1962 IRE International Convention Record

621.3806
SY68I
1962
pt 6 thru 10
c.1

PART 6

Sessions Sponsored by

IRE Professional Groups on

Component Parts

Industrial Electronics

Product Engineering and Production

Reliability and Quality Control

Ultrasonics Engineering

at

the IRE International Convention, New York, N.Y.

March 26-29, 1962

Retu 050814

The Institute of Radio Engineers

1962 IRE INTERNATIONAL CONVENTION RECORD

An annual publication devoted to papers presented at the IRE International Convention held in March of each year in New York City. Formerly published under the titles CONVENTION RECORD OF THE I.R.E. (1953 & 1954), IRE CONVENTION RECORD (1955 & 1956), and IRE NATIONAL CONVENTION RECORD (1957, 1958, & 1959).

Additional copies of the 1962 IRE INTERNATIONAL CONVENTION RECORD may be purchased from the Institute of Radio Engineers, 1 East 79 Street, New York 21, N.Y., at the prices listed below.

Part	Sessions	Subject and Sponsoring IRE Professional Group	Prices for Members of Sponsoring Professional Group (PG), IRE Members (M), Libraries and Sub. Agencies (L), and Nonmembers (NM)			
			PG	M	L	NM
1	8, 16, 23	Antennas & Propagation	$ .70	$ 1.05	$ 2.80	$ 3.50
2	10, 18, 26, 41, 48	Automatic Control Circuit Theory	1.00	1.50	4.00	5.00
3	1, 9, 17, 25, 28, 33	Electron Devices Microwave Theory & Techniques	1.00	1.50	4.00	5.00
4	4, 12, 20, 34, 49	Electronic Computers Information Theory	1.00	1.50	4.00	5.00
5	5, 13, 15, 22, 29, 47, 54	Aerospace & Navigational Electronics Military Electronics Radio Frequency Interference Space Electronics & Telemetry	1.20	1.80	4.80	6.00
6	3, 11, 31, 35, 42, 45, 50, 52	Component Parts Industrial Electronics Product Engineering & Production Reliability & Quality Control Ultrasonics Engineering	1.40	2.10	5.60	7.00
7	30, 37, 43, 51	Audio Broadcasting Broadcast & Television Receivers	.80	1.20	3.20	4.00
8	7, 24, 38, 46, 53	Communications Systems Vehicular Communications	1.00	1.50	4.00	5.00
9	2, 19, 27, 32, 39, 40, 44	Bio-Medical Electronics Human Factors in Electronics Instrumentation Nuclear Science	1.20	1.80	4.80	6.00
10	6, 14, 21, 36	Education Engineering Management Engineering Writing & Speech	.80	1.20	3.20	4.00
		Complete Set (10 Parts)	$10.10	$15.15	$40.40	$50.50

Responsibility for the contents of papers published in the IRE INTERNATIONAL CONVENTION RECORD rests solely upon the authors and not upon the IRE or its members.

Copyright © 1962 by The Institute of Radio Engineers, Inc., 1 East 79 Street, New York 21, N.Y.

1962 IRE INTERNATIONAL CONVENTION RECORD

PART 6 - COMPONENT PARTS; INDUSTRIAL ELECTRONICS; PRODUCT ENGINEERING AND PRODUCTION; RELIABILITY AND QUALITY CONTROL; ULTRASONICS ENGINEERING

TABLE OF CONTENTS

Page

Ultrasonics Engineering-I
(Session 3: sponsored by PGUE)

Ultrasonic Welding in Electronic Devices . . . . J. M. Peterson, H. L. McKaig, and C. F. DePrisco 3
Ultrasonics for Machining . . . . G. C. Brown 13
A Chemical Method for Measuring Relative Amounts of Cavitation in an Ultrasonic Cleaner . . . . A. Weissler 24
The Advantages of Ultrasonic Cleaning as a Production Tool . . . . . . R. C. Fencl and B. Hollerith 31
Evaluation and Applications of the Levavasseur Whistle . . . . I. Elias 36
Technical Aspects of the Cavitron Hyper Intense Proximal Scanning Ultrasonic Cleaner . . . . L. Balamuth 43

Ultrasonics Engineering-II
(Session 11: sponsored by PGUE)

Heat Treatment of Wire for Torsional Magnetostrictive Delay Lines . A. J. Brown and A. Rothbart 54
A Piezoelectric-Piezomagnetic Gyrator . . . . M. Onoe and M. Sawabe 59
Wide-Band, Instantaneous Spectrum Analyzers Employing Delay-Line Light Modulators . . . . L. B. Lambert 69
Electro-Optical Transfer Characteristics of Liquid Delay-Line Light Modulators . . . . M. Arm, L. B. Lambert, and B. Silverberg 79
Theoretical Power Limits of Sonar Transducers . . . . R. S. Woollett 90
High-Frequency Electrostatic Transducers for Use in Gases . . . . W. M. Wright 95

A Compendium of Significant Product Engineering and Production Trends
(Session 31: sponsored by PGPEP)

Thin-Film Circuit Packaging and the Numbers Game . . . . A. R. Meehan 101
Thermal Characteristics of Potted Electronic Modules . . . . J. I. Gonzales and C. E. Waugh 105
High-Density-Package Joining Techniques . . . . C. W. Johnson and S. G. Konsowski, Jr. 120
The Mathematical Basis of the Autonetics Etched Interconnection Design Program . J. Weissman 126
Needed: Wire Design Documentation for More Efficient Production . . . . J. T. LaForte 134

Page

New Views of Industrial Electronics
(Session 35: sponsored by PGIE)

Some Myths About Industrial Electronics . . . . S. Feldman 141
VWP: A New Approach to Small, Light, Efficient,
High Power Regulated Power Supplies . . . . V. Wouk 145
A Low Cost Industrial Television System . . . . H. R. Walker 158
Application of Audiofrequency for Anesthesia of Animals. . . . . J. E. Thompson 163

Space Age Components
(Session 42: sponsored by PGCP)

Solid-State Linear and Sinusoidal Synchros . . . . T. W. Parsons and D. R. Simon 169
A Brushless DC Motor with Solid-State Commutation . . . . G. Bauerlein 184
The System Concept of Precision Potentiometer Specification (Abstract) . . . . D. C. Hoos 191
Space Radiation Resistor Evaluation . . . . I. Doshay 192
The Effect of Radiation Environment on Film Resistors . . . . S. O. Dorst and L. H. Wurzel 206

Reliability and Quality Control
(Session 45: sponsored by PGRQC)

Reliability Studies of Mesa Transistors . . . . F. Aubin 215
Establishing Reliability Requirements . . . . L. R. Diamond 222
Distributions of Functions of Random Variables. . . . . D. L. Heck and L. H. Tung 232
A Quantitative Approach to the Evaluation of
System Reliability . . . . V. Lasewicz, S. Newman, and H. Thomas 241
Some Aspects of Test Equipment Reliability . . . . F. A. Applegate 255

Lumped and Distributed Microcircuit Components
(Session 50: sponsored by PGCP)

Synthesis of Solid State Distributed Parameter
Functions . . . . W. W. Happ, P. S. Castro, and W. D. Fuller 262
A Theoretical Comparison of Doubly Loaded Distributed
Bridged T and Lumped Twin T RC Notch Filters . . . . J. L. Ekstrom 279
The Electrical Properties of Porcelain Enamels and
Ceramic Coatings, and Factors Affecting Their Performance. . . . . G. Geltman 283
Properties of Thin Film and Silicon Solid State Components -
Their Effects on Microcircuit Performance . . . . . J. A. Ekiss, J. Roschen, and P. G. Thomas 291
Titanium Integrated Electronic Components . . . . W. D. Fuller 305

Reliability and Quality Control
(Session 52: sponsored by PGRQC)

Reliability in Real Time . . . . T. J. Scanlon 312
Individuality in Small-Plant Reliability. . . . . I. J. Bearer 321
Short Term Prediction of Hook-Up Wire Insulation Cold Flow . . . . H. W. Hicks 331
Flow Graph Techniques for Reliability Engineering . . . . J. L. Burroughs and W. W. Happ 338
Test Fixture Variance and Reliability . . . . L. G. Reynolds 367

ULTRASONIC WELDING IN ELECTRONIC DEVICES

J. M. Peterson, Sonobond Corp., West Chester, Pa.,
and H. L. McKaig and C. F. DePrisco, Aeroprojects Inc., West Chester, Pa.

Summary

The nature of the ultrasonic welding process and the solid state bond so obtained are described. The advantages of the process which are derived from the solid state character of the bond are discussed. There is a review of successful areas of application of the process in the electronics industry, with particular emphasis on semiconductor and microcircuitry manufacturing. A newly designed welder, intended primarily for electronics research and manufacturing is described.

The Ultrasonic Welding Process

In order to better discuss the application of ultrasonic welding techniques to electronic devices, it is well to first consider the fundamentals of the process.

The mechanism of ultrasonic welding, like many industrial processes, is not completely understood. However, there has been sufficient fundamental research to indicate the phenomena that occur during ultrasonic welding. This section describes the present theory and some of its limitations.

All welding processes are based upon the fact that metal atoms with unsatisfied bonds are capable of bonding to other atoms if they are brought into intimate contact. Atoms such as those adjacent to grain boundries or at the surface of a piece of metal, have unsatisfied bonds and it is just this which allows grains of metal to adhere to one another and which allows different pieces of metal to be joined by welding. Thus, if two pieces of metal with absolutely smooth, clean surfaces are brought together in intimate contact, the unsatisfied bonds of the atoms of the surfaces of both pieces will create a true metallurgical bond between the two pieces of metal.

However, if we consider the true nature of a clean, smooth metallic surface, it is evident that this kind of metallurgical bonding cannot normally occur. First, no matter how carefully a metal surface may be prepared, it is far from smooth on an atomic scale. A most carefully ground surface (super finished) still has irregularities with peak-to-valley vertical distances that average about 5×10^{-6} cm, which corresponds to a distance of approximately 200 atomic layers. The surface of an ordinary ground surface contains peaks which are as many as tens of thousands of atomic layers above the valleys. Since the attractive forces between atoms which have unsatisfied bonds decreases rapidly as the distance from the atom increases, very few atoms (only those on "peaks") will satisfy their bonds when two surfaces with such surface irregularities are brought into contact.

Not only are real metal surfaces not smooth, they are not clean. Since, as we have noted, the metallic atoms at the surface provide strong attractive forces, they are capable of attracting and holding molecules from the atmosphere which surrounds the surface. Oxygen molecules from the atmosphere react with the metal to form oxides. An apparently clean, bright steel surface will, in fact, have an iron oxide film on its surface approximately 200 molecule thick. At such thickness, the oxide films tend to behave as the bulk metal oxide. These oxide layers are also crystalline and the oxide molecules at the surface have unsatisfied bonds just as did the metal atoms of the free metallic surface. However, the attractive forces of these metallic oxide molecules exert only weak attractive forces upon symmetric molecules such as oxygen. Their attractive forces for asymmetric molecules such as water vapor are somewhat stronger. Hence, a film of condensed adsorbed moisture is formed on the oxidized metal surface. The thickness of this adsorbed moisture film depends somewhat upon the ambient relative humidity, but it is never less than 2 or 3 molecules thick.[1] See Figure 1.

Thus, ordinary metallic surfaces are characterized by three features which prevent the deformation of true metallurgical bonds. The adhered moisture layer and oxide film, both act as barriers which prevent intimate contact between atoms with incomplete molecular shells. Even in the absence of such films, the irregular surface provides a condition of hills and valleys which limits the area of metal to metal contact. Thus, the requirements for the creation of a welded joint is to remove these surface films and "smooth out" the surface irregularities so that a large area of uniform nascent metal contact is obtained.

The ultrasonic welding process overcomes the barriers to welding by plastically deforming the interface between the workpieces in such a way that the adhered moisture and oxide films are dispersed, and the irregular surfaces of the two workpieces are made to conform to each other, thereby causing a large area of intimate contact

between the nascent metal surfaces.

The typical ultrasonic welder which is used for spot welding in electronic devices incorporates the lateral drive welding system. Other types of ultrasonic welding systems are described elsewhere.[2] A schematic representation of this type welder is shown in Figure 2. In this machine, elastic vibration is created by the rapid expansion and contraction of a magnetostrictive transducer under the influence of a high frequency magnetic field. These elastic waves travel along a "transmission line" called the coupler, to the sonotrode tip which contacts the workpieces. A portion of this coupler is tapered to serve as a mechanical transformer to increase the amplitude of vibration, as needed to effect an impedance match. In this manner, the sonotrode tip undergoes excursions in a direction parallel to the surfaces of the weldment and induces a shear mode of vibration into the workpieces.

The weldment is clamped between the sonotrode and the anvil, so that workpieces experience the stresses resulting from the clamping force and the super-imposed vibration of the sonotrode. When the stresses reach a critical magnitude, such as the elastic limit of the material being welded, the material near the interface will undergo plastic deformation. As we have already noted, this plastic deformation of the parent metal will cause the moisture and organic films to fracture and disperse; and will cause the original irregularities of the two surfaces being welded to conform to one another over large areas. In this manner, a true metallurgical bond is created between the workpieces.

There are some supplementary phenomena which need to be discussed because of their effect upon the welding process. The first of these is a temperature rise that results from the elastic hysteresis of the highly stressed portion of the weld zone during the welding sequence. Just as when a paper clip is rapidly bent back and forth until it breaks, a sensible rise in temperature of the wire is experienced, so also the interface of the weldment undergoes a modest increase in temperature. The temperature rise has been measured by fine wire thermocouple and fusible insert techniques and considered in the light of extensive electromicrography. These investigations revealed that the maximum transient temperature was between 30% and 50% of the absolute melting point of the metal in a similar metal joint. Since this temperature rise is highly localized in the weld locale, there is, in fact, but a small average temperature rise in the weldment. This temperature rise does temporarily increase the ductility of the weld metal and promotes plastic deformations which are associated with the process.

A second supplementary phenomena may derive from the fact that metallic crystal structures can be temporarily plasticized by high frequency mechanical vibrations. This phenomena has been observed in ultrasonically accelerated processes, and in the nature and extent of the internal deformations in ultrasonic welds themselves. This plasticization occurs due to the acoustical excitation per se, and is independent of temperature and the thermal plasticization previously described. However, it does accomplish the same purpose, that is to facilitate the plastic deformation of the interface. Figure 3 summarizes the foregoing theoretical model of the ultrasonic welding process.

There are numerous shortcomings implicit in this theory. For example, it is known that the elastic limit is an over-simplification of the stress threshold required to initiate welding. A more inclusive criteria, the maximum supportable lateral shear stress, has been suggested and is reported elsewhere.[3]

Also, the resulting metallurgical structure of some ultrasonic welds belies the simple pressure welding analysis of plastic deformation effects. The concept of interfacial micro-slip has been proposed in this regard.[4] However, for all its faults, this model does explain and rationalize most of the conditions and effects of ultrasonic welding.

Advantages of Ultrasonic Welding in Electronic Applications

We have seen that ultrasonic welding is a means of creating true solid state metallurgical bonds, i.e. without melting of the parent metal. This characteristic of ultrasonic welding is the source of many of the advantages of this process that can be exploited by the electronics industry.

Thin gauges of materials can be welded.

There seems to be no lower limit to the thinness of materials that can be joined by ultrasonic welding. Fine wires of less than 0.0003 inch diameter have been satisfactorily welded and thin foils of 0.00017 inch thickness have been joined without rupture.

This ability results from the fact that since ultrasonic welding is not a melting process, the prohibitive thermal control problems required to avoid "burn through" with such fine components, is not a restriction on the process.

Thin gauges of material may be joined to very thick sections.

Again, since ultrasonic welding is not a thermal process, the heat sink potential of the heavy section does not require the application of such large amounts of heat that the thin section is damaged. Hence, there is no theoretical limit to the ratio of thickness that can be joined.

A wide variety of similar and dissimilar materials can be joined.

Figure 4 summarizes some of the similar and dissimilar material combinations that can be ultrasonically welded on a commercial basis.

In general, the blank spaces represent combinations in which welding has not yet been attempted, or in which higher powers appeared to be necessary at the time the tests were made, or in which tip problems had not yet been solved. Some of these combinations were evaluated several years ago when ultrasonic welding equipment was under development.

Even the remarkable results of welding aluminum to glass or copper to steel are available with the ultrasonic welding process. Joints such as these, with such wide disparities in physical properties, are easily understood when one realizes that joining by ultrasonic welding is accomplished by removing the oxide and adsorption coatings in the solid state.

Materials may be joined that are otherwise difficult to weld.

Materials which possess relatively large coefficients of electrical or thermal conductivity are inherently difficult to weld by means of the resistance welding process, due to the large amounts of electrical energy regained to create and maintain enough thermal energy in the weld area in order to accomplish fusion. If this magnitude of electrical energy is not properly controlled, "burn through" or insufficient fusion or other weld faults will result. Ultrasonic welding is not restricted by this phenomena.

Other materials, due to their composition or structure, respond to the elevated temperatures implicit in thermal welding processes with numerous undesirable results.

Oxidation, grain growth, recrystallization and other adverse metallurgical and mechanical effects of the thermal welding processes can be eliminated or minimized by utilization of ultrasonic welding. Thus, whole classes of applications, previously thought to be unsatisfactory for solution by welding, find a solution in the ultrasonic welding process.

Lack of change of parent material properties.

It is true that ultrasonic welding at room temperature is known to produce a local temperature rise from the combined effects of elastic hysteresis, localized slip, plastic deformation, etc. However, in no instance involving a monometallic weld made under normal conditions has there been any evidence of melting of the metal in the weld sections examined metallographically, although the transient heat sometimes results in some metallurgical changes such as phase transformation, recrystallization, diffusion or other phenomena.

Thus, the ultrasonic welding process accomplishes what is essentially a solid state metallurgical bond between relatively undisturbed parent material workpieces. In particular, the elimination of the cast structure with its typically inferior mechanical properties is of special interest when welding work hardened wrought materials.

Lack of contamination of weld area.

Since fluxes, electrode shielding coatings or adhesives are not required by the ultrasonic welding process, there is no contamination of the weld or its surroundings. This advantage has proven particularly valuable in packaging and encapsulating applications.

Ability to create hermetically sealed and pressure tight joints.

Ultrasonically welded continuous seams, whether produced by a series of overlapping spot-type welds by continuous seam roller or ring welding techniques are capable of producing hermetic seals and pressure tight joints in a wide variety of materials and thicknesses. Several types of aluminum packages sealed by continuous seam welds have proved to be helium leak tight to the limits of sensitivity of the mass spectrometer used for the tests (5×10^{-8} cc per second per inch of seam). Other packages containing volatile materials have shown no loss of weight after more than two years storage at ambient and elevated temperatures.

Cleaning not critical.

Surface cleaning is not highly critical when preparing most materials for ultrasonic welding. This is due to the fact that the vibratory displacements occurring during the welding process disrupt many normal oxide layers and other surface films on the mating surfaces. The more readily weldable materials, such as aluminum clad alloys, brass and copper, can be welded in the mill-finish condition and usually require only the removal of surface lubricants with a detergent reagent. Welding of heat-treated materials, such as structural aluminum alloys and the refractory metals containing a rather heavy heat-treat scale, is more readily accomplished after mechanical abrasion or descaling in a chemical etching solution. However, once the surface scale is removed, the elapsed time before welding is not important.

Also, it is possible to produce ultrasonic welds, even through surface deposits or coatings, although higher power levels are required. For example, heavily oxidized Inconel X sheet has been successfully welded without prior removal of the heavy oxide coatings, and sound welds have been produced in aluminum having anodized coatings up to 0.0001 inch thick.

Low ohmic resistance and low noise junctions.

Because the ultrasonic welding process produces void-free junctions with relatively little, if any, foreign material inclusions, it is a desirable means of creating low noise, low resistance junctions. This is particularly true in the case of dissimilar metal joints where the low temperature of ultrasonic welding eliminates or greatly minimizes the formation of brittle, high resistance intermetallic compounds.

Low Temperature joining.

The low temperature nature of ultrasonic welding permits it to be used in the immediate vicinity of temperature sensitive or even volatile materials without adverse affects. Further, there is no preheating required which may damage the assemblies or create problems of thermal distortion.

Early Applications of Ultrasonic Welding in the Electronics Industry

These, and other advantages of the ultrasonic welding process, were immediately grasped upon by some segments of the electronics industry as the basis for early applications of the process.

Some of these pioneering efforts are still of significant interest today, so that it is worthwhile to review them.

Bridgewire welding

The ultrasonic welding process has been successfully adapted to the manufacture of various ignitor devices which rely upon high resistance bridgewire elements to initiate various explosives or other chemical reactions. Figure 5 shows several of such components. Since most of these bridgewires are very small in diameter, the ability of ultrasonic welding to consistently join such fine wires to larger terminal posts is the principal reason for this application. Filament wires of 0.001 inch diameter of such materials as nickel-chromium and tungsten-platinum, have been joined to terminal posts of copper, iron, phosphor-bronze and other metals.

An important recent development in this area is the demonstrated capability of the ultrasonic welding process to weld combustible bridgewire materials, such as Pyrofuse® wire. Since this material does itself burn when heated to a relatively low temperature, it is very difficult to resistance weld it without causing it to ignite. Thus, the low temperature characteristic of the ultrasonic welding process provides a reliable means of joining to such wires without premature ignition.

® Registered Trademark, Pyrofuse Corp. Affiliate of Sigmund Cohn Corp.

Thermocouple junctions

It would be difficult, if not impossible, to catalogue all of the combinations of thermocouple junction and thermocouple wire to test piece junctions that have been made by means of ultrasonic welding. Iron-Constantan, Chromel-Constantan, Chromel-Alumel, and Copper-Constantan couples have all been welded to aluminum, steel, copper and other base metals. Recent work performed by the Martin-Marietta Co., Denver, Colorado, has demonstrated the feasibility of creating tungsten-rhenium couples for measuring temperatures above 4500°F. Related work, by this same organization, has also adapted the ultrasonic welding process as a means of attaching stainless steel and ceramic sheathed thermocouples to test structures.[5]

In applications such as these, the remarkable versatility of ultrasonic welding as a means of joining dissimilar metal combinations is immediately apparent. In many instances, it is even possible to weld thermocouple junctions directly to ceramic and other non-metallic test pieces.

Aluminum coil windings

Because of the weight savings possible by the adoption of aluminum as the winding material in inductive components, there has been considerable interest in the use of this metal as an electrical conductor. Aluminum foil windings have been applied to several coil components and ultrasonic welding has demonstrated its applicability to such manufacturing. This process is especially valuable where the foil is coated with various plastic films as a means of insulating the individual coil windings. The work of McCarthy, et al,[6] has shown that copper foil leads can be effectively and reproducibly joined to coated aluminum foil transformer windings, that such junctions demonstrate less variability in conductivity during long-life testing than joints produced by other methods, and that the process is practicable and readily adaptable to automated production.

Precious metal contacts

A very common problem in the electronics industry is to provide a means of joining contact materials such as silver, gold, palladium, and platinum to spring metals in the manufacture of switches and relays. Because of the differences in thermal and electrical conductivity of the contact metal and the spring material, these combinations are frequently very difficult to weld with resistance welding techniques. To overcome these difficulties it has often been necessary to prepare the contacts in the form of composite structure in which the contactor metal backed with a metal

which is suitable for welding to the spring material. Ultrasonic welding allows such contact metals to be welded directly to a wide variety of dissimilar spring metals. Also, since the ultrasonic weld is created in the solid state, the formation of brittle, high resistance intermetallic compounds which may ultimately result in failure or loss of conductivity, is minimized.

Tantalum capacitors

Another area of utilization of the ultrasonic welding process that has received considerable attention is its use as a means of attaching leads to tantalum electrolytic capacitor plates. The traditional method of fabricating these components is to resistance weld tantalum wire leads to tantalum foil plates and subsequently form an anodic film on the tantalum foil plate to serve as a dielectric. This method gives rise to several inherent shortcomings.

First, since the resistance welding process subjects the surface of the tantalum wire and foil to "near-melting" temperatures, a thick and irregular oxide film forms on both of these workpieces. This oxide film can affect the nature of the subsequent electrolytic process to such an extent that a uniform anodic film is not obtained and the leakage characteristics of the device may not be controlled. The low temperature nature of the ultrasonic welding does not cause this oxide film growth as a result of welding, hence, the joining of the lead to the plate does not result in leakage problems after electrolytic treatment.

Second, the resistance welds in tantalum are typically small and brittle, and display low mechanical strength. Because of their size, it is usually necessary to use several resistance welds to provide a sound junction and their poor mechanical properties often cause many failures in subsequent handling. Also, the resistance welding process usually results in grain growth in the lead wire with a resulting embrittlement of this area. With ultrasonic welding, it is possible to create one large area bond between the wire and the lead which has good mechanical properties and which does not result in an embrittling of the components.

Finally, due to oil films, oxide films and other inevitable surface contaminents, the surface resistance of the wrought tantalum products cannot be uniformly controlled. This results in a very high degree of resistance weld inconsistency. The ultrasonic welding process is totally insensitive to these variations and can be relied upon to provide a consistent joint in the face of these variable conditions.

From the scope and variety of these applications of ultrasonic welding, it is easy to grasp the importance of this unique process to the electronics industry. However, two more very significant applications have been singled out for separate consideration because of having become established as production techniques and/or because of their remarkable potential impact on the future process of the electronics manufacturing industry. These are the bonding of electrical conductors to semiconductor materials and to thin film microcircuitry devices.

Semiconductor Bonding

Ultrasonic welding is a practical method of bonding gold and aluminum wires to silicon and germanium. Such wires, ranging in diameter from 0.0005 inch to 0.025 inch, have been bonded to both plain and metallized semiconductor material surfaces.

The General Electric Company, Auburn, New York, applied and evolved the process to the point where it has been adopted as the means of attaching wires to silicon in that company's C-5, C-10, C-35, C-50 and C-204 silicon rectifiers. The manufacture of a new silicon rectifier, the C-226, soon to be released by General Electric, also utilizes ultrasonic welding. Figures 6 and 7 illustrate a typical bonded assembly (in this case, two 0.010 inch diameter aluminum wires bonded to silicon) and a portion of General Electric's ultrasonic welding facility for this type of work.

In addition to these rectifier components, which are already in production, many organizations are considering adopting the ultrasonic welding process as a means of lead wire bonding to transistor semiconductors. There are several important technical and economic motives behind these activities.

The first of these factors is the strong desire to improve the yield (number of useable units taken as a percent of total number of units processed) of all of the processes that are involved in the manufacture of transistors. The bonding of gold wires to the semiconductor elements has been a very pronounced problem with respect to yield. Although such information is naturally considered somewhat proprietary, instances of yields as low as 50% have been reported where thermocompression techniques are utilized. Similarly, the yields obtained with the ultrasonic welding process, based upon pilot run and laboratory evaluations, have been reported as remarkably high, with 100% yields being considered routine. Specifically, comments based upon the General Electric, Auburn, production experience, indicate that when the welding installation and set-up are properly established and maintained, the yield of the ultrasonic welding process is "beyond reproach."

Another impetus to the adoption of the ultrasonic welding process has been the recent notoriety of the so-called "purple plague." This nomenclature refers to the presence of a brittle,

porous, purple compound in the vicinity of gold wires which have been bonded to aluminum-silicon eutectic p-n junctions such as are found in planar transistors.[7] There appears to be a correlation between the mechanical failure and increased resistance of such junction where such purple staining is present. A successful means of eliminating such staining is to eliminate the gold from this region and to substitute aluminum wires in these connections. In this way, there is no gold available for the formation of this damaging Au Al_2 purple phase. Ultrasonic welding is by far the most successful means of joining the aluminum wire to the aluminized silicon stripe.

Some investigators have reported that even when purple staining is not apparent, there is still great danger in using the gold wire in the aluminum-silicon junction. Their evidence is to the effect that the brittle, high resistance Au Al_2 intermetallic will ultimately form via diffusion mechanisms and thus contribute to mechanical failure and/or altered performance. In this connection, it is worthy to note that it is possible to limit the width of the wire bond to the diameter of the wire in those cases where the width of the stripe needs to be as narrow as possible.

Another aspect of the aluminum-for-gold lead wire substitution, that is of significant importance, particularly for some military and space applications of transistors, concerns the improved shock and vibration resistance of the transistors that are assembled with aluminum leads. The lower density of the aluminum wire causes a marked reduction in the mechanical forces acting upon the wire to silicon bond. One investigator reports an immediate increase in shock resistance of from something less than 10,000 g's for thermocompression bonded gold wires to approximately 20,000 g's for ultrasonically welded aluminum wires.

Figure 8 illustrates a planar transistor in which the joint of the aluminum wire to aluminized stripes on silicon and the aluminum wire to gold plated kovar post bond have both been produced by means of the ultrasonic welding process.

Microcircuit Thin Film Junctions

In the last several years, there have been remarkable advances in the state-of-the-art in the field of microminiaturization. Many organizations have been active in this field to such an extent that many thin film circuit networks are already on the market. Soon modules with integrated semiconductor devices will be available. The rate of growth in this field has been outstanding.

However, through all of this progress, one problem has persisted. That problem is how to provide reliable junctions with these films, in order to connect them to the rest of the system. Ultrasonic welding has solved this problem for many applications.

Because it is a solid state joining process that does not require melting of the parent materials, ultrasonic welding causes relatively slight modifications in the characteristics of the film or the conductor element. Furthermore, very close dimensional tolerances can be maintained, since no preheating of the part is required and no thermal distortion can occur.

Because no electric arc occurs and no electric current passes through the joint, there is no contamination of the films and no problem of providing closed circuits for welding currents.

Because ultrasonic welding can create bonds between a very wide variety of dissimilar materials, it is an extremely flexible tool when applied to the problem of welding interconnections to microminiature circuit elements. Table 1 summarizes some of the combinations of films and wires and ribbons which have been successfully bonded. Though aluminum wires are the most versatile conductor elements, the process has been applied to gold, copper, and nickel conductors as well. Figure 9 presents a thin film circuit module to which 0.005 inch diameter aluminum wires have been ultrasonically bonded. The film in this case is gold, approximately 2000 Å thick deposited to glass.

This versatility, coupled with the simplicity and reliability of the process, make ultrasonic welding a likely candidate for expanded usage in joining conductors to thin film circuits and components.

Table 1

Electrical Conductors Which Have Been Ultrasonically Welded to Metallized Surfaces

Film	Conductor	
	Material	Thickness in.
On Glass Substrate		
Aluminum	Aluminum Wire	0.002
		.005
		.010
	Gold Wire	.003
Nickel	Aluminum Wire	.002
		.005
		.010
		.020
	Gold Wire	.002
		.003
Copper	Aluminum Wire	.002
		.005
		.010

Table 1 (Cont'd)

Film	Conductor Material	Thickness in.
On Glass Substrate		
Gold	Aluminum Wire	.002
		.005
		.010
	Gold Wire	.003
Tantalum	Aluminum Wire	.002
		.005
		.010
		.020
Chromel	Aluminum Wire	.002
		.005
		.010
	Gold Wire	.003
Nichrome	Aluminum Wire	.0025
		.010
		.020
Gold, 4 microinch	Aluminum Wire	.010
7 microinch		.010
12 microinch		0.010
Platinum, 1 microinch	Aluminum Wire	0.010
Gold-Platinum, 9 microinch	Aluminum Wire	.010
Palladium, 8 microinch	Aluminum Wire	.010
Silver, 8 microinch	Aluminum Wire	.010
100 microinch		.010
Copper, Electroplated, on Silver	Copper Ribbon	.028
On Alumina Substrate		
Molybdenum	Aluminum Ribbon	.003
		.005
Gold-Platinum, 7 microinch	Aluminum Wire	.010
Molybdenum-Lithium Metal-lized, Gold-Plated	Nickel Ribbon	.002
Copper-Plated	Nickel Ribbon	.002
Molybdenum-Manganese Metal-lized, Silver-Plated	Nickel Ribbon	.002
On Silicon Substrate		
Aluminum	Aluminum Wire	.010
		.018
		.020
	Gold Wire	.002
On Ceramic Substrate		
Silver	Aluminum Wire	0.010
Copper, Electroplated, on Molybdenum, Metallized with Manganese	Copper Foil	
On RCA Alumina Modules Substrate		
Copper, Electroplated, on Screened Silver	Copper Wire, bare, #44	

Ultrasonic Welders Developed for Electronics Applications

As with any new process, there must be a period of evaluation, based upon applications in the field. Ultrasonic welding equipment is no exception. As a result of such field experience with early models of ultrasonic welding equipment, many valuable improvements in both the design and function of the smaller units have been suggested.

As a result of these suggestions and subsequent developments in ultrasonic welding equipment design, new versions of welders are being developed for use in the electronics industry. The first of these new welders which is already in production use is illustrated in Figure 10. This welder is competent to perform most of the semiconductor and microcircuitry bonding operation to which ultrasonic welding should be applied.

This new welder includes a newly designed welding system and mounting assembly which is light and small, in order to provide precise control of the welding tip. The system is mounted in a precision bearing arrangement which provides complete repeatability in the positioning of the welding tip during the welding cycle. The clamping force has an extended lower range and greater reproducibility throughout the entire range.

Both the welding tip and the anvil are easily interchangeable to provide versatility in configurations that can be joined. The welding tip can be actuated by hand or by a foot pedal or can be adapted to any number of mechanical or pneumatic actuation systems when it is installed in automated or semiautomatic assembly machines. The welding head is adaptable to special installations and to micropositioning devices.

A revised ultrasonic power generator incorporates several improvements keyed to electronic

WRH

component fabrication. A visual tuning indicator is employed to simplify the proper set-up of the welder. An automatic control unit is included to correct frequency according to changes in transducer temperature. The power control, as well as the weld pulse time control, are step switches which prevent accidental changes of these parameters and facilitate reproducible set-up procedures.

The design objective of these features has been to combine the advantages of the ultrasonic welding process with the needs of electronic research and development.

References

1. Udin, H., E. R. Funk, J. Wulff, Welding for Engineers, John Wiley & Sons, 1954.

2. "Ultrasonic Welding," Chapter 52, Welding Handbook, Fourth Edition, Section Three, American Welding Society, 1960.

3. Jones, J. B., N. Maropis, J. G. Thomas, and D. Bancroft, "Fundamentals of Ultrasonic Welding, Phase II." Research Report No. 60-91, Aeroprojects Incorporated, Navy Contract NOas 59-6070-c, December 1960.

4. op. cit.

5. Martin-Denver Report No. D-89, "Ultrasonic Welding, Phase III and IV, Joining and Attaching Thermocouples," Sept. 8, 1961

6. McCarthy, D., V. Pirc, and W. Hannahs, "Ultra-Sonic Welded Aluminum-Copper Junctions as Electrical Connections." Reliable Electrical Connections, Third Electronics Industries Association Conference, Engineering Publishers, New York, 1958, pp. 113-119.

7. Bernstein, L., "Gold Alloying to Germanium, Silicon and Aluminum-Silicon Eutectic Surfaces," Parts I and II, Semi-conductor Products, July and August 1961.

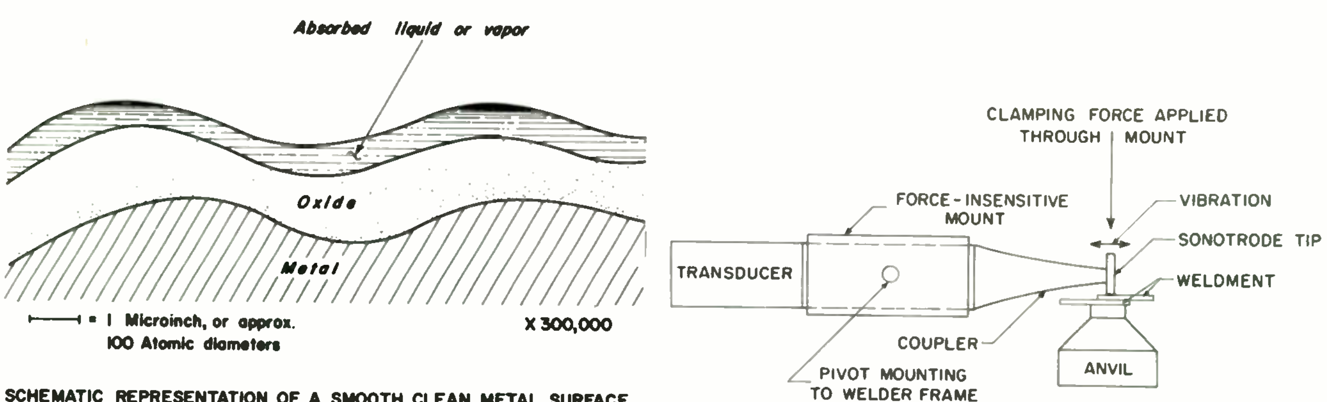


Fig. 1.

Fig. 2. Lateral drive ultrasonic welding system.

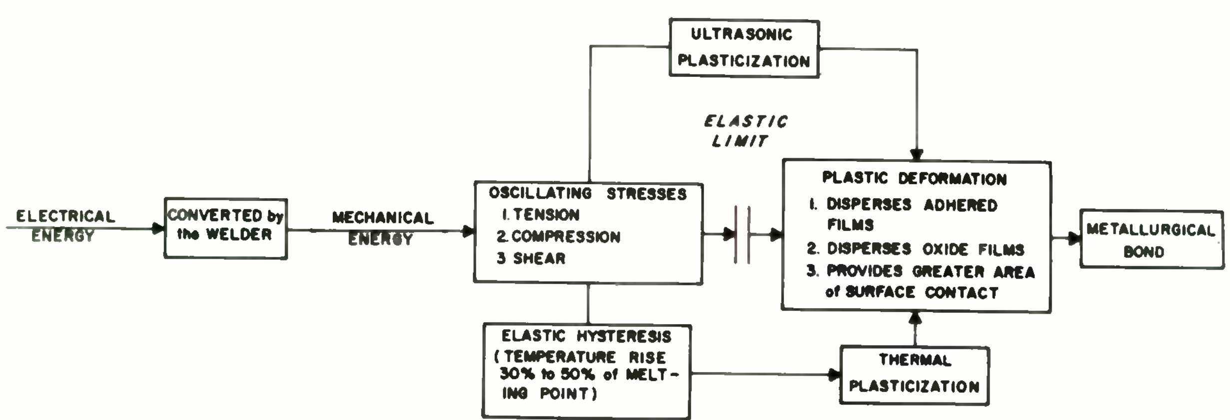


Fig. 3. Block diagram of the ultrasonic welding process.

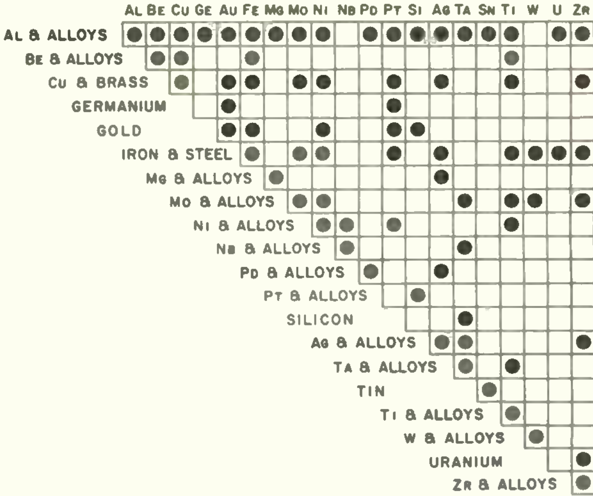


Fig. 4. Dissimilar material combinations which have been ultrasonically welded.

