) is seen from (33) and this expanded form of (34):

$$\mathfrak{U} \text{ div grad } \mathfrak{U} + \text{grad } \mathfrak{U} \cdot \text{grad } (\mathfrak{U} + \mathfrak{IC}) = 0. \tag{129}$$

Similarly, the validity of (126) follows from (33) together with a different expansion of (34):

$$\text{div grad } \mathfrak{U}^2 + 2 \text{ grad } \mathfrak{U} \cdot \text{grad } \mathfrak{IC} = 0. \tag{130}$$

It is evident that a given h_1 and h_2 can be interchanged in the above solutions to yield different solutions, and also that any given h_1 or h_2 can be replaced by an arbitrary constant multiple of itself plus a second arbitrary constant.

O. Construction of Solutions from Orthogonal Harmonic Fields, $N = 0$

We can write the differential equation system for the intrinsic semiconductor [(35) and (36)] in the form:

$$\text{div grad } \mathcal{P} = 0 \tag{131}$$

$$\mathcal{P} \text{ div grad } \mathcal{V} + \text{grad } \mathcal{P} \cdot \text{grad } \mathcal{V} = 0. \tag{132}$$

From these we verify the solution:

$$\mathcal{P} = h_1 ; \qquad \mathcal{U} = h_2 \tag{133}$$

for any harmonic h_1 and h_2 satisfying (124).

The solutions given by (127) and (133) have the property

$$\text{grad } \mathcal{P} \cdot \text{grad } \mathcal{U} = 0$$

and so may be considered, in a sense, complementary to the solutions in Sections L and M for which

$$\text{grad } \mathcal{P} \times \text{grad } \mathcal{U} = 0.$$

P. Superposition of a Harmonic $\mathfrak{C}$ Field, $N \neq 0$

Inspection of the equation system [(33), (130)] reveals the following superposition theorem for obtaining new solutions from some known solutions *for the case of no recombination or time variation:*

Theorem 12: If $[\tilde{\mathfrak{N}}, \tilde{\mathfrak{C}}]$ is a known solution and if h is any harmonic function such that $\text{grad } \tilde{\mathfrak{N}} \cdot \text{grad } h = 0$, then $[\mathfrak{N}, \tilde{\mathfrak{C}} + h]$ is also a solution.

Or, in terms of $\mathcal{P}$ and $\mathcal{U}$:

Theorem 12′: If $[\tilde{\mathcal{P}}, \tilde{\mathcal{U}}]$ is a known solution and if h is any harmonic function such that $\text{grad } \tilde{\mathcal{P}} \cdot \text{grad } h = 0$, then $[\tilde{\mathcal{P}}, \tilde{\mathcal{U}} + h]$ is also a solution.

In the latter form it is evident from Section O that the theorem holds also for $N = 0$, but does not extend the results of Section O.

Q. A Partial Differential Equation in Terms of $\mathfrak{C}$ Alone, $N \neq 0$

For $N = 0$, (21) provides a differential equation involving only one dependent variable—$\mathcal{P}$. We shall now derive an analogous—but vastly more complicated—differential equation *for the case* $N \neq 0$, $\dfrac{\partial \mathfrak{N}}{\partial t} = 0$.

For this case (30) and (32) become

$$\text{div grad } \mathfrak{C} = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{N})$$

and

$$\text{div}\left[\text{grad } \mathfrak{C} + \frac{1}{\gamma} \mathfrak{N} \text{ grad } (\mathfrak{N} + \mathfrak{C})\right] = 0,$$

or in terms of a familiar vector symbolism

$$\nabla^2 \mathfrak{IC} = -\frac{\alpha}{N} \mathfrak{R}(\mathfrak{U}) \tag{134}$$

and

$$\nabla \cdot \left[\nabla \mathfrak{IC} + \frac{1}{\gamma} \mathfrak{U} \nabla (\mathfrak{U} + \mathfrak{IC}) \right] = 0. \tag{135}$$

Now let $\mathfrak{S}(\mathfrak{U})$ be the inverse function to $\mathfrak{R}(\mathfrak{U})$, i.e. the function such that

$$\mathfrak{S}(\mathfrak{R}(\mathfrak{U})) \equiv \mathfrak{U}.$$

Then from (134) we have

$$\mathfrak{U} = \mathfrak{S}\left(-\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right). \tag{136}$$

Substitution of (136) into (135) yields after some computation

$$\begin{aligned} \mathfrak{S}\mathfrak{S}' \nabla^2 (\nabla^2 \mathfrak{IC}) - \frac{N}{\alpha} (\mathfrak{S}\mathfrak{S}'' + \mathfrak{S}'^2)(\nabla \nabla^2 \mathfrak{IC})^2 \\ + \mathfrak{S}' \nabla \mathfrak{IC} \cdot \nabla \nabla^2 \mathfrak{IC} - \frac{\alpha}{N} (\mathfrak{S} + \gamma) \nabla^2 \mathfrak{IC} = 0 \end{aligned} \tag{137}$$

where $\mathfrak{S}'(\psi) \equiv \dfrac{d}{d\psi} \mathfrak{S}(\psi)$, etc.

$\mathfrak{S}$, $\mathfrak{S}'$, $\mathfrak{S}''$ are considered as given functions of $\left(-\dfrac{N}{\alpha} \nabla^2 \mathfrak{IC} \right)$.

The simplest meaningful choice of $\mathfrak{S}$ is

$$\mathfrak{S}\left(-\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right) = \frac{\alpha}{N} \tilde{J} \cdot \left(-\frac{N}{\alpha} \nabla^2 \mathfrak{IC} \right) + \tilde{K} \tag{138}$$

(J, $\tilde{K}$: prescribed constants)

corresponding to constant mean lifetime recombination. For this $\mathfrak{S}$, (137) specializes to

$$\begin{aligned} \tilde{J}(\tilde{K} - J \nabla^2 \mathfrak{IC}) \nabla^2 (\nabla^2 \mathfrak{IC}) - \tilde{J}^2 (\nabla \nabla^2 \mathfrak{IC})^2 \\ + J \nabla \mathfrak{IC} \cdot \nabla \nabla^2 \mathfrak{IC} - (\gamma + \tilde{K} - J \nabla^2 \mathfrak{IC}) \nabla^2 \mathfrak{IC} = 0. \end{aligned} \tag{139}$$

If any $\mathfrak{IC}$ can be found satisfying (139), the associated $\mathfrak{U}$ is given (from (136)) by

$$\mathfrak{U} = \tilde{J} \nabla^2 \mathfrak{IC} + \tilde{K}.$$

R. Sample Application of the Results of Section L: Spherical Symmetry, $N \neq 0$

As an example of the solutions included in the results of Section L we consider the case of a spherically symmetric field about a point (or spherical) source of current.

We take, as the most general harmonic function having spherical symmetry,

$$\mathfrak{C} = \tilde{L}\frac{1}{r} + \tilde{M} \tag{140}$$

($\tilde{L}$, $\tilde{M}$: arbitrary constants).

For the time being we shall assume $\tilde{L} \neq 0$ and $\tilde{M} \neq 0$. Then from (100) and (28) and (29) we have

$$\mathcal{U} = \tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M} - \tilde{L}/r}{\tilde{A}}\right) + M + \frac{L}{r} \tag{141}$$

and

$$\mathcal{P} = \frac{Ne}{kT}\tilde{A}\Lambda\left(\frac{\tilde{B} - M - \tilde{L}/r}{\tilde{A}}\right). \tag{142}$$

In terms of $\mathcal{U}$ and $\mathcal{P}$, (3) and (4) can be written

$$\overset{\circ}{|}_p = \frac{-\mu_p e}{2}\left[(\mathcal{P} - N)\operatorname{grad}\mathcal{U} + \frac{kT}{e}\operatorname{grad}\mathcal{P}\right] \tag{143}$$

and

$$\overset{\circ}{|}_n = \frac{-\mu_n e}{2}\left[(\mathcal{P} + N)\operatorname{grad}\mathcal{U} - \frac{kT}{e}\operatorname{grad}\mathcal{P}\right]. \tag{144}$$

which yield upon substitution of (141) and (142):

$$\overset{\circ}{|}_p = \frac{1}{2}\mu_p e N\tilde{L}\left(\frac{e}{kT}\tilde{A} - 1\right)\frac{1}{r^2}\mathfrak{r}_1 \tag{145}$$

and

$$\overset{\circ}{|}_n = \frac{1}{2}\mu_n e N\tilde{L}\left(\frac{e}{kT}\tilde{A} + 1\right)\frac{1}{r^2}\mathfrak{r}_1 \tag{146}$$

where $\mathfrak{r}_1$ is the unit radial vector. The total current density is obtained by adding (151) and (152):

$$\overset{\circ}{|} = \frac{1}{2}eN\tilde{L}\left[(\mu_n + \mu_p)\frac{e}{kT}\tilde{A} + (\mu_n - \mu_p)\right]\frac{1}{r^2}\mathfrak{r}_1. \tag{147}$$

The currents flowing are obtained from the current densities from the relation

$$I = \Omega r^2 \overset{\circ}{||} \cdot \mathfrak{r}_1$$

where Ω is the solid angle (with respect to the origin) within which the flow field lies. (If the current source is surrounded by the homogeneous semiconductor, $\Omega = 4\pi$; if it lies on a flat surface of a large slab, $\Omega = 2\pi$, etc.) So we have

$$I_p = \frac{1}{2}\Omega\mu_p eN\tilde{L}\left(\frac{e}{kT}\tilde{A} - 1\right) \tag{148}$$

$$I_n = \frac{1}{2}\Omega\mu_n eN\tilde{L}\left(\frac{e}{kT}\tilde{A} + 1\right) \tag{149}$$

$$I = \frac{1}{2}\Omega eN\tilde{L}\left[(\mu_n + \mu_p)\frac{e}{kT}\tilde{A} + (\mu_n - \mu_p)\right]. \tag{150}$$

We shall now obtain expressions for the mathematical parameters $\tilde{B}$, $\tilde{A}$, $\tilde{L}$, and $\tilde{M}$ in terms of meaningful physical quantities: I_p , I_n , $\mathcal{V}_\infty$ and $\mathcal{P}_\infty$. (Subscript ∞ refers to values of variables as r becomes very large.) We shall take our reference voltage as the voltage "at infinity" so that $\mathcal{V}_\infty = 0$. Setting $1/r = 0$ in (141) and (142) we obtain

$$0 = \tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right) + \tilde{M}$$

and

$$\mathcal{P}_\infty = \frac{Ne}{kT}\tilde{A}\Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right)$$

from which follows (for $\tilde{A} \neq 0$)

$$\tilde{B} = \tilde{A}\ln\left|\frac{\tilde{M}}{\tilde{A}} + 1\right| \tag{151}$$

and

$$\tilde{M} = -\frac{kT}{eN}\mathcal{P}_\infty. \tag{152}$$

From (148) and (149) we readily find (for $\tilde{L} \neq 0$):

$$\tilde{A} = \frac{kT}{e}\frac{\mu_p I_n + \mu_n I_p}{\mu_p I_n - \mu_n I_p} \tag{153}$$

and

$$\tilde{L} = \frac{\mu_p I_n - \mu_n I_p}{\Omega \mu_n \mu_p eN}. \tag{154}$$

Finally we substitute (152) and (153) into (154) to get

$$\tilde{B} = \frac{kT}{e} \frac{\mu_p I_n + \mu_n I_p}{\mu_p I_n - \mu_n I_p} \ln \left| \frac{(N - \mathcal{P}_\infty)\mu_p I_n + (N + \mathcal{P}_\infty)\mu_n I_p}{N(\mu_p I_n - \mu_n I_p)} \right| \tag{155}$$

Equations (152)–(155) give the desired expressions for $\tilde{M}$, $\tilde{A}$, $\tilde{L}$ and $\tilde{B}$ in terms of I_p, I_n, and $\mathcal{P}_\infty$ for $\mathcal{U}_\infty = 0$ if $\tilde{A} \neq 0$ and $\tilde{L} \neq 0$.

For $\tilde{A} = 0$ we can repeat the above steps using

$$\mathcal{U} = \tilde{B} \tag{156}$$

and

$$\mathcal{P} = \frac{Ne}{kT}(\tilde{B} - \tilde{M} - \tilde{L}/r) \tag{157}$$

in place of (141) and (142). The result for $\tilde{L} \neq 0$ and $\mathcal{U}_\infty = 0$ is

$$I_p = -\tfrac{1}{2}\Omega \mu_p eN\tilde{L} \tag{158}$$

$$I_n = \tfrac{1}{2}\Omega \mu_n eN\tilde{L} \tag{159}$$

$$I = \tfrac{1}{2}\Omega eN(\mu_n - \mu_p)\tilde{L} \tag{160}$$

with

$$\tilde{B} = 0 \tag{161}$$

$$\tilde{M} = -\frac{kT}{eN}\mathcal{P}_\infty \tag{162}$$

and

$$\tilde{L} = \frac{-2I_p}{\Omega \mu_p eN} = \frac{2I_n}{\Omega \mu_n eN} = \frac{\mu_p I_n - \mu_n I_p}{\Omega \mu_p \mu_n eN}. \tag{163}$$

It is evident that $\tilde{A} = 0$ implies $\mathcal{U} =$ constant and $\mu_p I_n + \mu_n I_p = 0$.

The condition $\tilde{L} = 0$ makes $\mathcal{K} =$ constant, so we use (62b) and (63b) and set

$$\mathcal{U} = \tilde{Q} + \sqrt{\tilde{R} + \tilde{S}/r} \tag{164}$$

and

$$\mathcal{P} = \frac{Ne}{kT}\sqrt{\tilde{R} + \tilde{S}/r} \qquad (\tilde{Q}, \tilde{R}, \tilde{S}\text{: arbitrary constants}). \tag{165}$$

From (143) and (144) we obtain

$$I_p = -\frac{\Omega N e^2 \mu_p}{4kT}\tilde{S} \tag{166}$$

$$I_n = -\frac{\Omega N e^2 \mu_n}{4kT}\tilde{S} \tag{167}$$

and

$$I = -\frac{\Omega N e^2}{4kT}(\mu_n + \mu_p)\tilde{S}. \tag{168}$$

From (164) and (165) we readily obtain for $\mathcal{V}_\infty = 0$:

$$R = \left(\frac{kT}{Ne}\mathcal{P}_\infty\right)^2 \tag{169}$$

and

$$\tilde{Q} = -\frac{kT}{Ne}\mathcal{P}_\infty. \tag{170}$$

It is evident that $\tilde{L} = 0$ corresponds to the case $\mathcal{K} =$ constant and implies $\mu_p I_n - \mu_n I_p = 0$.