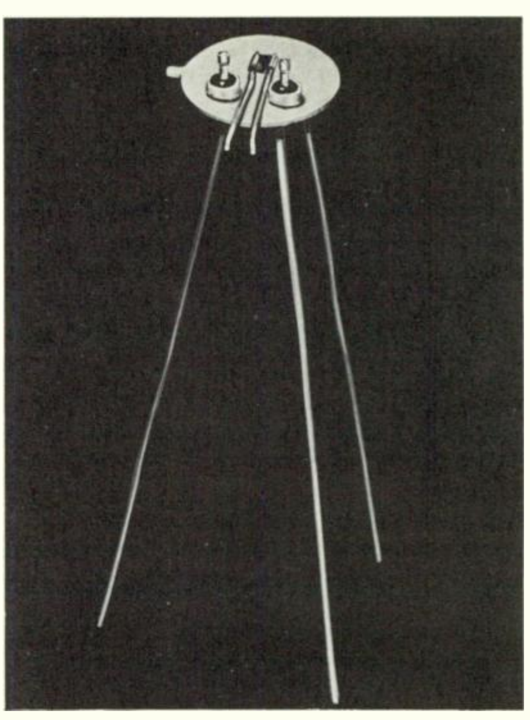


Fig. 6. General electric silicon rectifier in which two .010 inch diameter aluminum wires are ultrasonically welded to silicon.

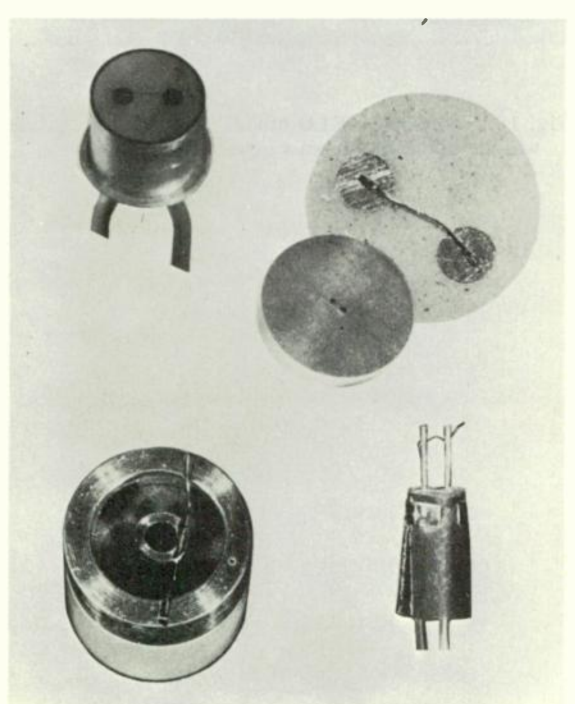


Fig. 5. Several types of ultrasonically welded bridge-wire assemblies.

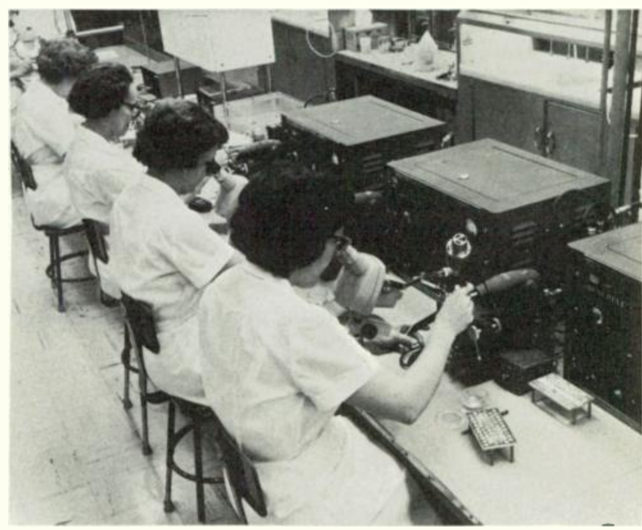


Fig. 7. A portion of the general electric ultrasonic welding facility for the production of silicon rectifiers.

WRH

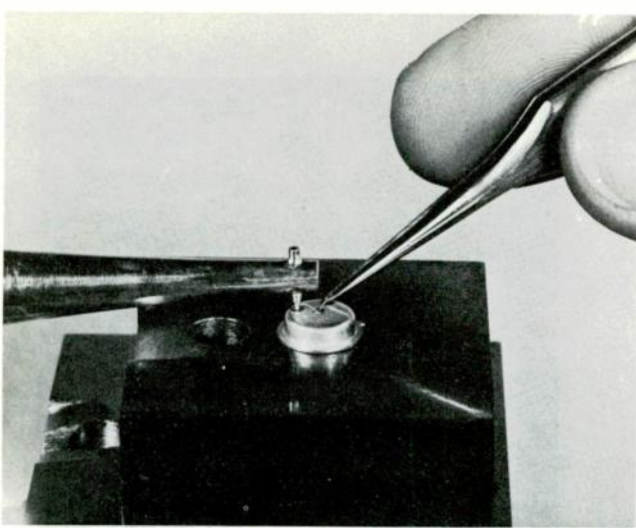


Fig. 8. A planar transistor in which aluminum wires are ultrasonically bonded to both the aluminized silicon stripes and the gold-plated kovar posts.

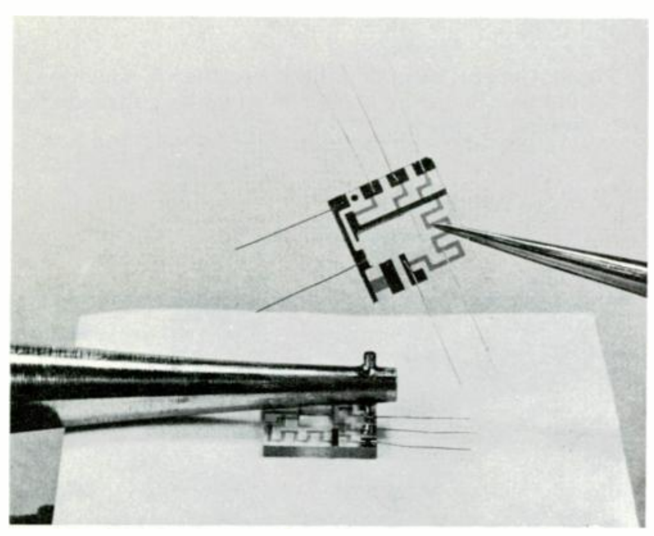


Fig. 9. The ultrasonic welding of a .005 inch diameter aluminum wire to 2000 Å thick gold film.

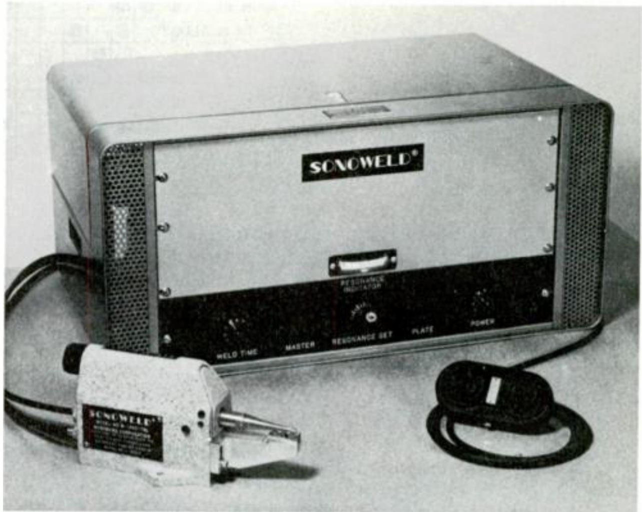


Fig. 10. The SONOWELD model W-1040-TSL ultrasonic welder and 100 r-f watt generator.

ULTRASONICS FOR MACHINING

George C. Brown

The Sheffield Corporation
Dayton 1, Ohio

Discussion Objectives

To those of you who are not completely aware of the nature of our subject or its connection with the present art of machining, we intend to define the word "Ultrasonics" as it is applied to the metal cutting techniques of today. As Ultrasonics is a science, a section will be devoted to the explanation of its use and conceptions, and then, with illustrations of successful applications in metal cutting, describe typical instances in which the technique can be applied.

The Ultrasonics field is in effect an industry today. It is, however, a youngster among the older established methods of machining. For that reason, changes and improvements are plentiful and appear in rapid fire order throughout each year. So that this paper will not become obsolete before its complete distribution, it is believed important that we add to the present day knowledge, descriptions of foundation work presently in progress for the near future and for periods beyond that point. Discussion will therefore be presented on a current development in the art of grinding wherein Ultrasonics has indicated benefits as an assisting medium. Additionally, as a research potential area, comments will also be made on its adaptation to the metals cutting tool field.

"Ultrasonics For Machining" Defined

Discussions in this paper for "Metal Cutting Today" will be limited to the use of Ultrasonics as a stock removal technique on ferrous and nonferrous materials. Other technical papers have been written covering the use of the Ultrasonic technique of machining in other materials; such as glass, ceramics, ferrites and a host of semiconductor materials.

Interpretations on the use of Ultrasonics as a motivating source in machining will be based on two different types of applications; namely, direct or indirect. In the former situation, energy of an Ultrasonic nature will be assumed to be directly applied to the cutting tool for stock removal purposes. In the latter instance, it will be assumed that Ultrasonic energy is again directly applied to the cutting tool for stock removal but would not be the only form of power input. For example, in the grinding application to be described later, rotational energy is provided to the wheel normally while Ultrasonic energy is simultaneously applied to the wheel, both power inputs being used concurrently to achieve a job of stock removal from the workpiece.

The word "Ultrasonics", as used in the title of this technical paper, denotes a science in which harnessed, inaudible sound energy is practically used for the job of stock removal from metals. The energy or source of power in this technique is actually a longitudinal or compression wave phenomenon produced in solids, such as the cutting tool of a machining operation. In affect, we are referring to a mechanical oscillatory type of operation that is imparted to the cutting tool and having a cyclic frequency of approximately 20,000 per second. As this frequency is normally above the hearing range of the human ear, the process is described as being inaudible. In another sense, Ultrasonics may be described as a system which, when practically used, produces mechanical motions of high frequencies in bodies that we desire to work with in the machining of metals. The heart of such a described system is an Ultrasonic transducer (see Fig. 1) and a generator of the required energy. In the majority of cases, an electronic type of generator is utilized from which a 20,000 cycle per second power output is derived having the necessary electrical characteristics to operate a receiver, namely the Ultrasonic transducer. This device produces at its output end mechanical motions of high frequency that can be transmitted to a load such as the cutting tool in a machining operation. The transducer unit is, as the name identifies, a unit for the conversion of one form of energy to another. In this situation, the energy transformation is electrical to mechanical. Sound waves are involved in this transformation as a go between the two forms of energy input and output. Consequently, the description of the technique sometimes utilizes the expression "sound energy device" and is, in effect, quite proper.

It is not the intent of this paper to describe the

technical aspects of how the mechanical motions are obtained from the transducer nor how they are transferred to the actual cutting tool tip. Other papers have covered this point adequately and it is not felt it should be repeated here.

It can be said, therefore, that Ultrasonics, as used in the title of this paper for machining purposes, means mechanical motivation of a cutting tool at Ultrasonic frequency during the machining operation. The tool motivation can be due to Ultrasonic input alone or to two or more power inputs, one of which is Ultrasonic. The cutting tool may contain cutting edges or may utilize loose abrasive grains, uniformly covering its cutting surface as tool cutting edges. Stock removal is strictly a mechanical type of operation, not based on any electrical or chemical phenomenon. (See Fig. 2 for typical machine tool illustration.) The technique is one of actual stock removal in the form of microscopic chips without noticeable metallurgical or physical property changes.

It is not intended that this paper explain the art of "Ultrasonics for Machining" as a competitive technique to any other process for the purpose of stock removal. Descriptive matter is based on accomplishments by the Ultrasonic technique that should not be practical in any other way. By "direct" application of the technique as a major stock removal method, the state of the art is limited to the smaller shaped cavities of high accuracy and quality in hardened steels or carbide. Again by "direct" application, as a minute stock removal technique (deburring), the art can include practically all of the metals used in industrial practices today.

As a future development, Ultrasonics may be applied to conventional grinding techniques to broaden the list of applications.

As a future research potential, Ultrasonics may be applied to conventional cutting-tool-operation types to broaden their scope of application.

Current Successful Applications Machining of Metals

The following sections cover applications which are successful uses of the "direct" method of Ultrasonic machining of metals. They have been selected to illustrate typical situations that may be duplicated in many other jobs as long as all of the requirements described are part of this desired result.

Basic Requirements

In order to successfully apply Ultrasonics for machining purposes in metals, such metals must be hard, having a minimum hardness of Rockwell "C" Scale-59. Metals having a hardness less than this amount become more and more impractical for the Ultrasonic technique. It can be said that the softer materials do not respond to cutting edge impacting of the abrasive grains. Softer materials have a tendency to absorb the bombarding effect and essentially deform with each abrasive edge load application instead of chipping or spawling away.

Proper application of the technique in this field denotes a machining operation to produce cavities in workpiece materials not possible by rotating or reciprocating type of conventional cutting tools or grinding wheels. Tolerances on the cavity profiles must be very closely controlled, generally .001 or less with locations generally .002 or less. Further, machining must not disturb previously established metallurgical or physical properties.

Further, cavity sizes may range from .005 in smallest dimension upwards.

Typical materials under this section for successful applications are 440C and 52100 steels, normally hardened to a Rockwell "C" Scale of 60-62, and all grades of carbide.

Typical Production Example - Valve

The following production example utilizes a servo valve sleeve of stainless grade 440C, having a hardness of Rockwell "C" Scale 60-62. The operation is one of machining a series of metering port orifices through the valve body wall. The machining operation is a practical one for the Ultrasonic technique.

Requirements of the machining operation are as follows (see Figs. 3 and 4): A total of 8 port configurations, having profile dimensions of .050 x .125 and profile tolerance of $\pm$.0005, must be machined through the hardened sleeve walls, having radial alignments within .0002 per port and an axial displacement tolerance of $\pm$.001. These ports must be machined through the valve bodies after 90% of the valve sleeve is complete. Maximum profile corner radii must not exceed .002 within a maximum corner radius at the

point of break out into the bore not exceeding .0002. Valve porting requirements of this type are normally found in the very high pressure servo valve systems and are, as you can see, critical in dimension. Finish specifications are also critical, requiring 10 microinch or better. Absolutely no physical or metallurgical disturbance in the machined cavities can be permitted. One minute spot in any of the machined surfaces, defective for either of these two characteristics, could spell failure of the valve.

Machining the port configuration in the servo valve sleeve is by the Ultrasonic technique. The process is of the "direct" type in which the cutting tool is directly coupled to the transducer and utilizes the abrasive slurry as cutting edges for stock removal. The operational frequency is 20,000 cycles per second. The tool has a rectangular shape, somewhat less than the desired shape of the cavity. Tool materials are generally a very tough grade of tool steel hardened to approximately Rockwell "C" Scale 40. No other power source is utilized for the tool and, therefore, resembles the conventional Ultrasonic machine tool application involving suitable tooling and fixturing. The machining time per port is approximately 10 minutes through a wall section of about .040.

Problems encountered by other methods of port machining are discussed in the next paragraph.

Port configurations are generally located diametrically opposite each other and originally made possible the following consideration for port machining. Broaches were used in the soft state of the valve body material. However, following the hardening operation and the final grinding of O.D. configurations and bore, it was never possible to obtain the accuracy of port profile and location due to heat treat distortions. Further, it was never possible to obtain the extremely accurate finishes required with sharp edge detail of the port openings at the break out point in the bore. Grinding of the port profiles after the hardening operation has always been a desired type of operation but as one can easily see, an impossible approach due to the port configuration and size. Other types of machining or grinding operations have been considered from time to time. However, in each there has always been some draw back eliminating it then from practical use. For example, some methods do promote metallurgical or physical disturbances, while others are not practical from the standpoint of constant control over the minimum corner radii required in the port profile.

Typical Production Example - Nozzle

This illustration is another production example in which the use of Ultrasonic machining techniques is practical. The operation is one of grinding the "V" notch, forming part of the spray orifice. The material is carbide.

The grinding operation for the "V" notch must be done after the small bore has been formed (see Figs. 5 and 6). The bore configuration has generally been accomplished by a pre-sintering operation, although in some cases the bore has been drilled by Ultrasonic techniques. However, the centerline of the bore and of the "V" notch must be exactly collinear and any departure from this characteristic affects the spray pattern of the nozzle and, naturally, has limitations. Normally, a displacement allowance of $\pm$.0005 is maximum. The depth of the "V", with respect to the break through point of the notch into the bore radius, is critical, as it produces an oval shaped opening and again affects the actual spray pattern. The shape of the "V" notch is also critical. Another critical requirement of the operation is that the "V" notch in the actual nozzle must have a point radius of no more than .001. The sharper this edge condition becomes, the greater is the control of the spray pattern. Finish of the notch profile should be 10 microinches or less.

Again, in this case, the cutting tool is directly coupled to an Ultrasonic transducer, receiving no other power type for the machining operation. A 20 kilocycle operational frequency is utilized. The actual tool shape is a sharply ground "V" profile in a material of tough tool steel, normally hardened to around a Rockwell "C" Scale 45. An abrasive slurry is used, resembling, therefore, the conventional Ultrasonic machine tool with a suitable fixture to hold and relate the work to the cutting tool. Approximate machining time from the solid is about 10 minutes.

Grinding of the notch form has been tried by conventional grinding practices. However, the minimum corner radius in the notch profile has been a draw back. Other methods have been considered from time to time, but, generally speaking, the maintenance of the small corner radius of notch profile again has been the draw back. Ruled out, therefore, have been the conventional or abrasive coated grinding wheel or directly applied cutting tools softer than Rockwell "C" 45.

WRH

Typical Tool Room Example - Plate

This application does not represent a production type operation. As a small quantity, typical tool room operation, it does feature advantages that can be gained by utilization of the Ultrasonic technique of machining over other methods of fabrication. A discussion of the requirements of the machining operation follows.

The material was a 52100 steel, hardened to approximately a Rockwell "C" Scale 60 (see Figs. 7 and 8). Thickness is .095. A totally enclosed profile of tooth forms had to be machined clear through the thickness where the tolerance on tooth profile and tooth spacing must not exceed .001. The critical characteristic of this machining operation is the minimum root and crest radius permitted. Corner radii could not exceed .001.

The "direct" method of applying the Ultrasonic technique to the cutting tool was again utilized, requiring an abrasive slurry and suitable tooling and fixturing for the operation. The tool material was a tough grade of tool steel, semi-hardened and having an extremely accurate tooth form ground in its periphery. In this instance, a quantity of two workpieces was required. Machining time was approximately 1 hour per plate.

The confined area in this workpiece configuration denied the use of conventional grinding practices. An attempt to form the part in a soft state and then hardening, produced such severe warping and distortions that it was considered an impractical technique. Other methods encountered draw backs in the control of the minimum root and crest radii control.

Typical Tool Room Example - Punch

Again, this part is not an example of production, but more of the tool room type of machining requirement. Only one punch body was required, and once more illustrates the advantage that the Ultrasonic technique of machining offered.

The material for this punch body was carbide (see Fig. 9). The tolerance on the profile was ±.0005. Finish required was 8 microinch or better and to be true of every surface of the fuse body. Corner radii in the profile was limited to a maximum of .002.

The "direct" method of Ultrasonic machining was again employed. An abrasive slurry with proper tooling and fixturing on an Ultrasonic machine tool was utilized. Tool material, in this case, was stainless and not hardened. Machining time was approximately 2 hours.

One other possible method of manufacture considered was to have the carbide pre-sintered to its final form. This presented two objections, the first being the high cost of the forming mold, and secondly the lack of predicted control of profile tolerance and minimum corner radius requirements. Grinding of the profile was considered impractical because of the recessed profile configuration. Other methods of fabrication considered were doubtful because of the minimum corner radius requirements and the absolute need for fine finishes on every part of the formed profile.

Ultrasonic Deburring of Metals

Ultrasonic deburring of metallic parts on a production scale has been gaining favor in manufacturing circles. Its proper application, however, has never been understood too well, nor have the benefits that can be derived from that type of operation been fully recognized. It is, therefore, the purpose of the following section to identify its proper application.

Basic Requirements

In order to offset the cost of equipment investment, the process should be devoted only to the deburring of large quantities of parts. Workpiece materials may be of most types of metals and may be either soft or hard. The burrs to be removed must have been previously considered impractical by any other conventional method, such as tumbling or blasting, etc. The actual size of burr must be minute. This, naturally, eliminates the type of burr that one would find after any metal cutting operation, such as lathe or drill presses. Deburring Ultrasonically becomes practical when you are considering the removal of feather edges, so to speak, from grinding operations. Further, this type of deburring becomes more practical when it is necessary to have a maximum of .005 corner radius between two intersecting surfaces. Resulting finishes from the deburring operation should be very good, generally 10 microinches or less and definitely without any resulting damage to adjoining finished surfaces.

Typical deburring examples by the Ultrasonic

technique are briefly discussed below (see Fig. 10).

Extremely fine pitch gears, generally used for instrument purposes, form the larger part of practical examples for this type of deburring operation. Gears of 120 pitch are common where the extent of burr removal must be complete and with minimum corner radii. Another application is one of absolute burr removal from parts found in compressor devices of the sealed type. Here, devices that are under warranty and subject to replacement free of charge, must operate continuously without failure in order to be practical devices. Microscopic sizes of foreign, contaminating materials do foreshorten the life of such equipment, and deburring represents a critical item in the course of component manufacturing. All surfaces generally of such parts are lapped to obtain good wear life characteristics and to maintain sharp intersecting faces. The Ultrasonic method of removing the fine feather type of burrs left even after a lapping operation is practical. Further, such work can be accomplished without damage to the lapped surfaces and most often can be accomplished without having to mask out other areas not to be burred.

Again, this is an example of "direct" application of the Ultrasonic transducer to a properly shaped cutting tool and is used in conjunction with an abrasive slurry. No other power source is required for the cutting tool. This, however, is offset by the fact that the equipment is designed for continuous, conveyorized type of operation, not requiring an operator to be in constant attendance. The equipment normally appears as a device that we have more frequently seen, the production degreaser. The powered cutting tool is located in a tank containing the proper abrasive slurry. The parts are automatically conveyed and properly positioned opposite the cutting tool for a planned period of time while undergoing the deburring operation. An attendant would only be required for loading and unloading equipment and inspection of deburred surfaces.

Generally speaking, simple forms of tooling are required, although complex workpiece shapes can be handled more efficiently through the use of correspondingly shaped cutting tools. The deburring, for example, of intersecting surfaces, internally located in the workpiece such as counterbores, can be handled by special tooling which probe the interior of such cavities. Operations of this type do require special fixturing and processing, but when set up, have automatic characteristics.

The deburring operation is accomplished by bombardment of intersecting faces with very fine grained abrasive particles. The feather like burrs that are to be removed are weak in their mechanical fastening to the workpiece itself, and, consequently, are quite easily removed by the closely controlled bombardment of small abrasive grains.

Ultrasonically Assisted Metal Cutting Tools

This following discussion fully recognizes the fact that the utilization of Ultrasonics with metal cutting tools is a future consideration, and not currently available to manufacturing today. It represents a research venture, which when undertaken, may require some two to three years before useful benefits can be derived. However, characteristics are interesting, and for that purpose are presented here for the benefit of furthering the knowledge of potentials that do exist.

Application Area

At times, with full utilization of the best conventional lathe, mill, etc., machining techniques, some materials offer complex problems in tool life commensurate with practical stock removal rates. In these instances, cutting edge break down on tools is rapid, irrespective of the lubrication or cooling medium, tool materials, etc. that may be used. Stock removal rates are limited. Greater rates are desired.

Future Research Relationship

It is generally agreed that the extreme, high, localized heat development at the tool cutting edge is the main problem concerned with cutting edge break down. The factors involved in producing this undesired feature are involved and it is not the point of this paper to go into those descriptions. As we all know, many different things have been devised to relieve this situation. Accomplishments have been made. However, we are suggesting that additional progress may be made by resorting to a technique that has not been fully investigated. Through other experimental results, it has been determined that an Ultrasonically energized cutting tool can reduce the frictional coefficient produced by the chips in a lathe application. A reduction in this characteristic is directly related to the degree of heat

concentration at the weak part of the cutting tool; namely, its cutting edge. Additionally, it is believed that the actual stress pattern occurring at the edge of the tool is also altered beneficially. In the Ultrasonic approach, an "indirect" method of application is involved. Although the Ultrasonic transducer may be coupled directly to the cutting tool, it only represents a part of the total input power used. The conventional feeding of the tool is still utilized with Ultrasonic energy superimposed upon that power input.

Basic results of experimental endeavors to date indicate that the Ultrasonic assist technique can be applied to rotating, multiple edge cutting tools as well as non-rotating, single edge tools. Frequency and mode of vibration variations can be designed to cover the many types of operations involved and potentially be effective.

Future Potential

Very much is unknown about the operating characteristics of this technique. This is the reason it is described as a research endeavor for the future. However, based on what is already known about the grinding application with Ultrasonic assist; it is believed that many comparable observations can be made and approaches to the problem duplicated. Future potential, therefore, appears very good for this technique.

Ultrasonically Assisted Grinding

Although the following discussion does not represent a technique that is currently available to manufacturing, it nevertheless represents a development that could be potentially available within the next one to two years. The new technique involves conventional grinding practices but Ultrasonically assisted so as to overcome problems in certain material applications.

Application Area

Utilizing the best of conventional grinding practices and equipment, some materials are not responsive. Basically, these problems involve the current high strength, high temperature types of materials, examples of which are Inconel, Titanium, 17-PH stainless, Unitemp 212, and others such as Beryllium, Columbium, Molybdenum and Tantalum. Stock removal rates by conventional grinding are low and wheel redressing is quite frequent with low wheel life. Extremely high and very localized temperatures produce warping or distortions in the workpiece itself with an accompanying high wheel loading condition.

In all applications, we are referring to the use of grinding as opposed to any of the metal cutting tool operations, such as lathes or mills. Grinders referred to may be center, centerless, internal or surface types.

Development Relationship

Ultrasonically assisted grinding appears to provide the potential needed to relieve the poor characteristic of wheel loading and the high localized workpiece temperatures resulting in distortions and warping. Tests have been run on surface grinders with most practical results gained by coupling of the Ultrasonic transducer to the rotating grinding wheel (see Fig. 11). In this case, the "indirect" type of approach is utilized, wherein the grinding wheel receives two types of power input; namely, Ultrasonic along with the rotational energy required for the wheel. It has been determined that the mode of vibration and the selection of proper vibrational frequencies can be matched to the type of grinder involved to gain definite benefits over the conventional use of the same equipment.

Future Potential

Although this technique is still in the development stage, the results are very encouraging. It also appears that this style of grinding can be applied to many of the types of grinding machine tools available today. Also possible is the fact that new wheel and grinding standards may be developed to further increase the presently successful and conventional techniques of grinding through utilization of the Ultrasonic assist medium.

Summarization

The following sections will summarize the highlights of the topics discussed in this technical paper on "Ultrasonics for Machining".

Present Day

In all cases where Ultrasonics has been successfully applied, the "direct" approach has been involved. The characteristic of this process is the direct coupling of an Ultrasonic transducer to a cutting tool, representing the only type of power input to the tool itself.

Normally, the mechanical vibrational energy available at the tool is used to drive abrasive grains in a slurry into the workpiece to accomplish the required job.

In those cases where major stock removal applications are involved, the workpiece materials may be hard steels (i.e. - Rockwell "C" Scale hardness 59 minimum) or carbides. Generally speaking, the smaller types of shaped cavities form the type of machining operation. Finish, corner radii and tolerance on profile are critical parameters. Further, metallurgical and physical disturbances cannot be allowed.

In Ultrasonic deburring applications, most metals may be considered, either in the hard or soft state. Burrs that are to be removed are of the minute type. The operation, further, is practical where minimum corner radii and an absolutely burr free condition prevails in the operation requirements. Adjoining finished surfaces must not be affected in any way nor can metallurgical or physical disturbances be allowed.

Future Development

As a technique that is currently in development, Ultrasonically assisted grinding appears to be potentially good for the grinding of those materials that do not respond under conventional practices. Additionally, the potential appears good for the improvement of present grinding techniques for currently successful applications. This technique falls under the "indirect" approach, wherein the grinding wheel receives power input from two or more sources; namely, Ultrasonic and at least rotational input power.

Future Research

This, again, is an "indirect" application involving Ultrasonic energy plus other types of power input to metal cutting tools, such as found on lathes, mills, etc. Research work is required, and as soon as such a program would be started, benefits could reach manufacturing practices in a period of some two to three years following.

Research work in this field could be rewarding because it is felt that the life of tool cutting edges can be vastly improved from the really rough jobs of today. It is also believed that potential benefits can be applied to many different types of metal cutting tool applications involving not only the non-rotating, single edge type but also the multiple edge, rotating type.

Future Potential Benefits

Ultrasonics as a science presents a brand new approach to many problem areas of today's industrial field. Ultrasonics provides, we believe, added impetus to the research technician for greater accomplishment in the art of metal cutting. It might also be said that it provides him with an entirely different outlook on problem solutions, completely removing him from a long line of smaller improvements on basically old line ideas or technologies. Further, Ultrasonics as already applied in other areas, provides many ways and means of evaluating problems never before possible. Currently available are Ultrasonic gage and flaw detectors along with devices for stress analysis. Another provides a unique way of feeding tools in machining operations with infinitely variable but accurately and uniformly controlled motion. In the grinding wheel field, a device provides a method of grading grinding wheels never before possible.

Encouraged wherever practical should be research and development work employing the use of Ultrasonics, either directly or indirectly in the metal machining field. It is hoped that this paper will point out the many opportunities for resolving problems that have never before been satisfactorily answered over a period of many years.

Bibliography

"Ultrasonic Grinding of Super Alloys" - Sheffield Corp. - Work reported made possible through support and sponsorship extended by the Manufacturing Technology Laboratory, AFSC, Aeronautical Systems Div., USAF, Wright-Patterson Air Force Base under Contract No. AF33(600)-40122 Project ASD 7-757.

"Ultra High Speed Machining" - Lockheed Aircraft Corp. - AMC Project 7-635 - Contract No. AF33(600)-36232.

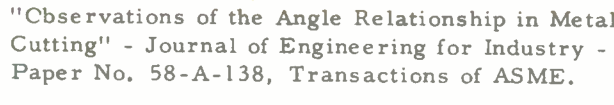
"Self-Excited Vibrations in Metal Cutting" - Journal of Engineering for Industry, Paper No. 58-A-189 - Transactions of ASME.

"Vibrations of Flexible Precision Grinding Spindles" - Journal of Engineering for Industry, Paper No. 58-A97 - Transactions of ASME.

"The Mechanism of Chip Formation" - Research Report #9 - May 1, 1958, ASTE Research Fund, Gideon, Simon and Grover.

"New Techniques in Metal Cutting Research" -

Transactions of the ASME, October, 1956, Backer and Krabacher.
"Present Knowledge of Cutting Fluids" - Research Report #11, May 1, 1958, ASTE Research Fund.


"Observations of the Angle Relationship in Metal Cutting" - Journal of Engineering for Industry - Paper No. 58-A-138, Transactions of ASME.

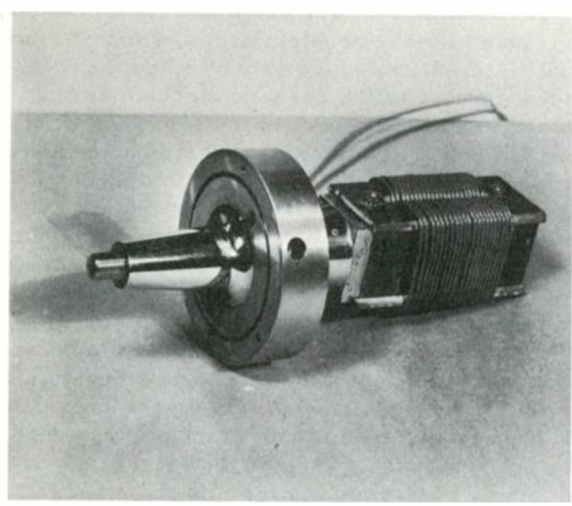


Fig. 1. Typical ultrasonic transducer. Housing, toolholder and cutting tool assembly-not shown. Power input end to toolholder at threaded section. Mechanical motion parallel to long axis - mounting in housing at large ring section.

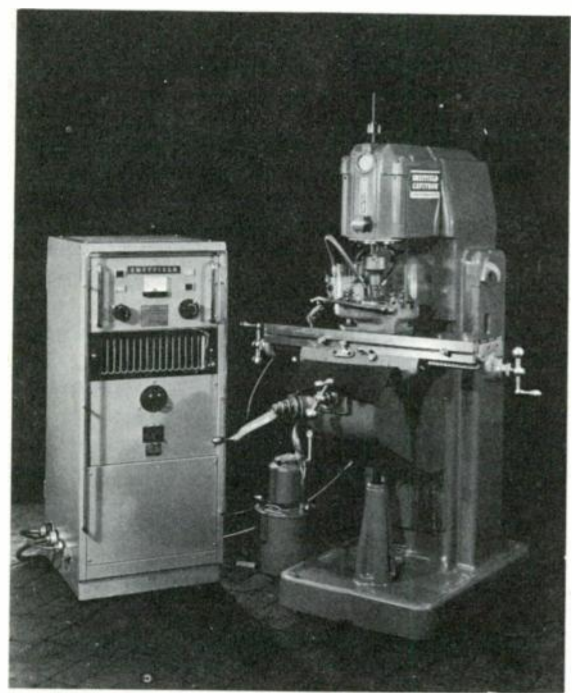


Fig. 2. Typical electronic generator with ultrasonic machine tool. Electrical output of generator (controllable) is to coil winding of transducer. Toolholder with cutting tool shown fastened to output end of transducer. Down feed of tool controlled by slide in head of machine. Pump and feed lines delivery abrasive slurry to work in table mounted chuck.

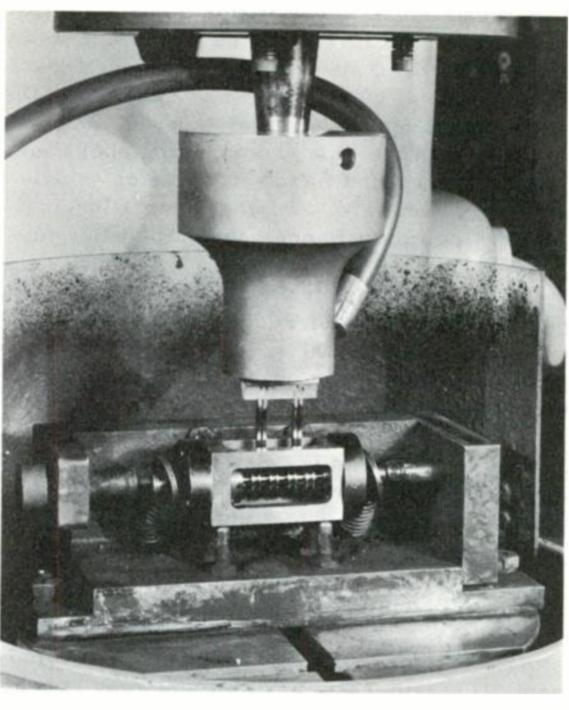


Fig. 3. Hydraulic servo valve machining setup. Valve body in supporting fixture is located in chuck of machine. Ultrasonic toolholder and cutting tools are shown in cutting position for one pair of parts. Hose shown behind tool ready to flood work area with abrasive slurry during machining operation.

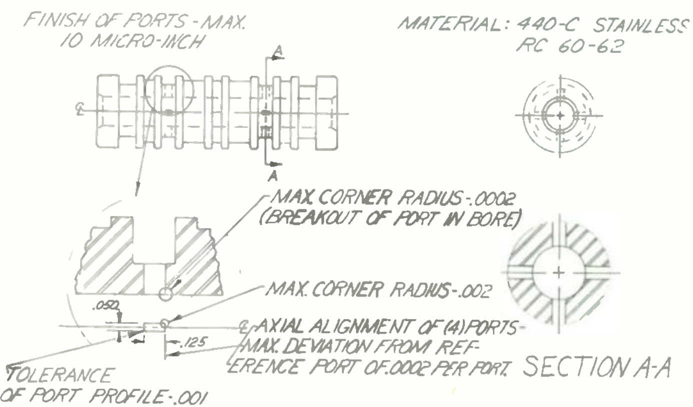


Fig. 4. Hydraulic servo valve body. Ports areas to be machined shown in sectional views.

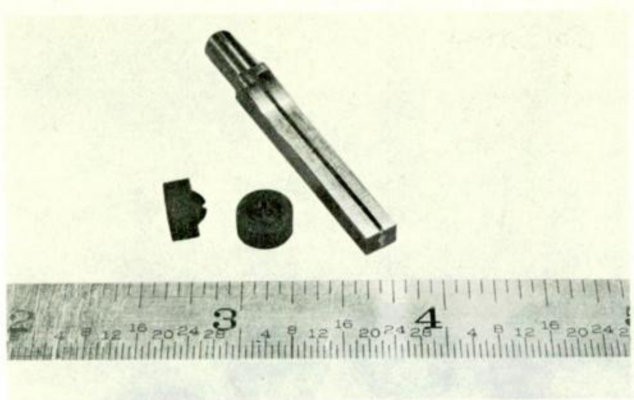


Fig. 5. Carbide spray nozzle and "V" notch cutting tool. Tool shown before mounting to ultrasonic toolholder.

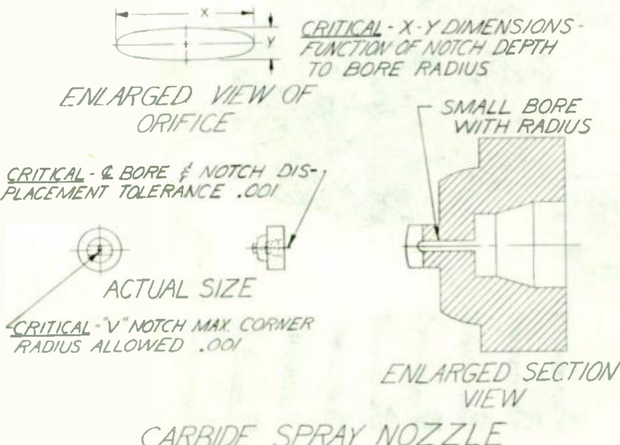


Fig. 6. "V" notch machining area of nozzle.