The foregoing now provides a formal solution with $\mathcal{V}_\infty = 0$ for every assignment of values to $\mathcal{P}_\infty$, I_p, and I_n. There remains the question of the requirements imposed by the condition

$$n, p \geq 0$$

which is equivalent to

$$\mathcal{P} \geq |N|. \tag{171}$$

This implies first of all that $\mathcal{P}_\infty$ must be chosen $\geq |N|$.

It is instructive to look first at the case $\tilde{L} = 0$. Equation (165) shows immediately that (171) requires the choice of the positive sign for the radical for $N > 0$ and the negative for $N < 0$ to avoid $\mathcal{P}_\infty < |N|$. We further find by substitution of (166) and (169) into (165) that (171) requires

$$r \geq \left[\frac{4}{\Omega kT\mu_p(\mathcal{P}_\infty^2 - |N|^2)}\right] NI_p. \tag{172}$$

The bracketed factor is positive. Since we are interested only in non-negative values of r, (172) imposes no restriction if I_p is zero or not of the same sign as N. However, for N and I_p of the same sign, (172) establishes an inner radius inside which the solution does not satisfy (171). This may be regarded as establishing the minimum radius for an inner spherical electrode for

prescribed I_p and $\mathcal{P}_\infty$, or alternatively as limiting the possible choices of I_p and $\mathcal{P}_\infty$ for prescribed inner electrode radius. Had we chosen the constants $\tilde{Q}$, $\tilde{R}$ and $\tilde{S}$ so as to obtain prescribed values of $\mathcal{P}$ and $\mathcal{U}$ at a preselected electrode radius r_0, restrictions analogous to (172) on the *maximum* radius would appear.

For the case $\tilde{A} = 0$ the restriction analogous to (172) is

$$r \geq -\left[\frac{2}{\Omega\mu_p kT(\mathcal{P}_\infty - |N|)}\right] I_p . \tag{173}$$

Since the bracketed factor is positive, (173) provides no restriction for $I_p \geq 0$, but for $I_p < 0$ establishes a minimum radius of the kind just discussed.

For $\tilde{L}$, $\tilde{A} \neq 0$, the analog of (172) and (173) is

$$r \geq \frac{\tilde{L}/\tilde{A}}{\Lambda^{-1}\left(\frac{kT\mathcal{P}_\infty}{Ne\tilde{A}}\right) - \Lambda^{-1}\left(\frac{kT|N|}{Ne\tilde{A}}\right)} \tag{174}$$

where $\tilde{A}$ and $\tilde{L}$ are given by (153) and (154) and Λ^{-1} denotes the inverse function of Λ—i.e.,

$$\Lambda^{-1}[\Lambda(x)] \equiv x$$

or

$$\Lambda^{-1}(\Lambda) = \Lambda + \ln|\Lambda - 1| .$$

Equation (174) is a minimum radius restriction of the same kind as those obtaining for $\tilde{A} = 0$, and $\tilde{L} = 0$, but the relationship between the minimum radius r_0 and $\mathcal{P}_\infty$, I_p and I_n is considerably more complicated than in the more degenerate cases.

It will be noted that the relation

$$\frac{kT}{eN}\frac{\mathcal{P}_\infty}{\tilde{A}} = \Lambda\left(\frac{\tilde{B} - \tilde{M}}{\tilde{A}}\right)$$

(with $\tilde{A}$, $\tilde{B}$, $\tilde{M}$ given in terms of $\mathcal{P}_\infty$, I_p, I_n by (152), (153), and (154)) determines which function (Λ_1, Λ_2, or Λ_3) is to be used for Λ in any given case, because any assigned value ($\neq 0$) is taken on by one and only one of (Λ_1, Λ_2, Λ_3).

If surface recombination is negligible as well as interior recombination, this spherically symmetric solution is of use in the study of "point" contacts on a plane surface of a semiconductor. [Fig. 3 and Ref. 2.]

The results of this section can easily be duplicated for any other choice of the harmonic function $\mathcal{H}$ to obtain a great variety of specimen solutions.

Solutions based on $\mathfrak{C}$'s having a single source singularity (such as the example above) will contain four mathematical parameters, and hence will permit arbitrary selection (subject to (6)) of the physical parameters, I_p, I_n, $\mathcal{P}_\infty$, and $\mathcal{V}_\infty$. However, solutions based on $\mathfrak{C}$'s having more than one source singularity will provide only a subset of the possible assignments of the physical parameters. For example, the harmonic function associated with the electrostatic field produced by two separate point charges each equidistant from two parallel infinite plane conductors provides solutions of

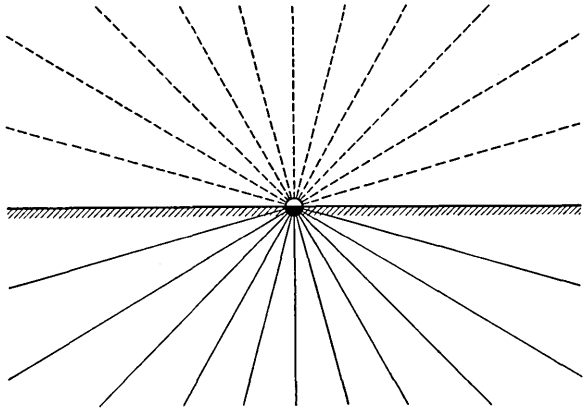


Fig. 3—Point source flow field, useful in connection with point contact theory

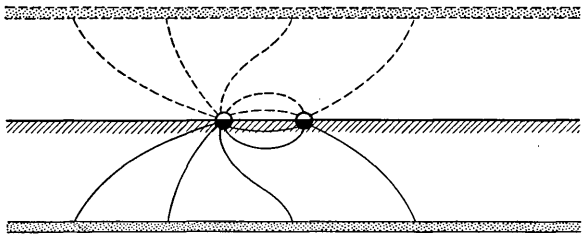


Fig. 4—Two-source flow field between conducting planes, useful in connection with Type A transistor theory.

interest in connection with the type A transistor configuration (Fig. 4). However, the family of solutions obtained contains only a five-parameter subset of the six-parameter family obtainable by arbitrary assignment of I_{p1}, I_{p2}, I_{n1}, I_{n2}, $\mathcal{P}_\infty$, and $\mathcal{V}_\infty$.

S. Sample Application of the Results of Section M: Spherical Symmetry, $N = 0$

We now round out the considerations of Section R by exhibiting the related solutions for $N = 0$ (i.e., the intrinsic semiconductor).

In accordance with the results of Section M, we choose for $\mathcal{P}$ the most general harmonic function with spherical symmetry:

$$\mathcal{P} = \hat{A}\frac{1}{r} + \hat{B} \qquad (\hat{A}, \hat{B}\text{: arbitrary constants}). \quad (175)$$

From (119) then, *for* $\hat{A} \neq 0$

$$\mathcal{U} = \tilde{H} \ln\left(\hat{A}\frac{1}{r} + \hat{B}\right) + \tilde{G} \quad (176)$$

and from (175), (176), (143), and (144)

$$I_p = \frac{1}{2}\Omega\mu_p e\hat{A}\left(\tilde{H} + \frac{kT}{e}\right) \quad (177)$$

$$I_n = \frac{1}{2}\Omega\mu_n e\hat{A}\left(\tilde{H} - \frac{kT}{e}\right) \quad (178)$$

$$I = \frac{1}{2}\Omega e\hat{A}\left[(\mu_n + \mu_p)\tilde{H} - (\mu_n - \mu_p)\frac{kT}{e}\right]. \quad (179)$$

From (177) and (178) we obtain

$$\hat{A} = \frac{\mu_n I_p - \mu_p I_n}{\Omega\mu_p\mu_n kT} \quad (180)$$

and

$$\tilde{H} = \frac{kT}{e}\frac{\mu_n I_p + \mu_p I_n}{\mu_n I_p - \mu_p I_n}, \quad (181)$$

and from (175) and (176) for $\mathcal{U}_\infty = 0$:

$$\hat{B} = \mathcal{P}_\infty \quad (182)$$

and

$$\tilde{G} = -\tilde{H}\ln\hat{B} = -\frac{kT}{e}\frac{\mu_n I_p + \mu_p I_n}{\mu_n I_p - \mu_p I_n}\ln\mathcal{P}_\infty. \quad (183)$$

The condition $\mathcal{P}_\infty \geq |N| = 0$ introduces the restriction (for $\hat{A} \neq 0$):

$$r \geq \left[\frac{1}{\Omega\mu_p\mu_n kT\mathcal{P}_\infty}\right](\mu_r I_p - \mu_p I_n). \quad (184)$$

Evidently this implies no real restriction for $\mu_n I_p - \mu_p I_n < 0$ (i.e., $\hat{A} < 0$), but introduces a minimum radius—of the same kind we have already discussed—when $\mu_n I_p - \mu_p I_n > 0$ (i.e., $\hat{A} > 0$).

For $\hat{A} = 0$, $\mathcal{P}$ is constant and, by Section I, $\mathcal{U}$ is harmonic. So we set

$$\mathcal{P} = \mathcal{P}_\infty > 0 \tag{185}$$

and

$$\mathcal{U} = \hat{C}\frac{1}{r} + \hat{D} \tag{186}$$

and obtain from (143) and (144)

$$I_p = \tfrac{1}{2}\Omega\mu_p e\hat{C}\mathcal{P}_\infty \tag{187}$$

and

$$I_n = \tfrac{1}{2}\Omega\mu_n e\hat{C}\mathcal{P}_\infty\,. \tag{188}$$

From (187) and (188):

$$\hat{C} = \frac{2I_p}{\Omega\mu_p e\mathcal{P}_\infty} = \frac{2I_n}{\Omega\mu_n e\mathcal{P}_\infty} = \frac{\mu_n I_p + \mu_p I_n}{\Omega\mu_n\mu_p e\mathcal{P}_\infty} \tag{189}$$

and from (186) for $\mathcal{U}_\infty = 0$,

$$\hat{D} = 0. \tag{190}$$

Evidently $\hat{A} = 0$ is associated with the condition

$$\mu_n I_p - \mu_p I_n = 0.$$

T. Summary List of Symbols

Coordinate Systems:

(x, y, z): ordinary rectangular cartesian coordinates.
(ρ, θ, z): ordinary circular cylindrical coordinates.
(r, θ, ϕ): ordinary spherical polar coordinates.
$\mathfrak{r}_1$: unit radial vector in (r, θ, ϕ).
t : time variable.

Physical Variables:

n : concentration of negative carriers (electrons).
p : concentration of positive carriers (holes).
$\mathcal{P}$: total carrier concentration $\equiv n + p$.
$\mathcal{U}$: $\equiv \dfrac{kT}{eN}\mathcal{P}$ $\quad (N \neq 0)$.
$\mathcal{R}$: recombination rate function.
$\mathcal{U}$: electrostatic potential.

$\mathfrak{JC}$: $\equiv \mathfrak{V} - \mathfrak{U} = \mathfrak{V} - \frac{kT}{eN}\mathfrak{P} \qquad (N \neq 0)$.

$\overset{\circ}{||}$: total current density vector.

$\overset{\circ}{||}_n$: electron current density vector.

$\overset{\circ}{||}_p$: hole current density vector.

subscript "0": designates thermal equilibrium values.

subscript "∞": designates values "at infinity".

Physical Constants:

T: absolute temperature.

e : magnitude of electronic charge.

k : Boltzmann's constant.

μ_n: electron mobility constant.

μ_p: hole mobility constant.

α : $\equiv 1/\mu_p + 1/\mu_n$.

β : $\equiv 1/\mu_p - 1/\mu_n$. (Assumed $\neq 0$)

γ : $\equiv \frac{\beta kT}{\alpha e}$

N: $\equiv n_0 - p_0$.

Other Constants and Functions:

$A, B, \cdots, Z$ ((except I, N, and T)),

$\tilde{A}, \tilde{B}, \cdots, \tilde{Z}$,

$\hat{A}, \hat{B}, \cdots, \hat{Z}$: arbitrary constants

$a, b, \cdots, z$ ((except e, h, k, n, p, r, t, x, y, z)),

$\tilde{a}, \tilde{b}, \cdots, \tilde{z}$: arbitrary functions of variables designated (e.g., $j(l)$).

h, h_1, h_2 : harmonic functions of variables designated at place of usage.

Λ: $\Lambda(x)$ is defined by the relation $\Lambda(x) + \ln|\Lambda(x) - 1| \equiv x$.
(See Figs. 1 and 2.)

$\mathfrak{S}$: $\mathfrak{S}(\mathfrak{U})$ is defined by $\mathfrak{S}[\mathfrak{R}(\mathfrak{U})] \equiv \mathfrak{U}$.

ACKNOWLEDGMENT

The author is indebted to J. Bardeen and W. van Roosbroeck for a critical reading of the manuscript and a number of valuable comments.

REFERENCES

1. W. van Roosbroeck, "Theory of the Flow of Electrons and Holes in Germanium and Other Semiconductors," Bell System Technical Journal, *29*, 4, 560–607 (October 1950).
2. J. Bardeen, "Theory of Relation Between Hole Concentration and Characteristics of Germanium Point Contacts," Bell System Technical Journal, *29*, 4, 469–495 (October 1950).
3. W. Shockley, *Electrons and Holes in Semiconductors*, New York, 1950.

Instantaneous Compandors on Narrow Band Speech Channels

By J. C. LOZIER

(*Manuscript Received Aug. 15, 1951*)

If speech is passed through an instantaneous compressor, the original speech frequency spectrum is substantially widened. It is known that instantaneously compressed speech can be transmitted over a medium with a passband no wider than that occupied by the uncompressed speech, and the original signals recovered without distortion. The conditions required for such distortionless transmission are examined. The analysis indicates that more severe requirements must be imposed on the attenuation and phase characteristics of the system when this reduced bandwidth mode of operation is used. The practical value of this exchange of transmission requirements is a matter for experimental determination.

Introduction

When a signal such as speech is instantaneously compressed in amplitude, harmonics and cross modulation products are generated which extend the frequency spectrum of the original signal by many octaves. It is proposed to demonstrate that the additional products thus generated are necessary for the distortionless recovery of the original signal. Then the conditions will be examined under which this broadband signal can be transmitted without distortion through a bandwidth no wider than that occupied by the spectrum of the uncompressed speech. Finally, some of the practical aspects of using intantaneous compandors on narrow band speech channels will be considered, with emphasis on the nature of the transmission requirements placed on the medium. The advantages to be obtained from the use of instantaneous compandors have already been presented by Mallinckrodt.[1]

Bandwidth vs Distortion

If a single frequency tone is compressed by a 2 to 1 compressor,[2] and then the fundamental alone is expanded, it can be shown that the resultant 3rd harmonic distortion is only 13 db below the fundamental. Expansion of both the fundamental and the 3rd harmonic output of the compressor will reduce this 3rd harmonic distortion to 29 db below the fundamental. Expansion of the fundamental plus the 3rd and 5th harmonics will reduce the 3rd harmonic distortion in the recovered signal to 45 db below the funda-

[1] C. O. Mallinckrodt "Instantaneous Compandors," *B.S.T.J.*, Vol. XXX, No. 3, July 1951.

[2] In a 2 to 1 compressor, the output amplitude is the square root of the input amplitude. The name comes from the fact that, in such a compressor, the output amplitude will increase 1 db for each 2 db increase in input amplitude.

mental. These results are indicative of the significance of such components in the compressor output spectrum to distortion in the recovered signal.

Frequency Analysis of Transmission under Reduced Bandwidth Conditions

It might be concluded from the results quoted above that a wideband channel is required for the distortionless transmission of instantaneously compressed speech. However, if the compressed speech is properly sampled

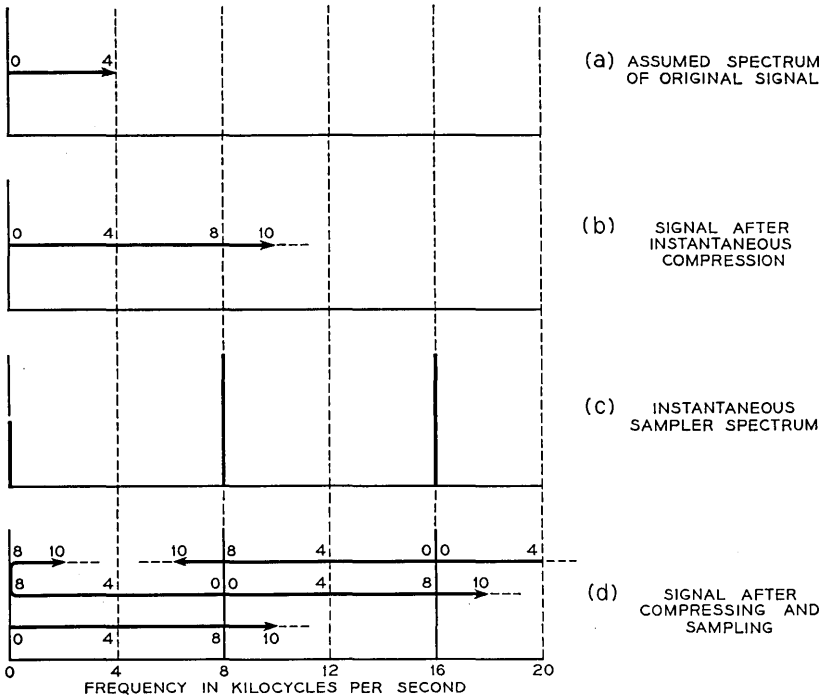


Fig. 1—Frequency analysis of instantaneous compressing and sampling of origina signal.

before transmission, and the received signals are again sampled at the receiver, the bandwidth of the intervening medium can be restricted to that of the original speech, and still the transmission can be distortionless. Hence, the sampling must transform the broadband spectrum of the compressed speech in such a way that it can be successfully transmitted over a relatively narrow band.

A steady state frequency analysis will serve to illustrate this phenomenon. Figure 1(a) shows the 4 kc frequency spectrum assumed for the original

signal, and Fig. 1(b) shows a 10 kc portion of this signal after instantaneous compression. Now the minimum sampling rate required to handle a 4 kc signal band is 8 kc. It is also the sampling rate that will allow the maximum band reduction in this case, as further analysis will show. The frequency spectrum of a sampling function with an 8 kc repetition rate has a d-c. component, an 8 kc fundamental, and all the harmonics of this repetition rate as shown in Fig. 1(c). These harmonics are all of equal amplitude and all are phased so as to add up every 125 microseconds to form the characteristic sampling waveform. Figure 1(d) shows the frequency spectrum formed by sampling the 10 kc portion of the compressed speech signal. It

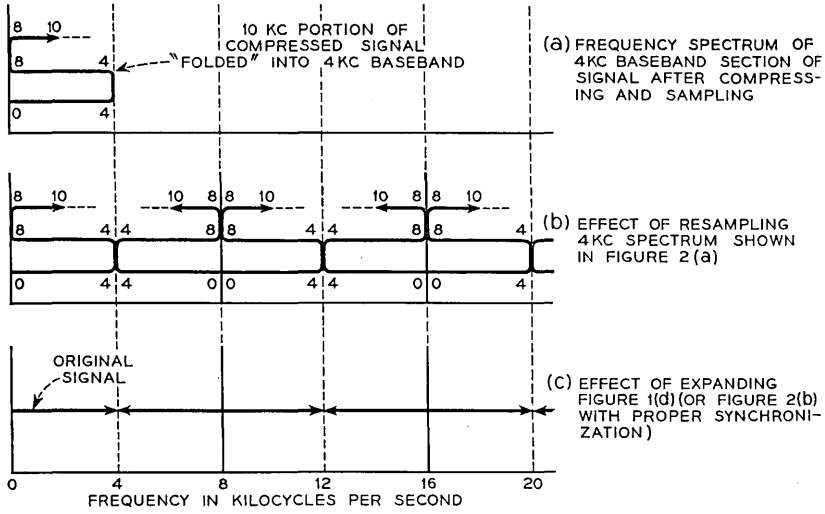


Fig. 2—Frequency analysis of instantaneous sampling and expanding of transmitted signal.

represents the product of the spectra of Fig. 1(b) and 1(c). As such, it is composed of the various component frequencies in the sampling spectrum as carrier frequencies, with the 10 kc portion of the compressed speech signals as amplitude modulated sidebands about these carriers.

Figure 2(a) shows the resulting spectrum that falls in the 4 kc baseband of Fig. 1(d). It represents that part of the compressed and sampled spectrum that would be received over an ideal 4 kc baseband channel. This spectrum is worth examining because it illustrates how the addition of instantaneous sampling makes it possible to transmit all the components in the compressed speech over a 4 kc channel. The effect may be described as a linear "folding" of the broadband spectrum back and forth over the 4 kc band. However,

although any broadband signal can be similarly folded into a 4 kc band by an instantaneous sampler with an 8 kc repetition rate, the process is not fully reversible. For example, there is no means of telling whether a 3 kc component in the folded signal comes from a 3 kc, a 5 kc, an 11 kc, or a 13 kc, etc. component in the original signal. Hence it is only a very special class of signals that can be recovered after their frequency spectra have been condensed in this fashion.

To recover the original speech in this case, the spectrum shown in Fig. 2(a) can be sampled at an 8 kc rate to produce the spectrum shown in Fig. 2(b). Now an examination of the spectra involved will show that when the second sampling is properly synchronized with the transmitting sampler, the two spectra shown in 1(d) and 2(b) will be identical. The spectrum in Fig. 1(d) represents the 8000 samples per second of the compressed speech generated at the transmitter. Thus, when the spectra of Figs. 1(d) and 2(b) are identical, samples will be recovered at the receiver which are identical to those that were generated at the transmitter. These can be converted to samples of the uncompressed speech by complementary instantaneous expansion. The spectrum of these samples is shown in Fig. 2(c). All that is necessary at this point to recover the original speech without distortion is to pass these samples through a 4 kc. low-pass filter.

Requirements for Distortionless Transmission on Reduced Bandwidth Basis

Thus the criterion for distortionless transmission of compressed and sampled speech under these conditions is that the samples recovered at the receiver be the same as the samples of compressed speech that were generated at the transmitter. This means sending 8000 pulses per second over a 4 kc band without intersymbol interference. Nyquist[3] has shown that this is the maximum rate at which independent pulses can be transmitted over a 4 kc band and still be recovered at the receiver. At this maximum rate, the bandwidth employed does not give the transient response of one pulse time to die out before the next pulse is received. Therefore the transient response in this case must be such that, when one pulse is at its peak, the transient responses of all other pulses will be going through zero. The infinitely sharp cut-off at 4 kc, which is required to separate out the spectrum shown in Fig. 2(a) from that in Fig. 1(d), will have the required zeros in its pulse response, provided the attenuation is constant and the phase is linear with frequency.

This is the familiar $\frac{\text{Sin } x}{x}$ shape of transient response. Nyquist has shown

[3] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Transactions*, Vol. 47, Pages 617 to 644, April 1928.

also that this is just one of a whole family of transmission characteristics with a specified symmetry about the cut-off frequency, all of which have the required transient zeros. However, there is no reason to suppose that any of them would prove less sensitive to variation of the transmission characteristics from the ideal than the one described above.

It is apparent that synchronization of the transmitting and receiving samplers is required to insure that the receiving sampling is done at the exact instant that all transient responses but the desired one are zero.

Effect of Variations from Ideal Transmission Characteristics on Distortion

In practice of course a certain amount of distortion is tolerable. To get some measure of the practicability of such reduced bandwidth transmission of compressed speech, the first step is to determine how much intersymbol interference can be tolerated in this type of signal, and then to translate this tolerance into allowable variations in the frequency characteristics of the transmission medium from the assumed ideal. However, it is hard to estimate what the allowable intersymbol interference might be in this case. In a single channel system, intersymbol interference produces a form of distortion, and the sensitivity of such signals to distortion is primarily a matter for subjective determination.

For computational purposes it will be assumed, however, that this intersymbol interference should be 20 db down in the output. It is apparent that a 5% variation in the amplitude of a sample before expansion will produce a 10% variation in the expanded sample. On this basis 5% intersymbol interference in the medium between transmitter and receiver is the maximum allowable. Using Wheeler's theory of paired echoes[4], it can readily be shown that a sinusoidal variation in the phase vs. frequency characteristics of the medium, with an amplitude of $\frac{1}{10}$ of a radian (5.7 degrees), will cause a pair of echoes each of which will have a peak equal to 5% of the original sample. Similarly a sinusoidal deviation in the attenuation vs. frequency characteristic of 0.9 db from the ideal will also cause a pair of echoes with an amplitude of 5%.

In estimating the average effect of such echoes, it cannot be expected that the intersymbol interference from a given echo will be appreciably less than its peak amplitude would indicate. The principal reason is that, in order to realize the savings in bandwidth, the pulses are 125 microseconds apart, which is as close together as the 4 kc band will permit. Reference to the $\frac{\text{Sin } x}{x}$

[4] H A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *I.R.E.*, June 1939.

form of transient response indicates that the width of pulses (and hence of echoes) received over a 4 kc band, is such that they will be within 65% of their peak amplitude for a full 125 microseconds. Thus such echoes will cause at least 65% of their peak interference to at least one subsequent pulse. This illustrates why it is so difficult to control intersymbol interference in pulse systems operating under minimum bandwidth conditions.

Assuming from this argument that the interference from echoes should be taken at their peak values, the tolerable phase deviations from linearity must be measured in tenths of a radian in this case. On ordinary speech channels the tolerable phase deviations from linearity are measured in radians, which represents a difference of one or two orders of magnitude.

Another estimate of the allowable intersymbol interference may be obtained by comparing it to quantizing noise on a PCM system. A 5-digit PCM system has 32 quantizing levels, and the average uncertainty in the recovered pulse amplitude is one half of a quantum step, or approximately 1.6%. The 10% intersymbol interference requirement chosen above represents approximately 6 times as much deviation in recovered pulse amplitude. Again only subjective measurements can tell whether intersymbol interference in this case is six times more tolerable than quantizing noise. However, a 5-digit PCM system is not a high quality circuit by Bell System standards.

The distortion effects due to lack of synchronization of the transmitting and receiving samplers have been ignored in the discussion so far, on the assumption that it would not prove too difficult in practice to make it a relatively negligible source of intersymbol interference. However, it may not prove to be a negligible factor from the economic standpoint, when an attempt is made to prove in a system of this type.

Multichannel Aspects

In the case of multichannel time division systems, the addition of instantaneous campandors seldom requires an increase in the transmission requirements of the medium. In multichannel PAM and PPM systems, for example, intersymbol interference causes intelligible crosstalk between channels, and the requirements on such crosstalk usually calls for the intersymbol interference to be some 60 db down in the recovered speech. In such cases the addition of an instantaneous compandor can serve to reduce this requirement on the line to some 40 db, through the so-called "Compandor Advantage"[5]. It is fair to point out, however, that such systems are seldom, if ever, operated as minimum band pulse systems.

[5] C. O. Mallinckrodt, "Instantaneous Compandors," *B.S.T.J.*, Vol. XXX, No. 3, July 1951.

Conclusions

It has been shown that distortionless transmission of instantaneously compressed speech over a frequency band no wider than that required for the uncompressed speech does involve the transmission of a broad-band signal over a relatively narrow-band channel. This is made possible by the use of an instantaneous sampler which serves to "fold" the spectrum of the compressed speech at the transmitting end so that the entire spectrum is contained within the desired bandwidth. The criterion for distortionless transmission of these "folded" signals is shown to be one of recovering at the receiving end the precise samples of compressed speech that were generated at the transmitter. To accomplish this distortionless recovery of the transmitted pulses it is necessary, first, that the transmission medium cause no intersymbol interference, and, second, that the signals at the receiver must be sampled in synchronism with the sampling at the transmitter.

It was also shown that the full reduction in bandwidth can be realized only by pulse operation under minimum bandwidth conditions. It was estimated that the accuracy of control of the steady state phase and attenuation vs. frequency characteristics that would be required to maintain the intersymbol interference below an acceptable level would be hard to meet in practice, primarily because of having to operate under such minimum bandwidth conditions.

The Evolution of Inductive Loading for Bell System Telephone Facilities

By THOMAS SHAW

(*Concluded from July 1951 issue*)

PART VI: CONTINUOUS LOADING

General

CONTINUOUS loading, i.e., the addition of uniformly distributed inductance, was studied theoretically in the Bell System several years before theoretical work started on coil loading. This early work of John Stone Stone, then a member of the headquarters technical staff of the American Bell Telephone Company, resulted in the issue to him on March 2, 1897 of a *U.S. Patent* (575,275) describing a "bi-metallic" wire cable.

Later on, when the commercial development was authorized, cost considerations made it desirable to start with laboratory experiments on an "electrically equivalent" artificial line using small lumped inductances. In planning these experiments, it soon became apparent that only a small amount of distributed inductance could be obtained with the best magnetic material then available, namely, iron. Recognition of the important advantages inherent in the use of large amounts of inductance, and of the absence of limitations regarding the magnitude of inductance that could be provided in coil form, then shifted the development emphasis to the as yet unsolved problem of spacing inductance coils in relation to wavelength. This theoretical problem was quickly solved by G. A. Campbell, who was then in charge of the project, and accordingly the laboratory artificial line was designed to demonstrate the practicability of coil-loading (early in 1899). The Bell System development work on continuous loading was then suspended for some time.