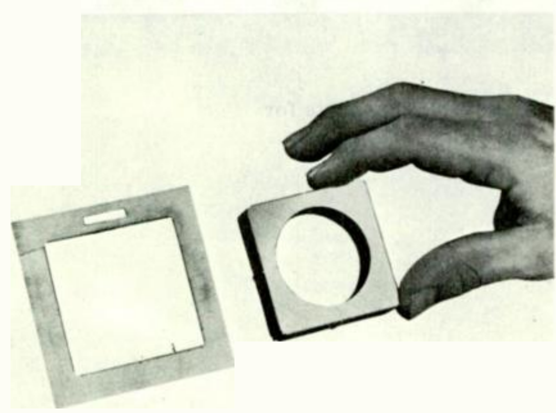


Fig. 7. Camera aperture plate and tool. Hand held object is tool prior to mounting on ultrasonic toolholder.

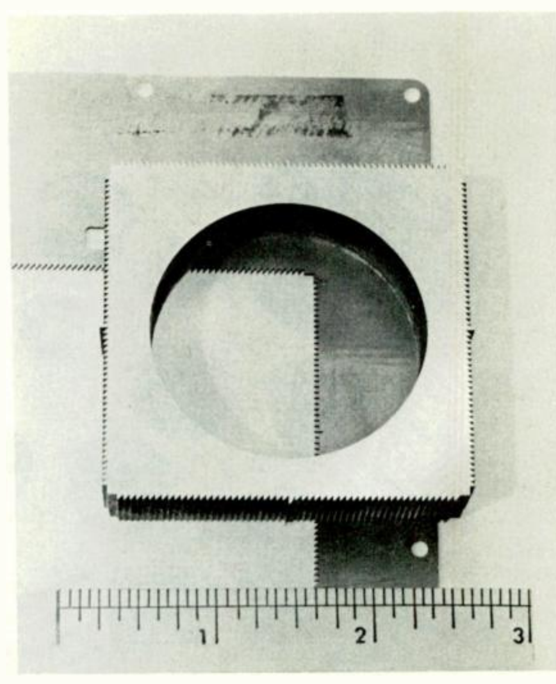


Fig. 8. Enlarged view of plate and tool.

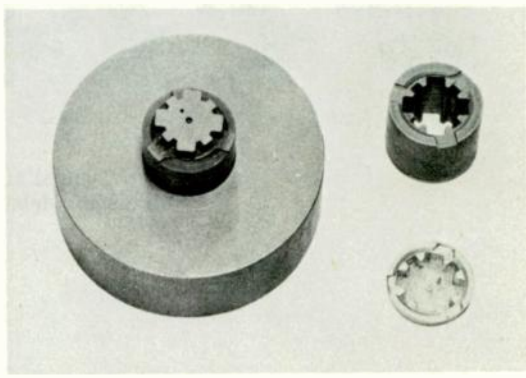


Fig. 9. Carbide fuse body punch (l.h.), cutting tool (upper r.h.) and punch product (lower r.h.).

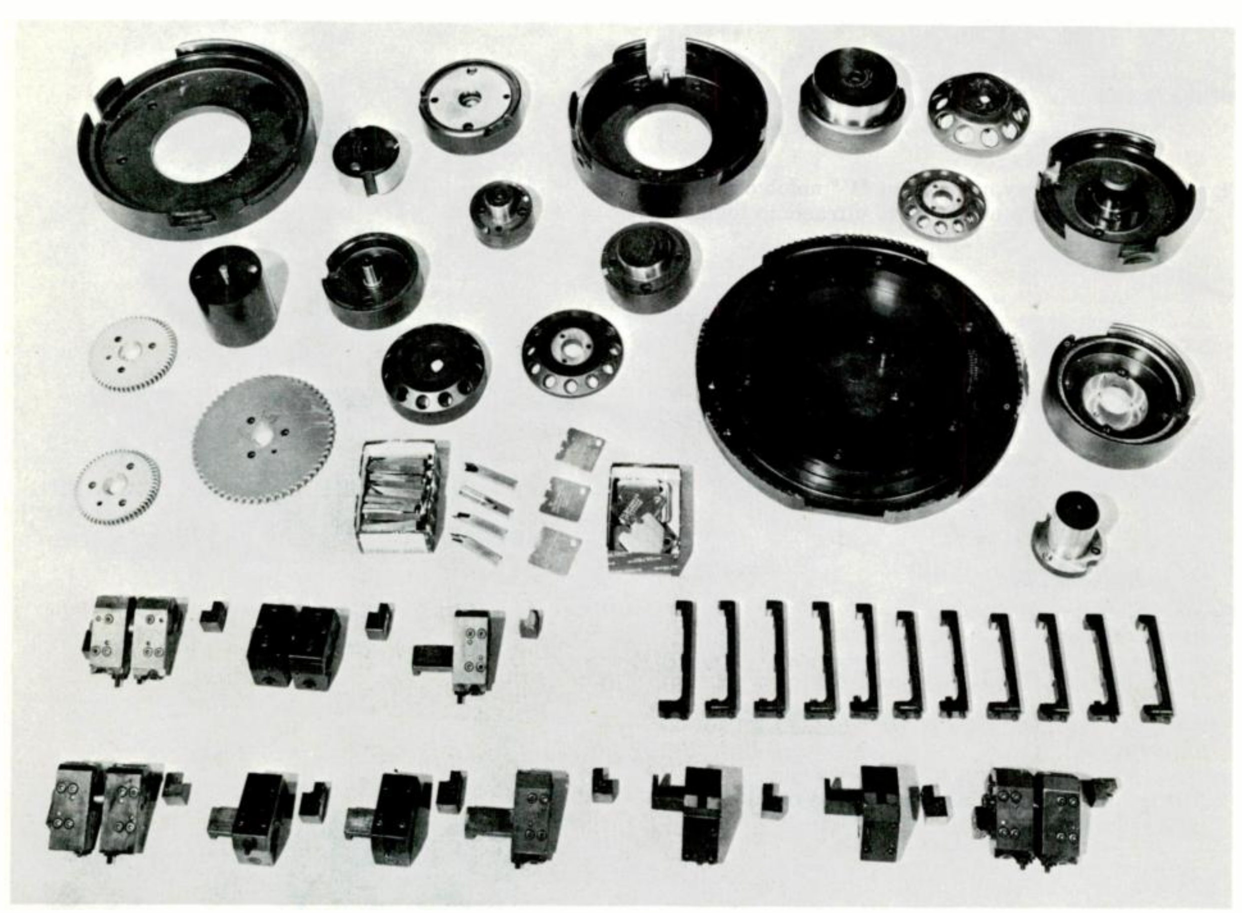


Fig. 10. Typical small gear and compressor parts for ultrasonic deburring.

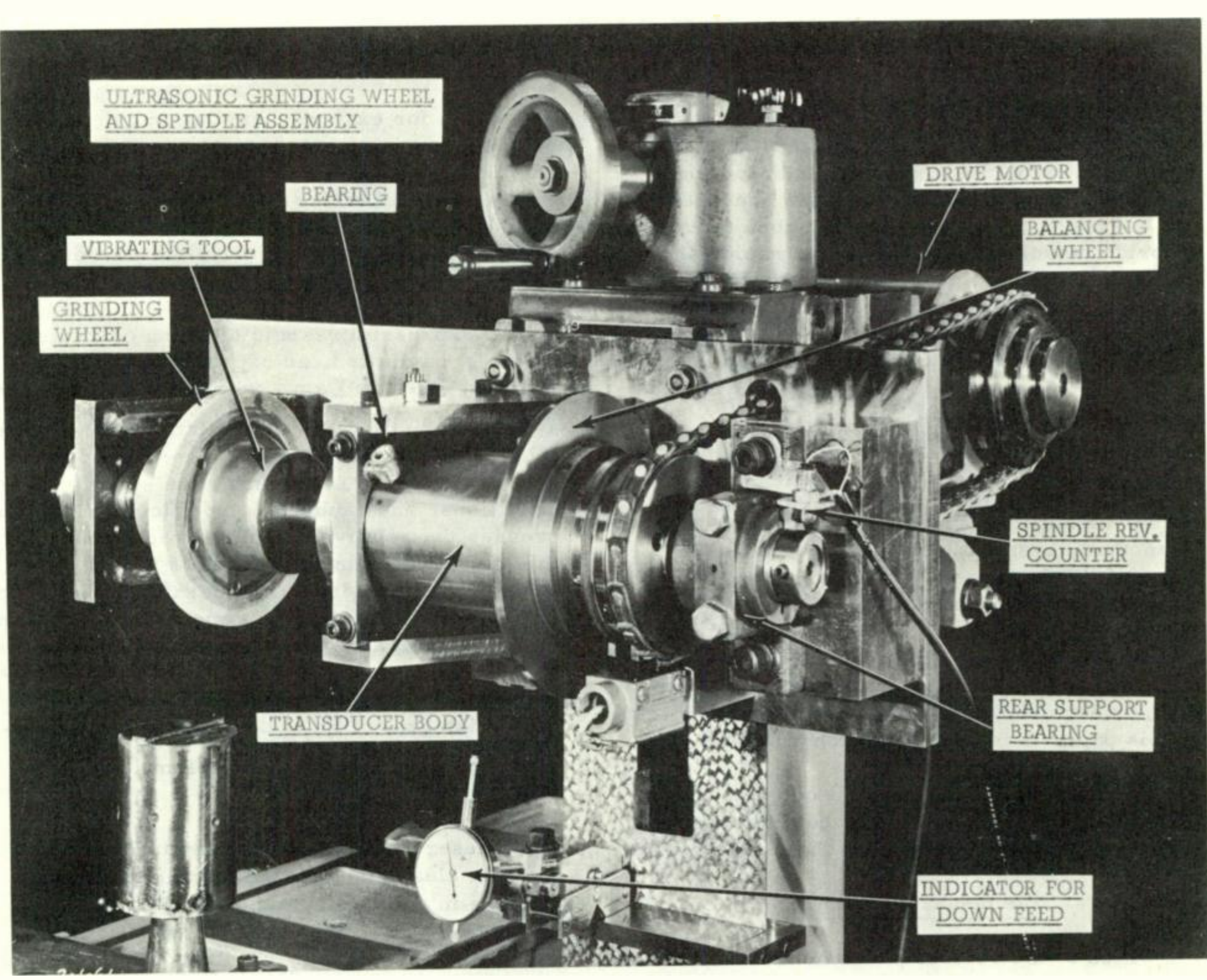


Fig. 11. Ultrasonically assisted grinder used in development work for grinding.

A CHEMICAL METHOD FOR MEASURING RELATIVE AMOUNTS OF CAVITATION IN AN ULTRASONIC CLEANER*

Alfred Weissler**
Laboratory of Technical Development
National Heart Institute
Bethesda 14, Md.

Summary

A simple colorimetric procedure is described for comparing the amount of cavitation in an ultrasonic cleaner under different conditions. The method is based on the sonochemical liberation of chlorine from carbon tetrachloride dissolved in water. It is rapid, and appears to have a linear response. With this method, some studies have been made of the effects of such variables as temperature, line voltage, and position in the cleaning tank.

- - - - - - - - - - - - - -

Introduction

A frequent problem in ultrasonic cleaning is the variability of the results from one time to another, even with the same equipment. Some of the factors which have been held responsible are tank overloading, poor degassing, and ageing of the equipment. Even so, it would be useful to have a simple method for measuring the cleaning efficacy, or something related to it, under different conditions. The purpose of this paper is to describe such a method.

The method is chemical: it consists of measuring the amount of sonochemical change in a test solution immersed in the cleaning tank.

Elementary Theory

Which of the various phenomena in an ultrasonic wave is responsible for the cleaning effect? The accepted answer[1] is "cavitation", which refers to the formation and violent collapse of cavities in the liquid as a result of the cyclical pressure changes. However, the concept of cavitation[2,3] is not defined with high precision: for example, cavitation can be gassy or vaporous, also steady-state or transient, and the bubbles can be of various diameters, including the size which is resonant at the operating frequency. The relative importance of different types of cavitation in the cleaning process is not yet known.

To express ultrasonic cleaning ability in terms of cavitation, an expression which suggests itself is $\sum a_i n_i$, where each i represents a characteristic class of cavitation events, each n is the number of events of this class per second per unit volume under the specific set of conditions, and each a is a coefficient which expresses the relative cleaning effectiveness for this class of cavitation. The n's will be functions of such factors as temperature, frequency, dissolved gas content, treatment time, location within the cleaning tank, and probably many others. A more elegant expression for the ultrasonic cleaning effectiveness would be the double integral of $\sum a_i n_i$ over the time of treatment and over the specified volume within the cleaner tank.

It deserves mention that cavitation causes not only cleaning, but also certain chemical changes in aqueous systems, which can be measured with good sensitivity and precision. The quantitative ability of ultrasonic cavitation to cause a particular chemical reaction could be represented, by analogy with the preceding paragraph, in terms of the expression $\sum b_i n_i$. Here, each b is the coefficient of relative sonochemical effectiveness for each class of cavitation.

It would be straightforward to get a valid measure of cleaning ability for an ultrasonic cleaner, in terms of the amount of chemical change caused by the cavitation, if it is true that there is a constant ratio between each a and its corresponding b. We do not yet know

* Part of the work described was performed at American University, with support from Office Of Naval Research grant No. GD-55-62.
** Present address: Air Force Office of Scientific Research, Washington 25, D.C.

whether this condition is in fact satisfied. But in the absence of contrary evidence, it seems reasonable to proceed farther (with caution) on the assumption that the ratio a_i/b_i is more or less constant, at least for a particular ultrasonic cleaner at a particular frequency.

Chemical Aspects

Examination of recent sonochemical literature [4,5] discloses a high-yield reaction which is suitable for the present purpose: the ultrasonic liberation of chlorine from carbon tetrachloride dissolved in water. A sensitive method of analysis for the amount of chlorine produced is to add ortho-tolidine reagent and measure the intensity of the resulting yellow color[6].

Details of Method

Two solutions are needed. One is a saturated solution of carbon tetrachloride in water, made by vigorously shaking distilled water with 2 ml. of carbon tetrachloride for each liter of water, and allowing to settle overnight or longer. The second solution, ortho-tolidine reagent, is prepared by dissolving 0.675 g. of ortho-tolidine dihydrochloride in a mixture of 75 ml. of hydrochloric acid and 425 ml. of distilled water. Both solutions should be stored in brown bottles, in order to minimize deterioration.

The following simple procedure was evolved for comparing the amounts of cavitation in an ultrasonic cleaning tank under various conditions. Twenty ml. of carbon tetrachloride solution and 1 ml. of ortho-tolidine reagent were pipetted into a convenient vessel. An ordinary glass test tube serves well, but most of the work to be described was done with a cell made by cementing a 0.002" Mylar sheet to the bottom of a length of 1" diameter glass tubing, for the sake of greater transparency to ultrasound waves. The cell was clamped in place in the center of the tank, at such a height that the water level was the same inside as outside the tube. Then the generator was turned on for a specified time interval, usually 10 seconds. The intensity of the yellow color was measured in a 1 cm. cell in a Beckman spectrophotometer using blue light of 436 mμ wavelength, and recorded as the optical density (D_{436}) which is the negative logarithm of the fraction of light transmitted.

Inasmuch as the color fades slowly, it should be measured within 15 minutes. Also, it is not possible to keep the colored solution around as a permanent standard for visual comparison with future runs. Instead, the comparison is made in terms of the numerical value of the optical density, as measured by a colorimeter or spectrophotometer. Experiments showed that D_{436} is a linear function of the concentration of dissolved chlorine, and that a D_{436} value of 0.30 corresponds to one part per million of chlorine in the water, under the conditions used here.

Tests of the Method

In order to find out more about the practicality of the proposed method as a measure of relative cavitation, several studies were made. A 26 kc commercial ultrasonic cleaner was employed for the experiments discussed in the next several sections, the water height being maintained at 6" in the tank (which was 10" deep by 10" long by 6" wide) . For the work described in the latter part of this paper, a 28 kc cleaner from a different manufacturer was used, under slightly changed conditions. With either apparatus, the reproducibility of duplicate runs was generally between 5% and 10%. The temperature was in the range from 21° to 24°C.

One point of interest was the potential error which might be caused by adding too much or too little of the ortho-tolidine reagent. A series of measurements was made, using various amounts of reagent per 20 ml. of CCl_4 solution. The results, shown in Fig. 1, were that the color intensity is independent of the amount of o-tolidine over wide limits. Below 0.2 ml. of reagent, however, an appreciable decrease in color takes place.

A second item was the possible effect of temperature variations on the depth of color produced by a specified concentration of chlorine. This was investigated by irradiating 20 ml. of CCl_4 solution in the absence of o-tolidine, then pipetting 5.0 ml. aliquots into three different 50 ml. volumetric flasks, which contained 40 ml. of water at 1°, 20°, and 40° C, respectively. Next, 1 ml. of o-tolidine reagent was added to each, and the contents diluted to the mark and mixed. The optical density at 436 mμ was found to be 5% higher at 1°, and 5% lower at 40°, than the value at 20°. It was therefore concluded that temperature variations have a negligibly small effect on the results of this colorimetric analysis for chlorine.

Another question was whether this method gives a linear response to linear increments of

cavitation. A partial answer was obtained by an experiment in which the color of the solution was measured after each of several consecutive treatments in the tank, for 2.0 seconds each; these successive irradiations were taken as being uniform increments of cavitation. For additional information, the experiment was repeated at three different levels of cavitation, obtained by changing the line voltage for the generator from 120 to 100 to 80 volts.

Fig. 2 shows the results obtained. At both 120 and 100 volts, it is clear that the method responds linearly to uniform increments of cavitation in successive periods of time. With the line voltage as low as 80 v., the initial treatment period produces less color than the later periods (possibly because of the longer time to get optimum degassing of the test solution at a low level of cavitation). But even in this case the response is linear after the first 2 seconds of irradiation.

In Fig. 2, the three lines have different slopes, with the highest voltage giving the greatest amount of cavitation, as expected. If one plots the three slopes against the corresponding line voltages, the result (Fig. 3) is roughly a straight line which goes through the origin. In other words, over the range of 80 to 120 volts the amount of chlorine produced per second is proportional to the generator line voltage, which in turn was found to be proportional to the driving voltage on the transducers. This constitutes a further bit of evidence for the validity of the proposed method as a measure of cavitation.

A related experiment on linearity of response was performed with a 400 kc Ultrasonorator, which has a quartz transducer and provides a higher level of cavitation than the ultrasonic cleaners. If the plate current in the driver output stage is varied by changing the voltage on the type 813 tubes, and the chlorine produced in the test solution is plotted against plate current as in Fig. 4, the line is straight except at the beginning. The curvature of the lowest portion is thought, as before, to be due to the longer time required for optimum degassing of the test solution when cavitation is weak. It should also be mentioned that, as the plate current was increased, the chemical change and the cavitation bubbles became visible at the same threshold value.

Further Results

Line Voltage

Next studied was the effect of a very wide range of line voltages (obtained with a Variac) on the amount of cavitation in the 26 kc ultrasonic cleaner which has transistorized circuits and a sound pressure sensor and feedback loop for automatic gain control.

The results obtained are depicted in Fig. 5. One interesting feature of the curve is the partial flattening in the region from 100 to 120 volts, presumably because of the operation of the automatic gain control which tends to minimize the effects of fluctuations in the power lines.

When the line supply is 30 volts or less, no cavitation streamers are visible in the tank, and no yellow color is produced in the test solution. At 40 volts, occasional cavitation streamers can be seen, and a small but perceptible amount of yellow color appears. As the voltage is raised, the amount of chlorine produced in 10 seconds increases gradually at first, up to 70 or 80 volts, and then more rapidly, up to the normal operating range. The initial slow increase is attributed at least in part to the long time needed for weak cavitation to degas the solution to the optimum value for either cleaning or chemical change.

The transition region where cavitation begins, between 30 and 40 volts, can be recognized in various ways. Not only is there an appearance of bubble streamers and an onset of chlorine liberation, but also a higher-pitch component is added to the audible sound produced by the generator. Still another way to detect the cavitation threshold is to observe on an oscilloscope the output of a barium titanate probe immersed in the tank. Below the threshold, the pattern is a smooth envelope of 120 cycles modulating the 26 kc carrier, as portrayed on the left side of Fig. 6. Above the threshold, the pattern shows clearly the superimposed irregular peaks of cavitation noise (Fig. 6, right side).

Temperature

Inasmuch as ultrasonic cleaning is often done in hot liquids, it was of interest to determine how temperature affects the amount of cavitation indicated by the present method. The procedure used was to bring the water in the tank to some specified temperature between 10° and 70° C,

and then to clamp the cell containing the test solution in place and leave it in the tank at least 5 minutes for thermal equilibration. Next, the generator was turned on for 10 seconds, as usual, and also the panel milliammeter reading (which is proportional to the voltage on the transducers) was recorded. This reading was taken in order to have a means for correcting for such factors as the alteration of transducer impedance and loading with temperature.

The results obtained for D_{436}, the milliammeter readings, and the ratio of the former to the latter are listed in Table I. The last column is considered to provide a better (i.e., corrected) index of the change in amount of cavitation with temperature.

Table I. Effect of Tank Water Temperature on Cavitation

Temperature	D_{436}	Milliamperes	D_{436}/ma.
10°	2.04	0.94	2.17
20°	1.58	0.86	1.84
30°	1.58	0.94	1.68
40°	1.41	0.98	1.44
50°	0.93	0.97	0.96
60°	0.54	0.81	0.67
70°	0.39	0.78	0.50

Table I shows roughly uniform decrease in cavitation as the temperature is raised; extrapolation of the figures gives the result that cavitation would be completely eliminated by increasing the temperature to about 90°. In view of the well-known effect of high vapor pressures in inhibiting cavitation phenomena, it is not surprising that ultrasonic cavitation should disappear as water is brought up close to its boiling temperature.

The work described in the next three sections was done with the other ultrasonic cleaner, at an operating frequency of 28 kc.

Geometric Position

The distribution of cavitation intensity in the horizontal plane of the tank of the second ultrasonic cleaner was studied next. A 32x200 mm. pyrex test tube containing the usual solution was clamped vertically in one of the eleven positions about the tank shown in Fig. 7, and irradiated for 10 seconds. The tube was immersed to such a depth that the liquid level inside was the same as outside in the tank; the tank had been filled to a height of 7.5" with distilled water containing 0.1% of Glim detergent. Then similar runs were repeated in each of the other positions. A typical set of results is set forth in Fig. 7, in which the figures inside each circle represent the optical density at 436 mμ, multiplied by 100. It is strikingly apparent that the cavitation intensity is greatest at the geometric center of the horizontal plane, and is relatively very small in the corners.

Water Height in Tank

Another variable investigated was the height of water in the tank, in its effect on the amount of cavitation. Starting with a water height of 9" in the tank, successive measurements of chlorine production were made (this time with a 20x150 mm. test tube as the cell) at heights down to 6", in 1/8 inch steps, by removing water as required. At each height, the vertical position of the tube was adjusted to maintain the same level of liquid inside the tube as outside. A periodic dependence of cavitation intensity on water height was found, as illustrated in Fig. 8. Inasmuch as the distance between consecutive maxima corresponds to a half-wavelength of 28 kc sound in water, the effect may be understood in terms of the cyclical variations in tuning of the water column which loads the transducer, giving a greater or lesser intensity of the standing wave.

Duration of Continuous Operation

The last factor looked into was the effect of continuous operation on the level of cavitation. Here, somewhat greater variability was found, among similar experiments performed on different days, than in the two preceding cases. The procedure was to insert the test tube containing the CCl_4 solution into a holder which positioned it in the tank as previously described, and then to withdraw it rapidly at the end of 10 seconds of treatment,add 1 ml of o-tolidine reagent, and measure the optical density. The generator was left operating continuously. Every few minutes, the entire process was repeated with a fresh 20 ml. portion of carbon tetrachloride solution. Fig. 9 shows that there was a generally declining trend in cavitation with increasing time of continuous operation. However, if the tank was filled with fresh distilled water (supersaturated with air) the amount of cavitation initially was low, then increased for the first 10 minutes, and later underwent a gradual decline.

Conclusion

A chemical method for measuring relative amounts of cavitation has been described, which is simple and rapid. Its response is approxi-

mately linear, and its usefulness has been shown in studies of the effects of certain variables. Even though no proof is now available that there is a constant relation between chemical-effect cavitation and cleaning-effect cavitation, such a relation might be established by comparative measurements in the future.

Acknowledgments

The assistance of Elizabeth J. Hine and Lenore E. Weissler in some of the experiments is acknowledged with thanks.

References

(1) L. D. Rosenberg, "On the Physics of Ultrasonic Cleaning," Ultrasonic News, vol. IV, no. 4, pp. 16-20; Winter 1960.

(2) Phillip Eisenberg, "Modern Developments in the Mechanics of Cavitation," Appl. Mech. Rev., vol. 10, no. 3, pp. 85-89; March 1957.

(3) M. Strasberg, "Onset of Ultrasonic Cavitation in Tap Water," J. Acoust. Soc. Am., vol. 31, pp. 163-176; February 1959.

(4) Alfred Weissler, "Some Sonochemical Reaction Yields," J. Acoust. Soc. Am., vol. 32, pp. 283-284; February 1960.

(5) Bojan H. Jennings and Suzanne N. Townsend, "The Sonochemical Reactions of Carbon Tetrachloride and Chloroform in Aqueous Suspensions in an Inert Atmosphere," J. Phys. Chem., vol. 65, pp. 1574-1579; October 1961.

(6) "Standard Methods for the Examination of Water and Wastewater," American Public Health Association, Inc., New York, N. Y., eleventh edition, pp. 85-92; 1960.

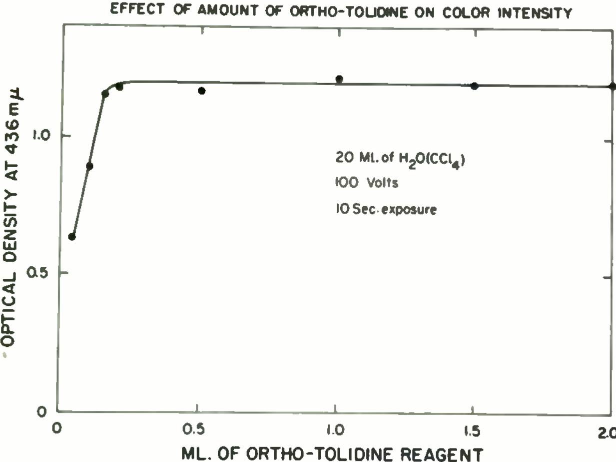


Fig. 1.

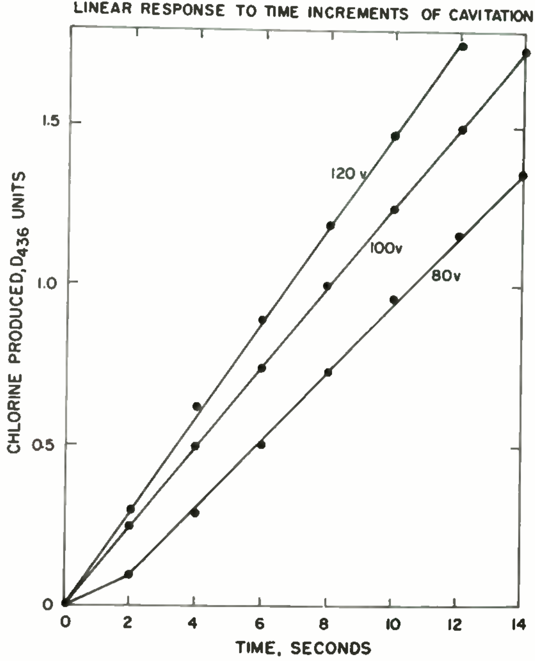


Fig. 2.

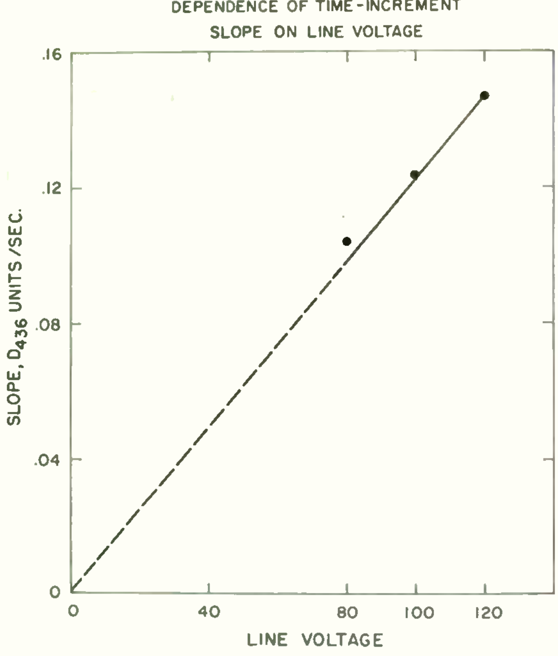


Fig. 3.

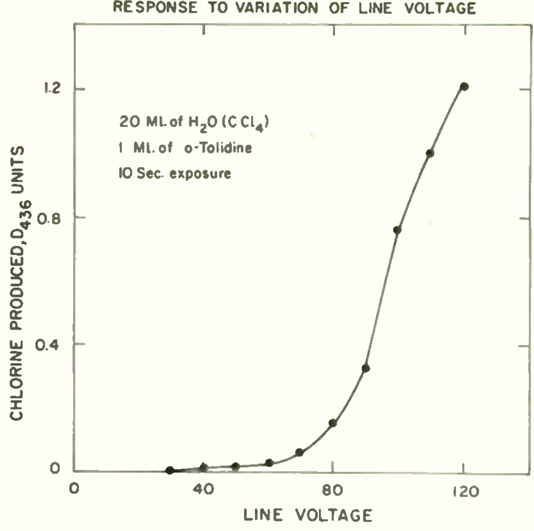


Fig. 5.

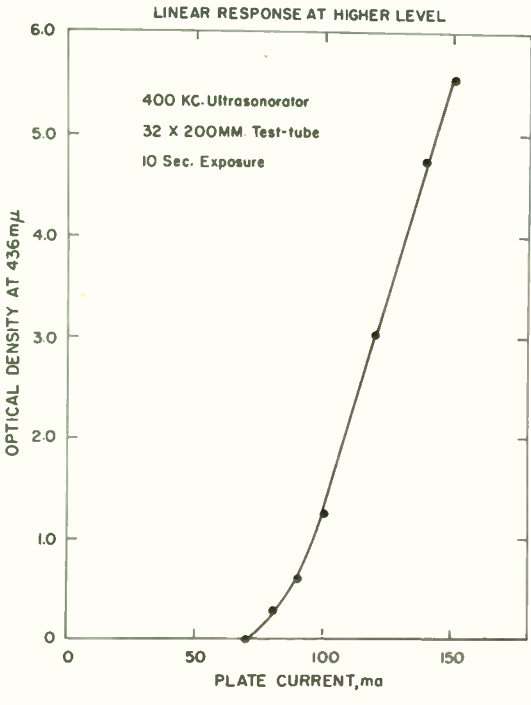
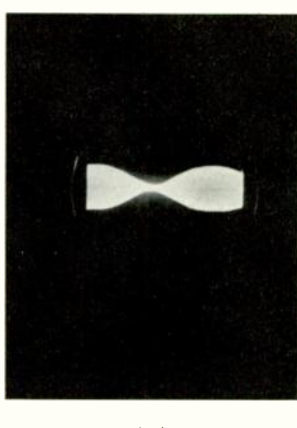


Fig. 4.



(a)

(b)

Fig. 6. Oscilloscope pattern. (a) Below cavitation threshold. (b) Above cavitation threshold.

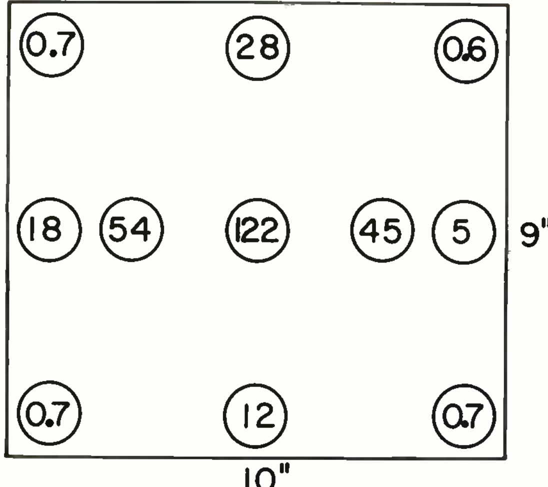


Fig. 7. Dependence of cavitation on position in horizontal plane.

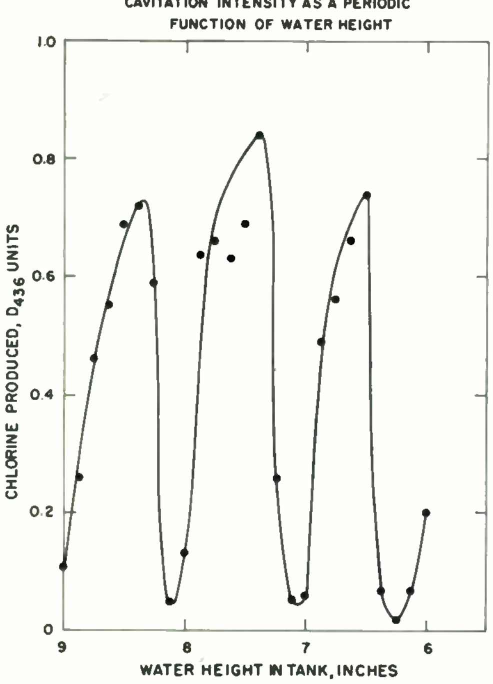


Fig. 8.

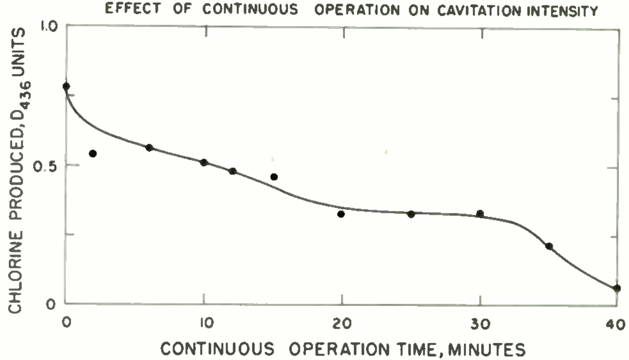


Fig. 9.

THE ADVANTAGES OF ULTRASONIC CLEANING AS A PRODUCTION TOOL

R. C. Fencl, Planning Engineer, Western Electric Company, Aurora, Illinois

Brett Hollerith, Sales Manager, Branson Instruments, Incorporated, Stamford, Connecticut

In industry to-day perhaps one of the most universal problems and one that is common to all manufacturing processes is that of cleaning. Whether the pieces being manufactured are missile components, ball bearings, printed circuit boards, steel castings or electronic assemblies, to mention only a few, the effectiveness of final and intermediate cleaning is of vital importance in achieving superiority of the ultimate product. It is the purpose of this paper to select a particular process in the manufacture of electronic equipment and show how ultrasonic cleaning is applied to it to make the cleaning process faster, less costly and more effective.

This process is the manufacture of a new class of telephone type relays. Slide #1 shows a typical relay which is characterized by card operation of pretensioned wire spring contact subassemblies. The major application of these relays is in automatic telephone switching systems. The basic design was developed by Bell Telephone Laboratories and manufactured through the cooperative efforts of the Western Electric Company. The basic parts of the relay consist of an armature, coil and core assembly, three molded wire spring blocks with a maximum of 24 contacts, and a moving card all held together by a spring clamp which keeps the parts in rigid alignment. A plastic molded dust cover, not shown, over the contacts keeps them free of dust particles and other insulating contamination which produces contact "opens". The expected life of this type of relay is 300 million operations. It is apparent that initial cleanness of the relay and particularly of the relay contacts is of vital importance in assuring this operating life. The final stage of manufacture must include a cleaning operation which will remove any dust particles, oils, finger marks and similar soils. It is also apparent that the cleaning process must be sufficiently thorough and yet of such a nature as to avoid damage to the contacts. Such a machine meeting these requirements has been in use for a number of years. It employs mechanical brushes in conjunction with an approved chlorinated solvent. Such machines will meet the cleanness requirements at the desired high production rate with an insignificant number of rejects.

There have been, however, a number of factors indicating the need for improved equipment as shown in slide #2.

The construction and configuration of the relay necessarily limits selection of the type of cleaning process. These same factors also make it undesirable to subject more than the contact area to the cleaning solvent. Because of these limitations ultrasonic cleaning was considered as the only practicable alternate to the process in use and, equally important, as the alternate that might achieve the benefits described in slide #2. At the same time it was apparent that large savings could be immediately effected if an aqueous type solvent could be used to replace the chlorinated solvent because of the inherent comparative costs of these two basic types of cleaning agents.

A series of investigations was made to determine the suitability of ultrasonic cleaning. Based on earlier investigations by Koontz and Amron[1] aqueous solutions of non-ionic detergents such as Igepal CO-710[2] and Triton X-100[3] were selected as having no harmful effects on electronic device components. They offered the further advantage of very low cost. However, use of water solutions would require deionized water rinsing but, more important, special drying techniques not needed with solvent cleaners would also be necessary.

To determine the comparative cleaning ability of the ultrasonic/aqueous method versus the brushing/solvent method, both electrostrictive and magnetostrictive ultrasonic systems were evaluated.

[1] An Ultrasonic System for Eliminating Physical Contaminants from Electron Devices D. E. Koontz and I. Amron, A.S.T.M. Special Technical Publication #246.

[2] Manufactured by Antara Chemical Division of General Aniline and Film Corporation.

[3] Manufactured by Rohm and Haas Company.

The degree of cleanness attained from each of the three cleaning methods was measured using the replica sampling technique. This technique consists of inserting a precleaned clear thermoplastic strip .015" thick between the relay contacts; closing the contacts and maintaining contact pressure while subjecting the contacts to infra red heat lamps in a protective enclosure to soften the plastic sufficiently (about 140° C) and enable any soils to become embedded in the plastic. A plastic such as vinylite is used. After cooling the contamination will be set in the plastic and the impression of the contact on the clear plastic may be viewed under a 100 power microscope using polarized light transmitted through the plastic. With a microscope eyepiece having ruled squares 100 microns x 100 microns the particles of contaminating soil may be readily viewed and their sizes determined. As shown in slide #3 weighted values are assigned depending upon the particle size.

The larger the particle the higher is the weighted value. For example, with a requirement maximum of 7, a reject would result from one particle over 100 microns plus two particles in the 50-99 micron range.

Based on this test the results shown in slide #4 were obtained.

It is apparent that ultrasonic cleaning gave not only very substantially superior cleaning but also offered the possibility of lowered maintenance and certainly reduced operating costs as a result of lowered solvent costs. Actual comparative costs will be reviewed later.

Based on further investigation two systems incorporating electrostrictive type transducers were installed. These systems replaced the production handling capacity of the existing three brushing/solvent systems. The ultrasonic systems were designed as complete units to handle relays on a continuously moving conveyor through the cleaning, rinsing and drying cycles so as to present a completely processed part requiring no intermediate handling. Slide #5 describes this cycle in schematic outline.

Relays are manually loaded on the conveyor and held in position by a specially designed relay holder, each designed to accommodate two relays. The conveyor is constructed to carry two rows of relay holders mounted back to back on the conveyor. In this way double the production capacity can be handled by having a second operator loading relays. After loading, relays and holders are moved to the first tank for a 45 second ultrasonic exposure. A non-ionic detergent as described earlier is used with deionized water that is continuously recirculated and filtered to maintain solution cleanness. All components in contact with the cleaning solution in all tanks are of stainless steel or teflon. The 45 second cleaning time allows a 50% safety factor over the necessary 30 second cleaning time determined by earlier investigations.

It is important to point out that in this cleaning, as in the successive rinsing tanks, only the contact area on the bottom 1/8" of the relay is immersed. The liquid height in the tank is also maintained by overflow weirs, and in all tanks the direction of the liquid flow is countercurrent to the direction of parts travel. Another feature is the use of a pulsing mechanism which permits opening and closing of the relay contacts during cleaning, rinsing and drying to augment these operations.

Following cleaning the relays are carried to the second tank for a 15 second rinse in continuously flowing tap water. Tap water is wasted after use and not recirculated.

Final rinsing in the next tank is with continuously flowing deionized water for a 30 second ultrasonic exposure to assure complete soil removal. Both the ultrasonic rinse and cleaning tanks are heated to approximately 140° F. which is the temperature shown by Hueter and Bolt[4] to be optimum for ultrasonic cleaning in water solutions.

Cleaned relays move next through the dryer section where filtered forced air heated to 140° F. by regulated steam coils gives a 3-1/2 minute drying cycle. Relays are unloaded on exit from the dryer section.

Because the utility of this entire system is essentially a function of the ultrasonic cleaning and made possible by the advantages of ultrasonic cleaning, some of the features of the ultrasonic components should be emphasized. Ultrasonic energy in the initial cleaning tank results from six immersible transducers each having a radiating area of 8" x 9". They are placed in the bottom of the tank and radiate energy upwards through the solution. These are powered by a 1000 watt ultrasonic generator and are activated by lead zirconate titanate type transducer elements selected because of their

[4] Hueter, T. F. and Bolt, R. H. Sonics by John Wiley and Sons, 1955.

high efficiency and low replacement cost. The use of immersible transducers over transducers bonded directly to the tank and forming part of the tank simplifies maintenance and replacement. Similarly the ultrasonic rinse section utilizes four transducers of the same size and also mounted in the bottom of the tank and powered by a 500 watt ultrasonic generator.

In slide #6 a complete system is pictured showing the cleaning and rinsing sections, drying section and load and unload zones. Operators may stand on either or both sides of the load and unload zone and access to the interior of the system is under the drying section where there is a clearance of over 6 feet.

To meet the production requirements of a single plant two of these systems replace three of the brushing/solvent type units. Slide #7 illustrates the saving in plant floor space achieved. Where the older systems required 450 square feet the ultrasonic systems need only 260 square feet. Of particular interest, however, are other comparative cost data. These are illustrated in slide #8. For example, the yearly cost for solvent in the old system is over 280 times that in the ultrasonic system. Maintenance and replacement costs are over 2-1/2 times as high. Depreciation is almost three times as high. Labor costs are slightly higher. There is no heating requirement in the brushing/solvent system as is true with the ultrasonic units and which necessitates an additional power cost. The total annual operating costs, therefore, are less than 1/2 for the ultrasonic system.

In conclusion I would like to emphasize that the cleaning application discussed here has been with specific reference to the manufacture of relays. The methods described, however, with some modification, will be applicable to many other varied processes. The important advantages of ultrasonic cleaning will still apply:

1. Faster cleaning.
2. More efficient cleaning.
3. Lower operating cost.
4. Lower initial investment.

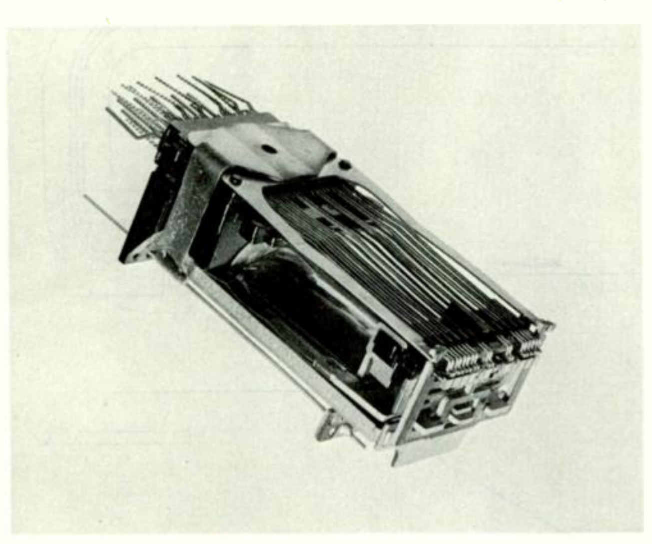


Fig. 1. Typical telephone relay.

REQUIREMENTS FOR IMPROVED CLEANING

(1) Improved cleaning.

(2) Lower Maintenance.

(3) Lower operating costs.

(4) Greater versatility.

(5) Smaller equipment size.

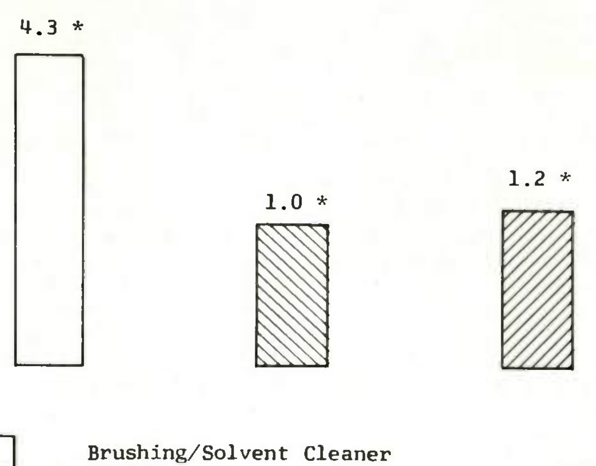
Fig. 2.

WEIGHTED SOIL VALUES ASSIGNED FOR REPLICA SAMPLING TECHNIQUE

Particle Size	Weighted Value
24-49 Microns	1
50-99 Microns	2
Over 100 Microns	4

Fig. 3.

Relative Cleaning Effectiveness



Brushing/Solvent Cleaner

Magnetostrictive Ultrasonic/Aqueous Cleaner
(Average Power Intensity 4.5 watts/in^2)

Electrostrictive Ultrasonic/Aqueous Cleaner
(Average Power Intensity 2.2 watts/in^2)

* Weighted value from replica sampling method.

Fig. 4.

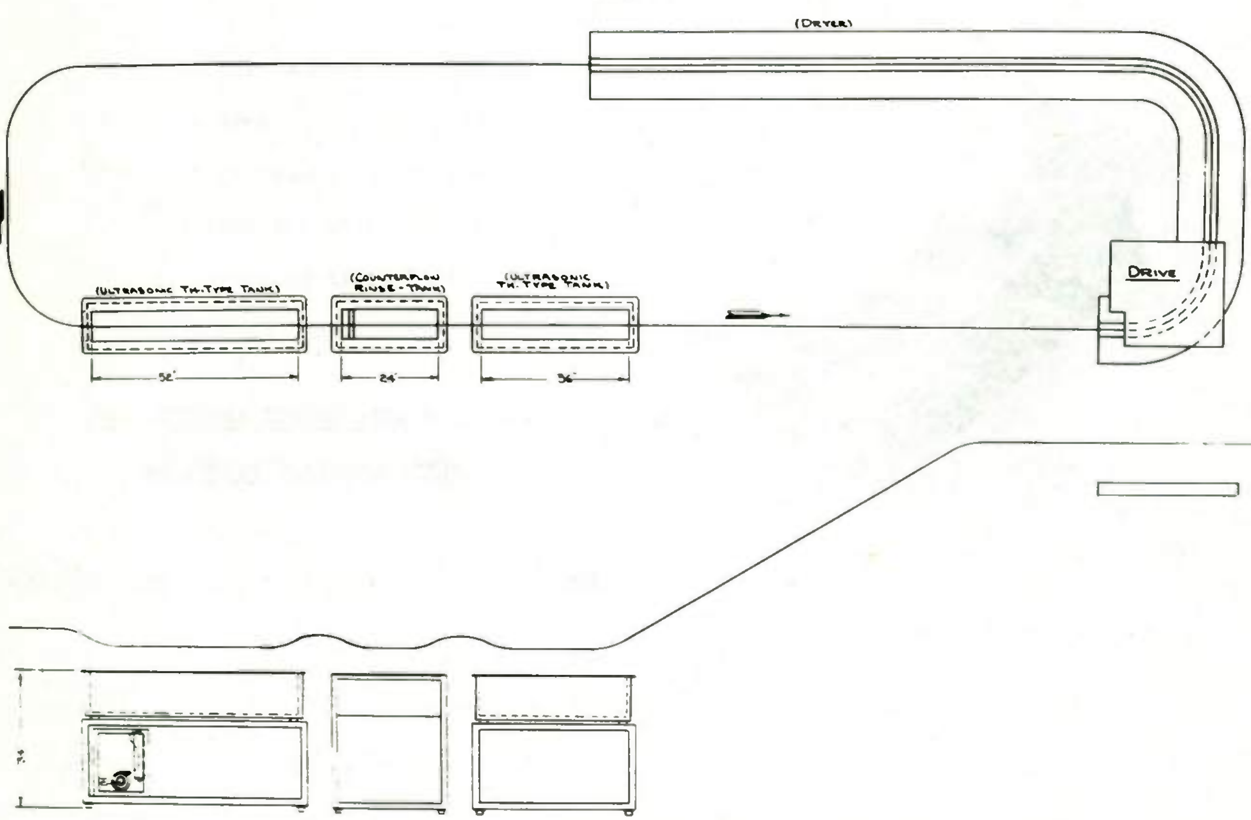


Fig. 5. Schematic diagram of ultrasonic cleaning system.

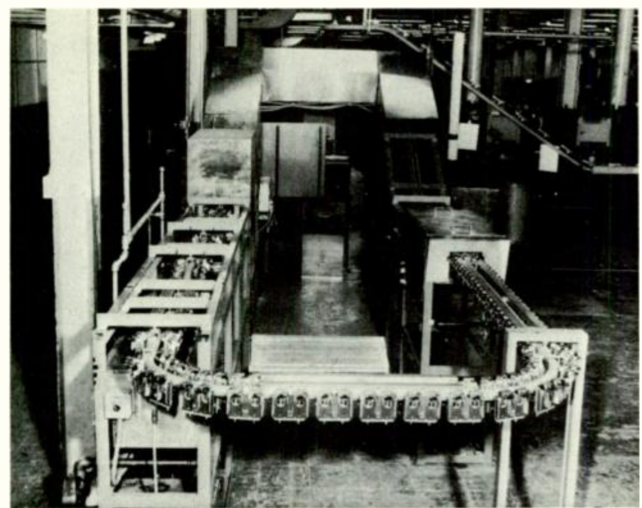


Fig. 6. Pictorial view of ultrasonic cleaning system.

Comparative Floor Space Requirements
To Meet Annual Production

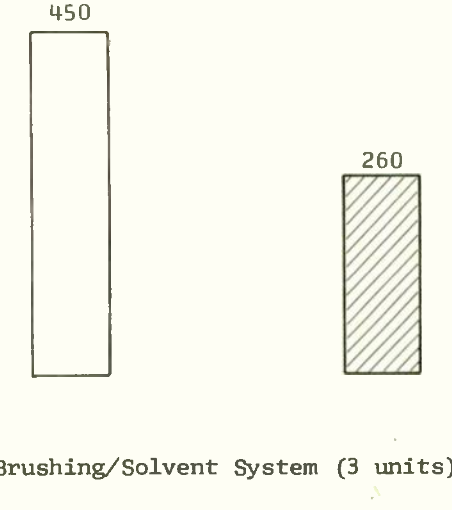


Fig. 7.

Comparative Annual Cost Figures for Brushing/Solvent Method
vs Ultrasonic/Aqueous Method

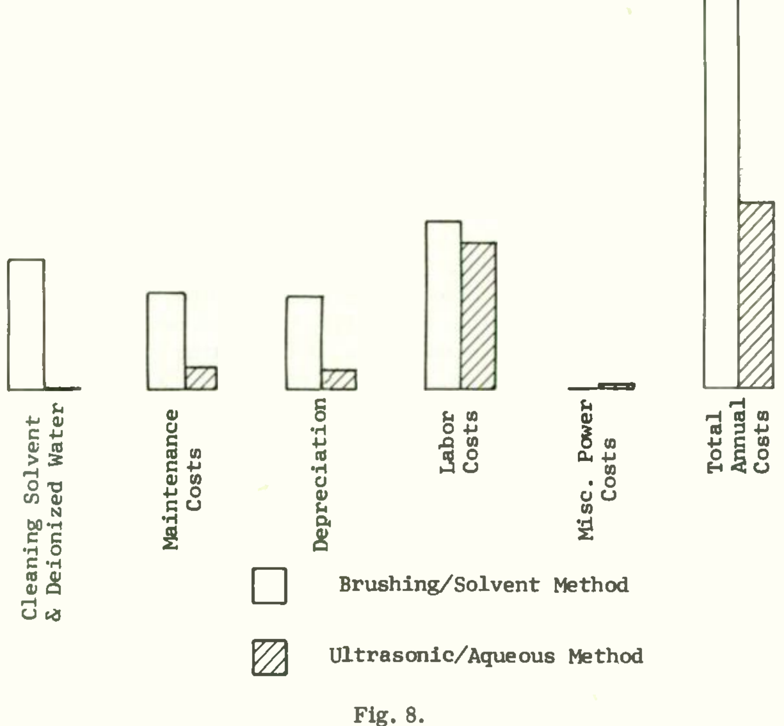


Fig. 8.

EVALUATION AND APPLICATIONS OF THE LEVAVASSEUR WHISTLE

Isidor Elias
Corporate Research Department
Acoustica Associates, Incorporated
Los Angeles 45, California

Summary

A Levavasseur Whistle of advanced design is described which is capable of producing high intensity sound fields at both sonic and ultrasonic frequencies in a gaseous atmosphere. The principle of operation is shown to depend on the interaction of the periodic deflection of a jet of gas with a resonant cavity. Siren-like behavior is thus achieved with no moving parts.

Results of free-field measurements show two distinct regimes of operation. For flows in which a pressure ratio across the whistle is sufficient to provide sonic flow, the tones produced are very intense and rich in higher frequency components. For flows of lower pressure, the sound field produced is more purely sinusoidal.

Frequencies of nominally 6, 10, and 21 K. C. are produced by variation of cavities. Intensities in the order of 160 db, and total power output as high as 80 watts are measured. Sound pressure level distributions at the mouth of the exit plane of the horn are presented. Auxiliary experiments designed to elucidate the sound generating mechanism of this device are described.

The Levavasseur Whistle being simple in design and construction, and capable of producing intense sound fields over wide ranges of frequencies offers many attractive applications to industry, and as a tool in scientific research. It is useful in aerosol coagulation, foam dispersion, and precipitation of fog, smoke, and dust. It is attractive as a catalytic agent in affecting chemical reactions. Above all, it is currently being used as a device for accelerating combustion in solid propellant rocket motors.

Introduction

There has been a need in both industry and in general scientific endeavors for a sound source capable of providing acoustic fields at controlled frequencies in both the sonic and ultrasonic ranges and at very high intensity levels. A number of devices have been produced which tend to satisfy some of these needs to varying extents. One particular sound generator, the Levavasseur Whistle, has some time ago been introduced to the industrial and scientific community (1). This device was particularly promising in that is was shown to be an efficient and inexpensive high power sound generator for use in air.

This paper reports on an effort conducted by Acoustica to evaluate a particular improved design of the original Levavasseur Whistle. This work is introductory to a more extensive effort designed to better document the behavior of this type sound generator. Considerations are given to the background of the device, to some aspects of its operation, and to performance results which were obtained from the testing conducted to date at Acoustica. Finally, a suitable applications of this device are discussed.

Physical Description

The Levavasseur Whistle is a sound generator capable of producing very high intensity tones at both sonic and ultrasonic frequencies. This device is a unique sound source in that it resembles both a siren and a conventinnal whistle in its sound-creating mechanism. In a siren, the sound is produced by a periodic interruption of an air jet. In the conventional whistle, the sound is produced by the interaction of a jet with a resonant cavity. In the Levavasseur Whistle, however, we have a combination of the two, there being a resonant cavity and an interrupted jet. As a result, the Levavasseur Whistle can be characterized as a "static siren", since a siren-like behavior is achieved without any moving parts.

Let us consider the physical construction of

the whistle in order to get a better understanding of the principles of operation of this device. Figure 1 shows in section a portion of the hardware items comprising the sound generating mechanism. Essentially there are two toroidal cavities (labeled A & B) separated by a circular slit. Air or any other desired gas is caused to flow under pressure through the slit. One of the cavities (A) has a knife edge which intersects the stream capturing the flow emanating from the slit and causes the flow to circulate along the circumference of the cavity. The directed air stream impinging upon the knife edge at sufficiently high velocities gives rise to edge tones which are a considerable contribution to the total acoustic power generated at the high output conditions. Upon emerging from cavity (A) the captured air jet strikes the primary air jet at right angles and deflects it toward cavity (B), causing the air to flow past the opening of (B), thus exciting (B) into resonance. The flow into cavity (A) is now interrupted and the primary jet, no longer being deflected by a transverse jet, is free to move back to its original path. The process now repeats itself, the periodicity being a function apparently of the time required for the air jet to traverse the circumference of the inner cavity. The frequency of this operation is given by an approximate expression $f = \frac{7}{d}$ (for sound generation in air) where f is in kilocycles and d in centimeters, is the diameter of the cross section of the toroid (1). Thus, we have a physical situation where a jet is chopped periodically and a cavity is excited by flow past its opening in a manner similar to that of a Helmholtz resonator.

Figure 2 shows a cross section of one particular design currently undergoing study at Acoustica. There are certain criteria for optimum operation which were suggested by the original design and which have been substantiated by Acoustica experiments. One of them is that cavity (B), the auxiliary cavity, have an opening larger than cavity (A), and furthermore, that the lip of cavity (B) be at a greater distance from the slit than the lip of cavity (A). The exact relative positioning of these cavities is not specified and, as will be shown later, is a function of the supply pressure of the gas being supplied to the whistle. Consequently, the design has incorporated in it a capability for moving the cavities relative to each other to determine experimentally the position producing the greatest sound output. The entire assembly is followed by an exponential horn to better couple the waves produced to the atmosphere into which it radiates.

As indicated above, the frequency of the fundamental of the sound produced is determined for a given working gas, by the diameter of the toroidal cavity cross-section. However, the power produced is proportional to the slit length which of course is equal to the circumference of the knife edge of cavity (A). It is this particular characteristic which makes the Levavasseur Whistle so attractive as a sound source of high acoustic power at high sonic and ultrasonic frequencies. Demands for high acoustic power outputs at very high frequencies can easily be satisfied by the Levavasseur Whistle by the simple expedient of going to a larger diameter circular slit.

Three different sized cavities, .57 inches, .36 inches, and .145 inches in diameter were tested. With air as the working medium, these cavities were computed to be capable of generating frequencies of 5 K. C., 8 K. C., and 19 K. C., respectively. For these tests a knife edge diameter of 3 inches was used.

Experimental Program

Procedures

Experiments were conducted with the whistle radiating into a free field. Nitrogen gas from pressurized bottles was used to excite the whistle. Frequencies produced were slightly higher than those calculated above for air, due in part to the difference of molecular weights. Figure 3 is a photograph of the experimental arrangement used. Measurements were made of the sound pressure level at various positions at the exit plane of the horn, thus providing for a pressure distribution characterizing the sound field at the whistle radiating surface. A Massa microphone, Model M-213, was used for the acoustic measurements. This microphone consists of an ADP crystal housed in a cylinder 1/4 inch in diameter x 1/2 inch long, and is calibrated to 50 K. C. at grazing incidence. The small diameter housing permitted a minimum of disturbance to the field at the frequencies measured. A Krohnhite variable band pass filter was used to explore particular regions of the spectrum. The output of the microphone was received by an oscilloscope, the screen being photographed to record wave forms or any other information for which documentation was desired. In addition to the sound pressure levels for each operating condition, the supply pressure of the gas actuating the whistle was measured and values for the mass flow were obtained with a Fisher and Porter Flowrator Meter.

The experimental procedure was to fit the whistle with a given pair of cavities and to drive the whistle over a range of pressures. The sound pressure level would be noted simultaneously with the whistle supply pressure and the mass flow.

Results

Fundamental frequencies of 5.8 K. C., 10 K. C., and 21 K. C. were produced with the .57, .36, and .145 inch diameter cavities, respectively. Sound intensities were generally 5 - 7 db lower for the .145 cavities than for the other two. A great deal of the experimentation was conducted with the .36 inch diameter cavities and the major part of the results will therefore be presented relative to this particular configuration.

There appeared to be two general regimes of operation; these being characterized by the whistle supply pressure. At low supply pressures, the signal produced and as measured at the horn exit plane is essentially sinusoidal with very little indication of any harmonic generation. As the supply pressure is increased, the sound intensity increases uniformly until a supply pressure of approximately 15 p. s. i. g. is reached, a pressure sufficient to aerodynamically choke the slit supplying the gas to the knife edge. A discontinuity in performance occurs at this point with the sound intensity jumping to very high values. Figure 4 shows the variation in total acoustic power radiated from the whistle as a function of the supply pressure for the .36 inch diameter cavity.

The wave forms produced at the higher supply pressures were considerably different from those seen at supply pressures below those values required to produce sonic flow at the slit. The wave form was rich in higher frequency components and became progressively distorted with increasing supply pressure. Furthermore, at very high pressures (60 p. s. i. g.) the signal tended to become increasingly unsteady with time. Figure 5 shows two oscillograms of the wave form recorded with the .36 cavity at a supply pressure of 30 p. s. i. g. One trace is that of the total signal while the other contains the fundamental at approximately 10 K. C. with information above 20 K. C. filtered out. Figure 6 is a record of the radiated signal with the same geometry at a supply pressure of 60 p. s. i. g. As can be seen, the total signal is considerably more distorted signifying greater higher frequency composition plus a general steepening of the wave due to high intensity effects. It is apparent that maximum efficiency in the high intensity regime will occur at the supply pressure which will just barely provide sonic flow in the slit.

Figure 7 is a plot of the pressure distribution of the sound field in db re .0002 dynes/cm^2 as measured at the exit plane of the horn for a whistle supply pressure of 20 p. s. i. g. with the .36 cavities. This distribution is typical of the nature of the field. The reference lines shown are 1/2 inch apart. A calculation was made for the total power radiated from the whistle by dividing the surface into circular zones 1/2 inch wide, the center of the zone containing the reference lines with its measured values as shown in Figure 7. Each zone was then divided into eight sub-zones and the power was calculated for each one based on the measured sound pressure for that zone. Plane wave motion was assumed in these calculations, which is a fairly reasonable assumption. A total of 45 acoustic watts was calculated for this particular point. This represents an efficiency of 10%.* With the supply pressure at 60 p. s. i. g., the whistle could provide 85 acoustic watts. The diameter of the horn at the exit plane is 5 inches.

One other interesting characteristic of the behavior of this whistle is the relative position of the cavities for the two regimes of operation. When the pressure is sufficient to choke the slits, the proper orientation of the cavities was such as to conform to the original design criteria of having the knife edge farther upstream than the resonant cavity trailing edge. In the sub-sonic regime, however, the orientation requirement is quite different, the actual position varying with supply pressure but with no apparent order.

Auxiliary Experiments

It is desirable to optimize the performance of this whistle. Central to this task is the need to understand clearly the factors which contribute to the noise generating mechanism. It is apparent that the intensity of the tones produced is related to a number of factors. These could be, to varying degrees, as follows: the jet edge-slit distance, the intensity with which the primary jet is chopped by the transverse jet, the percentage of the flow initially captured by the jet lip cavity, and the orientation of the resonant cavity lip to the deflected flow. A number of additional significant queries could probably be advanced.

In an effort to establish a better understanding of the mechanism of the whistle operation, a number of auxiliary experiments were conducted in which the various combinations of cavities were arranged. In addition to intermixing the cavities, configurations were provided in which one or both of the cavities were replaced by blank rings. Operation was maintained at a whistle supply pressure of 20 p. s. i. g. Results of these tests are contained in Figure 8. Only frequencies are presented. The total extent of the data obtained has still to be analyzed and digested. However, some

*Higher efficiencies may be obtained at lower operating pressures.

very interesting observations can be deduced from these results.

1. With the resonant cavity blank and regardless of the size of the jet lip cavity, the sound output was devoid of any tones, consisting mainly of white noise, characteristic of turbulence generation due to flow through a conduit. Apparently, the chopping function is not sufficient for tone generation in the Levavasseur Whistle.

2. Conversely, with the jet lip cavity blank and the resonator cavity functioning, discrete tones were produced, the frequencies being those shown in the table. The levels measured were generally ten db lower than the values recorded for the condition where the pair of cavities of the same size were used. The above two data suggest that there is a feedback activity of some sort between the resonant cavity and the jet lip cavity to provide the high levels experienced when the combined configuration was used. Of some interest is the fact that the frequencies produced did not vary inversely with the volume of the cavity as is expected in resonator behavior. Examination of the data shows that with the jet lip cavity blank, the frequencies generated were 26 K. C. , 17. 8 K. C. , and 22 K. C. for the . 145, . 36, and . 57 inch diameter cavities, respectively. This anomaly has not been resolved and may be clarified upon further study of the data.

It is felt that the information contained in Figure 8 is very significant in terms of understanding better the sound generating mechanism of the Levavasseur Whistle. It is anticipated that further review of this data and additional experimenting will provide the required information to optimize the design of this device.

Conclusions

An Acoustica design of a Levavasseur Whistle has been studied experimentally to better document its behavior and to explore techniques for optimizing the design. Results obtained from these tests demonstrate that very high intensity acoustic fields at controlled frequencies in the sonic and ultrasonic regime can be produced. There are a number of needs, both industrial and of a scientific tool nature, which this whistle can very well satisfy.

First of all, in any sort of application where pure tone signalling is required, this whistle lends itself admirably to that use. Directivity can be enhanced by appropriate design of the horn radiator. The levels generated are sufficiently high to insure propagation over reasonable distances consistant with the frequencies selected. The removal of smoke, dust, and aerosols in industrial environments are problems which can be attacked very readily by the use of this device. The formation of foams which are normal by-products of a number of industrial processes has long plagued manufacturers. At Acoustica, foam dispersion from the surface of liquids has been demonstrated with the Levavasseur Whistle. There are known chemical reactions such as synthesis of ammonia which lend themselves to being affected by application of high intensity acoustic energy. For these uses, the Levavasseur Whistle is very applicable. Combustion processes are known to become affected by the introduction of an acoustic field. For example, the phenomenon of unstable combustion in both air breathing jet engines and rocket motors is directly related to the acceleration in burning rate of the fuel when acoustic resonance of the chamber occurs. One of the most intriguing applications of the Levavasseur Whistle is in this very area of acceleration of burning rate of solid propellant by acoustic radiation of the propellant surface during burning. A Levavasseur Whistle has been used successfully by Acoustica in experiments to affect the burning rate. Substantial increases in burning rate were recorded, the exact quantity being classified.

There are many advantages which the Levavasseur Whistle offers as a high intensity source. It is simple, inexpensive, and has no moving parts. This avoids the exacting machining problems inherent in sirens. The Levavasseur Whistle can be used at high temperatures and in corrosive atmospheres, a degree of freedom which moving type generators cannot enjoy. The power output of the Levavasseur Whistle need not go down at higher frequencies since, although the diameter of the toroidal cross-section is reduced, the torus diameter itself can be increased, thus providing a greater radiating surface. Finally, the Levavasseur Whistle can be used with any gas or any mixture of gases to provide an even broader possibility for frequency variation.

Reference:

(1) "Generateurs Pneumatiques D'Ultra-sons Intenses", by V. Gavreau, Acustica, Vol. 8, 1958, Pages 121 - 130

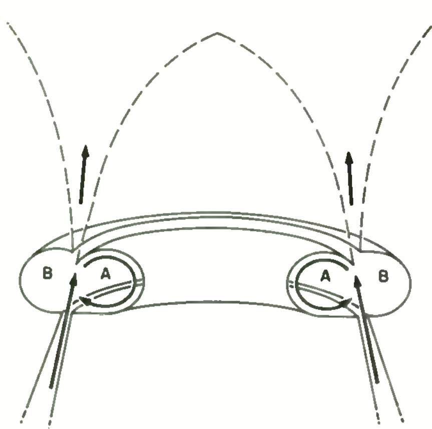


Fig. 1. Toroidal cavities.

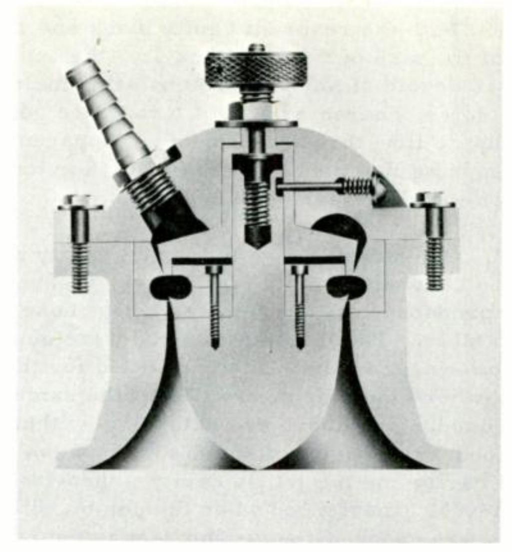


Fig. 2. Section of whistle.

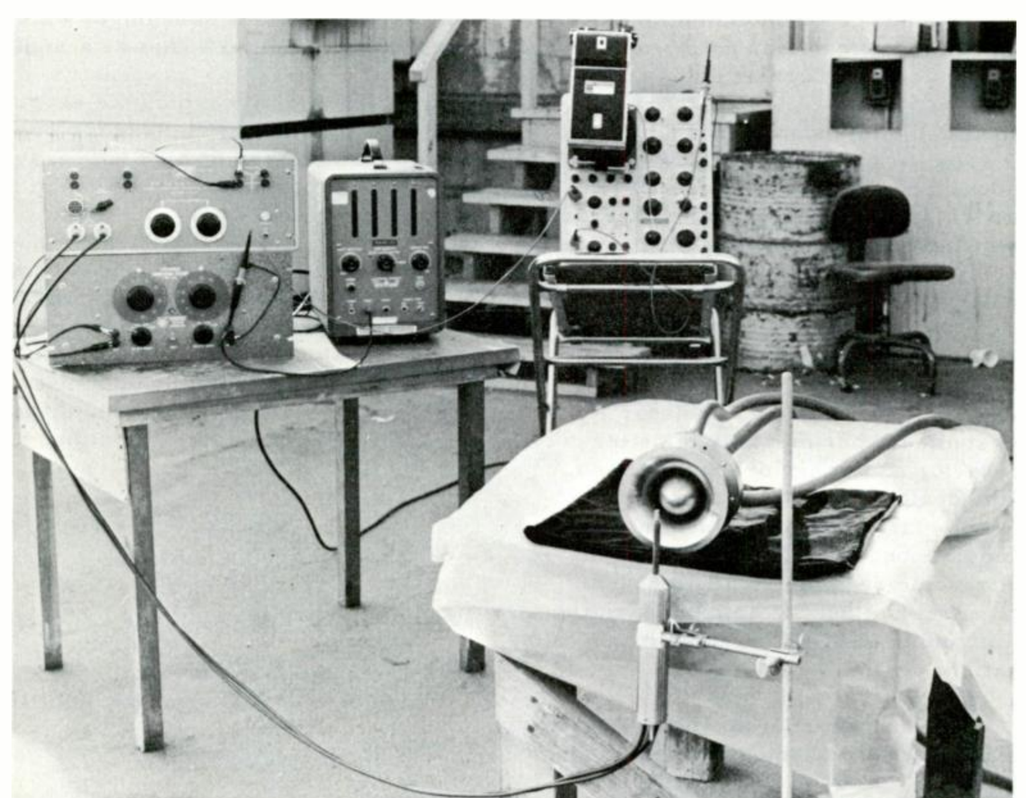


Fig. 3. Experimental arrangement for free field measurements.

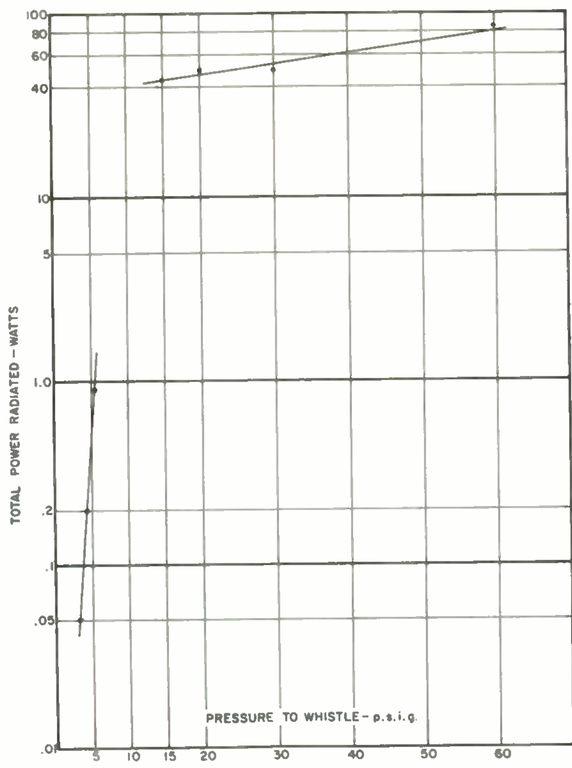


Fig. 4. Variation in power with pressure supplied to whistle-.36 cavity.

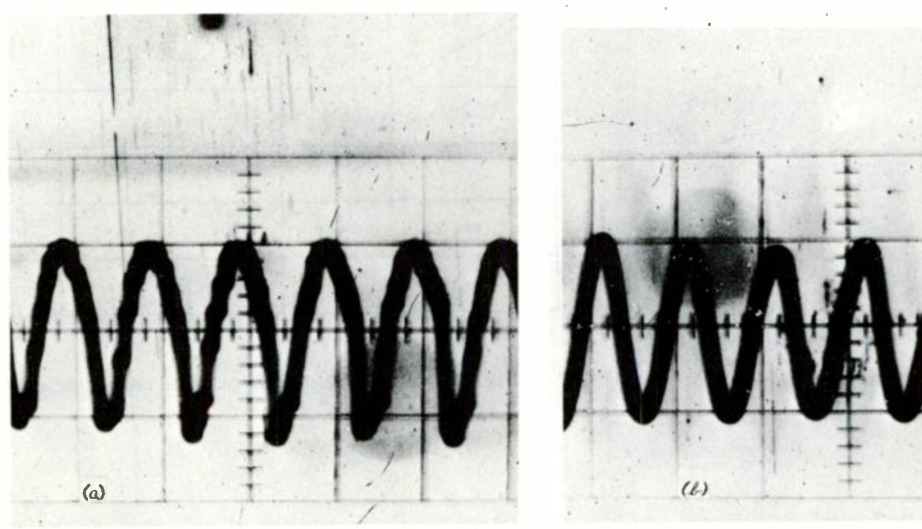
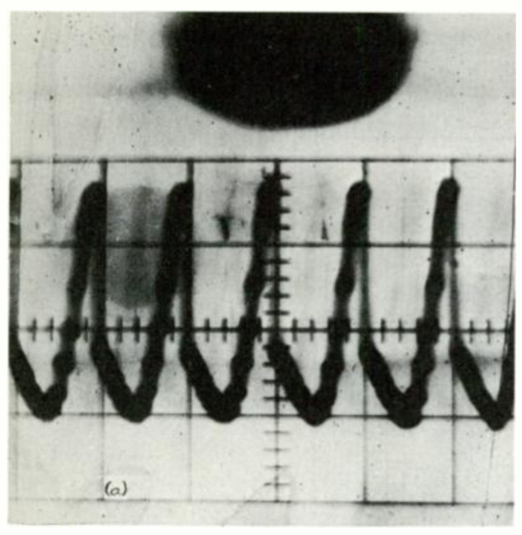


Fig. 5. Wave Form: 0.36″ cavity. Pw = 30 p.s.i.g. Sweep speed = 0.1×10^{-3} sec/cm. (a) Filter cut-off = 200 kc. (b) Filter cut-off = 20 kc.



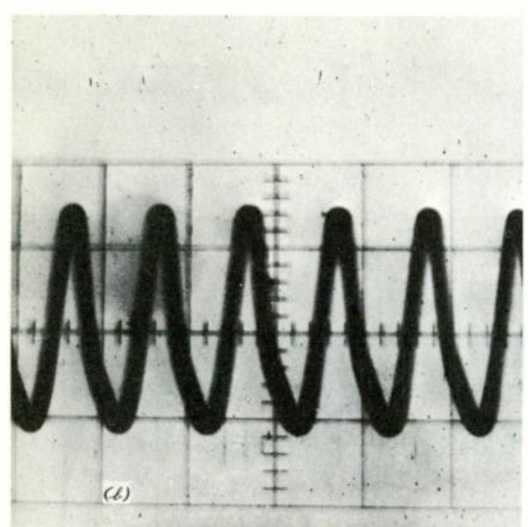


Fig. 6. Wave Form: 0.36" cavity. Pw = 60 p.s.i.g. Sweep speed = 0.1×10^{-3} sec/cm. (a) Filter cut-off = 200 kc. (b) Filter cut-off = 20 kc.

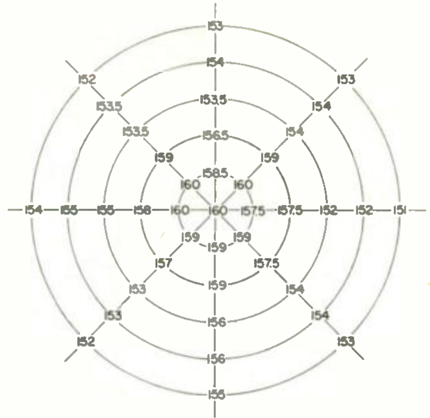


Fig. 7. Sound pressure distribution at mouth of horn (DB) .36 diameter cavities-P_W=20psig.

RESONATOR CAVITY B / JET LIP CAVITY A	.145	.36	.57	BLANK
.145	21 K.C.	--	--	WHITE NOISE NO TONE
.36	--	10 K.C.	7.4 K.C.	WHITE NOISE NO TONE
.57	34 K.C.	8.0	5.8 K.C.	WHITE NOISE NO TONE
BLANK	26 K.C.	17.8 K.C.	22 K.C.	WHITE NOISE NO TONE

Fig. 8. Frequencies with various cavity orientations.