During the next two decades, coil loading was found to be economically suited to all Bell System needs for inductive loading, even on short intermediate submarine cables required at shallow water crossings of rivers and bays. Shortly after the First World War, however, it became necessary to undertake the development of continuously loaded cable to meet an urgent demand for telephone communication with Cuba. Exploratory theoretical studies and laboratory investigations had been started shortly before the war, but were discontinued during the war. The exploratory work included consideration of the possible use of a new nickel-iron magnetic alloy which

was then under development by the Research Department of the Western Electric Company, and which later on became widely known as permalloy.[23]

Key West-Havana Submarine Telephone Cable System

This project required three different submarine cables ranging in length from about 100 to 105 nautical miles, each being a great deal longer than any previously designed for telephone transmission, and a large fraction of the route was in deep water, reaching a maximum depth somewhat over 6000 ft. The difficulties to be expected in protecting loading coils from injury under the great hydrostatic pressure involved, and the complications that would be encountered during installation and in subsequent maintenance work, prevented coil loading from receiving consideration. Moreover, the great water pressure also eliminated consideration of paper insulated cable.

Since the cables were intended for use in telephone circuits connecting remote points in the United States with Havana and remote points in Cuba, the over-all system design requirements were very formidable. In addition to a two-way telephone circuit in each cable, provision also was made for three carrier telegraph circuits above the voice range, and for direct current grounded telegraphy below the voice. These complex requirements brought in difficult problems regarding telegraph flutter interference and other types of non-linear distortion.

The fundamental design studies resulted in a decision to install single core, continuously loaded, cables using gutta-percha insulation, and having a concentric system of copper tapes wrapped around the insulated conductor, for use as a return conductor. (These cables were the first to be installed with this feature.) Iron-wire type continuous loading was chosen largely because the desired project in-service dates did not allow sufficient time for the additional research and development work, and the additional manufacturing preparations, that would have been necessary in order to use permalloy tape loading. The manufacturing situation presented serious problems, because it was necessary to plan for manufacture abroad, since no American company had facilities for making deep-sea submarine cable. Moreover, iron-wire type continuous loading (as proposed by C. E. Krarup* of Denmark) was old in the European telephone art, having been used in several short submarine cables, and some underground cables.

In the Cuban Straits cables under discussion, the central copper conductor had a diameter of about 0.140 inch. About this was closely wrapped a single layer of 0.008 inch soft iron wire and three layers of gutta-percha type insulation having a total thickness of about 0.135 inch. A thin copper tape directly on this core furnished protection against damage by the teredo,

* *E.T.Z.*, April 17, 1902.

and was part of the system of copper tapes previously mentioned which served as a return conductor.

The effective permeability of the iron-wire loading material was about 115. The distributed inductance of about 4.35 millihenrys per nautical mile resulted in a low nominal impedance of about 115 ohms. The energy losses in the loading material were the principal factors in limiting the top of the working frequency band to about 4000 cycles. At 1000 cycles per second, the bare line equivalent was of the order of about 22 db (for the mean value of the longest and shortest cable). At 4000 cycles it was about 2.2 times as great.

Space limitations prevent a more complete description and discussion here. Comprehensive information regarding all features of the project is given in a 1922 *A. I. E. E.* Paper[42] prepared by Messrs. W. H. Martin, G. A. Anderegg and B. W. Kendall. Engineers of the A. T. &. T. Co. and W. E. Co. were responsible for the electrical design of the cables, method of operation, and arrangement of the repeaters and other terminal apparatus. The cables were manufactured late in 1920 and installed early in 1921 by The Telegraph Construction and Maintenance Co. Ltd. of London, for the Cuban-American Telephone and Telegraph Company. The latter organization is jointly owned by the A. T. &. T. Co. and the Cuban Telephone Co. (a subsidiary of the International T. &. T. Co.)

1930 Non-Loaded Cable: Since the 1921 cables were not suitable for carrier telephone operation (largely because of excessive losses and non-linear distortion at high frequencies), it became necessary during 1930 to install a fourth cable between Key West and Havana in order to meet the demand for additional facilities. Advantage was taken of advances in the communication art, notably an improved cable insulation (paragutta), improved repeaters and carrier telephone systems, to design a non-loaded cable system which would be suitable for carrier operation. The initial carrier set-up provided three carrier telephone circuits, using a type C4 system which had originally been developed for open-wire lines. Early in 1942, a seven-channel system was substituted. Comprehensive information regarding the 1930 cable and its use of the 3-channel carrier telephone system is given in a 1932 *A. I. E. E.* paper by Messrs. Affel, Gorton and Chesnut.[43]

High Speed Transoceanic Loaded Telegraph Cables

During the First World War when the need for increasing the message-carrying capacity of existing non-loaded transoceanic telegraph cables became urgent, the Bell System engineers who worked on this problem finally came to the conclusion that to obtain a great advance in the existing art it would be necessary to have much better cables.

In July 1919 the continuing interest in this problem crystallized in a Western Electric proposal to use permalloy continuous loading in new transoceanic telegraph cables. Since this remarkable new magnetic alloy[23] had been invented and developed by Western Electric engineers, they were already familiar with its extraordinary high permeability characteristics, and had confidence in their ability to use it in providing a high impedance loading which would make practicable a great increase in message-carrying capacity. Loading with iron-wire would not have any advantage in telegraph speed, because of its low permeability. Intensive research work quickly started on the permalloy loaded cable design and installation problems, and on the related terminal apparatus and operating problems. The success attained in these efforts resulted in disclosures to the Western Union Telegraph Company regarding the great increase in telegraph signaling speed that could be obtained with the proposed new permalloy loading. In due course the Telegraph Company made arrangements with the Telegraph Construction and Maintenance Company Ltd. of London for the manufacture and installation of a 120-mile trial length, using loading material supplied by the Western Electric Company and applied and treated under the direction of Western Electric engineers. In October 1923 this experimental length was laid in deep water near the south shore of Bermuda. The trial installation tests were so satisfactory that the Western Union company arranged for the manufacture and installation of a 2300-mile cable to connect New York with Horta in the Azores. As with the trial length, the loading material was supplied by the Western Electric Company, and it was applied and treated under Western Electric supervision.

The new cable was laid during September 1924. After refined adjustments in the terminal apparatus, a speed of over 1900 letters per minute was obtained. This speed is about four times the carrying capacity of an ordinary non-loaded cable of the same length. At this point a brief statement of general theory is indicated: The effect of the inductance is to oppose the setting up of a current and to maintain it once it has been established, thus preserving a definite wave front as the signal impulse travels over the cable. The individuality of the signal impulses is retained, and thus the much higher speed becomes possible.

The permalloy loading material was applied in tape form in a close helix around a stranded copper conductor. The tape was 0.006 inch thick and 0.125 inch wide. The alloy was composed of about 79% of nickel and 21% of iron and a small amount of manganese, suitably heat treated. It provided an inductance of about 54 millihenrys per mile, slightly over 12 times that obtained by the use of iron wire in the Cuba cables previously described. The permeability of the loading was about 2300, or about 20 times that of

the iron wire used on the Cuba cables. An important feature of the cable not previously mentioned was a layer of viscous insulating material (under the regular gutta-percha insulation) which protected the strain-sensitive permalloy from the stresses caused by hydrostatic pressure in the great depths of the ocean.

Demand for other high-speed loaded submarine cables quickly followed the successful demonstration of the New York-Horta cable and several were installed during 1926, reaching a total of about 15,000 miles of high-speed cables. The new installations included the Horta-Emden cable manufactured and installed by the Norddeutsche-Seekabelwerke A.G. for the Deutsch Atlantische Telegraphengesellschaft, and the New York-Bay Roberts-Penzance cable manufactured and installed by the T. C. & M. Company for the Western Union Telegraph Company. These particular cables used an improved form of permalloy supplied by the Western Electric Company containing about 80% nickel, 17.5% iron, 2% chromium, and 0.5% manganese. This alloy had an initial permeability of about 3700 and provided a higher impedance loading than that used on the first high-speed cable. In consequence, the newer cables were capable of speeds of about 2500 letters per minute.

Other high-speed continuously loaded cables, installed in 1926 and subsequent years, used permalloy material manufactured under Western Electric Company patent license, in some instances under a special foreign trade name.

Comprehensive information regarding all features of the high-speed cable projects specifically mentioned above is given in two papers by O. E. Buckley, published in 1925[44] and 1928[45], respectively.

In passing, it should be observed that the permalloy loaded cables under discussion were not intended for, and were not suitable for telephone communication. For this purpose, a new family of magnetic alloys, the perminvars, was developed.[46] Their composition centered on 47% nickel, 25% cobalt, 20% iron, 7.5% molybdenum, and 0.5% manganese. When used as a thin loading tape, this alloy has electrical and magnetic properties especially suitable for telephone transmission, including very low hysteresis which is very advantageous in the control of all forms of non-linear distortion.

A Proposed Transatlantic Telephone Cable

During the late 1920's, there was worked out a design of a perminvar loaded cable suitable for voice frequency telephony between Newfoundland and Ireland (1800 nautical miles). It was of the single core type with a concentric return conductor. Four layers of very thin perminvar tape

provided the loading, and the loaded conductor was insulated with paragutta. The suitability of the design for use in deep water was verified by temporarily dropping a 20-mile length on the sea floor in a deep water section of the Bay of Biscay.

The general business depression of the early 1930's resulted in a postponement of the cable project because of its great cost. Later on the project was postponed indefinitely because, in the face of improvements in transatlantic radio telephone communicaion, so expensive a cable to carry a single conversation could no longer be justified.

Additional information regarding this cable project is included in Dr. O. E. Buckley's 1942 paper, "The Future of Transoceanic Telephony," constituting the 33rd Kelvin Lecture before the Institution of Electrical Engineers.[47]

Continuous Loading for Paper Insulated Telephone Cables

Tape and Wire Loading: When permalloy and perminvar first became available, theoretical studies were undertaken to determine the prospects of economic competition with coil loading on ordinary paper insulated telephone cables. Special consideration was given to the use of the magnetic alloys in situations where coil loading is most expensive, namely, in submarine intermediate cables at river crossings, many of which involve high-frequency carrier telephone operation. None of these studies, however, gave sufficient promise to warrant commercial development work.

Electroplated Permalloy Loading: During the middle 1920's, the Bell Telephone Laboratories started research work on a radical new concept of continuous loading using electroplated permalloy, which gave some promise of being less expensive than magnetic alloy tape or wire loading. The process involved the electrolytic deposition simultaneously of suitable proportions of nickel and iron on the copper conductor, and the use of special heat treatments to obtain the desired characteristic (magnetic and electrical) properties of permalloy. In due course, methods were devised for separating the concentric magnetic layer from the conductor, and for breaking it up into longitudinally discontinuous pieces, so as to secure the most advantageous properties for telephone transmission service, and to provide mechanical flexibility in handling.

The experimental work was concentrated on small copper conductors, partly because of the more simple process problems, and partly because such combinations appeared to have the best prospects of competing with coil loading from the plant cost standpoint. (N.B.—The amount of permalloy loading material required to provide a specified inductance per unit length, and its cost, is a direct function of the conductor diameter.)

The requirements for and the possibilities of using electroplated loading in the exchange area services were given priority in the theoretical cost studies—largely because of their extensive use of small conductor cables. These studies indicated some attractive possibilities of using light-weight electroplated loading on fine wires (26 and 24-gauge) as substitutes for larger size wires without loading, provided satisfactory solutions could be worked out for the circuit balance and magnetic instability problems. The balance problem arises from the difficulty of securing sufficient uniformity among the loaded conductors used as wire and mate in the individual pairs. This is complicated by the sensitivity of the permalloy continuous loading to magnetization by steady and intermittent superposed signaling currents. On the larger-size exchange cable wires that are not now used extensively without coil loading, the comparative cost estimates were not attractive for the electroplated loading.

The inflexibility of continuous loading is an adverse general factor, since it is not feasible to decrease or increase the weight of the loading after manufacture, in order to accommodate changes in transmission requirements made desirable by changes in performance standards or alterations in circuit layouts. Also, there would be inflexibility in conforming to changing requirements in complement sizes of loaded circuits in areas where it is necessary to have loaded and non-loaded circuits within the same cable sheath.

Theoretically, one of the flexibility limitations of the continuous loading could be reduced by using coil loading in combination with it, in order to extend its transmission range. However, this would reduce the width of the transmission band below that obtainable with the coil loading on a circuit not having continuous loading—the decrease in effective cutoff being a complex function of the ratio of distributed inductance to coil inductance. Combinations of high cutoff, low impedance, coil loading with low inductance continuous loading could be designed to have satisfactory band width properties. For a given grade of transmission performance, however, such combinations appear to be inherently more expensive than coil loading or continuous loading by themselves.

The experimental work on electroplated continuous loading for exchange area cables was carried on somewhat intermittently during the 1930's. At no time did the prospects of securing satisfactory over-all transmission performance, at costs which would encourage competition with coil loading, appear to be sufficient to warrant an all-out sustained attack on the many difficult technical problems involved. Although the development project has not been permanently abandoned, it had to be discontinued in the late 1930's on account of the great pressure of more urgent work.

The use of electroplated loading as a substitute for coil loading on toll cables, or on incidental cables in open-wire lines, did not appear to be attractive when the cost estimates and the complex requirements on circuit balance, stability, non-linear distortion and flexibility were taken into account.

Summary

Enough has been told in the preceding pages to support the earlier statements regarding the low importance of continuous loading in the growth of the Bell System, relative to that of coil loading. Obviously, the success attained by the intensive development and in the very extensive use of economical types of coil loading is an important factor in this situation. That these extent-of-use relations are not due to a lack of interest in continuous loading is well demonstrated by the Bell System initiative in developing the permalloy continuous loading that made high-speed telegraphy practicable in long submarine telegraph cables, and by the other development work summarized in this review.

PART VII: EXTENT OF USE AND ECONOMIC SIGNIFICANCE

Introduction

Up to now, this account of coil loading has been in terms of individual developments and their significance with respect to the prior art and current developments in related fields, with occasional information regarding their importance and extent of use.

It is now appropriate to supply and analyze some general statistics regarding the total amount of loading which has been used, in a rough appraisal of the importance of coil loading in the growth of the Bell Telephone System. Some important qualifications of the statistics are commented upon in advance of the presentation of actual figures.

The statistics here given and discussed are for the most extensive and most important applications of coil loading, namely, for voice-frequency loading over cable circuits. They are grouped in two principal categories: (1) non-phantom type coils used on non-quadded exchange area cables, and to a relatively very small extent on toll cables, and (2) side circuit and associated phantom coils used on quadded long-distance and interurban toll cables, and to a relatively very small extent on entrance cables in open-wire lines and on long quadded exchange cables.

The figures used are based on production statistics up to the end of 1949. The important significance of the production figures is that they measure at the time of manufacture the current demands for additional loaded facilities required by the growth of the telephone system, and the up-to-

then accumulated total demand. In general, the loading coils were manufactured to meet specific customers' orders; manufacture for merchandise stock in anticipation of future orders was seldom undertaken, except during periods of extraordinarily high, sustained, demand. On this basis practically all of the coils that were manufactured were installed in the telephone plant.