TECHNICAL ASPECTS OF THE CAVITRON HYPER INTENSE PROXIMAL SCANNING ULTRASONIC CLEANER

Lewis Balamuth, Ph.D.
Cavitron Ultrasonics Inc.
Long Island City, New York

Summary

The history of the Cavitron Ultrasonic Cleaning Unit is briefly presented, followed by description of the major parts of the equipment. The unit is presently used by many airlines as well as the Armed Forces. Operating details of the Ultrasonic Cleaning technique are accompanied by an analysis and comparison of the Ultrasonic Cleaning methods presently used. The superiority of the Cavitron method is clearly shown.

Introduction

This paper describes a new type of ultrasonic cleaner especially adapted to the cleaning and treating of aircraft filters. Primary emphasis has been given to ultrasonic tank cleaning operations and these are now well known in the art.

The filter cleaner herein described is based on a different approach, which can be described succintly as the Caviblast method. The characterizing features of the Caviblast process are the following:

A. Cavitation is generated in the smallest possible quantity of liquid. That means a minimum gap between the vibrating surface and the surface to be cleaned; biggest economy in acoustic power.

B. The transducer vibrations are transmitted to an elastic structure which can have any size and shape, perfectly adaptable to the requirements of the cleaning job to be done. This structure generally will vibrate in one of its higher flexural modes. It will be inexpensive, and easy to replace. Because of the complete separation of the transducer and the cavitation generating structure, high pressures, high temperatures and corrosive chemicals can be applied.

C. Vibrating bars, pipes or other ducts are ideal for feed-through operation, that means for fast cleaning of large numbers of pieces. Cleaning will be uniform, that means the statistical fluctuation of the degree of cleanliness will be a minimum, because the pieces are not allowed to cluster and to impede the necessary cavitation.

The particular form taken by the Hyper Intense Proximal Scanner ultrasonic filter cleaner came about as a result of a pressing need on the part of the manufacturers of such filters (Pall Corporation). The specifications of the currently used unit came directly from the experiences in this field. It is perhaps relevant to point out that the Caviblast solution came after the Pall Corporation engineers had exhausted practical tests with tank type cleaners without sufficient end results.

The cleaning of dirt-loaded metal filter elements is a specialized procedure which requires special equipment and close adherence to detailed methods. In the past, four types of cleaning methods have been used with varying degrees of success in the specific areas to which they have been applied. They are: brushing, back-flushing, chemical cleaning and ultrasonic tank cleaning. All have proved successful in limited applications, but none offered potential as a general

WRH

cleaning method due to inherent disadvantages of each method.

Various types and kinds of filtering screens and fabrics have been used to remove particle debris from gaseous or liquid streams in oil refining, in the manufacture of chemical and pharmaceutical products, and in hydraulic pressure systems. While the filter cleaning method and apparatus of this equipment has wide application to the rapid and thorough removal of pore clogging debris from numerous types and kinds of filters and filtering elements composed of various materials of selective porosity, this equipment has particular application to the cleaning and removal of particle debris from woven screen or fabric filters whose filtering pores are microscopic in size and in the order of ten microns or less, and which are particularly difficult to clean and condition for re-use, except by the method and apparatus of this equipment.

The entire cleaning and testing cycles can be divided into three (3) stages:

1. The filter element is rotated by means of a gloved hand, through the Caviblast action of the Hyperintense Proximal field so that the field may act upon the entire surface of the filter element. Varsol flows from the inside to the outside of the elements during the first stage of the cleaning cycle flushing out the particles as they are loosened by the ultrasonic energy.

2. The flow is then reversed for the last part of the cleaning cycle to remove any contaminant from the downstream side of the filter element. A scanning time of 10 minutes has been found to be sufficient on most filter elements tested to date, and most elements have been cleaned to their original differential pressure and close to, if not completely to, their original dirt-holding capacity.

3. The third stage of this cycle uses a compressed air system equipped with means and devices for testing the filtering screen to determine whether it has been cleaned to acceptable standards.

Description of the Hyper Intense Proximal Scanning Ultrasonic Cleaner

It is the purpose of this paper to discuss in detail the theory behind the equipment presently used in obtaining reusable filters.

Figure #1 shows schematically the arrangement required to convey ultrasonic vibrations to the edge of a cleaning tool. The apparatus includes a magnetostrictive transducer which is made to vibrate at high frequency and low amplitude when subjected to the influence of a corresponding high frequency alternating magnetic field generated by the winding. The longitudinal vibrations thus induced are transmitted to an acoustical transformer section, rigidly fixed to one end of the transducer and which provides a connecting body to which the cleaning tool is attached. The transformer section is designed to engender at the operating face of the cleaning shoe the type and kind of high frequency and low amplitude motion best suited to produce the most effective cleaning action at the surface of the filter.

The operating frequency is in the ultrasonic range, in the order of 20,000 to 40,000 cycles per second. At these frequencies, the transducer section may be formed of a stack of plates of a metal such as nickel, iron, cobalt, permanickel, permandur or other metal alloys which have high tensile strength and are highly magnetostrictive in character, so that the transducer section will longitudinally vibrate to a maximum degree when subjected to the influence of an alternating magnetic field. This is established when the wire wound around the magnetostrictive transducer is operated with a biasing DC current. This bias current places the quiescent point in the magnetizing curve and enables the transducer to expand and contract in synchronism with the frequency of the applied AC current and the acoustical characteristics of the metal from which the transducer section is made.

I - The Vibrator Assembly

The vibrator assembly is mounted so that the vibrating tool face is minutely spaced from the filter to be processed, while the filter unit is immersed in the cleaning fluid. The filter unit is

adjustably supported in a manner so that the filter unit may be rotated and longitudinally moved or reciprocated so as to position progressive cleaning areas of the filter in adjacent relation to the working face of the vibrating tool. The working face of the vibrating tool is vibrated at frequencies in the order of twenty thousand cycles per second and at relatively small amplitudes, and which vibrations produce hyperintense cavitation of a limited column of the cleaning fluid extending through the filtering screen of the filter and which covers a cleaning site area corresponding to the area of the contact face of the vibrating blade.

Figure #2 is a section of the vibrator assembly mounted in operating relation to a tubular filter unit to be processed while immersed in the cleaning fluid.

Since the transducer section generates heat during vibration, a coolant is provided in surrounding relation to the transducer section to maintain the transducer section and surrounding tubular housing in relatively cool condition. This is accomplished by using a transducer that is water filled and air cooled, thus avoiding the complications of a water cooling system. The stream of cooling air is propelled by the blower mounted on the outer housing of the vibrator assembly. The blade face is approximately 2 1/2 inches long by 1/2 inch wide. It is located between two filter element guides. The guides provide a small space between the transducer radiating face and the filter element being cleaned. The transducer input is 250 watts and its minimum output is 100 watts at a frequency of 20,000 ± 500 cycles per second. The 100 watt minimum output energy is concentrated at the surface of the filter element. Selection of the 100 watt figure for the 2 1/2 x 1/2 inch blade face was based on an experimental determination that an energy density higher than 80 watts per square inch at the radiating surface was not more effective in cleaning, and that a lower energy density was less effective in cleaning.

Figure 3 shows the transducer assembly as mounted in operative position on the unit. It is housed in a brass casing with holes in the housing for proper ventilation of the unit.

II - Amplitude Indicator

To produce the optimum hyperintense cavitational effect on the cleaning fluid interposed between the working face of the vibrating tool and the adjacent filter, the acoustical vibrator vibrates at resonance frequency, and the power generating circuit is designed to supply biased alternating current within a limited optimum frequency range to insure vibration of the vibrator unit at resonance frequency. In order to supply biased alternating current to the winding of the vibrator assembly (See Figure 1) at optimum frequency, the biased alternating current generating system is provided with tuning means under the control of the operator. To detect any malfunctioning of the power generating system or the vibrator assembly so that desired adjustments can be made to restore the system to optimum frequency and amplitude, some monitoring means for measuring the vibration amplitude or frequency at which the vibrator unit is vibrating, is a desirable requisite. Since the amplitude of vibration varies with the frequency of vibration, the frequency of vibration can be determined by monitoring any variations in the amplitude of vibration of the tool.

To enable the operator to adjust and monitor the amplitude of vibration of the tool, a pickup device, as shown in Figure 4, is provided. The pickup is disposed in operative relationship with the upper end of the tool to convert the vibratory motion of the latter into electrical signals which may be read on a meter. Such a pickup comprises a pin of magnetostrictive material fastened to the upper end of the tool member to be vibrated therewith and thus exhibit its magnetostrictive effect.

It has a length lying within the range between approximately 1/16 to 1/8 of the wavelength of the vibrations transmitted to work tool. Further, in order to minimize the loading of the work tool, the wire or pin forming the magnetostrictive member preferably has a small diameter. For example, a diameter of .10 inch and a length of .60 inch is employed in connection with a tool operated at a frequency of 20,000 cycles per second and having a wavelength of 9.4 inches. Arranged in operative relationship to the pin is a coil in which voltages proportional to the flux generated in the pin

are induced. The flux is proportional to the amplitude of the vibration imparted to the pin by the tool, and a meter arranged to sense the voltages induced in the coil may be calibrated to read amplitude of vibration. The operator merely adjusts the potentiometer tuning control until the meter, which is conveniently mounted on the instrument panel, indicates the desired amplitude.

Figure 5 shows the amplitude pickup as located in the transducer housing. Its output is fed to the meter referred above which indicates the double amplitude of the radiating face of the cleaning blade in microns. This meter is divided into a red area (unsatisfactory) and a green area (satisfactory) thus providing a go-no-go type of indication. The division is at 18 microns. Any malfunction in the generator or transducer is registered immediately. A normal output is represented by a double amplitude of 20 to 25 microns. Means are provided for measuring the amplitude of the radiating surface mechanically in order to check and calibrate the amplitude meter.

Figure 6 illustrates the electrical circuit used in connection with the amplitude pickup of Figures 4 and 5. The coil functions as a pickup coil and a permanent magnet is employed for polarizing the magnetostrictive pickup pin. The circuit includes a matching transformer having its primary connected to the coil. The secondary winding of the transformer is connected to a potentiometer with which the meter reading is adjusted. This voltage is rectified in order to use a linear scale instrument, and the circuit parameters adjusted so as to obtain a linear relationship between the meter readings and the amplitude of vibration over the frequency of interest.

III - Cleaning Fluid Circulation System

In accordance with the Caviblast principle, the object to be cleaned is supported in the cleaning fluid bath by an adapter which is connected to a cleaning fluid circulation system, and by means of which cleaning fluid under controlled pressure or suction is applied to one face of the adapter-supported object, while the opposite face is subjected to the hyperintense cavitational action of a limited column of cleaning fluid. The cavitational action produced is of such intensity as to penetrate into, and loosen the debris impacted within, the pores, and passages of the filter, while cleaning fluid pressure or cleaning fluid suction, or alternating cleaning fluid pressure and cleaning fluid suction, is applied to the opposite face of the object to effectuate the ejection or withdrawal of the debris or foreign material therefrom. The thorough cleaning of all the pores and passages of the filter is thus effected by the combined action of cavitational, pressure and/or suction forces. Seventy-five to ninety percent of the particle debris can thus be dislodged and outwardly removed from the filtering screen of the filtering unit, and which particle debris will then be deposited in the cleaning fluid bath in which the filter unit is immersed.

By way of example, Figure 7 is a section view showing the vibrating tool mounted in operative relation to a tubular filter unit to be processed while immersed in a cleaning fluid tank. It shows the tool which is vibrated at a high frequency and low amplitude in close proximity to the surface of a filter thereby generating a hyperintense cavitation of a limited column of cleaning fluid which penetrates the pores of the filter.

To remove any remaining particle debris from the filter, the cleaning fluid circulation system is provided with means which can be manipulated to apply suction force through the outer tube to thereby cause a counter-flow of the cleaning fluid through the filtering screen of the filter unit from the exterior to the interior surface thereof. This backwash or counter flow movement of the cleaning fluid proceeds while a localized column of cleaning fluid is under hyperintense cavitational action at the cleaning site area of the filter unit, as engendered by the vibrating tool face. The remaining particle debris thus removed from the filter unit is withdrawn through a reverse flow through the outer tube.

Figure 8 shows a tubular filter unit mounted in operating relation to the working end of the vibrating tool.

The element is cleaned by being rotated adjacent to the radiating face of the transducer for a 10 minute period. The element is attached to an adapter and is

rotated by hand. During this period the entire surface of the filter will have been exposed to the face of the transducer. For the first eight minutes, fluid flow is from inside out. During the last two minutes, flow is from the outside in. Varsol is used as the cleaning fluid and liquid-tight gloves are used to protect the operator's hands.

The cleaning fluid circulation system is also provided with means in association with the inner tube for measuring or metering the suction applied to the inside of the immersed filter unit when the vibrator assembly is not in operation, and which provides an indication as to the extent to which the filter unit has been cleaned. For example, a clean filter unit of known pore size and area would first be immersed in the cleaning fluid bath and the pressure drop between the opposite surfaces of its filter screen noted and registered as the maximum optimum pressure drop for a clean filter of this particular type. A used filter of the same type and with the same size pores would then be immersed in the cleaning fluid and the pressure drop between the opposite surfaces of its filtering screen noted. If this noted pressure drop of the previously used filter unit did not substantially exceed the known minimum optimum pressure drop indicated by the clean filter of this same type, then the assumption can be made that substantially all particle debris had been removed from the previously used filter unit, and that the previously used filter unit had been acceptably cleaned for re-use.

The filter unit cleaning operation takes place while the filter unit is fully immersed in a bath of cleaning fluid contained in the tank. Highly refined kerosene, such as an upper kerosene cut, and sold under the name of "Varsol", has been found to be satisfactory. Numerous other cleaning fluids may be used which possess high solvent capabilities and low corrosion characteristics, which are compatible with the gas or liquid from which the filter entrained particle debris has been removed, and which are substantially free of entrained water. The cleaning fluid must obviously be in liquid form, and adapted to be cavitated by the application of ultrasonic vibrations thereto.

IV - Compressed Air System

A compressed air system is provided which is used to test the filters either before or after cleaning, or both, to determine whether the filtering screen is defective or has been damaged and presents any pore or pores which are larger than the permitted maximum to which the filter unit has been designed, and which may result from a nick or other blow, or as a result of damaging stretching strains caused by a debris fragment. A part of this compressed air system is also used to test the cleanliness of the treated filtering screen of the filter unit and the degree to which its pores are still clogged with particle debris.

As shown in Figure 9, compressed air is supplied by a suitable compressed air line and which may be at a pressure in the order of six to two hundred and fifty pounds psi. The compressed air prior to reaching the first selector valve is processed as follows: The compressed air flows through a relief valve which is set to pass compressed air not exceeding two hundred and fifty pounds psi, and provides a protective device for the air system. The compressed air then flows from the relief valve through an air filter where the air is filtered and debris particles removed. The cleaned compressed air then flows through a pressure regulator valve which reduces and regulates the air pressure to a substantially uniform pressure of approximately five psi. Thus, the pressure regulator valve insures a constant flow of compressed air into the first selector valve at a regulated pressure which is maintained at about five pounds psi. An air pressure adjusting valve is tapped off at this point to control the amount of air necessary for the operation.

The compressed air after being processed enters the first selector valve through line A which is connected to one of the three ports of the two-way first selector valve having a control knob which may be adjusted to permit air flow through its opposite second port and into air line B. The air line B is connected to a manometer by means of which the relative pressure of the input air supplied to the manometer well by air line B and the pressure at the upper end of the manometer may be measured and indicated. The

manometer has a pressure scale and indicating fluid of the proper specific gravity, in order to obtain all pressure readings within the available scale.

The third port of this selector valve is connected to the atmosphere which permits venting the manometer well for calibration purposes.

The second selector valve controls the vacuum sensing circuit which is used to check pressure drop on the filter element being cleaned.

PRESSURE DROP MEASUREMENT

The filter element being cleaned is held in position by a specially designed adaptor which comprises an outer tube in which the cleaning fluid circulates, and an inner tube which extends to the center of the filter element and is connected to the vacuum sensing system (Figure 8). After the element has been cleaned, the flow of liquid is reversed so as to flow from outside to inside the filter element and a pressure gradient will be produced between the mesh and core. The resultant pressure drop is measured in the manometer which is connected via the second selector valve to a surge tank D, the level of which varies according to the pressure sensed by the probe. This renders a reliable measure of the suction force required to draw cleaning fluid at a predetermined rate from the bath through the pores of the filter. The pressure scale of the manometer is calibrated to indicate the normal expected pressure differential produced when a fully clean or new filter of the same type that has been tested simultaneously, to thus provide a standard of comparison.

If the vacuum level reading of the manometer produced by testing a used filter is substantially above the reference standard for that type of filter, it would be apparent that the filter under test has not been adequately cleaned and is still clogged with debris. In such case, the cleaning operation may be repeated, until the filter exhibits a pressure differential which compares favorably with the comparison standard for a fully cleaned filter unit of the same type.

The manometer is protected by fluid check valves C which prevent loss or contamination of the manometric fluid.

FIRST BUBBLE POINT PRESSURE MEASUREMENT

In addition to the need for assurance that the element is clean, it is also necessary to obtain assurance that elements remove particles with the desired efficiency. This is accomplished by measuring the "bubble point" of the element which correlates accurately with the diameter of the largest particle passed by the filter element.

The air pressure system is also used for the purpose of testing the filter to determine whether it has any filtering pores or openings larger than a passable minimum, and which enlarged pores or openings may have been caused by damage to the corrugated filter as by an exterior blow. As shown in Figure 10, this test is made when the filter unit is removed from its supporting assembly but nevertheless immersed in the bath of the cleaning fluid.

In making this test, the first selector valve is set to communicate air line A with B, so that air under pressure will flow both to the manometer well and the special adaptor which permits immersing the filter in the cleaning fluid for this test. The air pressure can be adjusted by means of a venting valve. With the filter immersed in the bath of the cleaning fluid, the pressure venting valve is gradually closed, which results in a corresponding increase in air pressure supplied to the interior of the tubular filter. The immersed filter is carefully observed for the appearance of the first air bubble on its outer immersed surface as indicated in Figure 10. When an air bubble first appears, the manometer will indicate the air pressure required to produce an air bubble on the surface of the immersed filter screen, and this air pressure is also applied to the manometer well, which will produce a reading of the manometric column.

An air pressure reading which is below the minimum optimum pressure reading for a clean or new filtering screen of the same type and filter pore size, would indicate that one or more of the pores of the filtering screen under test have become unduly enlarged beyond the uniform pore

size for which the filter unit is designed. The defective filter unit would then be carefully inspected to determine whether the enlarged pore or pores could be repaired and the filter unit placed in usable condition, or whether the filter unit has been damaged beyond practicable repair.

As before, the reading indicated by the manometer is compared with the similar reading obtained with a new filter subjected to the same test.

Conclusion

The above evidence and comment leads to the important conclusion that the Cavitron method shows up more than favorably when compared to current tank cleaning practices. Figure 11 is the completed assembly of the entire unit.

TOOL HOLDER INPUT STROKE .0005 IN.
VERTICAL STACK STROKE .0003 IN.
LOOP
NODE
GENERATOR
EXCITING CURRENT D.C. BIAS PLUS 20,000 CYCLE PER SECOND A C
LOOP
WINDING
NODE
FIRST MECHANICAL AMPLIFIER
MECHANICAL IMPEDANCE TRANSFORMER
LOOP
NODE
SECOND MECHANICAL AMPLIFIER
TOOL
LOOP
VERTICAL STROKE OUTPUT OF TOOL
VERTICAL STROKE OUTPUT OF TOOL .0010 INCHES

Fig. 1.

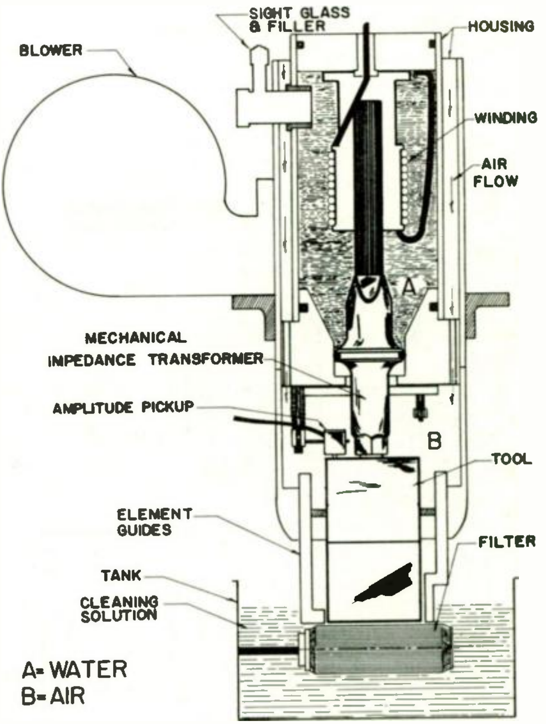


Fig. 2.

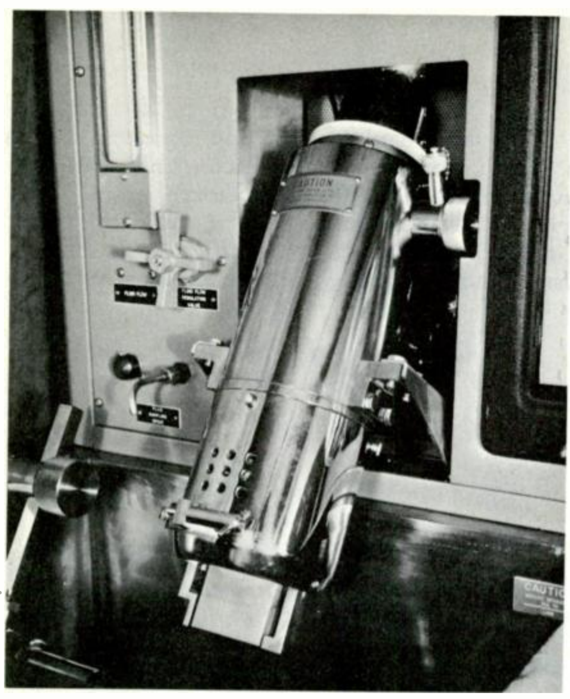


Fig. 3. Transducer assembly.

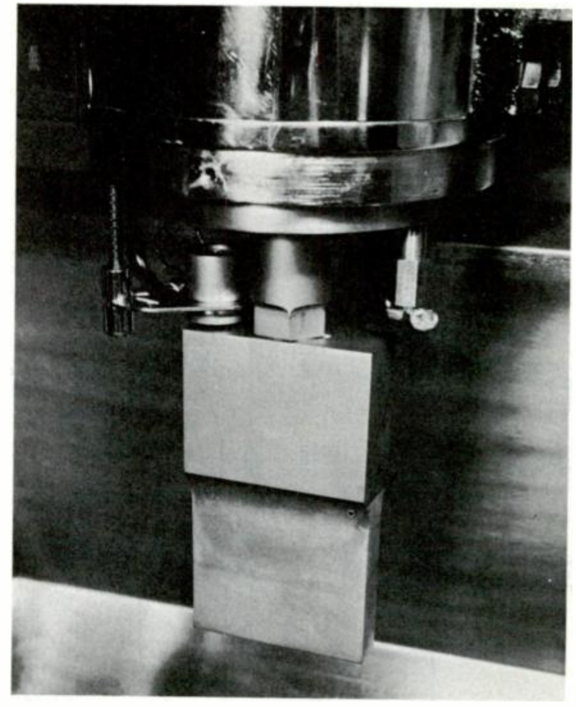


Fig. 5. Amplitude pickup.

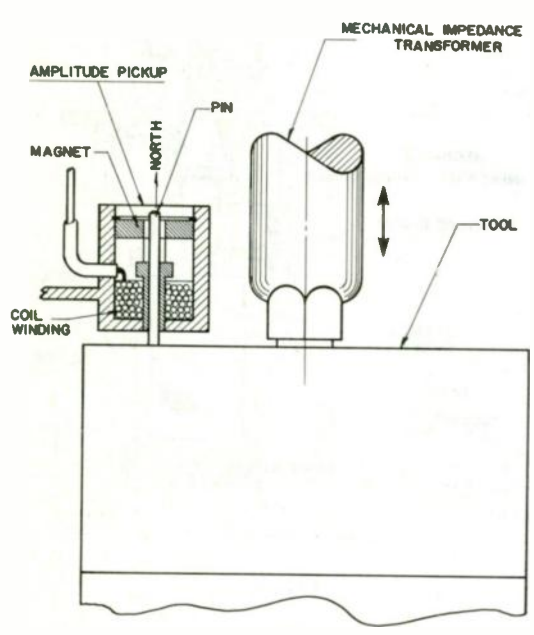


Fig. 4.

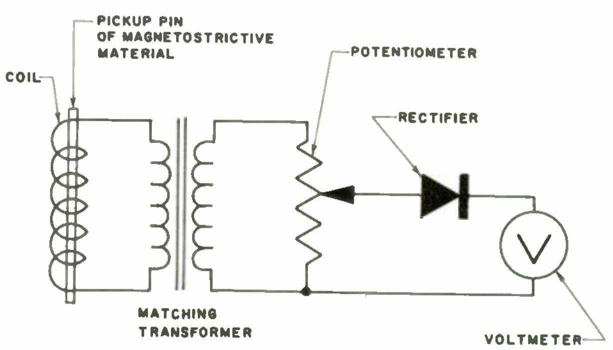


Fig. 6.

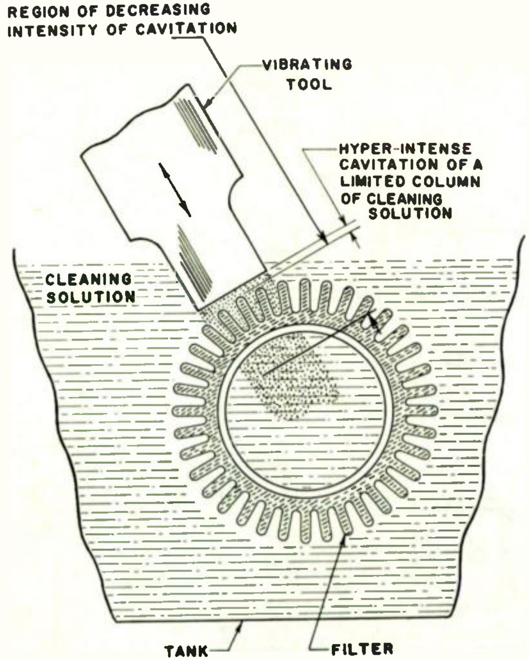


Fig. 7.

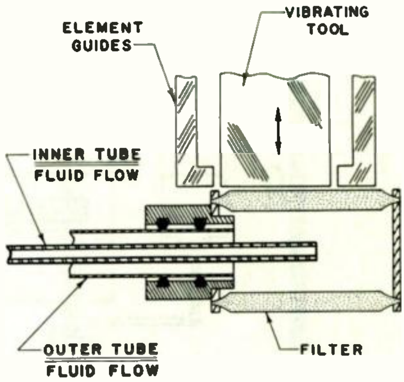


Fig. 8.

SCHEMATIC OF HYDRAULIC SYSTEM

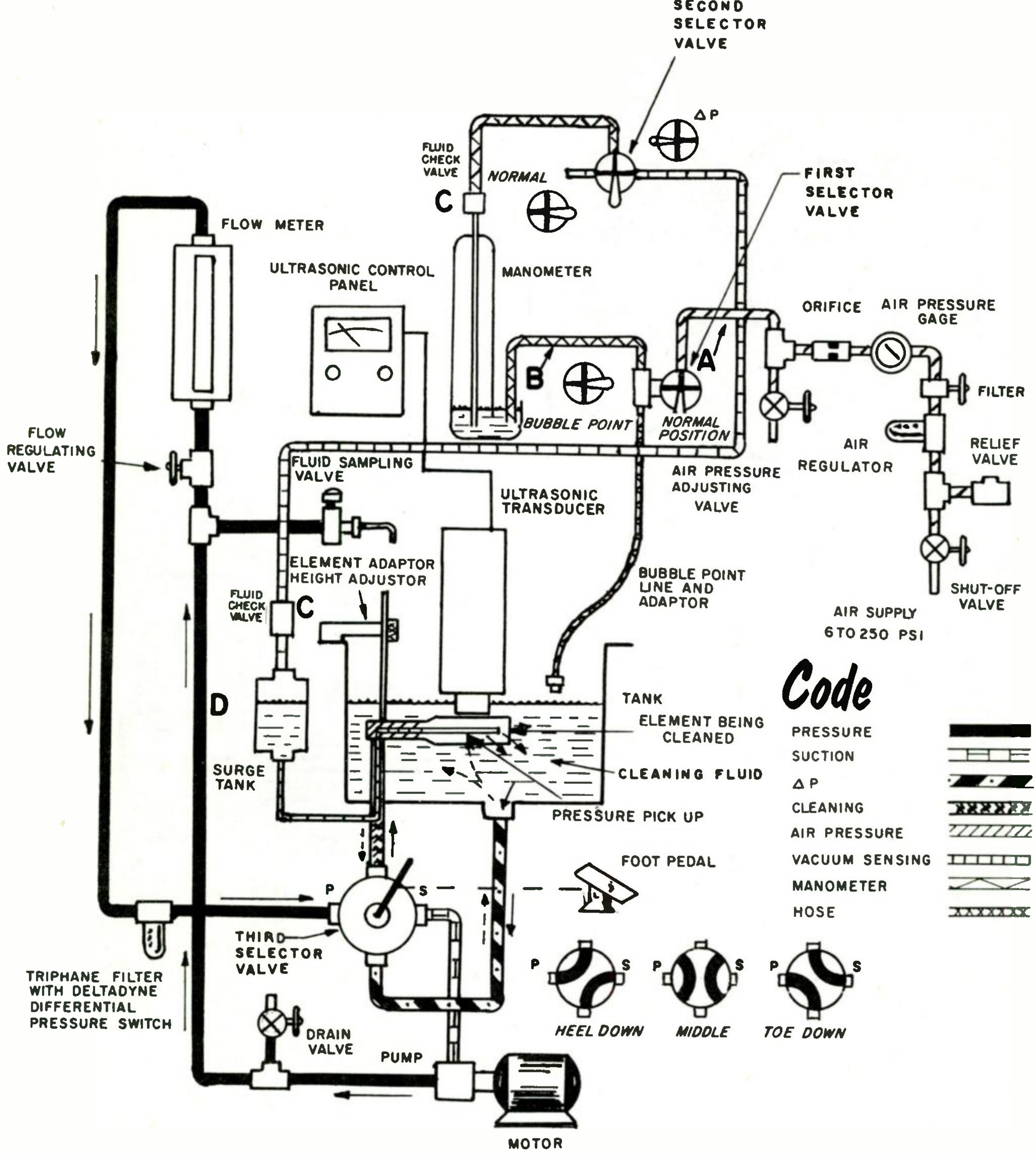


Fig. 9.

BUBBLE POINT TEST

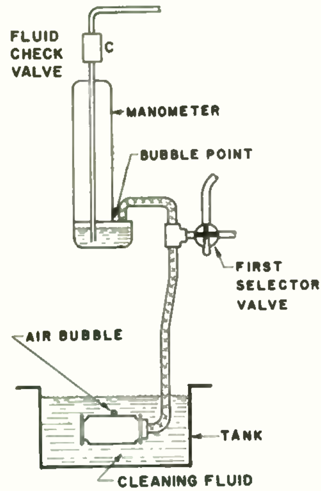


Fig. 10.

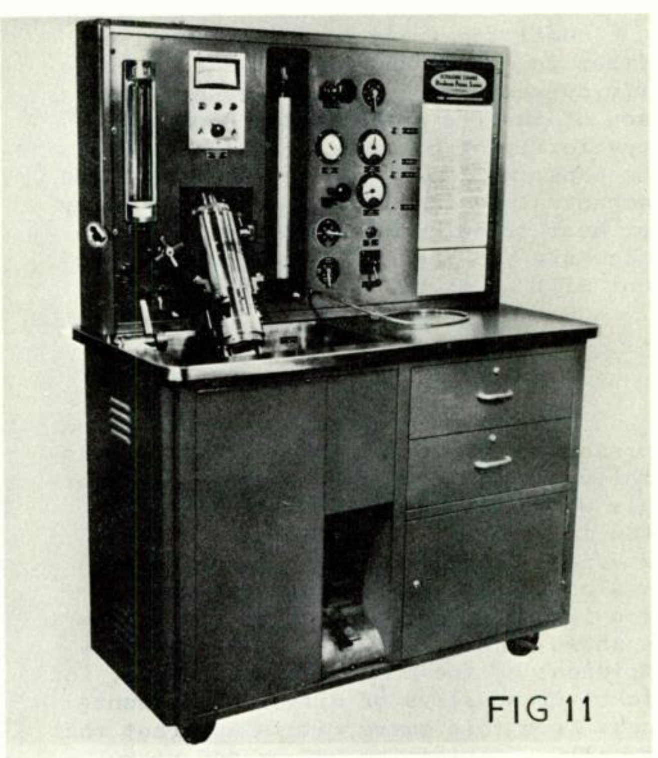


Fig. 11. Pall-Cavitron HIPS filter element cleaner. The Pall-Cavitron HIPS ultrasonic filter element cleaning system utilizes hyperintense proximal scanning to clean filter elements to their original differential pressures and dirt-holding capacities.

HEAT TREATMENT OF WIRE FOR TORSIONAL MAGNETOSTRICTIVE DELAY LINES

Alan J. Brown
Airborne Instruments Lab.
Deer Park, New York

and

Arthur Rothbart
Consulting Engineer
New York, New York

Summary

An important advantage of long torsional magnetostrictive delay lines over other forms of ultrasonic delay lines is their excellent thermal stability over extreme temperature ranges. This stability is achieved by using wire which contains an alloy of nickel and iron in the proper ratio. By adding titanium to this alloy and heat treating the wire, it is possible to vary this ratio over narrow limits. Strict control over this variation will produce a wire with a temperature coefficient of the modulus of elasticity which has the correct magnitude and polarity to cancel the effects of other factors on thermal stability. Thus, it is possible to obtain a delay line whose temperature coefficient of delay is near zero.

Heat treatment may be accomplished either in a furnace or by passing an electric current through the wire. A comparison of the two methods will show the superiority of electrical heat treatment. The paper will present the electrical and mechanical design of a practical facility for heat treating wire electrically. The procedure and results of a typical treatment will be given.

Introduction

The thermal stability of the long torsional magnetostrictive delay lines is achieved by using wires composed of materials whose modulus of elasticity can be made independent of temperature. These materials belong to a group of constant modulus alloys which contain nickel and iron in the proper proportion. In Figure 1 is shown a curve[1] of the temperature coefficient of the modulus of elasticity for nickel-iron alloys of different percentages. From this curve, it is apparent that the alloys containing 27% or 44% nickel have zero temperature coefficients.

Since these percentages are quite critical, a small amount of chromium can be added to the alloy to shift the peak of the curve downwards as shown by the dashed section of curve. This addition provides a flat composition range of zero temperature coefficient alloys and is the basis of the constant modulus alloys known as Iso-elastic and Elinvar. In these alloys, the temperature coefficient of the modulus is determined solely by their chemical composition and may differ from one batch of material to another due to commercial variations in melting and alloying. Therefore, this particular class of alloys is unsuitable for wire delay lines.

However, if a small amount of titanium [2], is added to a nickel-iron alloy of the proper proportion, the effective nickel content of the alloy can be controlled by heat treatment. During the heat treatment, nickel is precipitated out of the alloy in the form of an intermetallic compound of nickel and titanium. Thus, it is possible to obtain a wire whose temperature coefficient of the modulus of elasticity has the proper magnitude and polarity to cancel the effects of other factors on thermal delay-line stability. This results in a wire delay line[3] whose temperature coefficient of delay is near zero.

Another advantageous effect on delay line performance occurs during heat treatment. As the nickel precipitates, it becomes very hard, thus increasing its mechanical Q[3]. The higher the Q, the greater is the bandwidth and output of the delay line. These precipitation-hardening constant modulus alloys are known as the Nickel-Iron-Titanium alloys. If chromium is added to one of these alloys to make the heat treatment less critical, the alloy, Ni-Span C, is obtained. However, chromium acts as a damping agent to lower the Q of the wire.

Heat Treatment of Nickel-Iron-Titanium Alloys

Due to their titanium content, the Nickel-Iron-Titanium alloys respond to a

precipitation-hardening heat treatment (artificial aging).[1,4] This heat treatment is similar to that used with beryllium-copper and some of the aluminum alloys. Precipitation can occur when an alloy in its equilibrium condition at room temperature consists of two or more phases, and when a portion of one or more of these phases can be dissolved in the matrix of the alloy by raising its temperature.

Precipitation-hardening is accomplished in a two-step process. The first step, solution annealing, consists of heating the alloy to a high temperature, at which one phase is dissolved in the matrix. The alloy is then quickly cooled from the solution annealing temperature (quenched) to prevent precipitation of this phase which would normally occur during slow cooling. In the Nickel-Iron-Titanium alloys, it is believed that an intermetallic compound of nickel and titanium (Ni_2Ti) is formed for those alloys containing approximately 42% nickel and 2.5% titanium. This compound is dissolved in the nickel-iron matrix by heating the alloy to a solution temperature from 1750°F to 1950°F. By quenching, the compound is retained in a super-saturated solid solution. Solution annealing is normally performed by the processor of the wire.

Following solution annealing, it is customary to perform a certain amount of cold work on the alloy before the second step of the heat treatment. The cold reduction increases both the hardness and the stiffness coefficient of the material, and causes the alloy to respond more rapidly to the precipitation-hardening heat treatment. The increase in hardness is shown in Figure 2 as a function of cold reduction.[1] The second step of the treatment, which is done under the direction of the delay-line manufacturer, is to heat the alloy to an intermediate temperature, about 1100°F to 1350°F, at which time the desired amount of nickel is allowed to precipitate.

Generally, wires are supplied by the processors on 9 inch diameter spools. Since the wire must be straightened for non-dispersive delay-line performance[3], such straightening must be performed mechanically if the wire is to be heated in a furnace. Where heat treatment is accomplished electrically by using the wire as a self-heating element, straightening can be achieved by stretching out the wire in a straight line during heating.

Heat Treatment in a Furnace

For heat treating in a furnace, the wire should be supplied in the straightened condition in coils having a minimum diameter of 18 inches. It is not advisable to use a furnace whose minimum dimension is less than 18 inches, since the heat treatment will cause the wire to take a permanent set in the diameter in which it is treated. This set is not due to the material being heated to its true plastic range, but rather to a slow creep in its "semi-plastic" condition. The characteristics resulting from cold reduction are not materially changed during subsequent heat treatment[1], as shown in Figure 3. The furnace temperature should vary between 1100°F and 1350°F for about 4 hours, depending on the composition of the wire. For those Nickel-Iron-Titanium alloys which contain chromium, such as Ni-Span C, the temperature of heat treatment is a function of the percentage of titanium and chromium[1], as shown in Figure 4.

The furnace atmosphere may be one of these types:

1. A vacuum.
2. A reducing atmosphere of cracked ammonia or hydrogen.
3. An inert atmosphere of argon or neon.

Furnace treatment is not carried out in air since the furnace muffle (usually carbon steel) would oxidize. Except in the case of a high vacuum (10^{-3} mm Hg), the wire usually acquires a blue-gray oxide coating which is not detrimental to delay-line performance.

Electrical Heat Treatment

In electrical heat treatment, the wire is suspended between two points in a straight line. The heat treatment will bring the wire into its "semi-plastic" range where it will be permanently straightened at the conclusion of the treatment. The cycle may be specified by the time and temperature required in treatment, or alternatively, by specifying the current through the wire and the time. Temperature and electrical power may be related by using the Stefan-Boltzmann equation for radiant emission

(neglecting convection terms present when the wire is treated in air).[5]

$$P_A = \epsilon_T \sigma (T^4 - T_o^4).$$

P_A is the power radiated per unit area in watts per cm^2.
ϵ_T is the thermal emissivity (about 0.1 for most metals).
σ is the Stefan-Boltzmann constant.
T is the temperature of the radiating body in degrees Kelvin.
T_o is the surrounding temperature in degrees Kelvin, which is normally negligible.

Using the above equation, it is theoretically possible to relate the temperatures experienced in heat treatment to the power dissipated in the wire. However, it should be noted that difficulties arise in determining accurate values for the terms of the expression and the effects of convection.

For normal electrical heat treatment, the process may be carried out in air since the initial formation of oxides on the wire surface insulates the interior of the wire. However, the temperatures are deliberately lower than in a furnace treatment to prevent excessive oxide formation. Should any other type of atmosphere be desired, the wire may be strung inside of an insulated muffle (glass pipe, or metallic pipe with a ceramic liner).

Successful electrical heat treatment of wires has been achieved with the following two-stage process:

1. In the first stage which may take from 5 to 30 minutes, a current of 4.5 to 5 amperes is passed through a 20 mil diameter wire, and a current of 7.5 to 8 amperes is passed through a 30 mil wire.

During this stage, the wire is brought into its semi-plastic range for straightening. Precipitation of the nickel from the nickel-iron matrix does occur, but because of the short cycle this precipitation is not carried to completion.

2. In the second stage which may take from 4 to 24 hours, a current of 3 amperes is used for the 20 mil wire, and a current of 5 amperes for the 30 mil wire.

During this stage, the wire is slowly precipitation-hardened. The lower temperatures are sufficient to sustain the precipitation of nickel at a reduced rate. Small changes in current make it possible to control accurately the final properties of the wire material. The reduced temperatures are too low to cause significant crystal growth, and permit completion of the process in air below the critical range of oxidation.

A Comparison of Furnace and Electrical Heat Treatment

The following presentation, which lists the features of each type of heat treatment, clearly indicates the superiority of the electrical process from both an engineering and an economic point of view.

Furnace Treatment

The cost of a furnace facility may vary from $5,000 to $20,000, depending on the type desired. To provide the necessary atmosphere, the operation involves the purchase of tanked gas. Power for heating must be supplied from a heavy gas or electrical source. The usual capacity per heat treatment is about 10 to 15 lbs. of wire, 4,000 feet. A complete cycle, including warm-up and cool-down time, takes about 8 hours for a wire treatment of 4 hours. Following treatment, the wire is permanently formed into the diameter at which it is treated. Except for high vacuum operation, a mild oxide coating forms on the wire.

Electrical Treatment

The cost of an electrical facility is less than $2,000. The heating power of 2.5 to 5 kilowatts may be supplied from a 3 phase 220 volt ac source having a current rating of 15 amperes. Operation can be automatic except for loading and unloading. A typical capacity is about 100 feet of wire per treatment. The cycle time varies from 5 minutes to 24 hours, depending on the wire characteristics desired. Following treatment, the wire is permanently straightened. Although treated in air, a mild oxide coating forms on the wire.

An Electrical Heat Treating Facility

A practical facility, which permits full control over time and current cycles, may be constructed using commercially available autotransformers and other components. The power supply should be capable of delivering a maximum output cur-

rent of 10 amperes at 540 volts. Suitable control circuits should provide for adjustable time settings for the heating cycles. Proper interlocks should be inserted wherever safety considerations so dictate. A simplified diagram of a practical installation with typical autotransformer connections is shown in Figure 5.

Typical Heat Treatment Procedure

In a typical heat run, a length of wire (about 50 feet long) is suspended between the end feeders of the power supply. The wire is then cleaned and degreased with a suitable solvent. Heat treatment then proceeds in accordance with the two-stage process described earlier in this paper. After treatment, the wire is removed and wound on a large diameter (not less than 18 inches) before test.

Results of a Typical Heat Treatment

In order to measure the temperature-delay characteristics of a treated wire, it is customary to wind the wire into a delay-line test fixture to the diameter of its intended use. In this fixture, torsional mode converters can be quickly welded to the wire. Transducer assemblies are then slipped over the end tapes. The assembled delay line is then placed into a temperature chamber for the desired test run.

From the recorded test data of change in delay versus temperature, it is desirable to plot a normalized curve whose ordinate is the change in delay from the reference delay at 20°C divided by the reference delay. A typical normalized test curve for a line with a delay of about 4,000 microseconds is shown in Figure 6. In the region around 60°C, the temperature coefficient is practically zero. At lower temperatures, the coefficient becomes negative. At higher temperatures, it becomes positive.

From this curve, it is easy to calculate an average temperature coefficient of delay in parts per million per degree C (PPM/°C) over a specified temperature range. Several of these average coefficients are shown with their corresponding temperature ranges.

Average Temperature Coefficient of Delay PPM/°C	Temperature Range °C
0.75	40 to 80
1.33	20 to 90
2.09	0 to 100

Conclusion

Proper identification and control is mandatory for both incoming and heat treated stocks of wire. The heat treatment for a given temperature coefficient of delay may vary from batch to batch of wire as received from the supplier, although the wire may have been ordered with identical specifications. Each length of heat treated wire should be labeled with a stock control number which identifies the batch and the heat treatment.

The proper maintenance and correlation of control records and test data will enable the delay-line manufacturer to eliminate faulty materials, attain consistent performance, and predict correct heat treatment procedure for any application.

References

1. H. A. Wilson Co., Union, New Jersey.

2. N. B. Pilling and A. M. Talbot, Age-hardening of Metals. Am. Soc. Metals 231-61. Dispersion-hardening alloys of Ni and Fe-Ni-Ti 394

3. G. G. Scarrott and R. Naylor, "Wire-Type Acoustic Delay Lines for Digital Storage", Proc. IEE, vol. 103, pt. B, suppl. no. 3, pp. 497-508; March, 1956.

4. W. H. Clapp and D. S. Clark, Engineering Materials and Processes, p. 183, International Textbook Co., Scranton, Pa., May,1949.

5. M. W. Zemansky, Heat and Thermodynamics, pp. 101-106, McGraw-Hill Book Co., New York, N. Y., 1957.

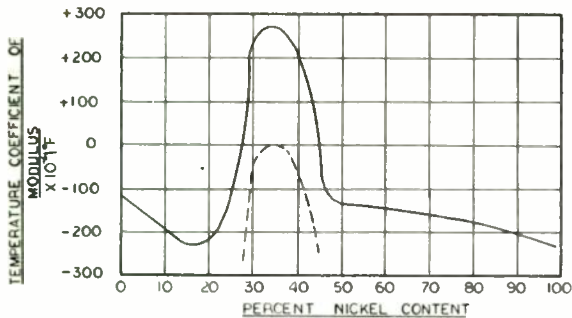


Fig. 1. Temperature coefficient of the modulus of elasticity of a nickel-iron alloy versus percentage nickel content.

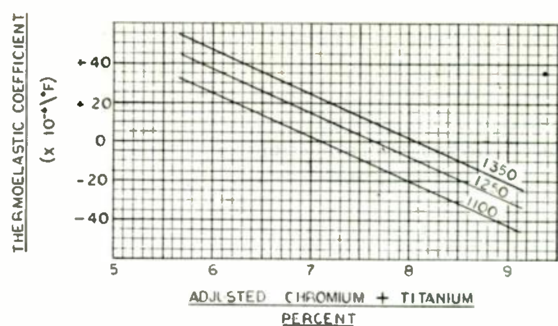


Fig. 4. Thermoelastic coefficient versus adjusted chromium plus titanium percentage for 4 hour aging treatment. (Temperatures are in °F.)

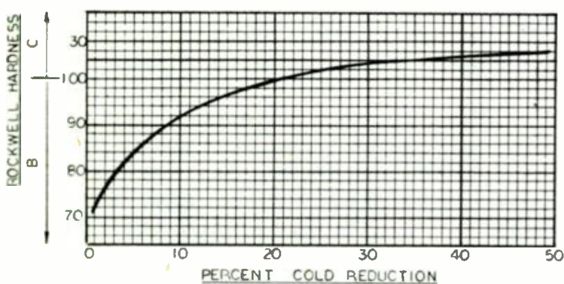


Fig. 2. Work-hardening rate of Ni-Span C Alloy.

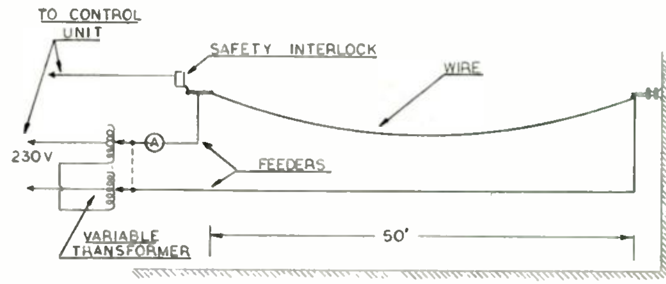


Fig. 5. Simplified diagram of an electrical heat treatment facility.

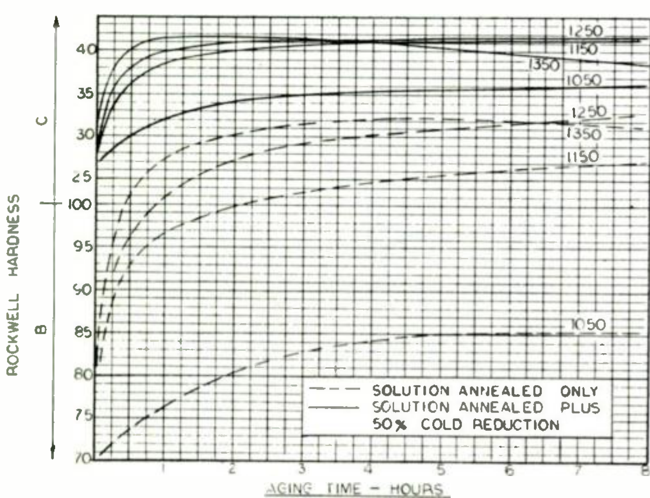


Fig. 3. Hardness versus aging time for Ni-Span C Alloy. (Temperatures are in °F.)

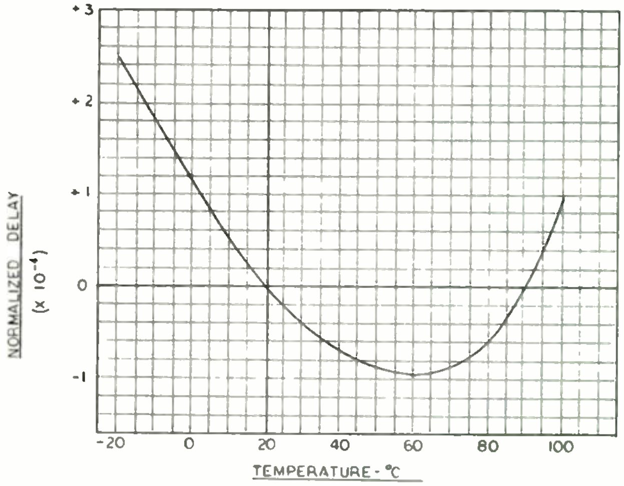


Fig. 6. Typical curve of normalized delay versus temperature.

A PIEZOELECTRIC-PIEZOMAGNETIC GYRATOR

Morio Onoe* and Mitsuo Sawabe
Institute of Industrial Science
University of Tokyo
Azabu, Tokyo, Japan

SUMMARY

A linear passive unilateral element, which has a high forward/backward transmission ratio at all frequencies, is realized by combining two resistances and an electromechanical gyrator according to Gamo's theory. The gyrator consists of three, mechanically coupled, ceramic elements, two of which may be piezomagnetic and the third piezoelectric or two of which may be piezoelectric and the third piezomagnetic. In either arrangement a unilateral coupling between mechanical and electrical systems is provided by using one piezoelectric and one piezomagnetic element. The third element is merely for driving the mechanical system. Such a design makes manufacturing considerably easier. Experimental models based on this design have been constructed and tested and the measured performance characteristics show good agreement with the theory.

I. INTRODUCTION

A linear, passive, electric circuit which violates the reciprocity theorem has been a subject of much study in recent years. Tellegen introduced the ideal gyrator as the fifth fundamental circuit element, besides the existing four elements, namely, resistance, capacitance, inductance and the ideal transformer.(1) An ideal gyrator has circuit equations shown in Fig. 1. Any linear, passive, nonreciprocal network can be represented by a suitable combination of ideal gyrators and reciprocal circuits.

Two features of a gyrator are especially important for practical applications. First a gyrator is an immittance converter as shown in Fig. 2, which transforms any circuit into its dual. Second, a gyrator can be an isolator, which allows the transmission of signal in one direction only, by a suitable combination with a resistance as shown in Fig. 3.

In practice, a gyrator has been realized by using the Hall effect in a semiconductor,(2) the Faraday effect in a ferrite(3) or the nonreciprocal nature in a combination of two dual types of electromechanical transducers. These practical gyrators consist of an ideal gyrator and extra circuits, which usually limit their range of application. This is especially true for an electromechanical gyrator because the band-pass characteristics of the transducers make the gyrator a relatively narrow band device. Hence the most promising application of electromechanical gyrators seems to be in isolators, which will be the subject of the present paper. McMillan first noted that if an electric transducer is mechanically coupled to a magnetic transducer, the resultant system is an anti-reciprocal four-terminal circuit, which can be an isolator by a suitable combination with a circuit component (usually a resistor) as shown in Fig. 4(a).(4) Fig. 4(b) shows a dual of McMillan's isolator. Black and Scott realized such a system by a combined piezoelectric-piezomagnetic vibrator made of a bar of x-cut quartz bonded to a bar of stainless steel.(5) More recently, piezoelectric ceramics and piezomagnetic ferrites with high electromechanical coupling have been used to obtain greatly improved characteristics by Germano, Curran, et al.(6,7) and by the authors.(8)

The isolator described by McMillan has one severe drawback in comparison with the ideal isolator. Unilateral transmission can be obtained only for a single frequency or a group of frequencies, because the transfer immittance of combined vibrators vary widely with frequency. Gamo established, however, a general method for obtaining electromechanical systems, unilateral at all frequencies.(9) These systems are named "achromatic" isolators and are shown in Fig. 5. In the figure black boxes are the same reciprocal four-terminal circuits. Cascaded connection of the box with an ideal gyrator as shown in the upper branch yields an anti-reciprocal circuit, which may be realized by a piezoelectric-piezomagnetic combined vibrator similar to that previously discussed. The box in the lower branch may be realized by either a purely piezoelectric or a purely piezomagnetic vibrator. It should be noted that these isolators need two resistors instead of one as used in the ideal isolator. This increases the loss in the forward transmission but provides the advantages of achromatic, unilateral, transmission characteristics. Realization of Gamo's isolator in practice, however, is extremely difficult, because it is necessary to match the frequency characteristics of the two vibrators in the upper and lower branches.

*Presently visiting Bell Telephone Laboratories, Whippany, New Jersey, on leave of absence.

In this paper a novel type of electromechanical, achromatic isolator is presented. The isolator consists of three, mechanically coupled transducers, two of which may be piezoelectric and the third piezomagnetic or two of which may be piezomagnetic and the third piezoelectric. In either arrangement a unilateral coupling between mechanical and electrical systems is provided using one piezoelectric and one piezomagnetic transducer. The third transducer is merely for driving the mechanical system. Such a design makes manufacturing considerably easier, because there is only one mechanical system and because adjustment by purely electrical means is possible after fabrication of the mechanical parts.

II. THE ELECTROMECHANICAL GYRATOR

Fig. 6 shows a few configurations of electromechanical gyrators consisting of three transducers. The hatched portion is made of piezoelectric ceramic material and bonded to the rest of system made of piezomagnetic ferrite. The terminals 1-1' are for a piezomagnetic transducer and the terminals 2-2' are for a piezoelectric transducer. Combination of both transducers provides an achromatic unilateral transmission between electrical and mechanical systems. The terminals 3-3' are for the third transducer, which may be piezoelectric or piezomagnetic.

In practice the piezoelectric third transducer is preferred. This is because there will be less stray coupling between the third transducer and the other transducers and because a higher electromechanical coupling coefficient is obtainable in piezoelectric ceramics than in ferrites. Hence all the examples except (a) in Fig. 6 have piezoelectric third transducers. Configurations (a), (b) and (c) may be called cascaded structures and (d) and (e) sandwiched structures. Configurations (c) and (e) use a piezoelectric transducer with divided electrodes and, therefore, need only one bonding operation.

III. EQUIVALENT CIRCUIT

Although Gamo's theorem can be extended to cover unilateral transmission between electrical and mechanical systems, the following analysis based on an equivalent circuit is much simpler. Fig. 7 shows the equivalent circuit based on the voltage force analogy near a mechanical resonance of composite structure. L_1 is the clamped inductance of the piezomagnetic transducer, while C_2 is the clamped capacitance of the piezoelectric transducer. Z_M represents a mechanical resonance of the composite structure and is a series resonant circuit as shown in the figure. It should be noted, however, that a more elaboratoe structure can be inserted between the first two transducers and the third transducer, so that much more complicated frequency characteristics like those of a mechanical filter may be obtained.

The condition for suppressing the transmission from the first two transducers to the third transducer can be described as follows:

$$\varphi_1 I_1 + \varphi_2 E_2 = 0 \qquad (1)$$

where φ_1 and φ_2 are the transformer ratios of the piezomagnetic and piezoelectric transducers, respectively. I_1 is a driving current for the piezomagnetic transducer and E_2 is a driving voltage for the piezoelectric transducer, both are to be derived from the same signal source. Such distributions of current and voltage are possible in the two circuits shown in Fig. 8, where

$$\frac{L_1}{C_2} = R_1 R_2 \qquad (2)$$

$$-\frac{\varphi_1}{\varphi_2} = R_1 \qquad (3)$$

In this figure the mechanical portion is omitted because there is no current in that portion under the condition (1). It can be seen that the conditions (2) and (3) are independent of frequency and therefore, an achromatic suppression of transmission is achieved. The need for two resistors is consistent with the Gamo's theorem.

The values of L_1, C_2 can be measured at a frequency well removed from mechanical resonances. The ratio φ_1/φ_2 is conveniently obtained by measuring first, the output voltage of the third transducer for a given input current I_1 into the piezomagnetic transducer with a short circuit of the terminals 2-2', and second, the output voltage for a given input voltage E_2 into the piezoelectric transducer with an open circuit of the terminals 1-1'. The other parameters can be determined by a conventional method for measuring electromechanical vibrators as described in Appendix.

IV. CONDITION FOR CONSTANT INPUT RESISTANCE

It can be seen that if the two resistances, determined by equation (2) and (3) happen to be equal, the input impedance looking into the

terminals 0-0' becomes a constant resistance R, which is also equal to these two resistances,

$$R = R_1 = R_2 \tag{4}$$

This condition is interesting in practice because it simplifies the matching at the terminals 0-0'. The condition can be realized by a proper design of the physical dimensions of a vibrator. The realization is also possible by purely electrical means even after the fabrication of the vibrator.

Roughly speaking φ_1, and by virtue of (3) also R_1, are proportional to the number of turns of the coil, N, while L_1 is proportional to N^2. This yields R_2 proportional to N according to equation (2). Hence changing the number of turns of the coil in order to obtain a constant input resistance has little effect. On the other hand, the magnetic biasing field affects φ_1 considerably but L_1 only slightly. Hence the condition (4) is easily obtained by adjusting the biasing field, especially when the original R_1 is larger than R_2. When the original R_2 is larger than R_1, the condition (4) is achieved by merely adding some capacitance parallel to C_2.

V. FORWARD TRANSMISSION CHARACTERISTICS

The transmission characteristics in the forward direction, namely from the terminals 3-3' to the terminals 0-0', can be conveniently described based on the equivalent circuit shown in Fig. 7. The portion left of the terminals m-m' is simplified under conditions (2) and (3) as shown in Fig. 9 (a) and (b). Further simplification is possible as shown in Fig. 10 if condition (4) is satisfied. Since this condition is preferable in practical applications and can always be achieved as discussed in the previous section, the present discussion is limited to this special case. The right half section of the equivalent circuits shown in Fig. 10 is a conventional filter circuit, characteristics of which have been well discussed in literature. If the mechanical loss is negligible, a perfect matching at the resonant frequency is possible by the use of a tuning inductance parallel to C_3. This yields the available power of the source, P_o,

$$P_o = \frac{E^2}{4\left(2\varphi_2^2 R\right)} \tag{5}$$

while the available power of the left half section, P_L, is,

$$P_L = \frac{E_M^2}{4\varphi_2^2 R}, \tag{6}$$

where $E_M = E/2$ at the resonance. Hence, at the resonance, the ratio P_o/P_L is two, and this yields the minimum insertion loss of 3db. Any mechanical loss or any mismatching increases the insertion loss.

VI. EXPERIMENTS

An achromatic isolator operating at 135 kc/s was made of barium titanate and nickel-ferrite and reported in a previous paper.[8] The cascaded structure shown in Fig. 6(b) was used. The frequency characteristics are shown in Fig. 11. The 3 db bandwidth is 1.2kc/s and unilateral discrimination of more than 50 db is obtained. An irregularity at 141 kc/s is due to an unwanted mode of vibration and may be removed by an improved mechanical design. The high forward insertion loss of 12db is due to a poor electromechanical coupling of material and a mechanical loss.

In order to improve the forward insertion loss, Clevite PZT-5 ceramic (k_{33} = 0.68) and Kearfoot N-51 ferrite (k = 0.40) have been used. Table 1 shows equivalent parameters for three models. Gyrator A uses the sandwiched structure shown in Fig. 6(d). Resistance values experimentally obtained for the best suppression of backward transmission are compared with the calculated values from equations (2) and (3). The discrepancy seems chiefly due to an error in determining the ratio φ_1/φ_2 (about 10%). The condition for constant input resistance discussed in section 4 was achieved by adding a capacitance of about 375 pF parallel to C_2 and adjusting slightly the bias magnet. This yields experimental values shown in Table 1. Fig. 12 shows the open circuit impedances of the terminals 0-0' and 3-3' measured by a Wayne-Kerr bridge, type B601. It can be seen the former is almost constant and equal to R_1 and R_2 as the theory predicts. Insertion loss characteristics under matched termination conditions are shown in Fig. 13. A fairly flat suppression of backward transmission over the measured range of frequency was obtained as expected. The forward loss was reduced to 5db, of which 3db is the theoretical minimum and 2db is an excess loss due to mechanical losses.

Gyrator B uses the sandwiched structure with divided electrodes shown in Fig. 6(e). Dividing the electrodes reduces the coupling of the piezoelectric transducer (2-2') relative to that of the piezomagnetic transducer. This increases the ratio φ_1/φ_2 and achieves the condition for constant input resistance without any additional adjustment as seen in Table 1. Fig. 14 shows the impedance characteristics and Fig. 15 the insertion loss characteristics under matched termination conditions. An irregularity in the backward insertion loss seems due to the off-nodal support of this particular model and can probably be removed by a slight change of the electrode configuration which allows the usual nodel support. Fig. 16 and 17 are the impedance and the insertion loss characteristics, respectively, of Gyrator C which uses the cascaded structure shown in Fig. 6(b). A capacitance of about 2270 pF is added parallel to C_2 in order to achieve the condition for constant input resistance.

VII. CONCLUSION

A linear passive isolator, which has an achromatic suppression of backward transmission has been realized by combining two resistances and an electromechanical gyrator according to Gamo's theory. The gyrator consists of three, mechanically coupled, ceramic elements, two of which may be piezomagnetic and the third piezoelectric or two of which may be piezoelectric and the third piezomagnetic. In either arrangement a unilateral coupling between the mechanical and the electrical system is provided by using one piezoelectric and one piezomagnetic element. The third element is merely for driving mechanical system.

Characteristics of the gyrator have been discussed based on an equivalent circuit. A constant input resistance can be obtained at a terminal pair of the isolator by purely electrical means even after the fabrication of gyrator. This simplifies the matching at this terminal. The theoretical minimum insertion loss is 3db under the matched termination conditions. A few models have been made using both the sandwiched and the cascaded structures. A highly achromatic suppression of backward transmission has been obtained in agreement with theoretical predictions. Minimum forward insertion loss consists of the theoretical minimum of 3 db and an excess loss of a few db due principally to mechanical losses.

ACKNOWLEDGMENT

The authors are grateful for the advice of Professor N. Takagi. The models shown in Fig. 12 to 17 were made through the courtesy of Bell Telephone Laboratories, where one of the authors is visiting presently. Thanks are due to W. J. Nowotarski for fabrication and J. J. Gallo for measurement. Drs. A. H. Meitzler and J. E. May, Jr., kindly read the English text.

REFERENCES

(1) B. D. H. Tellegen, Philip. Res. Rep. Vol. 3, 81-101, 1948.

(2) W. P. Mason, W. H. Hewitt and R. F. Wick, J. Appl. Phys. Vol. 24, 166-175, 1953.

(3) C. L. Hogan, Bell Sys. Tech. J. Vol. 31, 1-31, 1952.

(4) E. M. McMillan, J. Acoust. Soc. Am. Vol. 18, 344-347, 1946.

(5) L. J. Black and H. J. Scott, J. Acoust. Soc. Am. Vol. 25, 1137-1140, 1953.

(6) C. P. Germano and D. R. Curran, Electronic Components Conference (San Francisco), No. 24-1, May, 1961.

(7) D. R. Curran, J. H. Silverman and J. Schoeffler (private communication).

(8) Morio Onoe and Mitsuo Sawabe, Proc. Spring Meeting Acoust. Soc. Japan, No. 91, May, 1961.

(9) H. Gamo, Proc. 26th Joint Conference of Elec. Eng. Societies of Japan, No. 9-1, 1952; J. Acoust. Soc. Japan, Vol. 10, 65-76, 1954; Trans. International Symposium on Circuit and Information Theory, 283-298, 1959.

APPENDIX A

Determination of Equivalent Circuit Parameters

The parameters of the equivalent circuit shown in Fig. 7 can be determined in the following manner. φ_3 can be set equal to one without loss of general validity.

(1) Clamped parameters, L_1, C_2 and C_3 are measured at a frequency well above the mechanical resonance.

(2) Admittance measurement at the terminals 3-3' with the terminals 2-2' shorted and the terminals 1-1' open yields the resonant frequency f_o and the resonance resistance R_M.

(3) Admittance measurement at the terminals 2-2' with the terminals 3-3' shorted and the terminals 1-1' open yields the same resonant frequency and another resonance resistance R'_M. The transformer ratio φ_2 is given by the following equation;

$$\varphi_2 = \left(R_M/R'_M\right)^{1/2} \tag{A1}$$

(4) The above two admittance measurements determine the quadrant frequencies, f_1 and f_2, which correspond to conductances equal to half of the resonance conductance. Then

$$Q_M = \frac{f_o}{f_1 - f_2} \tag{A2}$$

$$L_M = \frac{Q_M R_M}{2\pi f_o} \tag{A3}$$

$$C_M = \frac{1}{(2\pi f_o)^2 L_M} \tag{A4}$$

It may be noted that the clamped capacitances C_2 and C_3 can also be obtained from the algebraic mean of susceptances at the quadrant frequencies.

(5) The ratio φ_1/φ_2 is determined as described in the text. Since φ_2 is already known, φ_1 can be determined. It should be noted that impedance measurements at the terminals 1-1' are possible but the separation of parameters is much more complicated because of the loss of the coil.

APPENDIX B

Derivation Of Equivalent Circuits For Achromatic Isolators

The portion left of the terminals m-m' in Fig. 7, associated with resistances in the manner shown in fig. 8(a) yields a four terminal network with terminals o-o' and m-m'. Let voltages across and currents flowing into these terminals be E, I, E_M and I_M, respectively, then the following equations hold;

$$E_M = \varphi_1 I_1 + \varphi_2 E_2, \tag{B1}$$

$$E + \varphi_1 I_M = (R_1 + j\omega L_1) I_1, \tag{B2}$$

$$E - E_2 = R_2 (I - I_1) \tag{B3}$$

$$\varphi_2 I_M = j\omega C_2 E_2 - I + I_1, \tag{B4}$$

$$R_2 \varphi_2 I_M = (1 + j\omega C_2 R_2) E_2 - E. \tag{B5}$$

These equations together with Equations (2) and (3) in the text yield the following equations;

$$E_M = \frac{\varphi_2^2 (R_1 + R_2)}{1 + j\omega C_2 R_2} I_M, \tag{B6}$$

$$I + \frac{2\varphi_2}{1+j\omega C_2 R_2} I_M = \left(\frac{1}{R_1+j\omega L_1} + \frac{j\omega C_2}{1+j\omega C_2 R_2}\right) E, \tag{B7}$$

which correspond to the left portion of equivalent circuits shown in Fig. 9(a). If the condition of constant input resistance holds, as given by Equation (4) in the text, Equations (B6) and (B7) can be further reduced to the following,

$$E_M = \frac{2\varphi_2^2 R}{1 + j\omega C_2 R} I_M, \tag{B8}$$

$$I = \frac{E}{R} - \frac{E_M}{\varphi_2}, \tag{B9}$$

which correspond to the left portion of the equivalent circuit shown in Fig. 10.

In the case of Fig. 8(b) the following similar set of equations apply instead of Equations (B1) to (B5)

$$E_M = \varphi_1 I_1 + \varphi_2 E_2, \tag{B10}$$

$$I = \frac{E_1}{R_2} + \frac{E_1 + \varphi_1 I_M}{j\omega L_1} \tag{B11}$$

$$= \left(\frac{1}{R_1} + j\omega C_2\right) E_2 - \varphi_2 I_M, \tag{B12}$$

$$E = E_1 + E_2, \tag{B13}$$

where E_1 is the voltage across the resistance R_2. These equations together with the Equations (2) and (3) in the text yield the following equations,

$$E_M = \frac{\varphi_2^2 (R_1 + R_2)}{1 + j\omega C_2 R_2} I_M, \qquad \text{(B14)}$$

$$E = \left(\frac{j\omega L_2 R_2}{R_2 + j\omega L_1} + \frac{R_1}{1 + j\omega C_2 R_1}\right) I + \frac{2R_2}{\varphi_2 (R_1 + R_2)} E_M, \qquad \text{(B15)}$$

which correspond to the left portion of the equivalent circuit shown in Fig. 9(b). If the condition of constant input resistance holds, these equations also reduce to Equations (B8) and (B9).

TABLE 1. Equivalent Circuit Parameters of Gyrators and Their Characteristics

Gyrator	A	B	C	Unit
Type	Fig. 6(d)	Fig. 6(e)	Fig. 6(b)	—
L_1	103	120	140	μH
C_2	720	330	440	pF
C_3	720	330	440	pF
f_0	152.33	164.32	113.42	kc/sec
Q_M	246	299	157	—
L_M	165	522	340	mH
C_M	6.66	1.81	5.77	pF
R_M	640	1800	1.55	Ω
R'_M	630	1800	1.53	Ω
φ_2	1.01	1.00	1.01	-
$-\varphi_1/\varphi_2$	266	626	250	Ω
Calculated resistances from Equations (2) and (3)				
R_1	266	626	250	Ω
R_2	537	580	1270	Ω
Resistances experimentally obtained for the best backward suppression				
R_1	281	565	—	Ω
R_2	460	578	—	Ω
Resistances experimentally obtained for the condition of constant input resistance				
R_1	308	—	236	Ω
R_2	304	-	236	Ω
Minimum insertion loss				
forward	5.0	6.0	6.0	db
backward	46	43	49.5	db
3 db bandwidth	2.9	1.5	2.0	kc/sec

IDEAL GYRATOR

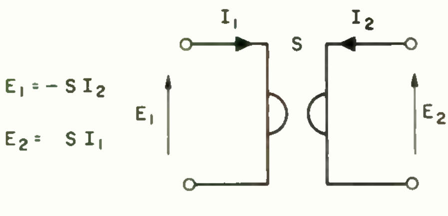


Fig. 1.

IMMITTANCE CONVERTER

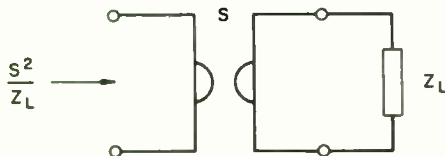


Fig. 2.

IDEAL ISOLATOR

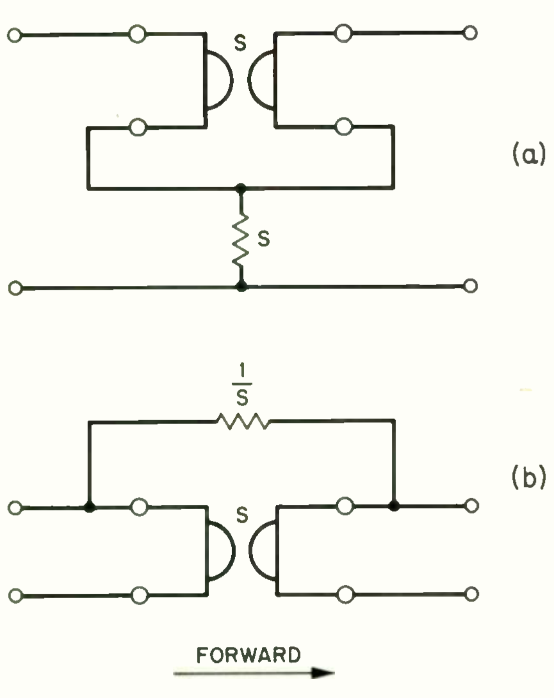
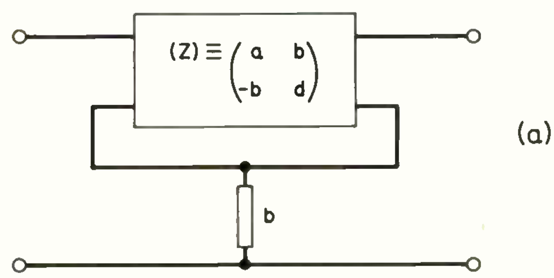


Fig. 3.

ISOLATOR AFTER MCMILLAN AND ITS DUAL



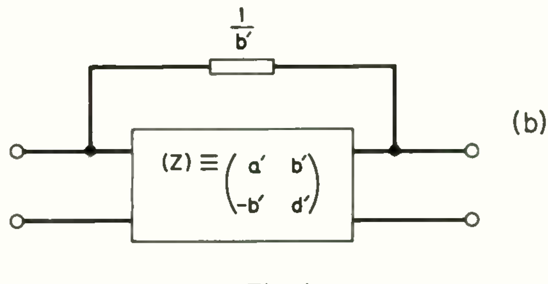


Fig. 4.

ACHROMATIC ISOLATORS AFTER GAMO

DOTTED LINES SHOW THE ALTERNATIVE SERIES CONNECTION FOR THE OUTPUT

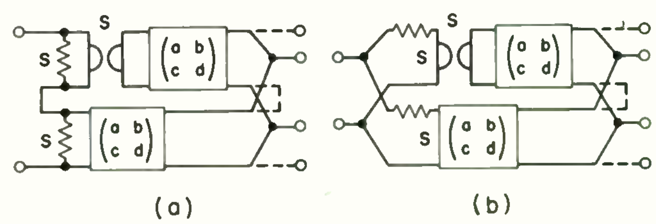


Fig. 5.

ELECTROMECHANICAL GYRATORS CONSISTING OF THREE TRANSDUCERS

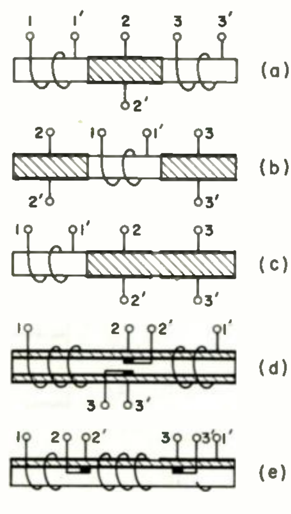


Fig. 6.

EQUIVALENT CIRCUIT OF THE GYRATOR

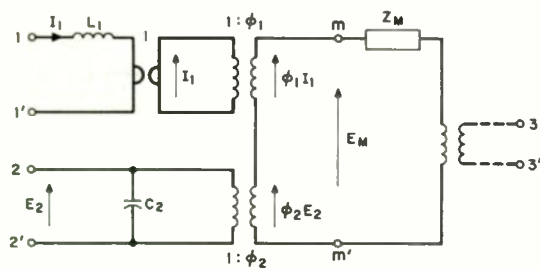


Fig. 7.

REALIZATION OF ACHROMATIC SUPPRESSION TRANSMISSION BY MEANS OF ASSOCIATED RESISTANCES

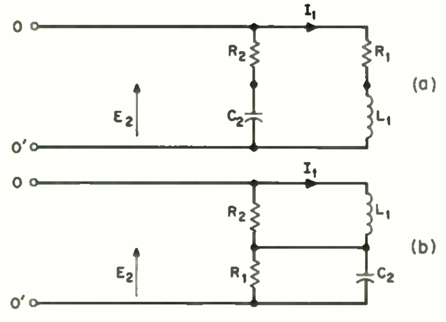
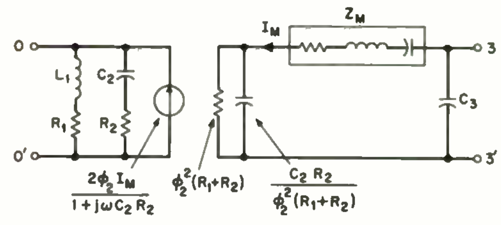


Fig. 8.

EQUIVALENT CIRCUITS FOR ACHROMATIC ISOLATORS



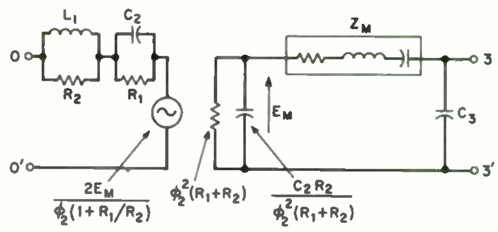
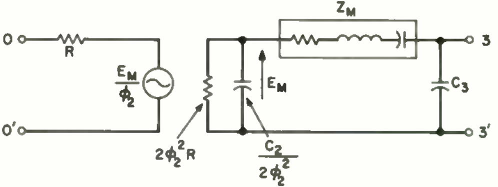


Fig. 9.