The production statistics of course include a considerable number of coils which were installed shortly after manufacture and which were taken out of service many years later to facilitate the use of improved transmission systems that required different types of coils, or to permit the use of carrier systems on the unloaded toll cable circuits. In general, complete potting complements were not taken out of service in preparation for carrier systems operation; i.e., a large fraction of the disconnected loading coils remain in the cases in which they were originally potted and installed, and the other coils in the same cases are still in service. It is important to remember that the displaced loading coils played an important part in the improvement and growth of telephone service in their own period of commercial use. The unavailability of statistics regarding displaced loading makes it impossible to supply accurate information regarding the total number of loading coils now being used for regular telephone service. It seems probable, however, that about 80% or more of all the toll cable coils that have been manufactured are in service, or installed in circuits which will be used as soon as traffic growth requires them. The corresponding percentage figure for exchange area coils is probably higher. The number of loading coils taken out of service because of incipient defects that were not detected in the factory inspection tests, or which became unserviceable in consequence of service injuries, or which have been junked because of obsolescence, is a very small fraction of the total number of coils that have been manufactured for Bell System use

General Production Statistics, Voice-Frequency Cable Loading

Total Production

The grand total production figure (up to the end of 1949) for all types of voice-frequency loading coils for Bell System use is of the order of 20.7 million. Approximately 54% of this total (about 11,270,000 coils) are non-phantom type coils, used almost entirely on exchange area non-quadded cables. Nearly 9,500,000 coils are side circuit or phantom loading coils used on quadded toll and toll entrance cables. Approximately three-quarters of the grand total have been manufactured during the last two decades.

The greatly varying rates in the growth of loading coil production are shown, (a) in terms of accumulated total production through 1949 in

Table XIX and (b) in annual totals during the period 1920–1949, plotted in Fig. 35.

Annual Production Totals

In general, the average and peak figures of annual production prior to 1920 were very small relative to those in the 1920–1949 period covered by the chart. For example, the maximum annual production of side circuit and phantom toll cable coils prior to 1925 was in the war year 1918,* and the maximum annual production of non-phantom exchange area cable coils

TABLE XIX
ACCUMULATED TOTAL PRODUCTION(1)—VOICE FREQUENCY CABLE LOADING COILS (IN MILLIONS OF COILS)

At End of Year	Side Circuit(2) and Phantom Coils	Non-Phantom Coils	Total
1915	0.31	0.22	0.53
18	0.52	0.30	0.82
20	0.64	0.35	0.99
22	0.73	0.39	1.12
1924	0.95	0.53	1.48
26	1.49	0.79	2.28
28	2.69	1.32	4.01
30	5.59	2.06	7.65
1934	6.44	2.60	9.04
38	6.65	3.21	9.86
40	7.04	3.81	10.85
42	7.82	5.15	12.97
1944	8.14	5.48	13.62
46	8.49	6.69	15.18
48	9.33	9.76	19.09
49	9.46	11.27	20.72

Notes: (1) All production figures are approximate values.
(2) Commercial production of side and phantom coils did not start until 1910. Up to that time non-phantom coils were used for toll cable loading (and for exchange area cables).

prior to 1923 was in the war year 1917.† Thus, with occasional exceptions, the production data for the years prior to 1920 could not be accurately plotted on the chart without using a confusing scale.

In the beginning, the use of loading was small relative to its subsequent use because the Bell System cable plant was small. For nearly a decade the expanding toll cable plant used fewer coils than the exchange plant. From then on, in the two-decade period 1913–1932, toll cable loading dominated

* 117,000 coil peak in 1918; 187,000 coil total in 1925.
† 33,000 coil peak in 1917; 38,000 coil total in 1923.

in the extent of use, reaching its all-time peak in growth during 1930. The four-year period of most rapid expansion of toll cable loading coincided with: (a) the full scale introduction of four-wire repeatered loaded (H44-25) circuits for long haul long-distance facilities, (b) the introduction of permalloy-core loading coils which resulted in large loading economies, and (c) the planned use of relatively large circuit-groups in order to speed up the long-distance service.

The business depression of the early 1930's terminated the rapid expansion period in all types of loading. Several years later, when business conditions

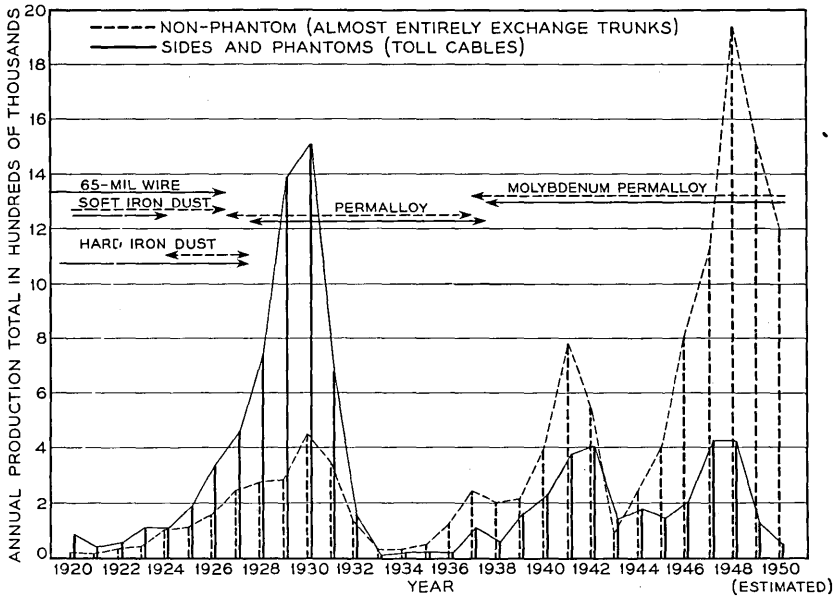


Fig. 35—Annual production totals of voice-frequency cable loading coils for Bell Telephone facilities.

improved sufficiently to require another large expansion in the toll cable facilities, the demand for new long-haul circuits was taken care of generally by the use of Type K carrier systems on non-loaded cable pairs and pairs from which loading was removed; and the use of new toll cable loading was largely restricted to short-haul repeatered and non-repeatered circuits. Thus it happened that, during the 1939–1942 period of rapid plant expansion, the production of exchange area loading coils substantially exceeded that for toll cable loading in the struggle to meet the demands for additional facilities required by the war effort.

The post-war drive to meet the greatly increased demands for long-

distance telephone service, and the provision of a tremendous amount of new exchange plant to take care of more than eleven million new Bell System telephone stations, made it necessary to build up the production rates to higher values than those during the war period. An important factor in the new heavy demands was the desire to restore the speed of service to the pre-war standards.

The post-war demand for exchange area loading has been greatly in excess of that in any previous spurt in demand, reaching its peak value during 1948, and has been very large in relation to the toll cable loading requirements. The post-war rapid build-up of a backbone network of coaxial cables, together with the expanding use of carrier systems in existing and new cables of the conventional types, and the introduction of microwave radio relay systems have held down the demand for new toll cable loading to relatively small quantities for use on relatively short circuits.

Relative Costs, Toll and Exchange Loading

Production statistics by themselves do not indicate the relative economic importance of exchange area and toll cable loading. Except in the early years when coils of the same size were used for both types of loading, the toll cable loading coils have been considerably more expensive than the exchange area loading coils. During the periods of maximum production and use portrayed in Fig. 35, the average prices per potted toll cable loading coil have ranged up to about twice or three times as large as those per potted exchange area coil. Consequently, the total plant investment in toll cable loading is substantially greater than the total investment in exchange area loading, notwithstanding the somewhat greater total use of exchange area loading, as indicated by the production statistics. This is consistent with the fact that more expensive types of cable are used for the toll circuits and the service requirements are more difficult.

Analysis in Relation to Core Materials

There now follows a rough breakdown of total production in terms of core materials, in recognition of the importance of the cores in determining the coil performance characteristics and costs:

In general, the production percentage figures in Table XX do not discriminate between types of facilities (toll or exchange area). If separate percentage-of-total figures should be derived for toll facilities and for exchange area facilities, those for toll facilities would substantially exceed the tabulated figures for total iron-wire, iron-dust, and permalloy-powder core loading coils, especially in the case of the latter, and the percentage-of-total figure for exchange area molybdenum-permalloy core coils would greatly exceed that for toll cable loading.

In considering the two different permeability types of iron-wire and of iron-dust core-materials, it is important to note that in each case the lower permeability material had a much more extensive total use than the higher permeability material, and that it was used in the more important facilities.

It is of special interest from the plant-cost standpoint that nearly two-thirds of the compressed molybdenum-permalloy powder core coils (up to the end of 1949) are the reduced cost designs using Formex-insulated conductors in their windings, this being an important factor in coil-size reduction. The other molybdenum-permalloy core coils are larger-size coils using a combination of textile and old type of enamel conductor-insulation.

It is highly significant with respect to the economics of the Bell System plant growth that over one-third of all voice-frequency loading coils manufactured up to the end of 1949 are of the lowest-cost types ever standardized

TABLE XX

ESTIMATED DISTRIBUTION OF ACCUMULATED TOTAL LOADING COIL PRODUCTION UP TO END OF 1949 IN TERMS OF CORE MATERIALS

Core Material	Percentage of Total Loading Coil Production	Approx. Period(1) of Commercial Manufacture
Fine Iron-Wire	1.5	1901–1927
Compressed Powdered-Iron	10.5	1916–1928
Compressed Powdered-Permalloy	33.	1927–1938
Compressed Powdered Molybdenum-Permalloy	54.	1937–
Non-Magnetic (Carrier loading)	2.	1920–

Note (1): For more definite dates in relation to different types of facilities, and in relation to the two different permeability values of the iron-wire and iron-dust materials, reference should be made to Table III (page 158).

for general use. This total includes about 60% of the total production (through 1949) of all types of exchange area loading coils.

Loaded Circuit Mileage Estimates

To add some substance to the significance of the production statistics on voice-frequency loading, it is desirable to record some rough estimates regarding the aggregate length of the cable circuits which have been loaded.

For exchange area loading, a weighted average coil-spacing between the 6000 ft. and 3000 ft. values now standard can be assumed. Considering the time elements in the evolution of loading practices, as discussed in Part III of this review, it is reasonable to assume an average coil-spacing somewhat longer than the mean value of the two standard spacings, say about 5000 ft. On this assumption, the aggregate loaded cable-mileage which corresponds with an assumed production total of 11,200,000 coils is of the general order of 10,500,000 pair miles.

The 3000-ft. spacing has been used much less extensively on toll cable circuits than in the exchange plant, on which basis the weighted average coil-spacing for quadded toll cable loading is somewhat longer than the weighted average value for exchange area loading. Within the accuracy required for the present general estimates, 5500 ft. seems to be a reasonable estimate for the average coil spacing in quadded toll cable loading. On this basis, and assuming a production total of about 9,500,000 side circuit and phantom coils, the aggregate loaded toll cable circuit-mileage is of the order of 9,900,000 miles. Keeping in mind the substantially universal use of quadded cables and of phantom group loading for long-distance and inter-urban toll cables, the aggregate mileage of loaded toll cable quads is of the order of 3,300,000 miles. Because of the extensive installation of loaded H 44-25 four-wire repeatered circuits during the period 1925–1931, the *loaded "facility"* mileage-aggregate is considerably less than the *loaded "circuit"* mileage-figure above given. Meanwhile, much of the loaded H 44-25 4-wire circuit mileage has been converted for short haul two-wire circuit usage, and much has been unloaded to permit the operation of Type K carrier systems. The available data on these plant changes do not permit accurate estimates regarding the mileage of loaded four-wire and two-wire types of toll cable circuits now in commercial use. It is again appropriate, however, to call attention to the important part in the growth and improvement of the telephone service which the displaced loading coils played in their own period of commercial use.

Economic Significance

Since loading has been used only when it permitted the use of cheaper facilities than would otherwise have been feasible, the great economic value of loading in the growth of the Bell Telephone System is indicated by the circuit mileage-figures given above. Other factors, however, would have to receive consideration in a complete appraisal, namely, the contributions of loading to nation-wide customer satisfaction that have resulted from improved transmission performance and higher speed of service. In turn, these factors themselves have been greatly influenced by the unit plant-cost reductions made possible by the use of loading.

For example, if loading had not been available when new or additional facilities became desirable, it is highly questionable as to whether it would have been economically feasible to work to the high-grade transmission-performance standards that have been readily achieved at reasonable costs with the cheaper loaded facilities. Moreover, it is even more questionable whether it would have been economically feasible to provide as many facilities without loading as were actually installed on a loaded basis.

Because of the speculative uncertainties involved in making assumptions regarding relative transmission-performance and relative plant-size, with and without loading, and because of the practical difficulties involved in evaluating in monetary terms the differences in transmission performance and in speed of service, no complete appraisal of the economic value of coil loading has ever been attempted for the exchange area plant. These, and additional special complications subsequently discussed, have also prevented accurate appraisals of the economic value of toll cable loading.

Exchange Area Loading

During the first two decades or so of the use of exchange area loading, rough estimates of its economic significance were sometimes made by comparing the total costs of the loaded facilities with the much higher cost of the non-loaded cable plant which otherwise would have been required to meet the same trunk-loss limits at 800 or 1000 cycles. Depending on the period under study, the estimated aggregate plant-cost reduction figures ranged up to and beyond $100,000,000. These estimates included the plant-cost reductions that resulted from the use of less expensive pole lines for aerial cables, and less expensive conduit systems made possible by utilizing a smaller total number of cables, each having a larger number of pairs. If similar studies should be made now, the corresponding hypothetical plant-cost reduction figure would probably be many times as large as the figure previously mentioned. These figures ignore the superior over-all transmission in loaded trunk plant that results from the much more favorable distortion characteristics. Also they assume equal sizes of trunk plant, with and without loading. Because of these qualifications, and because of the magnitude of the cost-reduction estimates, it is difficult to define their real significance.

A better understanding may perhaps be obtained from consideration of the cable data given in Table XXI, following. This compares some of the most important types of cable on which loading has been used with the types which would probably have been required for transmission reasons, if loading had not been available.

The large savings which loading permitted in the use of cable copper and in the amount of lead sheath per cable pair, are indirectly indicated by the tabulated data. Moreover, with loading on finer-wire cables a given total number of facilities can be provided with a much smaller total number of cables, thus permitting the use of less expensive conduit systems. This factor is extremely important in some routes of congested sections of large metropolitan areas such as Manhattan and the loop section in Chicago, where there might well be a question as to the *physical practicability*, dis-

regarding costs, of installing enough large-conductor, non-loaded cables to provide as many facilities as those made available in existing loaded small-conductor cables.