EQUIVALENT CIRCUIT IN THE CASE $R_1 = R_2 = R$



AVAILABLE POWER FROM THE SOURCE UNDER MATCHED CONDITION

$$P_0 = \frac{E^2}{4(2\phi_2^2 R)}$$

AVAILABLE POWER FROM THE OUTPUT TERMINAL 0–0′

$$P_A = \frac{E_M^2}{4R\phi_2^2} \qquad E_M = \frac{E}{2}$$

INSERTION LOSS UNDER MATCHED CONDITION

$$\frac{P_0}{P_A} = 2 \longrightarrow 3\text{db}$$

Fig. 10.

CHARACTERISTICS OF AN ACHROMATIC ISOLATOR

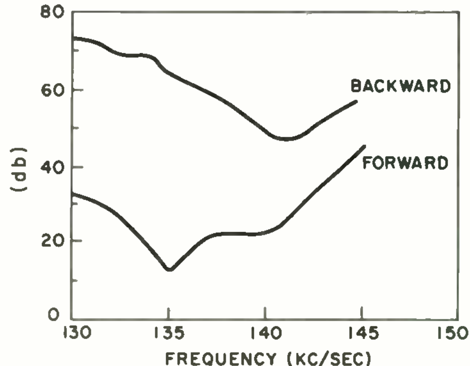


Fig. 11.

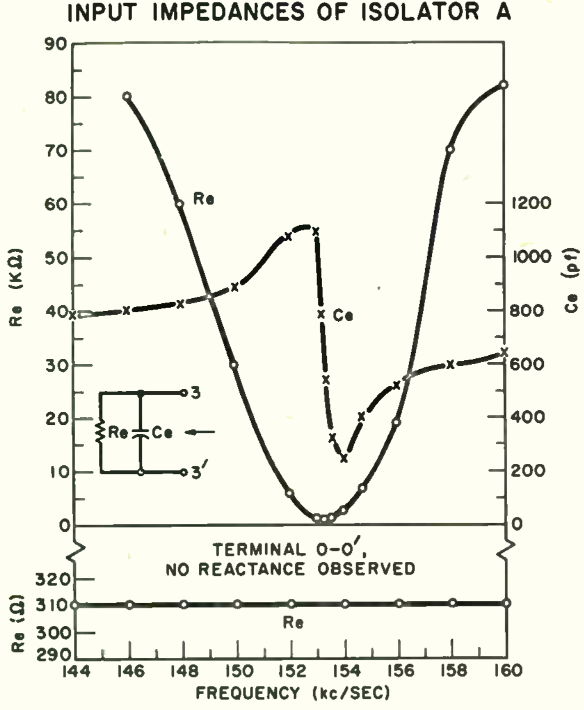


Fig. 12.

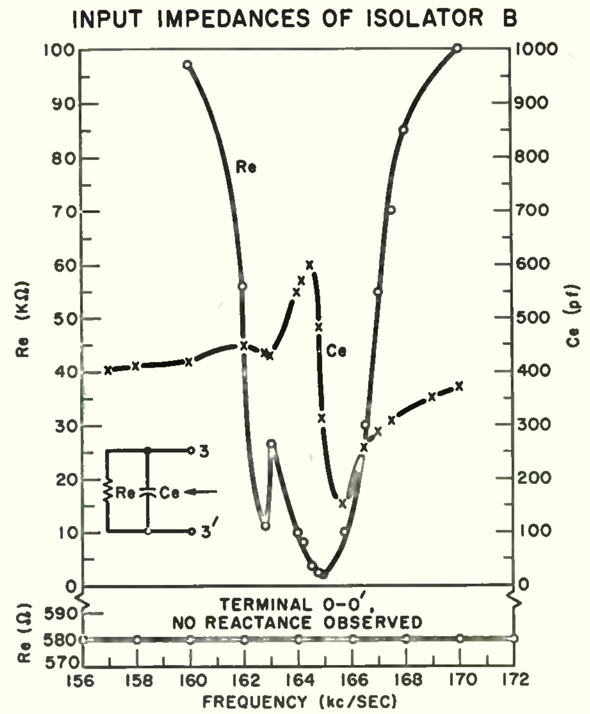


Fig. 14.

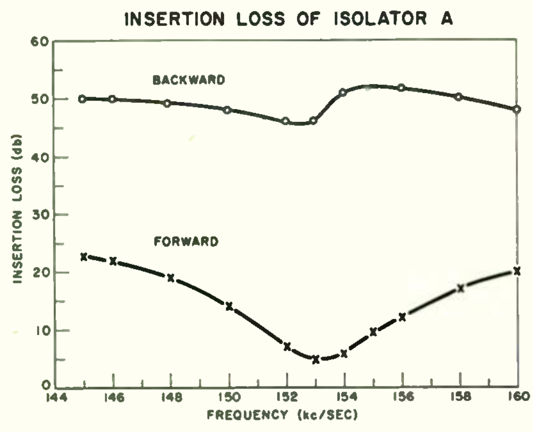


Fig. 13.

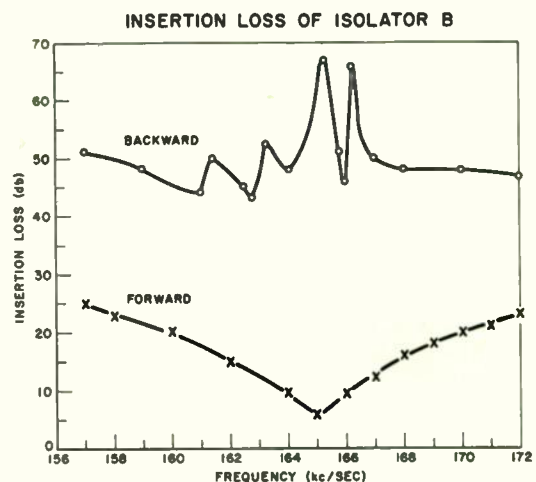


Fig. 15.

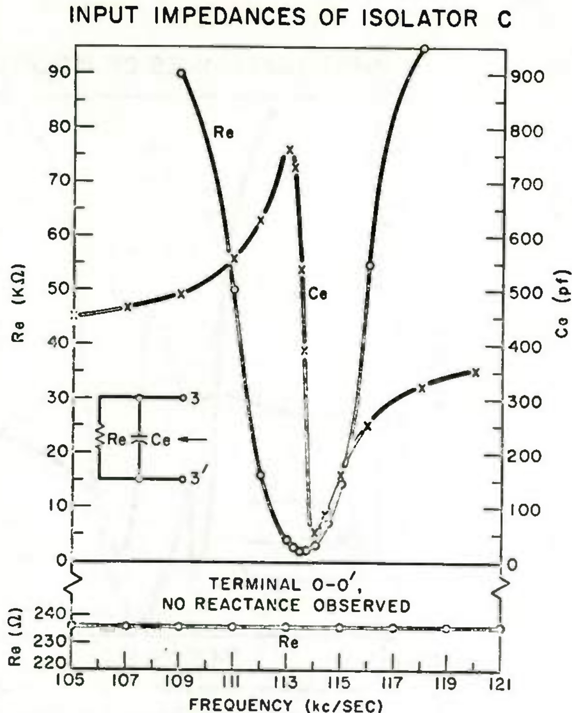


Fig. 16.

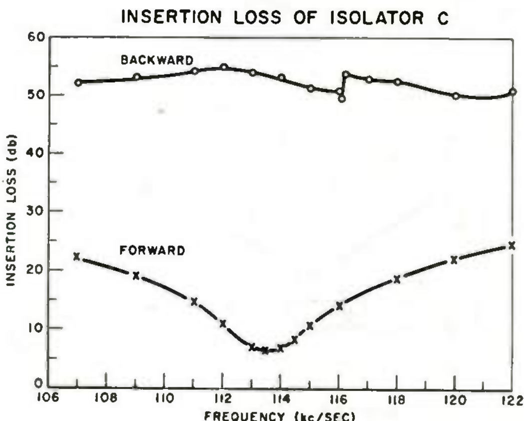


Fig. 17.

WIDE-BAND, INSTANTANEOUS SPECTRUM ANALYZERS EMPLOYING DELAY-LINE LIGHT MODULATORS

Louis B. Lambert
Electronics Research Laboratories
School of Engineering and Applied Science
Columbia University
New York 27, N. Y.

Summary

Spectrum analyzers are devices which are usually used to measure the power spectrum, as a function of frequency, of real signals. Our attention shall be focused on fast acting analyzers which provide a wide frequency coverage (10 to 100 mc) and a large number of resolution elements (1000 to 10,000).

The synthesis technique that will be discussed employs the Debye-Sears effect in a transparent ultransonic delay line to convert the electrical input signal to a proportional spatial pattern in the form of a traveling pressure wave. Partially coherent light is spatially modulated by this pattern, and a lens is used to spatially integrate the resulting light distribution. This provides the Fourier transform of the input signal as a distribution of light in the focal plane of the lens. A suitable photosensitor detects this light pattern, producing an electrical output signal which is proportional to the instantaneous power spectrum of the input signal.

In this paper, a first order theory of operation is reviewed along with certain device limitations. Specific analysis and design problems will be presented in separate papers similar to the companion paper in this issue.*

I. Spectra

Consider the real waveform $v(t)$ which is the result of some physical process. Subject to the usual constraints,** we can express the waveform by the Fourier transform pair;

$$v(t) = \int_{-\infty}^{\infty} V(f)\exp(j2\pi ft)df \ , \qquad (1a)$$

$$V(f) = \int_{-\infty}^{\infty} v(t)\exp(-j2\pi ft)dt \ . \qquad (1b)$$

Several spectra of the waveform $v(t)$ can now be defined. Typically, $V(f)$ is defined as the complex spectral density or "spectrum" of $v(t)$, $|V(f)|$ is the "amplitude spectrum," and $|V(f)|^2$ is the "power spectrum." Since we shall be concerned with spectrum analyzers rather than the process which causes the waveform $v(t)$, we should define spectra which are essentially determined by the analyzer and do not require integration over all (past and future) time. Definitions of this type have been stated previously by Karkevich (1), Fano (2) and Page (3) where they take the form

$$V_T(f,t) \equiv \int_{t-T}^{t} v(\tau)\exp(-j2\pi f\tau)d\tau \ . \qquad (2)$$

Here, $V_T(f,t)$ is defined as the sectioned spectrum, or the <u>instantaneous spectrum</u> of the "input" waveform $v(\tau)$, where τ represents input time (t now represents output time). This instantaneous spectrum can be seen to be the spectrum of a segment of the waveform. This segment has a duration T which is sometimes referred to as the integration time. Also, the integration time immediately precedes the instant t, at which time the spectrum is evaluated. Thus, the integration time is constant but it constantly "slides" along the time axis.

An ideal spectrum analyzer can now be defined as a signal processor which operates on the input waveform $v(\tau)$ to produce at its output, a signal which is proportional to the <u>instantaneous spectrum</u> $V_T(f,t)$, or the <u>instantaneous amplitude</u> spectrum $|V_T(f,t)|$, or the <u>instantaneous power spectrum</u> $|V_T(f,t)|^2$.

For certain applications, it may be desirable to purposefully modify the instantaneous spectrum. Also, actual networks can only approximate the instantaneous spectrum since their impulsive

---

* Arm, M., Lambert, L., Silverberg, B., "Electro-Optical Transfer Characteristics of Liquid Delay-Line Light Modulators," in this issue.

** The waveform must contain a finite number of discontinuities and a finite number of maxima and minima in any interval. Also, the waveform must satisfy the inequality

$$\int_{-\infty}^{\infty} |v(t)|\, dt < \infty \ .$$

response may not be uniform during the integration time. A more general definition is given by the weighted spectrum,

$$V_w(f,t) \equiv \int_{-\infty}^{\infty} w(\tau-t)v(\tau)\exp(-j2\pi f\tau)d\tau \quad .(3)$$

The term $w(\tau-t)$ is called the weighting function. If the weighting function is equal to unity, we obtain the spectrum $V(f)$. When the weighting function is composed of unit steps* such that $w(\tau-t) = u(\tau-t+T)-u(\tau-t)$, we obtain the instantaneous spectrum $V_T(f,t)$. Thus, an ideal spectrum analyzer implies a constant, finite weighting during the integration time (T) and zero weight for all other times.

II. Optical Processors

As was originally implied by Abbe (4) when he developed his theory of optical resolution for coherent illumination, optical systems are capable of processing information. Filter theoretic descriptions of optical systems are described in the text by Born and Wolf (5), in the text edited by Wolf (6) and in articles by Elias (7, 8), Rhodes (9, 10), Cheatham (11), O'Neill (12, 13, 14), Cowley (15), Cutrona (16) and Gamo (17). In order to synthesize a spectrum analyzer which employs optical techniques, the input signal $v(t)$, must first be transformed to a signal $v(x)$ or $v^2(x)$ where x is a variable in physical space. This transformation may be accomplished by first converting $v(t)$ to a proportional time variation of light amplitude or intensity, and then employing a photographic process to obtain the spatial variation. Synthesis techniques of this general type are described by Howell (18, 19) as applied to "Optical Analog Computers," and by Cutrona (16) as applied to "Data Processing and Filtering Systems."

In order to circumvent the time delays, film transport problems, etc., that are associated with a photographic process, other spatial light modulation devices have been employed. An investigation of various instantaneous light modulation techniques was performed by Wilmotte (20). Ultrasonic diffraction techniques have been employed by Wilmotte (20) and Airborne (21) to synthesize a spectrum analyzer but only a limited number of resolution elements have been realized.

There have been other investigations and applications of "instantaneous" spatial light modulators. Major successes have been achieved by employing the Debye-Sears effect (22)* in a transparent ultransonic delay line which is directly incorporated into an optical configuration. Such an instantaneous spatial light modulator was successfully employed as a light valve in the Scophony television receiver (25) and as a pulse shaper by Terry (26). Levi (27) employed this device to provide video recording, Lambert (28) employed it as a "Two-Dimensional Filter" and Rosenthal (29) presented a description of the functions that could be obtained from other optical configurations. The use of the Debye-Sears effect to spatially modulate light in a partially coherent, multiplexed, optical configuration presently appears promising as a means of synthesized fast acting wide bandwidth spectrum analyzers having a large number of resolution elements.

III. Optical Transfer Functions

Consider a plane, linearly polarized, sinusoidal, electromagnetic wave traveling in free space with its Poynting vector parallel to the z axis as shown in Fig. 1. Let the peak amplitude of the electric vector be E_c and let this wave be incident upon an aperture which contains an object in the x-y plane. The object can modify both the amplitude and the phase of the incident plane wave of light, producing an electric field at $z = 0^+$ given by

$$E_a(x,y) \equiv [E_c][T(x,y)] \quad . \qquad (4)$$

$T(x,y)$ is defined as the complex transmission (or transparency) of the object which can be expressed by

$$T(x,y) \equiv T_o(x,y)\exp[j\psi(x,y)] \qquad (5)$$

where $T_o(x,y)$ is the relative amplitude, and $\psi(x,y)$ is the relative phase of the light emanating from the aperture. We can refer to the object as having spatially modulated the incident light.

As has been shown by Abbe (4), Born (5), Rhodes (9) and Cutrona (16), the electric field distribution in the focal plane of a lens can be (approximately) expressed by

---

* $$u(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases} .$$

* As first predicted by Brillouin (23) in 1921, and experimentally verified by Debye and Sears (U.S.A.), and Lucas and Biquard (24) in 1932, this phenomena can be stated as follows: If a medium (such as a liquid) is traversed by a compressional wave of short wavelength, then, when the wave is irradiated by light, a diffraction phenomena occurs similar to that of a diffraction grating.

$$E(\xi,\eta) \cong K_1 \iint_A \left\{E_a(x,y)\right\} \cdot$$

$$\left\{\exp\left[-j\frac{2\pi}{\lambda_L}\left(\frac{\xi}{S}x + \frac{\eta}{S}y\right)\right]\right\}dxdy$$

where ξ and η are space variables in the focal plane, K_1 is a complex constant, λ_L is the wavelength of the incident plane wave, and S is a slant distance as shown in Fig. 1. For small diffraction angles θ and ϕ, $S \simeq F$. If we now let

$$f_x \equiv \xi/F\lambda_L\ ,\quad f_y \equiv \eta/F\lambda_L\ , \tag{6a}$$

and use Eq. 4, we can write the electric field distribution in the focal plane as

$$E(f_x,f_y) \cong K_2 \iint_A \left\{T(x,y)\right\} \cdot$$

$$\left\{\exp\left[-j2\pi\left(f_x x + f_y y\right)\right]\right\}dxdy\ . \tag{6b}$$

Thus, within the constant factor K_2, we see that the distribution of light in the focal plane $(\xi\ ;\ \eta)$ is given by the Fourier Transform of the transmission function of the object. Also note, that since photodetectors respond to light intensity rather than light amplitude, we can obtain an (electrical) output signal $V(f_x,f_y)$ which is given by

$$V(f_x,f_y) = K_3\,|\,E(f_x,f_y)\,|^2\ . \tag{6c}$$

Restricting our attention to a one-dimensional aperture and object as shown in Fig. 2, we can write

$$E(f_x) = K_4 \int_{-p/2}^{+p/2} T(x)\exp[-j2\pi f_x x]dx\ . \tag{7}$$

If $T(x)$ is considered the analog of an input time waveform $v(t)$, then $E(f_x)$ forms its instantaneous amplitude spectrum (Eq. (6b)) and $V(f_x)$ forms its instantaneous power spectrum (Eq. (6c)). The variable f_x is entirely analogous to the frequency variable of the spectrum of the input signal $v(t)$.

As an example, consider an input signal of the form $v(t) = \cos(2\pi f_s t)$ which has a duration of T seconds from $t = 0$ to $t = T$. Let this signal be recorded on film so as to provide a transparency function $T(x) = 1/2[1 + \cos(2\pi x/\lambda_s)]$ over a film length of p meters. The "written" wavelength on the film is, therefore, $\lambda_s = p/Tf_s$. From Eq. (7) we obtain the (normalized) output amplitude

$$A(f_x) \equiv [E(f_x)]\,\frac{2}{Kp} \tag{8}$$

$$= A_o + A_{+1} + A_{-1}$$

where $A_o = \sin(\pi p f_x)/(\pi p f_x)$

$$A_{\pm 1} = \frac{1}{2}\sin(\pi p f_x')/(\pi p f_x')$$

and $f_x' \equiv f_x \pm 1/\lambda_s$.

This result is shown in Fig. 3, and it is, as expected, the spectrum for an amplitude modulated signal. Near $f_x = 0$, we observe $A_o(f_x)$ which is the response due to the bias level in $T(x)$. This response is called the <u>zero order fringe</u>, the 3 db width of its main lobe being approximately $1/p$. A <u>positive first order fringe</u>, A_{+1}, appears at $f_x = 1/\lambda_s$ and a <u>negative first order fringe</u>, A_{-1}, appears at $f_x = -1/\lambda_s$.

Remembering that $\lambda_s = p/Tf_s$ and that $f_x = \xi/F\lambda_L$, we observe that the positive first order fringe location (ξ_{+1}) is proportional to the input frequency, i.e.

$$\xi_{+1} = |\,\xi_{-1}\,| = F\lambda_L/\lambda_s = [F\lambda_L T/p]f_s\ . \tag{9}$$

Typically, a lens could be used such that its focal length $F = 20\ p$, where p is the aperture length. For $\lambda_L = 5000$ Å $= 1/2\ \mu$, a signal duration of 100 μsec, and a signal frequency of 20 mc, we would obtain a displacement ξ_{+1} of 2 cm. The 3 db width would occupy a length along ξ of $\Delta\xi \cong F\lambda_L/p = 10\ \mu$. This fine detail can be detected either by a scanning slit and a photomultiplier in the focal plane, by a mosaic of photodetectors as would be obtained from a television camera tube (see Refs. 30 and 31), or by photographic film (provided that sufficient light is available).

IV. <u>Debye-Sears Spatial Light Modulator</u>

The optical spectrum analyzer example that was used above, employed film as the modulating medium. By utilizing the Debye-Sears effect in a transparent, ultrasonic delay line, we can also accomplish a spatial modulation function instantaneously. As a simplified example, consider the configuration shown in Fig. 4. The input signal, $v(t)$, excites the piezoelectric

transducer which has been cut to vibrate in a longitudinal (compressional) mode. These mechanical vibrations cause a pressure wave to propagate in the delay medium with a characteristic speed s. For an aperture length of p, the signal duration which can just fill the aperture is $T = p/s$, corresponding to a time interval from $(t - T)$ to t.

Now, consider the plane electromagnetic wave which is incident upon the delay line as shown in Fig. 4. To a first approximation, a single electromagnetic "ray" experiences a uniform refractive index as it propagates through the delay medium but adjacent rays experience slightly different indices. For a piezoelectric crystal of length L (along z), the phase of the electromagnetic wave can be shown to change by an amount

$$\psi(x) = \frac{2\pi L}{\lambda_L} n(x) \quad , \tag{10}$$

where λ_L is the wavelength of the electromagnetic wave and $n(x)$ is the index of refraction of the delay medium as a function of x (the aperture dimension).

As an example, if the input signal is $v(t) = v_m \cos(2\pi f_s t)$, the light wave will be phase modulated with a relative phase given by

$$\psi(x) = \psi_o + \psi_m \cos(2\pi f_s x/s + \psi_i) \tag{11}$$

where $\psi_o = 2\pi n_o L/\lambda_L$, $\psi_m = 2\pi n_m L/\lambda_L$, and ψ_i is some initial phase related to t, the time chosen for the examination of the wavefront. The index of refraction n_o is the average (or unexcited value) for the delay medium, and n_m is the peak value. It can also be shown that the change in refractive index is proportional to the input voltage so that the peak phase deviation (modulation index) is given by $\psi_m = c_1 v_m$ where c_1 is a constant. Thus, to a first approximation, the transparent delay line produces a transfer (transparency) function which contains only phase modulation, i.e., $T_o(x) = 1$ and $T(x) = \exp[j\psi(x)]$.

The "first order" theory that is presented above is based on the assumption that the incident electromagnetic wave behaves as simple "rays" which maintain a straight-line path. Actually, bending (and reflection) does occur, the true problem being that of obtaining a solution to Maxwell's equations for an electromagnetic wave incident upon a medium which contains a variable index of refraction. Approximate solutions have been obtained by Raman and Nath (32), Mertens (33), Bhatia and Noble (34), Phariseau (35) and others as is summarized in the text by Born and Wolf (5, Chapter 12). These theoretical results and their experimental verifications could be employed to determine the constraints on the first order theory presented above. Such an analysis will not be attempted here.

V. An Electro-Optical Spectrum Analyzer

The spatial phase modulation produced by a transparent, ultrasonic delay line (Fig. 4) can be combined into the optical configuration shown in Fig. 2 to produce a distribution of light amplitude in the focal plane as shown in Fig. 5.

Each fringe has the same shape (and the same resolution) as obtained from the amplitude modulated film (Fig. 3) except all harmonic orders are present as occurs for the spectrum of a phase (or frequency) modulated signal. The nth order fringe has an amplitude $J_n(\psi_m)$ where J_n is the nth order Bessel function and this fringe is located at $f_x \equiv f_{\pm n} = \pm n/\lambda_s$ (positive and negative orders). Now, $\lambda_s = p/Tf_s = s/f_s$, giving $f_{\pm n} = \pm (n/s) f_s$ so that the position of the first order fringe is given by

$$\xi_{+1} = [F\lambda_L/s][f_s] \quad . \tag{12}$$

Thus, the location of this fringe is proportional to input frequency for a system having a constant focal length (F), a constant light wavelength (λ_L) and a constant sonic speed (s).

If the input frequency (f_s) is changed by an amount Δf_s, the position of the first order fringe will shift by an amount $\Delta\xi_{+1} = [F\lambda_L/s][\Delta f_s]$. For this fringe to move by an amount equal to its 3 db width $\Delta\xi_{3db} \simeq F\lambda_L/p$. Thus, the input frequency change necessary to move the fringe one 3 db width ($\Delta\xi_{+1} = \Delta\xi_{3db}$) is given by

$$\Delta f_s \simeq 1/T \tag{13a}$$

where T is the integration or delay time and Δf_s is the frequency resolution. Also, for a delay line having a bandpass of B cps, the number of resolution elements obtained is

$$N = B/\Delta f_s \simeq BT \quad . \tag{13b}$$

So far, we have considered time as being fixed and we assumed that the signal of interest filled the aperture. Actually, the Debye-Sears effect is obtained from a traveling pressure wave so that new sections of the waveform are continuously being processed. For an aperture which extends from $x = -p/2$ to $x = +p/2$ and for a wave

propagating in the positive x direction, Eq. (11) becomes a traveling (phase) wave given, typically, by

$$\psi(x,t) = \psi(t - x/s - p/2s) \tag{14}$$
$$= \psi_o + \psi_m \cos[2\pi f_s(t - x/s - p/2s)],$$

where t is now running time, s is the speed of propagation of the ultrasonic wave, and f_s is the input signal frequency. The spectrum (light amplitude distribution in the focal plane, Eq. (7)) is now

$$E(f_x,t) = K_4 \int_{-p/2}^{p/2} [T(t - x/s - p/2s)]$$
$$\times [\exp[-j2\pi f_x x]dx] \tag{15a}$$

where the transmission function of the light modulator is

$$T(t - x/s - p/2s) = \exp[j\psi(t - x/s - p/2s)] \ . \tag{15b}$$

Equations (15) express the physical fact that the input signal continuously moves past the aperture of length p. Thus, as time t progresses, new sections of the input waveform are being integrated and the instantaneous spectrum is continuously obtained. To see this more clearly, let $\tau = t - (x + p/2)/s$, then Eq. (15) can be shown to be given by

$$E(f,\tau) = K' \int_{t-T}^{t} T(\tau)\exp[-j2\pi f\tau]d\tau \tag{16}$$

where K' is a new constant and $f \equiv - f_x s$. This is the instantaneous spectrum that was defined previously.

Note that a weighting function $W(\tau-t)$ may be introduced into Eq. (16) by introducing a second object or shading mask $W(- x + p/2)$ in the object plane. This could provide a net transmission function $T_W(x,t)$ given by

$$T_W(x,t) = W(- x+ p/2)T(t - x/s - p/2s) \ .$$

The weighting function is fixed in space, automatically performing the desired weighting. This method can be employed to modify the transfer function of the spectrum analyzer. An arbitrary, time-invariant filter can be realized in this manner.

VI. Spatial Multiplexing

An optical system has two degrees of freedom (the x and y dimensions). So far, we have utilized one-dimension (x) to process the input signal, the second dimension (y) being used only to allow a finite amount of electromagnetic energy to be processed by the optical system. As has been shown by Cheatam (11), O'Neill (13), Cutrona (16) and others, the second dimension can be used to process a second variable or to simultaneously process multiple channels. The optical configuration for the latter function is shown schematically in Fig. 6.

The object plane now consists of separate channels, each channel containing the same "time-slice" of the input signal but different portions of its frequency content as indicated in Fig. 7. Letting each channel cover a frequency band $B = \Delta F_1 = \Delta F_2 =$ etc. the total frequency coverage for M channels becomes $B_T = MB$. A cylindrical lens which integrates (focuses) only in the x dimension is employed to obtain the Fourier transform of the aperture (object) function along the ξ axis. The focal plane in Fig. 6 shows a typical fringe distribution which would occur if the input signal contained the sum of the frequencies f_{o1}, f_{o2}, f_{o3} and f_{o4} as shown in Fig. 7. Note that since the y and η dimensions are only used to preserve channel separation, complete coherence of the light in this direction is not required.

VII. Typical Parameters and Conclusions

A. Frequency Coverage

In order to avoid detecting the ambiguities introduced by the higher order fringes (above the first order, see Fig. 5), the input frequency band must be limited so that the highest frequency is no more than twice the lowest frequency. This allows a maximum bandwidth which is 67 per cent of the center frequency of the delay line light modulator. In practice, fractional bandwidths between 50 per cent and 65 per cent can be achieved for both liquid and solid delay media (see aforementioned companion paper).

B. Frequency Resolution

The frequency resolution Δf_s is minimized when the integration or delay time T is maximized ($\Delta f_s \simeq 1/T$). To maximize the delay time, the aperture length p must be maximized and the sonic speed of propagation s must be minimized ($T = p/s$). In general, liquids provide lower sonic speeds than solids but liquids absorb more

ultrasonic energy per unit length compared to solids. The essential effect of ultrasonic attenuation in the delay medium is to cause an exponential weighting function to appear superimposed on the phase modulation at the output of light modulator. This monotonic weighting of the phase modulation results in a broadening of the fringe in the focal plane, i.e., the resolution width increases so that $\Delta f_s > 1/T$. It can be shown that if the attenuation constant for the delay medium is α and the aperture length is p, then for $\alpha p \leq 2.5$, the fringe broadening will be less than 20 per cent ($\Delta f_s \leq 1.2/T$).

For many liquids, the acoustic attenuation α is proportional to the square of the input frequency, so that

$$\alpha = af_s^2 \tag{17}$$

where a is a constant for a given medium and f_s is the input frequency. If the maximum allowable fringe-width broadening is specified, then

$$\alpha p \leq b \tag{18}$$

where b is a design constant. Combining Eqs. (17) and (18) and using $p = sT$, we obtain

$$T \leq (b)\left(\frac{1}{as}\right)\left(\frac{1}{f_s^2}\right) \quad . \tag{19}$$

Thus, by reducing the maximum input frequency (therefore reducing the frequency coverage) the integration time can be increased, resulting in a decrease in the frequency resolution-width. The lenses in the optical system do, however, place an upper limit on the maximum aperture size p that can be implemented. Presently, for light wavelengths of the order of 5000 Å, "diffraction limited" lenses having diameters up to 15 cm are readily obtainable. For water as the delay medium (s = 1500 m/s), a 15 cm lens would allow an integration time of 100 µsec and permit a highest input frequency near 25 mc with a broadening of less than 20 per cent (a resolution width less than 12 kc). Liquids having sonic speeds below that in water are available but they usually exhibit much higher attenuation constants.

C. <u>Number of Resolution Elements</u>

As was previously shown, the maximum number of resolution elements or equivalent filter channels (N) that can be obtained is approximately equal to the time-bandwidth product (BT) of the delay line light modulator. From Eq. (19) we can obtain the result

$$N \equiv BT \leq \left(\frac{B}{f_s}\right)\left(\frac{1}{s}\right)\sqrt{\frac{b}{a}\cdot p} \quad . \tag{20}$$

For a given delay medium, both a and s are constants; for a specified, maximum allowable fringe broadening, b is a constant; the ratio B/f_s is proportional to the maximum fractional bandwidth which cannot exceed 67 per cent. Thus, the maximum number of resolution elements N increases only as the square root of the increase in the aperture p . Also, for a given aperture size and delay medium, there is an optimum bandwidth and hence an optimum delay line center frequency which maximizes the number of resolution elements.

Typically, for a water delay line with a 15 centimeter aperture, a 10 mc bandwidth could be maintained without significantly degrading the frequency resolution of 10 kc. This provides 1000 resolution elements. Spatial multiplexing can then be employed to increase the number of resolution elements. For example, ten ultrasonic beams in a 15 cm water delay medium would provide 10,000 resolution elements every 100 µsec across a 100 mc band. Additional filtering and heterodyning would be required in order to maintain the maximum time-bandwidth product in each ultrasonic channel. Such a scheme is shown schematically in Fig. 8 (also see Fig. 7).

Light modulators which employ quartz as the delay medium are not limited by acoustic attenuation over a 15 cm length. They are presently limited by the absolute bandwidth that can be achieved with a bonded transducer and by the electrical drive problem which exists at high frequencies (see Refs. 36 and 37). Typically, a quartz delay line would provide 40 mc of frequency coverage with a useable delay of 25 µsec in a 15 cm aperture. Again, 1000 resolution elements would be obtained, the frequency resolution would be 40 kc across a 40 mc band. Spatial multiplexing could also be employed to increase the number of resolution elements.

D. <u>Dynamic Range</u>

In addition to the problems of maximizing the frequency coverage, the integration time, and the number of resolution elements, we must concern ourselves with the problem of maximizing the dynamic range of these spectrum analyzers. The dynamic range can be defined as the ratio of the maximum output signal to the peak output spurious response level. Major spurious responses may appear at the output due to such causes as intermodulation distortion and photodetector noise.

Referring back to Fig. 5, we note that for a single frequency input signal to a delay line light modulator, the fringes are located as harmonics of the input signal, as would occur for the spectrum of any signal which is sinusoidally phase modulated with a peak deviation of ψ_m radians. Also, the relative amplitudes vary as Bessel functions J_o, J_1, J_2, etc., corresponding to the amplitude of the zero order fringe, first order fringe, second order fringe, etc. For a peak phase deviation less than a radian, these Bessel functions can be approximated by

$$J_n(\psi_m) \simeq \psi_m^{\ n}/2^n n! \qquad (21)$$

so that $J_o \simeq 1$, $J_1 \simeq \frac{\psi_m}{2}$, $J_2 \simeq \frac{\psi_m^2}{8}$, etc.

The modulation index ψ_m must be maintained small in order to avoid the introduction of large spurious (intermodulation) signals. It can be shown that if the peak modulation index is maintained below 0.2 radians then the intermodulation fringes would be at least 40 db below the first order fringe. However, if $\psi_m \simeq 0.2$ radians, then the first order fringe would have a maximum amplitude of $J_1 \simeq 0.1$, relative to the zero order fringe. Thus, less than 1 per cent of the collimated light which is incident upon the light modulator is actually detected. In addition, in order to obtain the required monochromatic, collimated light from ordinary noncoherent light sources, narrow slits, collimating optics, and narrow band color filters must be employed. This drastically reduces the light power that is available in the collimated region. As a result of these large losses, the number of photons that are available for detection in a first order fringe is quite small and photodetector noise (see Refs. 38, 39, 40) can be shown to limit the dynamic range for spectrum analyzers having a large time-bandwidth product.

The optical maser is a device which produces essentially monochromatic, collimated light at a power level which is several orders-of-magnitude greater than that obtainable from other (classical) light sources. A ruby optical maser has been successfully employed as a pulsed light source for an electro-optical (delay line) spectrum analyzer.* However, the data rate was limited by the repetition rate of the optical maser rather than by the integration time. It can be shown, that an optical maser which continuously provides an output somewhat less than one watt in the visible region would allow a dynamic range of 40 db in a 15 cm aperture system. A properly designed spatially multiplexed water delay line light modulator would provide 10,000 resolution elements covering a 100 mc band every 100 μsec for this 15 cm aperture.

* Electronics Research Laboratories, Columbia University.

VIII. Acknowledgment

This research is sponsored by the Advanced Research Projects Agency of the Department of Defense and is currently administered by the Air Force Office of Scientific Research under Contract AF 49(638)-1113. This work was initially directed by the Rome Air Development Center, Air Force Systems Command, USAF under Contract AF 30(602)-1971.

IX. References

1. Kharkevich, A. A., "Spectra and Analysis," (Translated from Russian), Consultants Bureau, New York, 1960.

2. Fano, R. M., "Short Time Autocorrelation Functions and Power Spectra," J. Acoust. Soc., Vol. 22, No. 5, pp. 546-550, 1950.

3. Page, C. H., "Instantaneous Power Spectra," J. Appl. Phys., Vol. 23, No. 1, pp. 103-106, 1952.

4. Abbe, Ernst, Archiv. f. Mikroskopische Anat., Vol. 9, p. 413, 1873.

5. Born, M. and Wolf. E., "Principles of Optics," Pergammon Press, 1959 (Text).

6. Wolf, E., Progress in Optics, Vol. I, North-Holland Pub. Co. - Amsterdam and Interscience Pub. Inc., New York, 1961, A Collection of Review Articles.

7. Elias, P., Grey, D., Robinson, D., "Fourier Treatment of Optical Processes," J. Opt. Soc. Amer., Vol. 42, pp. 127-134, February 1952.

8. Elias, P., "Optics and Communication Theory," J. Opt. Soc. Amer., Vol. 43, pp. 229-232, 1953.

9. Rhodes, J. E. Jr., "Analysis and Synthesis of Optical Images," Amer. J. Physics, Vol. 21, pp. 337-431, May 1953.

10. Rhodes, J. E. Jr., "Microscope Imagery as Carrier Communication," J. Opt. Soc. Amer., Vol. 43, No. 10, October 1953.

11. Cheatham, T. P. and Kohlenberg, A., "Optical Filters - Their Equivalence to and Differences from Electrical Networks," Conv. Rec. IRE Natl. Conv., 1954.

12. O'Neill, E. L., "The Analysis and Synthesis of Linear Coherent and Incoherent Optical Systems," Tech. Note 122, Boston Univ. Optical Res. Labs., September 1955.

13. O'Neill, E. L., "Spatial Filtering in Optics," IRE Trans. Info. Theory, Vol. IT-2, pp. 56-65, June 1956.

14. O'Neill, E. L., "Selected Topics in Optics and Communications," ITEK Corp., September 1958.

15. Cowley, J. M. and Moodie, A. F., "Fourier Images," Parts I-IV, Proc. Phys. Soc. (Grt. Britain), B, 70, 1957.

16. Cutrona, L. J., Leith, E. N., Palermo, C. J. and Porcello, L. J., "Optical Data Processing and Filtering Systems," IRE Trans. Info. Theory, Vol. IT-6, No. 3, pp. 386-400, 1960.

17. Gamo, H., "An Aspect of Information Theory in Optics," IBM Res. Center Report RC-188, February 15, 1960.

18. Howell, B. J., "Optical Analog Computers," Ph.D. Thesis, Univ. of Utah, 1954.

19. Howell, B. J., "Optical Analog Computers," J. Opt. Soc. Amer., Vol. 49, No. 10, October 1959.

20. Wilmotte, R. M., "Ultrasonic Spectrum Analyzer," Final Report: Summary and Recommendations," Report 14, Raymond M. Wilmotte, Inc., AD 88698, December 20, 1955, Unclassified.

21. "Ultrasonic Spectrum Analyzer," Report 202-1, Airborne Inst. Lab., Inc., Mineola, New York, April 1950.

22. Debye, P. and Sears, F. W., "On the Scattering of Light by Supersonic Waves," Proc. Natl. Acad. Sci., Vol. 18, No. 6, June 15, 1932, Washington, D. C.

23. Brillouin, L., Ann. de Physique, Vol. 17, p. 103, 1921.

24. Lucas, R. and Biquard, P., "Nouvelles Proprietes Optiques des Liquides Soumis a des Ondes Ultra-sonores," (New Optical Properties of Liquids Excited by Ultrasonic Waves), in French, Comptes Rendus - Acad. Sci. Paris, 194, 1932.

25. Robinson, D. M., "The Supersonic Light Control and Its Application to Television with Special Reference to the Scophony Television Receiver," Proc. of IRE, August 1939.

26. Terry, N. B., Mumford, H. and Holloway, D. G., "Novel Uses of the Ultrasonic Cell," Automatic Telephone and Electric Co. J. (Grt. Britain), Vol. 11, No. 1, pp. 2-16, January 1955.