Toll Cable Loading: An accurate appraisal of the economic value of toll cable loading would have the specific complications mentioned above in the discussion of exchange area loading, and in addition certain intricate difficulties briefly discussed below.

In the aggregate, a very much larger amount of loading has been used on repeatered facilities than on non-repeatered voice-frequency circuits. The over-all plant-cost reduction and the transmission and speed of service

TABLE XXI
LOADED AND NON-LOADED EXCHANGE AREA CABLES
RELATIVE USE OF DIFFERENT TYPES

Degree of Use[a]	Loaded Exchange Area Cable			Alternative Types of Non-Loaded Cables		
	Conductor Size B & S ga.	Weight-Lbs.[1] Copper Pair-Mile	No. Pairs Full Size Cable	Conductor Size B & S ga.	Weight-Lbs.[1] Copper Pair-Mile	No. Pairs Full Size Cable
Very Extensive	22	21.0	909[2]	19	42.0	455[2]
				16	84.3	152[3]
Very Extensive	24	13.2	1515[2]	22	21.0	909[2]
				19	42.0	455[2]
Substantial	19	42.0	455[2]	16	84.3	152[3]
				13	168.8	75[3]
Small	26	8.3	2121[2]	24	13.2	1515[2]
				22	21.0	909[2]

Notes: (1) These weights include a small allowance for the effect of pair-twist and stranding, in increasing the conductor length, relative to the cable sheath-length.
(2) High-capacitance cables–(approx.) 0.082 (±) mf/mi.
(3) Low-capacitance cables–(approx.) 0.066 mf/mi.
(a) In the very extensive installations of exchange area loading during the 1928–1949 period, a very large fraction of the total use was on 22 and 24-gauge cables in nearly equal quantities.

improvements that have resulted from the use of loading in combination with voice-frequency repeaters must of course be jointly credited to the repeaters and the loading. Since as yet no rationally acceptable procedure for allocating the pro-ratio credits has evolved, very questionable arbitrary allocations would become necessary. Moreover, very debatable uncertainties would be involved in making assumptions regarding the types of facilities which would have been employed if loading and repeaters could not have been jointly used on small-gauge toll cable conductors.

In appraising the economic importance of toll cable loading it is therefore necessary to revert to general terms, namely, its great extent of use as

indicated by the previously discussed production and circuit-mileage statistics.

In short-haul, non-repeatered, toll cable circuits, loaded 19 ga. conductors are generally used for service which would have required 16 or 13 gauge conductors without loading. The plant-cost savings in cable, copper, and lead are much greater per unit length than the average savings realized in the loaded exchange area cables. The aggregate mileage in this type of toll plant, however, is but a small fraction of that in the loaded exchange area plant.

Until the commercial exploitations of lower-cost carrier telephone systems started during the late 1930's, the loaded repeatered voice-frequency cable facilities satisfactorily met the quantitative and qualitative needs for the rapidly expanding long-distance telephone services along dense traffic routes where the use or the extension and expansion of the open-wire plant would have been unduly expensive, even on a carrier basis. In such backbone routes, and also along slow-growing tributary routes, and for short-haul toll facilities, the repeatered and non-repeatered loaded toll cables have provided more economical service than could have been obtained in an open-wire plant, and with increased dependability. Also, as previously indicated, larger circuit groups have been economically feasible, with valuable results as regards the speed of service.

In concluding this part of the review, it is noteworthy that the phantom-group loading almost universally used on voice-frequency repeatered and non-repeatered toll cable facilities is a major factor in the plant economies that have resulted from the commercial exploitation of the phantom working principle. These particular plant-cost savings constitute an important contribution to the aggregate economies achieved by toll cable loading.

PART VIII: SUMMARY AND CONCLUSION

General

The story of coil loading told in the present review is one of continuing evolution whereby its inherent capabilities have been substantially realized in its adaptation to the growing and changing needs of exchange area facilities and of interurban and long-distance communications by wires, throughout the Bell Telephone System. Also, full advantage has been taken of the opportunities offered by the development of better core-materials and new manufacturing techniques and tools to improve the loading apparatus and reduce its cost.

It was inevitable that by far the most important uses of coil loading would be for voice-frequency telephony over cable circuits. The very low

ratio of distributed inductance to distributed capacitance, incidentally resulting in low impedances, and the relatively high conductor resistances of cable circuits, gave loading its greatest opportunities in exercising its natural functions of reducing the circuit attenuation and attenuation-frequency distortion. Clearly appreciated from the beginning, these possibilities have been advantageously realized to a very great extent, and they still have substantial economic importance for future voice-frequency applications in the continuing growth of the exchange area non-quadded cable plant, and short, quadded interurban toll cables.

Open-Wire Loading

The higher ratios of distributed inductance to distributed capacitance in the open-wire lines made the reduction of attenuation-frequency distortion a relatively minor objective in the use of loading, attenuation reduction being the primary objective. Incidentally, the relatively high impedances of the non-loaded lines that resulted from their higher ratios of inductance to capacitance limited the attenuation reduction obtainable by coil loading to smaller percentage values than those obtainable on cable circuits. However, full advantage of these important, though limited, possibilities was realized in the expanding open-wire plant during the decade that preceded the commercial introduction of vacuum-tube repeaters. The early uses of these repeaters on open-wire lines were on circuits having improved loading designed especially for use in conjunction with repeaters. In 1915, this combination of loading and repeaters made transcontinental telephony economically feasible, and for several years greatly increased the demand for loading. The importance of open-wire loading soon started to decline, however, as a result of improvements in the repeaters, their circuits, and auxiliary networks, which made it possible to secure considerably better voice-frequency transmission on long lines at a lower total cost by discarding loading and using more repeaters. The climactic event in this new trend was the beginning of the operation of the first transcontinental circuits on a non-loaded basis during 1920. During the middle and late 1920's the general removal of open-wire loading was expedited to increase the plant flexibility and facilitate the commercial exploitation of carrier telephone and telegraph systems over non-loaded lines.

Since, for transmission-cost reasons, it is not feasible to develop suitable loading for long lines over which carrier systems are operated, there is no reason to expect any new leases of life for open-wire coil loading. Notwithstanding its small extent of use relative to that for cable loading, and the relatively short period during which it was standard practice, open-wire loading was a necessary and a vitally important factor in the rapid expansion of long-distance telephony that began nearly five decades ago.

Toll Cable Loading

The pattern of the commercial evolution of loading practices for long-distance cable systems has been generally similar to that for open-wire loading, but with important quantitative and qualitative differences, and especially in the relative time-elements. These various differences have been mainly due to the previously mentioned inherent differences in the basic transmission properties of non-loaded cables and non-loaded open-wire lines.

Prior to the availability of vacuum-tube repeaters, loading was an essential factor in the establishment of a very important expanding network of storm-proof, intercity, toll cables; coarse-gauge conductors and expensive coils were used for distances ranging up to about 250 miles, 16 ga. conductors and less expensive coils being satisfactory for terminal business over short distances. Without using loading, these early toll cable systems would not have been economically feasible.

In the early uses of repeaters on toll cables the cable circuits also used loading. These combinations permitted improved transmission performance and important extensions in transmission range. In this general connection, it is of interest to note that it was not economical to use non-loaded conductors for toll cable transmission until cable carrier telephone systems became available about two decades after the commercial introduction of the vacuum-tube repeater. For voice-frequency transmission, the use of repeaters without loading would have been unduly expensive, due to the high costs of the additional repeaters and the much more expensive distortion-correcting networks and regulating networks that would have been required.

In the early part of the period that intervened between the introduction of vacuum-tube repeaters and of cable carrier systems, the substantially continuous development of improved loading, and of improved repeaters and auxiliary equalizing and regulating networks, provided improved facilities of several different types especially proportioned on a minimum cost basis to meet the transmission-service needs of different geographical distances.

High-velocity, four-wire, H 44-25 19 gauge circuits were very extensively used for long-haul facilities ranging up to about 2000 miles in length. It is of interest that the timely completion of the development of the first cable-carrier system stopped the contemporary efforts to make additional improvements in the H 44-25 voice-frequency loaded four-wire circuits so that they would be suitable for transcontinental distances. These improvements would have involved the use of velocity distortion corrective networks.

Nineteen gauge two-wire circuits having lower-velocity, higher-impedance, loading than that employed on the above mentioned four-wire

circuits were very extensively used for short-haul repeatered and non-repeatered facilities.

A large curtailment in the demand for loading on new long cable circuits immediately followed the commercial exploitation of the Type K cable-carrier system, which started during the middle 1930's. The drastic nature of this impact was subsequently increased by the standardization of a still more economical (K2) cable carrier system,[48] and by the post-war extensive installation of coaxial cable systems. The very recent development of a relatively inexpensive short-haul carrier system (Type N), which uses two pairs in the same cable for its opposite-direction paths, promises an additional substantial reduction in the need for new loaded toll cable facilities, even for short distances. However, it seems probable that the demand for new loading may continue indefinitely on a low-level basis for more or less special short-haul situations where carrier telephony may be more expensive.

During the past two decades or so, loading cost-reduction has been carried so far that the prospects of further substantial cost-reductions are not now in sight. It seems improbable that any further design cost-reduction could be large enough to reverse the present general trend towards a large dependence upon carrier telephony for new short-haul toll cable facilities.

Exchange Area Loading

During the period covered by the present review, telephone transmission over exchange area cables has been entirely on a voice-frequency basis. Moreover, the use of vacuum-tube repeaters in conjunction with loading (or on non-loaded cables) has been statistically insignificant in comparison with the very extensive use of loading. In consequence, exchange area loading does not have to share with developments in repeaters and in carrier systems the great credit which it has earned with respect to the improvement of exchange area transmission performance and the reduction of plant cost.

The simple pattern in the evolution of exchange area loading practices, relative to those for toll cable loading, is of course basically due to the shortness of the circuits and the relatively uncomplicated service-requirements.

In certain important respects, the improvements achieved by the nearly continuous development work are generally similar in the two types of loading, notably: (1) the improvement in transmission quality obtained by increasing the transmission band-width, and (2) the successive facility-cost reductions resulting from the successive developments of lower-cost loading apparatus. These plant-cost reduction activities were carried out to a greater degree in the exchange area loading. It is especially noteworthy

that the most important apparatus-cost reduction developments were completed in time for exploitation during periods of peak demand for new coils.

With respect to the effects of other developments in reducing the demand for exchange area loading, the introduction of improved subscriber sets during the 1930's warrants special mention. By permitting higher losses in the trunks, somewhat longer non-loaded trunks could be used.

Looking towards the future, the prospective use of a new low-cost repeater of an entirely new type (El telephone repeater) is expected to reduce the demand for the heavier weights of loading. Also, the new Type N short-haul cable carrier system, referred to on page 1240, may have some considerable use on relatively long non-loaded exchange trunks along heavy traffic routes. It is also of interest that a greatly improved telephone set (500-type) now in the final stages of development will probably reduce the need for loading on long subscriber loops.

Although it is not possible at present to make accurate quantitative estimates of the ultimate effects of the just mentioned new developments upon the future demand for new exchange area loading, there is no reason to believe they will be so drastic as the effects of carrier system developments upon the ultimate future demand for toll cable loading. It seems especially probable that the low-cost H-spaced loading will continue indefinitely to be an important factor in the economy of design of new exchange area cable plant to provide telephone service for a continually increasing number of subscribers.

Loading for Incidental Cables in Open-Wire Lines

The impedance-matching loading systems used on entrance and intermediate cables have made vital contributions to the excellence of the over-all performance of the open-wire transmission systems. These are of great importance relative to the amount and the cost of the loading actually used.

In consequence of the increasing utilization of open-wire carrier systems, the voice-frequency loading is much less important than it was two to three decades ago. However, an indefinitely continuing, though small, demand seems certain, because of the valuable transmission improvements which the loading makes available at low cost.

The demand for additional carrier loading is expected to continue in a somewhat rough proportion to the number of additional open-wire carrier systems that are installed. However, in consequence of the high cost of the loading for multi-channel systems (which is much higher than formerly in consequence of greatly increased labor and material costs), it seems probable that more and more consideration will be given (especially on "long"

incidental cables) to the use of lower-cost transmission-improvement treatments, even though they are not so good as loading in certain respects.

Cable Program Circuit Loading

During the 1930's and early 1940's, there were extensive applications of loading on the cable sections of nation-wide chain networks used for transmitting AM broadcast program material. Now that high-grade program transmission circuits may be obtained by carrier methods on broadband cable carrier systems, the future demand for 8-kc loaded cable program circuits will be largely limited to special situations where the carrier program circuits are not economical.

It is expected also that there will be a moderate, continuing demand for the recently developed loading that provides a 15-kc band for the transmission of FM program broadcast material, principally on studio-transmitter circuits, and on end links in toll cable networks, where carrier program circuits may be uneconomical.

Continuous Loading

Over the years, a substantial amount of exploratory development work on continuous loading for ordinary types of paper-insulated cable has been done, but with negative results so far as commercial applications in the Bell System are concerned; it has not yet been found feasible to compete with coil loading in service performance and cost.

However, continuous loading has had a few applications in single core submarine cables, in deep water installations where coil loading is not feasible. The three 1921 cables between Key West and Havana are the only continuously loaded cables to become a part of the Bell System. They use iron wire as the loading material. Several years later, permalloy tape continuous loading developed by the Bell Telephone Laboratories made possible a great increase in the message-carrying capacity of transoceanic telegraph cables. During the middle 1920's, an aggregate of about 15,000 nautical miles of the new type, high speed, cable was installed for use by non-affiliated telegraph and cable companies.

Late in the 1920's, a perminvar type loaded cable suitable for voice-frequency telephony between Newfoundland and Ireland was developed by the Bell Telephone Laboratories. The business depression of the early 1930's intervened to cause a temporary postponement of the project; later on, an indefinite postponement resulted from improvements in transatlantic radio-telephony.

From the foregoing, it is clear that the importance of continuous loading has been low relative to that of coil loading in the growth of the Bell Telephone System.

Conclusion

During the half century that has intervened since its invention, coil loading has played a very important part in making nation-wide telephony possible and in helping to make possible the great growth in the business which has occurred. Although the application of coil loading to new circuits has now been greatly curtailed, due in large part to the development of carrier systems, coil loading still has an important field of application in exchange area telephone plant and for some rather special circuit applications.

The reader may take it for granted that the organization which has developed and used loading to the maximum degree of utility in the present telephone plant will be on the alert in the future to make full use of loading in situations wherever loaded circuits provide a more economical solution of the transmission service needs than the other available procedures. It is also reasonable to expect that new types of loading and new loading apparatus will be developed to the extent that may be economically warranted.

Bibliography (*Concluded*)

23. H. D. Arnold and G. W. Elmen, "Permalloy, An Alloy of Remarkable Magnetic Properties," *Journal of the Franklin Institute*, Vol. 195, 1923.
42. W. H. Martin, G. A. Anderegg, and B. W. Kendall, "Key West-Havana Submarine Cable System," *Trans. A.I.E.E.*, Vol. XLI, 1922.
43. H. A. Affel, W. S. Gorton, and R. W. Chesnut, "A New Key West-Havana Carrier Telephone Cable," *B.S.T.J.*, Vol. XI, April 1932.
44. O. E. Buckley, "The Loaded Submarine Telegraph Cable," *B.S.T.J.*, Vol. IV, July 1925; *Electrical Communication*, Vol. 4, No. 1, 1925, *Journal A.I.E.E.*, Vol. XLIV, No. 8, 1925.
45. O. E. Buckley, "High Speed Ocean Cable Telegraphy," *B.S.T.J.*, Vol. VII, April 1928. Presented at the International Congress of Telegraphy and Telephony in Commemoration of Volta, Lake Como, Italy, September 1927.
46. G. W. Elmen, "Magnetic Alloys of Iron, Nickel and Cobalt," *Jl. Franklin Institute* Vol. 207, p. 583, 1929.
47. O. E. Buckley, "The Future of Transoceanic Telephony," The Thirty-third Kelvin Lecture of the Institution of Electrical Engineers, April 23, 1942; *The Journal of The Institution of Electrical Engineers*, Vol. 89, Part 1, 1942. (Bell Laboratories reprint, Monograph B-1346).
48. H. S. Black, F. A. Brooks, A. J. Wier and J. G. Wilson, "An Improved Cable Carrier System," *Trans. A.I.E.E.*, Vol. 66, 1947.

In addition to the published articles referred to in the text or footnotes, the following will be of interest:

W. Fondiller, "Commercial Loading of Telephone Cable," *Electrical Communication*, Vol. 4, No. 1, July 1925.

George Crisson, "Irregularities in Loaded Telephone Circuits," *B.S.T.J.*, Vol. IV, October 1925.

F. L. Rhodes, "Beginnings of Telephony," Harper and Brothers, New York, 1929.

L. G. Abraham, "Circulating Currents and Singing on Two-Wire Cable Circuits," *B.S.T.J.*, Vol. XIV, October 1935.

L. L. Bouton, "Four-Wire Circuits in Retrospect," *Bell Lab. Record*, December 1938.

S. G. Hale, "Splice Loading Developments," *Bell Lab. Record*, January 1951.

Abstracts of Bell System Technical Papers Not Published in This Journal

*A Full Automatic Private Line Teletypewriter Switching System.** W. M. BACON[1] and G. A. LOCKE.[1] *Elec. Engg.*, v. 70, pp. 408–413, May, 1951.

ABSTRACT—A full automatic teletypewriter message switching system has been developed for use in private line networks involving one or more switching centers and a multiplicity of local or long distance lines, each of which may have one or more stations. This system provides fast teletypewriter communication from any station to any other station or group of stations in the network.

*Crossbar Tandem System.** R. E. COLLIS.[1] *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 997–1004, 1950.

*A Study of Nuclear and Electronic Magnetic Resonance.** K. K. DARROW.[1] *Elec. Engg.*, v. 70, pp. 401–404, May, 1951.

ABSTRACT—Since the discovery of magnetic resonance in solids, liquids, and gases in 1945, the phenomenon has been used in the determination of nuclear magnetic moments and magnetic field strengths, as well as in the study of crystal structure and relaxation times.

The Genesis of Submarine Cables. L. ESPENSCHIED.[1] Bibliography. *Elec. Engg.*, 70, pp. 379–383, May, 1951.

ABSTRACT—It was a century ago that the first submarine cable was laid between Dover and Calais. To mark this centenary the author reviews some of the events leading up to this achievement which made possible further advances in the communications field, such as laying of the transatlantic cable by the Great Eastern escorted by four ships, as shown in the picture.

Borocarbon Film Resistors. R. O. GRISDALE,[1] A. C. PFISTER[1], and G. K. TEAL.[1] *Natl. Electronics Conference, Proc.*. v. 6, pp. 441–442, 1950.

ABSTRACT—The carbon film type of resistor is particularly useful at high frequencies, for not only can it be made to have small reactance but it is, in effect, all skin so that there is no increase in resistance at high frequencies due to skin effect. The film is also well cooled through its intimate contact with the core and this makes possible the dissipation of large amounts of power per unit area. While primarily developed for high frequency applications in this country, the pyrolytic carbon resistor possesses other

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

characteristics which have led and are leading to greatly expanded fields of application. Principal among these are the tolerances of one per cent or better attainable in production, the stability in use, the relatively small and predictable temperature coefficient of resistance, and the low noise level. These properties result in large part from the ultimate crystalline structure of the carbon films.

Some Methods of Solving Hyperbolic and Parabolic Partial Differential Equations. R. W. HAMMING.[1] International Business Machines Corp. Computation seminar. *Proceedings, Dec., 1949, Ed. by C. C. Hurd. N. Y., I.B.M.*, pp. 14–23, 1951.

ABSTRACT—The main purpose of this paper is to present a broad, non-mathematical introduction to the general field of computing the solutions of partial differential equations of the hyperbolic and parabolic types, as well as some related classes of equations. I hope to show that there exist methods for reducing such problems to a form suitable for formal computation, with a reasonable expectation of arriving at a usable answer.

I have selected four particular problems to discuss. These have been chosen and arranged to bring out certain points which I feel are important. The first problem is almost trivial as there exist well-known analytical methods for solving it, while the last is a rather complicated partial differential-integral equation for which there is practically no known mathematical theory.

*Electrography and Electro-Spot Testing.**H. W. HERMANCE[1] and H. V. WADLOW.[1] *Physical Methods in Chemical Analysis; Ed. by W. G. Berl. N. Y., Academic Press*, v. 2, pp. 155–228, 1951.

Correlation Energy and the Heat of Sublimation of Lithium. C. HERRING.[1] Letter to the editor. References. *Phys. Rev.*, v. 82, pp. 282–283, Apr. 15, 1951.

*Some Theorems on the Free Energies of Crystal Surfaces.** C. HERRING.[1] References. *Phys. Rev.*, v. 82, pp. 87–93, Apr. 1, 1951.

ABSTRACT—Although the interpretation of experiments in such fields as the shapes of small particles and the thermal etching of surfaces usually involves problems of kinetics rather than mere equilibrium considerations, it is suggested that a knowledge of the relative free energies of different shapes or surface configurations may provide a useful perspective. This paper presents some theorems on these relative free energies which follow from the Wulff construction for the equilibrium shape of a small particle, and some relations between atomic models of crystal surfaces and the surface free energy function used in this construction. Equilibrium shapes of

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

crystals and of non-crystalline anisotropic media are classified, and it is pointed out that the possibilities for crystals include smoothly rounded as well as sharp-cornered forms. The condition is formulated for thermodynamic stability of a flat crystal face with respect to formation of hill-and-valley structure. A discussion is presented of the limitations on the applicability of the results imposed by the dependence of surface free energy on curvature; and it is concluded that these limitations are not likely to be serious for most real substances, though they are serious for certain idealized theoretical models.

*The Crystal Structures of NiO·3BaO, NiO·BaO, $BaNiO_3$ and Intermediate Phases With Composition Near $Ba_2Ni_2O_5$; With a Note on NiO.** J. J. LANDER.[1] References. *Acta Cryst.*, v. 4, pp. 148–156, Mar., 1951.

ABSTRACT—The crystal structures of NiO·3BaO, NiO·BaO and $BaNiO_3$ have been determined from X-ray diffraction data, and data are given for phases with composition near that represented by $Ba_2Ni_2O_5$. In each of these structures nickel behaves in a novel fashion. A coplanar triangular arrangement of oxygen around nickel is found in NiO·3BaO. In $BaNiO_3$ nickel has a valence of four and the structure is a close-packed hexagonal stacking of planar arrangements found in perovskite 111 planes. The compound NiO·BaO has a magnetic moment corresponding to two unpaired electrons, whereas the deduced coplanar square arrangement of oxygen around nickel suggests that there should be no unpaired electrons. Compounds with composition near $Ba_2Ni_2O_5$ contain an amount of oxygen which is a continuous function of temperature and possibly contain mixtures of bi- and tetravalent nickel.

The problem of NiO having octahedral co-ordination of oxygen is considered.

New Ferroelectric Tartrates. B. T. MATTHIAS[1] and J. K. HULM. Letter to the editor. *Phys. Rev.*, v. 82, pp. 108–109, April 1, 1951.

*A Negative Impedance Repeater.** J. L. MERRILL, JR.[1]. *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 1461–1466, 1950.

Interexchange Tandem Trunking in the Los Angeles Metropolitan Area. W. F. PFEIFFER[1]. *A.I.E.E., Trans.*, v. 69, pt. 2, pp. 1071–1079, 1950.

ABSTRACT—Twenty-four years have elapsed since the first large-scale machine switching tandem system was designed and installed for service in Los Angeles. As an intermediate switching center, the tandem office enabled operators to use the dial method of operation for establishing interexchange telephone connections over the associated trunking network. During the intervening years, it has facilitated the rapid handling of tele-

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

phone calls between the various communities in and around the city. Step-by-step tandem equipment was employed and, as the volume of calls grew, the trunk capacity was increased by installing additional switching equipment. In 1946 it became evident that the abnormal rate of growth required additions substantially beyond the practical size limit of the step-by-step tandem unit. To solve the resulting problem, it became necessary to reorganize the tandem trunking system and select a multiunit tandem switching plan. It also provided an opportunity to consider the application of the more recently developed crossbar tandem switching system. This paper reviews the factors affecting the general problem of interexchange trunking which have led to the development of the present tandem network in the Los Angeles metropolitan area. It describes the major elements of a system which now employs a total of five tandem switching units, three of which are crossbar tandem offices.

p-n Junction Rectifier and Photo-cell. W. J. PIETENPOL[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 120–121, Apr. 1, 1951.

*Formulas for the Determination of Residual Stress in Wires by the Layer Removal Method.** W. T. READ, JR.[1]. *Jl. Applied Phys.*, v. 22, pp. 415–416, Apr., 1951.

ABSTRACT—The distribution of residual axial stress in a beam or wire of circular cross section is derived as a function of the moment required to straighten the wire after removal of successive layers of material. Application of the formulas involves two graphical differentiations and integrations of experimental curves.

Observation of Magnetic Domains by the Kerr Effect. H. J. WILLIAMS[1], F. G. FOSTER[1], and E. A. WOOD[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 119–120, Apr. 1, 1951.

*Particle Size in Suspension Polymerization.** F. H. WINSLOW[1] and W. MATREYEK[1]. Bibliography. *Ind. & Engg. Chem.*, v. 43, pp. 1108–1112, May, 1951.

ABSTRACT—Control of size and geometrical form of densely cross-linked hydrocarbon polymers yields fluid spherical powders useful as dielectrics and in rheological studies. Such studies also bear on polymer forms important in ion exchange resins.

Several significant factors influencing the preparation of polymer spheroids have been established on a semi-quantitative basis: Polyvinyl alcohol proved to be a highly efficient stabilizer for polymer spheroid preparations. Under comparable conditions, (a) high molecular weight grades, (b) partially hydrolyzed grades, and (c) high concentrations of stabilizer were

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

associated with spheroids of lower mean diameters. These generalizations cover suspension stabilization down to roughly 0.1% stabilizer. The concentration limits where suspending action begins are, however, of special interest. Here it was found that the number of polyvinyl alcohol molecules present became important—that is, for equal weight concentrations in the vicinity of 0.005%, low molecular weight polymer (19,000) produced stabilized (although large) spheres whereas the usual high molecular weight polymer (95,000) was ineffective.

Close to the maximum possible yield of well-formed spheroids was reproducibly obtained in narrow size distribution and with average spheroid diameters ranging from 5 microns to several millimeters in diameter—a thousand-fold variation in dimensions.

Elastic and Electromechanical Coupling Coefficients of Single-Crystal Barium Titanate. W. L. BOND[1], W. P. MASON[1], and H. J. MCSKIMIN[1]. Letter to the editor. *Phys. Rev.*, v. 82, pp. 442–443, May 1, 1951.

Making Small Spheres. W. L. BOND[1]. *Rev. Sci. Instruments*, v. 22, pp. 344–345, May, 1951.

Submarine Telephone Cable With Submerged Repeaters. J. J. GILBERT[1]. *Electronics*, v. 24, pp. 164, 168, 172+, June, 1951.

*Electrode Reactions in the Glow Discharge.** F. E. HAWORTH[1]. References. *Jl. Applied Phys.*, v. 22, pp. 606–609, May, 1951.

ABSTRACT—The reactions which occur at silver electrodes in a normal glow discharge in air have been determined. These are: (1) formation of $AgNO_2$ and some Ag_2O at the anode at the rate of 3.4 μg/coulomb; (2) loss of metal from the cathode by chemical action at the rate of 3.5 μg/coulomb (probably the same reaction as (1) with subsequent loss of the reaction products by the greater heating of the cathode, but this hypothesis has not been established); and (3) normal sputtering loss at the cathode at the rate of 0.4 μg/coulomg. These processes result in building a conducting layer on the anode. If the electrode separation is so small that the anode extends into the region of the cathode fall, then the high electric field pulls the newly formed and not very coherent growth upon the anode across into a bridge between the electrodes.

*Storing Video Information.**A. L. HOPPER[1]. *Electronics*, v. 24, pp. 122–125, June, 1951.

ABSTRACT—Comparison of signal amplitudes along adjacent television scanning lines can be made by storing the video information of one line for 63.5 microseconds. Storage is done in an ultrasonic delay line employing a fused silica bar with quartz transducers.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

Cross Sections for Ion-Atom Collisions in He, Ne, and A. J. A. HORNBECK[1] and G. H. WANNIER[1]. Letter to the editor. *Phys. Rev.*, v. 82, p. 458, May 1, 1951.

*Ferromagnetic Resonance.**C. KITTEL[1]. Bibliography. *Jl. de Physique*, v. 12, pp. 291–302, Mar., 1951.

Theory of Antiferroelectric Crystals. C. KITTEL[1]. References. *Phys. Rev.*, v. 82, pp. 729–732, June 1, 1951.

ABSTRACT—An antiferroelectric state is defined as one in which lines of ions in the crystal are spontaneously polarized, but with neighboring lines polarized in antiparallel directions. In simple cubic lattices the antiferroelectric state is likely to be more stable than the ferroelectric state. The dielectric constant above and below the antiferroelectric curie point is investigated for both first- and second-order transitions. In either case the dielectric constant need not be very high; but if the transition is second order, e is continuous across the Curie point. The antiferroelectric state will not be piezoelectric. The thermal anomaly near the Curie point will be of the same nature and magnitude as in ferroelectrics. A susceptibility variation of the form $C/(T + Z)$ as found in strontium titanate is not indicative of antiferroelectricity, unlike the corresponding situation in antiferromagnetism.

Theory of Antiferromagnetic Resonance C. KITTEL[1]. Letter to the editor. *Phys. Rev.*, v. 82, p. 565, May 15, 1951.

*Barium-Nickel Oxides With Tri- and Tetravalent Nickel.** J. J. LANDER[1] and L. A. WOOTEN[1]. *Am. Chem. Soc., Jl.*, v. 73, pp. 2452–2454, June, 1951.

ABSTRACT—The compound $BaNiO_3$ and intermediates with composition ranging between $Ba_3Ni_3O_8$ and $Ba_2Ni_2O_5$ have been prepared. $BaNiO_3$ is black, stable in alkali, and has a structure made up of layers identical with the 111 planes of a perovskite but stacked in a close-packed hexagonal fashion. At 730° in 730 mm. of oxygen, the structure changes to that associated with the series $Ba_3Ni_3O_8$ to $Ba_2Ni_2O_5$ in which the oxygen content appears to decrease continuously with temperature increasing to 1200°, at which point sharp melting is observed. These materials are black and stable in alkali with an hexagonal structure for which the details have not been determined. Resistivities and magnetic susceptibilities are reported. A wide range in composition, temperature and reaction atmosphere was studied but only one additional compound was observed. Attempts to isolate this compound were not successful.

*The Phase System BaO–NiO.** J. J. LANDER[1]. *Am. Chem. Soc., Jl.*, v. 73, pp. 2450–2452, June, 1951.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

ABSTRACT—The phase system BaO–NiO has been studied largely by means of X-ray diffraction. The two compounds NiO BaO and NiO 3BaO occur in the system. Their preparation and properties are described. NiO BaO is black, stable in air, orthorhombic, and melts at 1240°. NiO 3BaO is gray-green, unstable in air, hexagonal, and melts at 1160°. A eutectic melting at 1080° is observed between these compounds, but none between NiO 3BaO and BaO. Intersolubility of all solid phases in the system is small, even at high temperatures, but quantitative data have not been obtained.

*A Phenomenological Derivation of the First- and Second-Order Magnetostriction and Morphic Effects for a Nickel Crystal.** W. P. MASON[1]. References. *Phys. Rev.*, v. 82, pp. 715–723, June 1, 1951.

ABSTRACT—In order to account for experimental results which showed that the saturation elastic constants of a single nickel crystal varied with the direction of magnetization, a phenomenological investigation has been made of the stress, strain, and magnetic relations for single nickel crystals. The variation in elastic constants is shown to be a "morphic" effect caused by the change in the crystal symmetry due to the magnetostriction effect. In the energy equation this effect is represented by additional terms which involve squares and products of both the magnetic intensities and stresses. These terms are as large as the magnetostrictive terms when the stresses are of the order of 10^{10} dynes/cm^2. The energy equation has been used to derive the first- and second-order magnetostrictive effect, and the resulting terms agree with Becker and Döring's empirical constants for saturation conditions. For smaller magnetic intensities the terms divide up into first- and second-order terms which vary differently with magnetic field intensity. It is shown that the morphic effects involve six measurable constants, and some of these are evaluated experimentally.

*Dielectric Properties of Sodium and Potassium Niobates.** B. T. MATTHIAS[1] and J. P. REMEIKA[1]. *Phys. Rev.*, v. 82, pp. 727–729, June 1, 1951.

ABSTRACT—The following paper deals with evidence of ferroelectricity in $KNbO_3$ and $NaNbO_3$. Temperatures at which both materials undergo crystallographic changes and corresponding changes in dielectric constant and loss tangent are reported. Photographs of dielectric hysteresis loops and values of saturation polarization taken at various points over a temperature range are given for $KNbO_3$.

Ferroelectricity. B. T. MATTHIAS[1]. Bibliography. *Science*, v. 113, pp. 591–596, May 25, 1951.

ABSTRACT—Under the name of Ferroelectrics one classifies those materials which exhibit dielectric anomalies phenomenologically similar to the mag-

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

netic behavior of the ferromagnetics. Perhaps it would have been more logical to use the term Rochelle electrics, thus emphasizing the similarity in the dielectric behavior to that of Rochelle salt, for which this behavior was first discovered by J. Valasek.

In this discussion the known ferroelectrics will be listed, and the various theories that have been created to explain them will be examined.

Theory of Ferroelectric Behavior of Barium Titanate. P. W. ANDERSON[1]. References. *Ceramic Age*, v. 57, pp. 29–30, 33+, April, 1951.

Criterion for Superconductivity. J. BARDEEN[1]. Letter to the Editor. *Phys. Rev.*, v. 82, pp. 978–979, June 15, 1951.

*Magnetic Domain Patterns.** R. M. BOZORTH[1]. Bibliography. *Jl. de Physique*, v. 12, pp. 308–321, March, 1951.

Electron Temperature vs Noise Temperature in Low Pressure Mercury-Argon Discharges. M. A. EASLEY[1] and W. W. MUMFORD[1]. Letter to the Editor. *Jl. Applied Phys.*, v. 22, pp. 846–847, June, 1951.

*The Origin of Bombardment-Enhanced Thermionic Emission.** J. B. JOHNSON[1]. References. *Phys. Rev.*, v. 83, pp. 49–53, July 1, 1951.

ABSTRACT—Measurements on bombardment-enhanced thermionic emission from oxide cathodes show that (a) the effect is not related to normal fading and recovery of thermionic emission; (b) the emitted electrons have energies in the thermal range rather than in the secondary range. Calculations indicate that the electron bombardment releases more than enough internal secondaries to account for the effect as increased thermionic emission. A more comprehensive theory is needed for explaining why the observed effect is not even larger.

Dipolar Domains in Paramagnetic Crystals at Low Temperatures. C. KITTELL[1]. Letter to the Editor. *Phys. Rev.*, v. 82, pp. 965–966, June 15, 1951.

*Methods of Measuring Adjacent-Band Radiation from Radio Transmitters.** N. LUND[1]. *I.R.E. Proc.*, v. 39, pp. 653–656, June, 1951.

ABSTRACT—A review of three possible methods of measuring or estimating adjacent-band radiation characteristics of a radio transmitter is given. These three methods differ in the type of signal applied to the transmitter and may be termed the two-tone, normal signal, and thermal noise methods. Measurements on a multichannel single-sideband transmitter using each of these methods are presented to show that there is a good correlation between the normal signal and thermal noise methods.

An empirical method for calculating the slope of the adjacent-band radiation as a function of frequency from the measured two-tone distortion values

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

is given, and the measured and calculated slopes are shown to be in fairly good agreement.

Microwave Spectrum in NO_2. K. B. Mc AFEE, JR.[1]. Letter to the Editor. *Phys. Rev.*, v. 82, p. 971, June 15, 1951.

A Simple Electronic Differential Analyzer as a Demonstration and Laboratory Aid to Instruction in Engineering. M. H. NICHOLS[1] and D. W. HAGELBARGER[1]. *Jl. Engg. Education*, v. 41, pp. 621–630, June, 1951.

Telecommunications. H. S. OSBORNE[1]. *Ordnance*, v. 36, pp. 87–90, July–August, 1951.

Triangular Permutation Numbers. J. RIORDAN[1]. References. *Am. Math. Soc., Proc.*, v. 2, pp. 429–432, June, 1951.

*Measurements of Dynamic Internal Dissipation and Elasticity of Soft Plastics.** H. C. RORDEN[1] and A. GRIECO[1]. *Jl. Applied Phys.*, v. 22, pp. 842–845, June, 1951.

ABSTRACT—In order to measure the mechanical properties of soft plastics over wide frequency and temperature ranges two new techniques have been devised. The first one, which operates in the frequency range of a few cycles, uses a horizontal oscillating pendulum. The shear impedance of the sample is measured by mounting a small pad of the material between the vibrating pendulum and a fixed platform and determining the change in frequency and the change in the decrement caused by the sample. From these measurements the shear mechanical resistance and reactance of the specimen can be determined. The other technique, which is applicable in the frequency range from 100 cycles to 10,000 cycles, makes use of a vibrating tuning fork. Two identical samples are mounted between a stationary weight and the moving tines, and the shear mechanical impedance is determined by determining the change in frequency and change in decrement caused by the specimen. These two techniques have been applied to measuring the shear properties of a number of soft plastics including Pyralin, Koroseal, Keldur, polyvinyl butyral, Thiokol, and gum rubber. All of these show relaxation effects. The polyvinyl butyral appears to be approaching a crystalline elastic stage at the low frequency of 1000 cycles, while gum rubber remains in a quasi-configurational stage from 2 cycles to 1000 cycles.

*The Mobility of Electrons in Silver Chloride.** J. R. HAYNES[1] and W. SHOCKLEY[1]. References. *Phys. Rev.*, v. 82, pp. 935–943, June 15, 1951.

ABSTRACT—Techniques are described which utilize the "print out effect" to obtain both the direction and velocity of photoelectrons in silver chloride crystals in an electric field. Hall mobility of the electrons is calculated from their change in direction produced by crossed electric and magnetic fields.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

Drift mobility of the electrons is obtained by measurement of their velocity in known electric fields. The value obtained for the Hall mobility ($R\sigma$) multiplied by $8/3\pi$ is 51 cm^2/volt sec at 25°C. The values obtained for the drift mobility are shown to be a function of temperature. A value of 49.5 cm^2/volt sec was obtained at 25°C, which is within experimental error of $(8/3\pi)R\sigma$, indicating that acoustical scattering is the principal mechanism and that temporary trapping is unimportant. A summary of the behavior of conduction electrons in silver chloride, calculated from the results of these experiments, is included.

*p-n Junction Transistors.** W. SHOCKLEY[1], M. SPARKS[1], and G. K. TEAL[1]. References. *Phys. Rev.*, v. 83, pp. 151–162, July 1, 1951.

ABSTRACT—The effects of diffusion of electrons through a thin p-type layer of germanium have been studied in specimens consisting of two n-type regions with the p-type region interposed. It is found that potentials applied to one n-type region are transmitted by diffusing electrons through the p-type layer although the latter is grounded through an ohmic contact. When one of the p-n junctions is biased to saturation, power gain can be obtained through the device. Used as "n-p-n transistors" these units will operate on currents as low as 10 microamperes and voltages as low as 0.1 volt, have power gains of 50 db, and noise figures of about 10 db at 1000 cps. Their current-voltage characteristics are in good agreement with the diffusion theory.

* A reprint of this article may be obtained on request.
[1] Bell Tel. Labs.

Contributors to This Issue

B. S. BIGGS, B.A., Southwest Texas Teachers College, 1927; M.A., University of Texas, 1931, Ph.D., 1933; Civil Research Laboratory, Carnegie Institute of Technology, 1933–1936. Bell Telephone Laboratories, 1936–. With the Laboratories he has worked chiefly on the synthesis of wood preservatives, on dielectric materials and on other phases of organic chemistry. He is a member of the American Chemical Society and of Sigma Xi.

G. T. FORD, B.S., Michigan State College, 1929; M.A., Columbia, 1936. Bell Telephone Laboratories, 1929–. With the Laboratories he has worked on gas tubes, thermistors, general vacuum tube development, and electron tubes for broad band amplifiers. He is a member of the Institute of Radio Engineers.

R. W. FRIIS, B.E.E., University of Minnesota, 1930. Bell Telephone Laboratories, 1930–. With the Laboratories Mr. Friis has been concerned with transoceanic and ship-to-shore radio telephone, fire-control radio transmitters, and the microwave radio-relay system. He is a Senior Member of the Institute of Radio Engineers.

J. C. LOZIER, B.A., Columbia, 1934. R.C.A. Mfg. Co., 1935–1936. Bell Telephone Laboratories, 1936–. Mr. Lozier's work with the Laboratories has been principally transmission development for radio and carrier telephone systems, the theory and design of servomechanisms, and the theory of feedback systems such as compandors and regulators. He is a Senior Member of the Institute of Radio Engineers.

R. C. PRIM, III, B.S.E.E., University of Texas, 1941; M.A. and Ph.D., Princeton, 1949. General Electric Company, 1941–44; Naval Ordnance Laboratory, 1944–49. Bell Telephone Laboratories, 1949–. Here his work has been chiefly mathematical research on non-linear partial differential equations and as a consultant on military projects. Dr. Prim is a member of the Amer. Math. Soc., the Amer. Phys. Soc., Sigma Xi and Tau Beta Pi.

JOHN RIORDAN, B.S., Yale, 1923. Amer. Tel. and Tel., 1926–34; Bell Telephone Laboratories, 1934–. With the American Company and subsequently with the Laboratories, Mr. Riordan has been concerned chiefly with

transmission theory, the application of Boolean algebra to switching, number theory in cable splicing, and combinatorial and probability studies of traffic. He is a member of the Amer. Math. Soc., Math. Assoc. of America, Inst. of Math. Statistics, and Fellow of the Amer. Assoc. for the Advancement of Science.

A. A. ROETKIN, B.E.E., Ohio State University, 1927; M.Sc., 1929. Bell Telephone Laboratories, 1929–. With the Laboratories Mr. Roetkin has worked on overseas radio telephone receivers, ultra-high frequency, point-to-point radio telephone service, pulse multiplex microwave radio repeaters for the armed forces, and microwave radio-relay systems. He is a member of the Institute of Radio Engineers.

THOMAS SHAW, S.B., Massachusetts Institute of Technology, 1905. American Telephone and Telegraph Company, Engineering Department, 1905–19; Department of Development and Research, 1919–33. Bell Telephone Laboratories, 1933–48. Mr. Shaw's active telephone career was mainly concerned with loading problems in telephone circuits, including the transmission and economic features of the loading apparatus. The article which is concluded in this issue was started shortly before his retirement in 1948.

K. D. SMITH, B.A., Pomona College, 1928; M.A., Dartmouth, 1930. Bell Telephone Laboratories, 1930–. Consultant to National Defense Research Council, 1941–44. Awarded Joint Army-Navy Certificate of Appreciation for Scientific Achievement following World War II. With the Laboratories Mr. Smith has been concerned with the coaxial cable system, radar bombing equipment, broad band microwave radio system, and transistors. He is a Senior Member of the Institute of Radio Engineers.

R. L. WALLACE, JR., B.A. summa cum laude, physics and mathematics, University of Texas, 1936; M.A., physics, 1939; Special Research Associate, Harvard, 1941–45. Bell Telephone Laboratories, 1946–. Mr. Wallace's work with the Laboratories has been chiefly concerned with magnetic recording and transistors. He is a member of the Acoustical Society of America, Phi Beta Kappa, and Sigma Chi.

E. J. WALSH, Bell Telephone Laboratories, 1928–. Mr. Walsh's work with the Laboratories has been chiefly on vacuum tube design, magnetrons, proximity fuse tubes, reflex oscillators and close-spaced fine-wire grid tubes.