27. Levi, L., "High-Fidelity Video Recording Using Ultrasonic Light Modulation," J. SMPTE, Vol. 67, pp. 657-661, October 1958.

28. Lambert, L., Arm, M. and Weissman, I., "Development of a Two-Dimensional Filter," Technical Report P-2/153, Electronics Research Laboratories, School of Engineering and Applied Science, Columbia University, July 1959.

29. Rosenthal, A. H., "Application of Ultrasonic Light Modulation to Signal Recording, Display, Analysis and Communication," IRE Trans. Ultrasonic Eng., Vol UE-8, No. 1, pp. 1-5, March 1961.

30. McGee, J. D. and Wilcock, W. L., "Photo-Electronic Image Devices, Advances in Electronics and Electron Physics, Academic Press, New York, Vol. 12, 1960.

31. Marton, L., Advances in Electronics and Electron Physics, Academic Press, New York, Vol. 13, 1960.

32. Raman, C. V. and Nath, N. S., "The Diffraction of Light by High Frequency Sound Waves," Proc. Ind. Acad. Sci., Part I: Normal Incidence, Sec. A, Vol. 2, p. 406, 1935; Part II: Oblique Incidence, Sec. A, Vol. 2, p. 413, 1935; Part III: Time Variations, Sec. A, Vol. 3, p. 75, 1936; Part IV: General Periodic Wave, Sec. A, Vol. 3, p. 119, 1936; Part V: Oblique Incidence, Sec. A, Vol. 3, p. 459, 1936; Generalized Theory, Sec. A, Vol. 4, p. 222, 1936.

33. Mertens, R., "On the Theory of the Diffraction of Light by Supersonic Waves," Simon Steven, Vol. 27, p. 212-230, 1951.

34. Bhatia, A. B. and Noble, W. J., "Diffraction of Light by Ultrasonic Waves," I: General Theory; Proc. Royal Soc., A 200, pp. 356-368, December 8, 1953; II: Approximate Expressions for the Intensities and Comparison with Experiment, A 200, pp. 369-385, December 8, 1953.

35. Phariseau, P., "The Diffraction of Light by Ultrasonic Waves," Simon Steven, Vol. 33, No. 2, pp. 72-78, June 1959.

36. Konig, W., Lambert, L. and Schilling, D., "The Effects of Bonding and Backing Materials on the Characteristics of Ultrasonic Delay Lines," Wescon Conv. Rec., August 1961.

37. Konig, W. and Lambert, L., "Research on Ultrasonic Delay Lines," Research Progress Reports P-1/152, P-2/152, P-3/152, P-4/152, P-5/152, Electronics Research Laboratories, School of Engineering and Applied Science, Columbia University, 1959 through 1961.

38. Braddick, H. J., "Photoelectric Photometry," Phys. Soc., Vol. 23, p. 154, 1960.

39. Engstrom, R. W., "Absolute Spectral Response Characteristics of Photosensitive Devices," RCA Review, Vol. 21, No. 2, pp. 184-190, June 1960.

40. McGee, J. D. and Wilcock, W. L., "Photo-Electronic Image Devices," Advances in Electronics and Electron Physics, Academic Press, New York, Vol. 12, 1960.

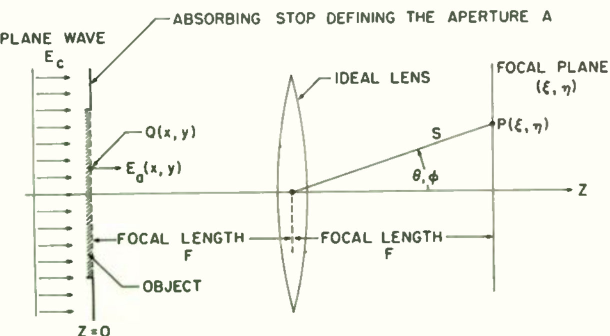


Fig. 1. Fraunhofer diffraction.

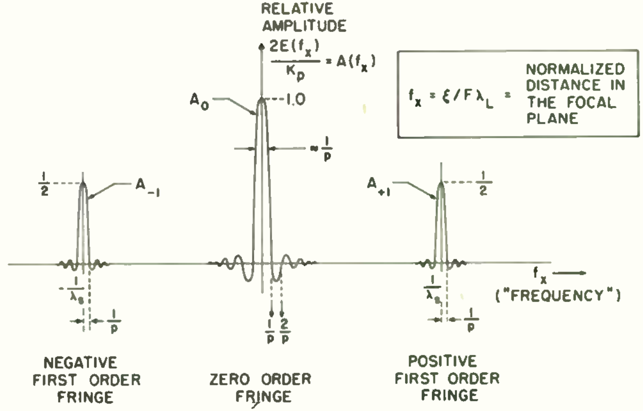


Fig. 3. Relative amplitude distribution in the focal plane due to sinusoidal amplitude modulation.

$$T(x) = \frac{1}{2}(1 + \cos 2\pi x/\lambda_s) \quad -P/2 < x < +P/2.$$

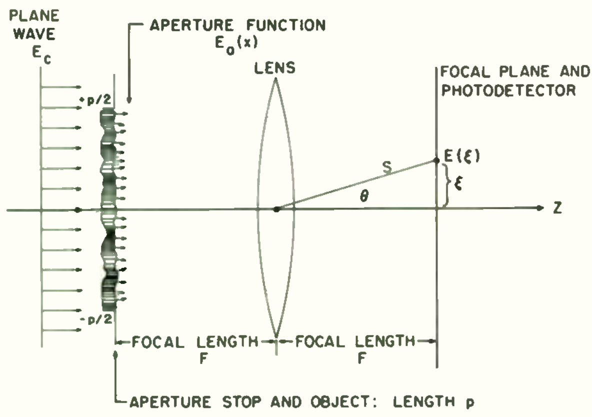


Fig. 2. Obtaining the spectrum of a one-dimensional object.

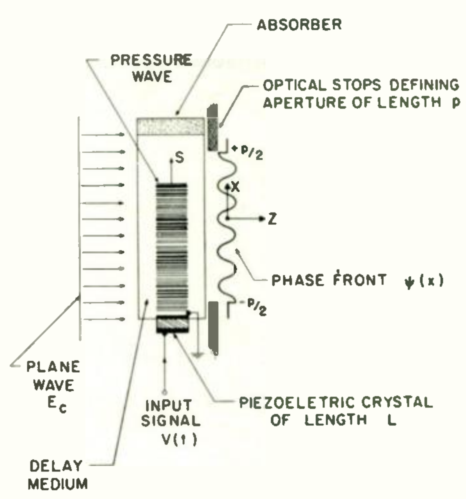


Fig. 4. A Debye-Sears, spatial light modulator.

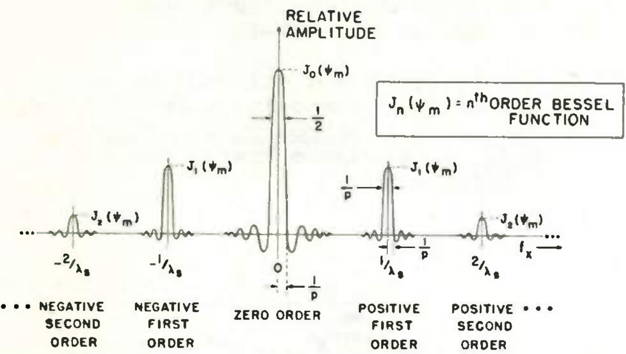


Fig. 5. Relative distribution in the focal plane due to sinusoidal phase modulation.

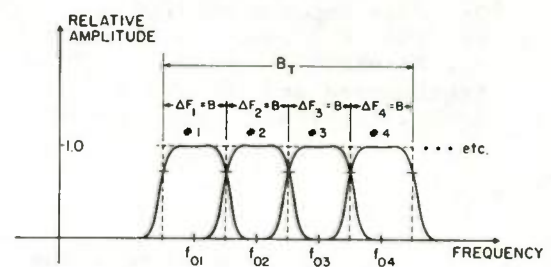


Fig. 7. Channel frequency coverage and total frequency coverage.

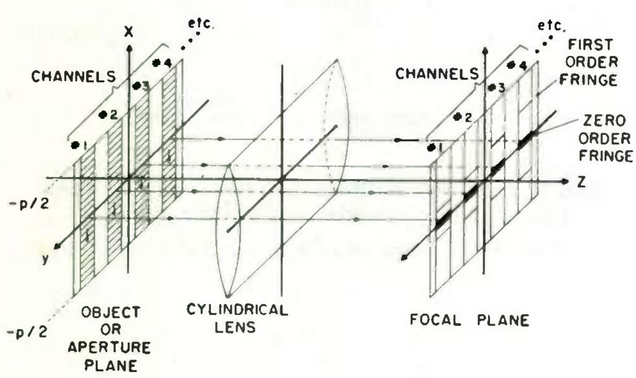


Fig. 6. Spatially multiplexed spectrum analyzer.

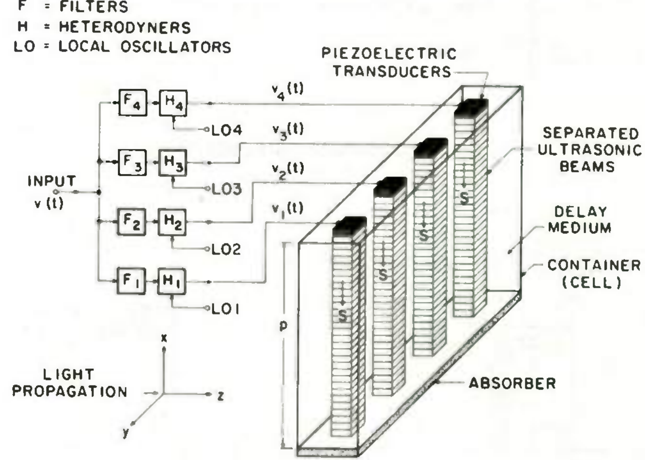


Fig. 8. Spatially multiplexed light modulator maintaining a maximum BT.

ELECTRO-OPTICAL TRANSFER CHARACTERISTICS OF LIQUID DELAY-LINE LIGHT MODULATORS

Moses Arm, Louis B. Lambert, and Barbara Silverberg

Columbia University, Electronics Research Laboratories
of the School of Engineering and Applied Science
New York 27, New York

Abstract

The Debye-Sears effect in a water delay line is utilized to obtain spatial light modulation in real time. The design, implementation, and measurement techniques which have been employed to optimize the electro-optical transfer characteristic of this light modulator are presented.

Insertion loss, bandwidth and phase response are theoretically determined for the case of the back-loaded transducer and the case of the quarter-wave matching section. These results are then compared with corresponding experimental results. It is shown that a low insertion loss can be obtained with a 50 percent bandwidth by employing a quarter-wave matching section technique.

Techniques which have been successful in producing quarter-wavelength matched transducers are described, and experimental results are presented.

I. Introduction

The synthesis of many different and useful data-processing devices may be realized by employing electro-optical techniques which rely upon the spatial modulation of light in a coherent optical configuration. The principle of operation of the light modulator to be described briefly below is based on the Debye-Sears[1] effect. A piezoelectric transducer is placed in contact with a transparent medium chosen to possess certain characteristics. When the transducer is electrically excited, a longitudinal pressure wave traverses the medium, causing it to become stratified into parallel layers of varying refractive index. A collimated beam of monochromatic light, passing through the medium perpendicular to the direction of propagation of the pressure wave, emerges from the medium spatially phase modulated in accordance with the excitation applied to the transducer. This spatially phase modulated light may be considered as a superposition of plane waves travelling in different directions. In the case of a sinusoidal excitation signal, the far-field light pattern consists of sharp fringes whose angular displacement "θ" relative to the direction of the incident light is given by[2]:

$$\sin \theta = \pm \frac{n\lambda_L}{\lambda_S} \qquad n(\text{integer}) \geq 0 \tag{1}$$

where λ_L = light wavelength

$\lambda_S = \frac{v_S}{f_o}$ = ultrasonic wavelength in medium

v_S = ultrasonic velocity of propagation in medium

f_o = frequency of excitation signal.

The relative intensity of the fringes is given by:

$$\frac{I_n}{I_o} = \frac{J_n^2(\psi_m)}{J_o^2(\psi_m)} \tag{2}$$

where J_n = Bessel function of the nth order

$\psi_m = \frac{2\pi q \mu_m}{\lambda_L}$ = peak light phase deviation

q = thickness of ultrasonic beam

μ_m = peak deviation of refractive index.

Most applications of the light modulator make use of only one of the first-order fringes, since the zero-order fringe contains no information and the remaining fringes contain only redundant information. To avoid spurious intermodulation products, it is necessary to maintain linear operation in the light modulator. This requires that the peak phase deviation of the signal

be limited to a value of less than one radian. The relative intensity of the first-order fringe may be obtained from Eq. (2) by expanding the Bessel function and ignoring the higher order terms. The resultant equation

$$\frac{I_1}{I_o} \approx \frac{\psi^2}{4} \tag{3}$$

relates the available light to the peak phase deviation. We will next consider the relationship between the peak phase deviation and the input power to the transducer.

II. Drive Requirements

In deriving the relationship between the peak phase deviation and the power delivered to the ultrasonic medium, certain assumptions must be made. First, it is assumed that the piezoelectric transducer is in intimate contact with the medium. Second, it is assumed that the transducer is sufficiently large (approximately 50 wavelengths in diameter) to insure that the spreading of the ultrasonic beam due to diffraction is negligible. Third, it is assumed that the ultrasonic medium is a lossless liquid and has a perfect acoustic termination (there are no reflections). Under these conditions, the power delivered to the medium is given by:

$$W = \left[\frac{3\mu_o\lambda_L}{(\mu_o^2 - 1)(\mu_o^2 + 2)q}\right]^2 \left[\frac{\rho_o v_S^3 A_t}{2}\right] \left[\frac{\psi_m}{\pi}\right]^2 \tag{4}$$

where μ_o = static refractive index of medium

ρ_o = static density of medium

A_t = effective area of transducer.

The derivation of Eq. (4) is presented in the Appendix. Let us assume a typical situation, where the liquid medium is water and the transducer is x-cut quartz; then the parameters are

$$\mu_o = 1.33$$

$$\rho_o = 997.2 \text{ Kg/m}^3$$

$$\lambda_L = 5461 \times 10^{-10}\text{m}$$

$$v_S = 1500 \text{ m/sec}$$

$$A_t = 10^{-4}\text{m}$$

$$q = 10^{-2}\text{m},$$

and the power delivered to the transducer is

$$W = 0.0962\ \psi_m^2 \text{ watts}.$$

This is the power which is actually delivered to the ultrasonic medium. It should be emphasized that the required power capability of the driver must generally be much larger, since some of the output power is dissipated before it reaches the medium. In the case of a backed transducer, power is dissipated in the backing material; in the case of a transducer with a matching section, there is no power dissipation in in the transducer. However, in both cases power must be dissipated in the driver in order to achieve a broad frequency band. These considerations will be treated in more detail in the next two sections.

III. Wide-Band Light Modulator

For a large number of data-processing applications, the light modulator must have a transfer characteristic which provides a large fractional bandwidth with low insertion loss, and in addition, a linear phase response. These characteristics are determined mainly by the transducer design. The transducer, being a resonant element, has an inherently narrow bandwidth. However, there are techniques whereby the bandwidth of a transducer can be broadened. Two of these techniques, to be discussed in detail herein, are: first, backing the transducer with a material having a relatively large acoustic impedance; and second, matching the transducer to the medium which it drives.

A. Theory of Backed Transducer

The electrical equivalent circuits of electromechanical transducers and the analysis of systems employing such transducers have been developed by Mason[3] and other investigators.[4] The transducer to be considered in this paper is an x-cut quartz crystal which vibrates in the thickness mode. Its thickness is equal to a half-wavelength at the resonant frequency. The equivalent electrical circuit, due to Mason, is given in Fig. 1.

In this analysis, it is assumed that the quartz crystal behaves as if it were an infinite slab, i.e., that plane waves are propagated from its faces, and the faces are electrical equipotentials. In practice, the infinite slab is well approximated by making the crystal faces at least 30 times larger than an ultrasonic wavelength. The parameters of the equivalent circuit are as follows:

$Z_x = \rho_x v_x A_t$ = mechanical impedance of the crystal

$\Phi = \frac{hA_t}{\beta t}$ = ideal transformer voltage ratio

$C_o = A_t/\beta t$ = static capacitance of the crystal

ρ_x = density of crystal

v_x = ultrasonic velocity in the crystal

h = piezoelectric stress constant

β = dielectric impermeability at constant strain

A_t = effective area of the crystal

$t = \frac{v_x}{2f_o}$ = crystal thickness

f_o = resonant frequency

$a = \frac{\pi f}{f_o}$ = fractional frequency deviation

F and F' = forces exerted by the crystal faces 1 and 2, respectively.

The MKS system of units is used throughout.

Consider now the case of a crystal transducer loaded on both faces; one face is loaded by the delay medium and the other by some backing material. It is assumed that the bonds to the loading media are ideal, and of zero thickness. Practically, this can be achieved by employing as loading media liquids which wet the transducer or materials which can be made to adhere. It is further assumed that the loading media are of infinite length or else properly terminated acoustically; then they can be represented by real mechanical impedances. Placing the mechanical impedance of the delay medium across one set of the mechanical terminals of the equivalent circuit of Fig. 1, and the mechanical impedance of the backing material across the other, the circuit can then be redrawn in the form of Fig. 2a. Making use of the circuit identity of Fig. 2b, the equivalent circuit of a crystal loaded on both faces becomes that shown in Fig. 2c. The parameters are:

$Z_q = \rho_o v_s A_t$ = mechanical impedance of the delay medium

$Z_b = \rho_b v_b A_t$ = mechanical impedance of the backing material

where ρ_o, v_s and ρ_b, v_b are the densities and ultrasonic velocities in the delay and backing media, respectively. Near the resonant frequency, the term $-j2Z_x \cot\frac{a}{2}$ approximates a series resonant circuit whose capacitance in the case of quartz can be shown to be smaller than $\frac{C_E}{4}$ by approximately a factor of 10^{15}; therefore, $\frac{C_E}{4}$ in the circuit is negligible. It is further assumed that the input voltage is independent of frequency, i.e., the frequency response of the driving circuit is not considered. Consequently, the equivalent circuit of Fig. 2c can be redrawn in the form shown in Fig. 3a, where F is the force exerted by the crystal face on the delay medium.

The circuit of Fig. 3a can be approximated near resonance by that shown in Fig. 3b, and in this form the circuit immediately indicates the role of the backing material. As the impedance (Z_b) of the backing material is increased, the series loading of the circuit increases; consequently, the bandwidth increases, but at the expense of a higher insertion loss. The frequency response of the circuit in Fig. 3a is given by:

$$\frac{F(f)}{F(f_o)} = \left\{ \frac{1 + \left(\frac{m}{n}\right)^2 \cot^2 \frac{a}{2}}{1 + \left[\frac{4}{n^2}\left(\frac{m+n^2}{1+m}\right)^2 - 2\right]\cot^2\frac{a}{2} + \cot^4\frac{a}{2}} \right\}^{\frac{1}{2}} \quad (5)$$

where $m = \frac{Z_b}{Z_q}$

$n = \frac{Z_x}{Z_q}$.

Eq. (5) is plotted in Fig. 4 for five different backing materials, with water as the delay medium in all cases. It is of some interest to note that for a backing material of high acoustic impedance, such as mercury, the bandpass characteristic displays a double peak.

The midband attenuation of the circuit shown in Fig. 3a is given by:

$$\frac{F}{2\Phi V} = \frac{1}{1+m} . \quad (6)$$

Normalizing with respect to the air-backed (unbacked) case ($Z_b = 0$),

$$\frac{F(Z_b)}{F(air)} = \frac{1}{1+m} . \quad (7)$$

An examination of the circuit in Fig. 2c shows the electrical input admittance of the transducer at resonance to be

$$Y_{t,in} = \frac{4\Phi^2}{Z_b + Z_q} + j2\pi f_o C_o \quad . \qquad (8)$$

The theoretical results indicate that by backing a transducer with a material of finite acoustic impedance, it is possible to achieve a fractional bandwidth of greater than 100 percent of center frequency (mercury backing produces a fractional bandwidth of 135 percent). This is achieved, however, only at the expense of insertion loss.

We will now proceed to examine the phase characteristics of the backed transducer. It is evident from the circuit shown in Fig. 3b that, for the case of backing materials which produce relatively narrow bandwidths, the phase response is essentially that of a simple tuned circuit. Making use of the approximate equivalent circuit of Fig. 3a the relative phase shift of the output across Z_q is found to be:

$$\theta = \text{arc tan}\left\{-\frac{m}{n}\cot\frac{a}{2}\right\} - \text{arc tan}\left\{\frac{2(m + n^2)\cot\frac{a}{2}}{n(1 + m)(\cot^2\frac{a}{2} - 1)}\right\} . \qquad (9)$$

The relative phase shift is plotted in Fig. 5 for various backing materials. The relative time delay of the output may be found by differentiating the phase shift with respect to frequency. This delay is plotted in Fig. 6 for various backing materials.

B. Theory of Matched Transducer

The matching technique consists of placing a quarter-wavelength plate of appropriate mechanical impedance between the transducer and delay medium. The quarter-wavelength plate, acting as a mechanical impedance transformer, causes a larger load to be reflected to the transducer thus producing a broad bandwidth as in the case of backing. Now, however, instead of most of the power being absorbed by the backing material and not utilized, it is transmitted to the delay medium through the matching section, resulting in a considerable reduction of the insertion loss.

Making the same assumptions as were made previously in Sec. III-A, the equivalent circuit of a matched transducer is shown in Fig. 7. The matching section is analogous to a lossless quarter-wavelength transmission line. The input impedance to such a line, terminated in an impedance Z_q , is given by:

$$Z_{in} = Z_q \frac{1 + \tan\frac{a}{2}}{1 + k^2\tan^2\frac{a}{2}} + jZ_q \frac{\left(\frac{1}{k} - k\right)\tan\frac{a}{2}}{1 + k^2\tan^2\frac{a}{2}} , \qquad (10)$$

and the transfer function is given by:

$$\frac{F_o}{F_1} = \frac{1}{\cos\frac{a}{2} + j\frac{1}{k}\sin\frac{a}{2}} \qquad (11)$$

where $k = \frac{Z_q}{Z_o}$

Z_o = mechanical impedance of matching section.

The assumption that the line is lossless is reasonable since in a quarter-wave section at 20 mc there is less than 0.5 db loss in plastics and much less in metals.

The transfer function of the electrical network shown in Fig. 7 is given by:

$$\frac{F_o}{2\Phi V} = \frac{A + jB}{E + jD}\left[\frac{1}{\cos\frac{a}{2} + j\frac{1}{k}\sin\frac{a}{2}}\right] , \qquad (12)$$

where $A = n(k - \frac{1}{k})C$

$B = n(C^2 + 1)$

$E = n(\frac{1}{k} - k + 2n)C^3 + n(k - \frac{1}{k} + 2nk^2)C$

$D = n(1 - C^4)$

in which $C = \cot\frac{a}{2}$. At midband ($f = f_o$) the transfer function reduces to

$$\left|\frac{F_o}{2\Phi V}\right|_{f = f_o} = k \qquad (13)$$

which is equivalent to the voltage attenuation of the matching section terminated by the impedance Z_q .

Normalizing with respect to the midband response, the frequency response is given by:

$$\frac{F_o(f)}{F_o(f_o)} = \left\{ \frac{(A^2 + B^2)(C^2 + 1)}{k(E^2 + D^2)(C^2 + \frac{1}{k^2})} \right\}^{\frac{1}{2}} . \quad (14)$$

Equation (14) is plotted in Fig. 8, with k as the parameter. From these curves it can be seen that the maximally flat response is obtained for $k = 0.316$. This corresponds to

$$Z_o = \sqrt{Z_x Z_q} \quad ,$$

i.e., the matching section impedance is the geometric mean of the impedances of quartz and water.

It is of interest to compare the relative midband power insertion loss and voltage attenuation of the backed and matched transducers. For any given peak phase deviation, the relative power insertion loss is given by:

$$\frac{P_{backed}}{P_{matched}} = 20 \log (1 + m) \quad (15)$$

and the relative voltage attenuation is given by:

$$\frac{V_{backed}}{V_{matched}} = 20 \log[(1 + m)k] \; . \quad (16)$$

Assuming a fractional bandwidth of 50 percent for both the matched and backed transducers, the required parameters are $m = 5.80$ and $k = 0.316$; then

$$\frac{P_{backed}}{P_{matched}} = 16.6 \text{ db}$$

and

$$\frac{V_{backed}}{V_{matched}} = 6.65 \text{ db} \; .$$

Thus, by using a matched transducer in place of a backed transducer, one gains 16.6 db in the required power and only 6.65 db in the required voltage across the transducer. This difference between the voltage and power requirements results from the fact that the matching section acts as a step-down voltage transformer.

In addition to providing a broad bandwidth, it is also required that the transducer introduce a phase shift that varies linearly with frequency within the bandpass of interest. The normalized phase shift of a quarter-wavelength matched transducer, as obtained from Eq. (12), is given by

$$\theta = \arc\tan\left\{-\frac{1}{k}\tan\frac{a}{2}\right\} \quad (17)$$

$$-\arc\tan\left\{\frac{2n(k-\frac{1}{k})\tan^2\frac{a}{2} + (1+\frac{1}{k^2}\tan^2\frac{a}{2})(\tan^2\frac{a}{2}-1)}{2n\tan\frac{a}{2}(1+\tan^2\frac{a}{2})}\right\} .$$

Equation (17) is plotted in Fig. 9 with k as the parameter. The derivative of the phase shift with respect to frequency is a measure of the time delay distortion; this is plotted in Fig. 10 with k as the parameter. From Fig. 10, we see that for a fractional bandwidth of 50 percent at a midband frequency of 20 mc, the time delay distortion is 0.03 μsec.

In the implementation of the quarter-wavelength matching sections, curves showing the effect of errors in the thickness of the section on the frequency response of the transducer were needed. The equivalent circuit of Fig. 7 may be used in the analysis of this effect. It is only necessary to consider the matching section to be resonant at a different frequency from the transducer. Accordingly, in the equations for the input impedance and transfer function of the lossless matching section the argument of the tangent, $\frac{a}{2}$, is replaced by $\frac{a}{2(1+\delta)}$, where δ represents the percent error in the thickness or resonant frequency of the matching section. Thus the frequency response, normalized with respect to the transducer resonant response, is given by:

$$\frac{V_o'(f)}{V_o'(f_o)} = \left\{ \left(\frac{G^2+1}{G^2+\frac{1}{k^2}}\right)\left(\frac{G_o^2+\frac{1}{k^2}}{G_o^2+1}\right) \right. \quad (18)$$

$$\left. \frac{1}{4n^2C^2\left(\frac{G^2+1}{G^2+\frac{1}{k^2}}\right)^2 + \left[\frac{2mCG(k-\frac{1}{k})}{G^2+\frac{1}{k^2}} + (1-C^2)\right]^2} \right\}^{\frac{1}{2}}$$

where
$$C = \cot\frac{a}{2}$$
$$G = \cot\left[\frac{a}{2(1+\delta)}\right]$$
$$G_o = \cot\left[\frac{\pi}{2(1+\delta)}\right] .$$

Equation (18) is plotted in Fig. 11 for $k = 0.316$, with positive values of δ as the parameter, denoting that the matching section is too thin. The error curves for negative values of δ are very nearly mirror images of the curves shown in Fig. 11.

IV. Experimental Results

The frequency response of the backed and matched transducers was measured by means of a coherent optical system simulating an actual processor employing a light modulator. The technique is based on the fact that, for sufficiently small excitation, the light content of the first-order fringe is proportional to the square of the peak change in the refractive index of the delay medium, which in turn is related to the square of the force exerted by the transducer face. Thus, by measuring the light content of the first-order fringe as a function of frequency, at constant drive level, the frequency response of the transducer is obtained.[5]

A block diagram of the experimental set up is shown in Fig. 12. The transducers used were x-cut quartz crystals with a nominal resonant frequency of 20 mc. They were circular in cross section, with an effective diameter of 0.9 cm, and were gold plated on both faces. The transducer retainer and cell were especially designed to accommodate interchangeable units with various backing materials and matching sections.

There are certain effects in the light modulation process which are quite frequency dependent. These are: (1) ultrasonic attenuation, (2) internal refraction, and (3) oblique incidence. These effects had to be taken into account in the frequency response measurements, but because of space limitations they cannot be discussed in detail here.

A. Backed Transducers

Broadband backed transducers were implemented using various backing materials including several liquids, epoxies, and mixtures of epoxy and metal powder. The common characteristics of these materials were their relatively high acoustic impedance and their natural adherence to quartz. The mixture of epoxy and metal powder was particularly useful since the acoustic impedance could be controlled by varying the ratio of the two materials in the mixture. In the case of the epoxies, the backing was simply molded on to the back face of the transducer. Since epoxies have a rather high acoustic attenuation, a backing one-half inch in length was sufficient to eliminate reflections. The maximum bandwidth obtained using epoxy mixtures was 70 percent, corresponding to a backing material with an acoustic impedance of about 13×10^6 Kg/sec-m^2. This was obtained with a backing composed of an 87 percent tungsten powder-13 percent epoxy mixture, with an insertion loss of 19 db. The frequency response of a typical transducer backed with this mixture is shown in Fig. 13. With a higher ratio of tungsten to epoxy, the mixture would not adhere to the crystal. A fractional bandwidth as high as 135 percent was obtained with mercury backing, with an insertion loss of 22 db.

B. Matched Transducer

It was shown in Sec. III-B that the required acoustic impedance of the matching section material is given by $Z_o = \sqrt{Z_x Z_q}$. For a water delay medium and a quartz transducer this corresponds to $Z_o = 4.75 \times 10^6$ Kg/m^2-sec. From backed transducer measurements, it was determined that a mixture consisting of 62 percent aluminum powder and 38 percent Araldite #502 epoxy by weight gave approximately the desired acoustic impedance. However, variations of 10 percent were found from mixture to mixture and even between different samples of the same mixture. This is due to the fact that the acoustic impedance is critically related to the rate of setting of the epoxy and the homogeneity of the mixture. Since the acoustic velocity of the medium is related to its acoustic impedance, it was necessary to tailor the thickness of each matching section in order to obtain a thickness of exactly a quarter-wavelength.

In the implementation of the matching section, a standard microtome was employed with a modified head to hold the transducer, and a specially designed cutting knife. This device essentially consists of a reciprocating head which can be advanced against a cutting knife in steps of one micron. The procedure used to make a matching section was as follows: a thin layer of the epoxy mixture was spread on the transducer face and allowed to set at a controlled temperature. The transducer, vacuum-held in a jig, was then attached to the microtome head and the epoxy was cut to a thickness of 35 microns. This corresponds to a thickness about 10 percent higher than calculated. A frequency response of the transducer was then measured by the optical technique described previously. The frequency response curve was compared with the theoretical error curves of Fig. 11 and the exact error in the thickness of the matching section was established. The transducer was then placed in the microtome and the thickness of the matching section was corrected by the desired amount.

The measured frequency response of a typical matched transducer and the corresponding theoretical curve are shown in Fig. 14, for the case where $k = 0.35$ and the error in thickness is zero. As can be seen, there is good agreement between the experimental results and the theory.

V. Conclusion

The results presented illustrate that a broadband spatial light modulator can be implemented using either backed or matched transducers. Using backed transducers, fractional bandwidths of over 100 percent were obtained, but at the expense of high insertion loss. With matching techniques the fractional bandwidth was limited to 50 percent, with essentially zero insertion loss.

The phase responses of both the matched and backed transducers were given. It was shown that for a 20 mc matched transducer the delay distortion within the passband is 0.03 μsec, which is small compared to the reciprocal of the bandwidth. The implementation technique that has been successful in producing quarter-wavelength matching sections at 20 mc has been discussed. It appears likely that multiple layer matching sections can be implemented using this technique, thus extending the bandwidth of the matched transducer.

Appendix

In order to calculate the power delivered to an air-backed transducer in contact with a liquid medium, it is first necessary to find the relationship between the modulation index (ψ_m) and the peak voltage across the transducer (V_m).

In the ultrasonic cell the piezoelectric transducer radiates acoustic waves into the liquid medium. As the sound waves propagate through the liquid cell, there are small increases and decreases in pressure accompanied by changes in temperature, volume, density and hence refractive index. All these variations are interrelated in a manner determined by the equation of state of the fluid and the type of thermodynamic process that takes place.

Assuming that the process is adiabatic (though this assumption is not generally correct, it provides a reasonably close approximation), the relation between the density, ρ, under a pressure P and volume compressibility β is given by:

$$\frac{d\rho}{dP} = \beta\rho \ . \tag{1}$$

The velocity of sound in the medium, v_S, is given by:

$$v_S^2 = \frac{1}{\beta\rho} \tag{2}$$

so that

$$\frac{d\rho}{dP} = \frac{1}{v_S^2} \ . \tag{3}$$

The relationship between the refractive index of light and density is found from the law of molecular refraction and is given by

$$\frac{\mu^2 - 1}{(\mu^2 + 2)\rho} = \text{constant} \ . \tag{4}$$

For small deviations in μ and ρ about their quiescent values, Eq. (4) may be differentiated, giving

$$d\mu = \frac{(\mu^2 - 1)(\mu^2 + 2)}{6\mu\rho} d\rho \ . \tag{5}$$

Substituting for $d\rho$ from Eq. (3) gives

$$d\mu = \frac{(\mu^2 - 1)(\mu^2 + 2)}{6\mu\rho \, v_S^2} dP \ . \tag{6}$$

Since the variations are small, $\mu \approx \mu_o$, $\rho \approx \rho_o$, the static density, the refractive index, and the pressure are linearly related so that

$$\frac{\partial\mu}{\partial P} = \frac{\mu_m}{P_m} \tag{7}$$

where P_m is the peak pressure change. Then,

$$\mu_m = \frac{(\mu_o^2 - 1)(\mu_o^2 + 2)}{6\mu_o\rho_o v_S^2} P_m \tag{8}$$

and substituting for the peak change of the refractive index,

$$\psi_m = \frac{\pi q}{3\lambda_L} \cdot \frac{(\mu_o^2 - 1)(\mu_o^2 + 2)}{\mu_o\rho_o v_S^2} P_m \ . \tag{9}$$

Assume that the liquid has perfect acoustic termination and is lossless. Also assume a half-wave resonant air backed

lossless transducer. Then the relationship between the peak pressure (P_m) and the peak transducer voltage (V_m) is given by:

$$P_m = \frac{4hV_m}{\beta\lambda_t} \qquad (10)$$

where h = piezoelectric stress constant

β = dielectric impermeability at constant strain

λ_t = wavelength of sound in transducer.

The mechanical impedance of the liquid medium is

$$Z_q = \rho_o v_s A_t \ . \qquad (11)$$

The impedance reflected back to the electrical side of the transducer is

$$R_e = \frac{Z_q \lambda_t^2 \beta^2}{16h^2 A_t^2} = \frac{\rho_o v_s \lambda_t^2 \beta^2}{16h^2 A_t} \ , \qquad (12)$$

and the power delivered to the transducer is given by:

$$W = \frac{V_m^2}{2R_e} \ . \qquad (13)$$

Solving Eqs. (9), (10), (12) and (13) for the power delivered to obtain a peak phase deviation ψ_m, results in

$$W = \left[\frac{3\mu_o \lambda_L}{(\mu_o^2 - 1)(\mu_o^2 + 2)q}\right]^2 \left[\frac{\rho_o v_s^3 A_t}{2}\right] \left[\frac{\psi_m}{\pi}\right]^2 . \qquad (14)$$

References

1. Debye and Sears, Proc. National Acad. Sciences, Vol. 18, No. 6, June 15, 1932.

2. Raman and Nath, "The Diffraction of Light by High Frequency Sound Waves," Pts. I-V, Proc. Indian Acad. of Science, 1935-1936.

3. Mason, W. P., Electro-Mechanical Transducers and Wave Filters, Second Edition, D. Van Nostrand Co., New York, N. Y., 1958.

4. McSkimin, H. J., "Transducer Design for Ultrasonic Delay-Lines," Jour. Acous. Soc. of America, Vol. 22, No. 2, p. 302, March, 1955.

5. "Development of Two-Dimensional Filter," Technical Report T-1/153, Electronics Research Laboratories, School of Engineering, Columbia University, New York 27, New York, July 1, 1959. Confidential.

Acknowledgement

This research is sponsored by The Advanced Research Projects Agency of the Department of Defense and is currently administered by the Air Force Office of Scientific Research under Contract AF 49(638)-1113. This work was initially directed by the Rome Air Development Center, Air Force Systems Command, United States Air Force, New York, under Contract AF 30(602)-1971.

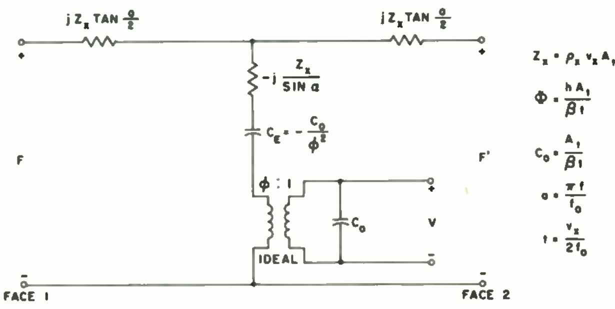
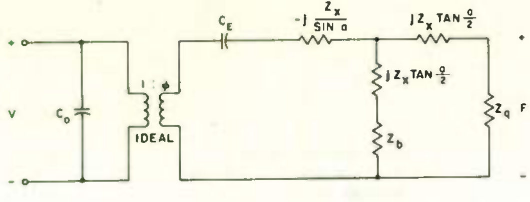
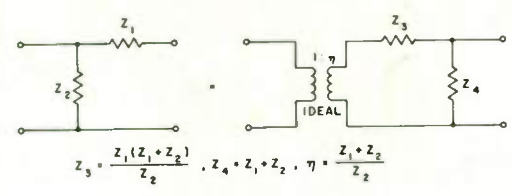


Fig. 1. General equivalent circuit for a thickness vibrating crystal.

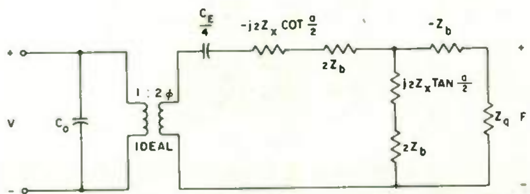


a) EQUIVALENT CIRCUIT FOR CRYSTAL LOADED ON BOTH FACES



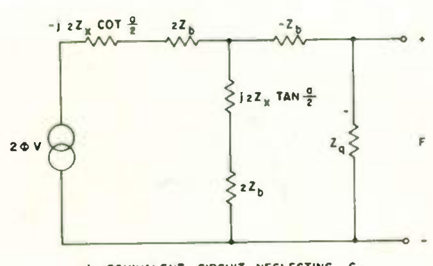
$$Z_3 = \frac{Z_1(Z_1+Z_2)}{Z_2},\ Z_4 = Z_1+Z_2,\ \eta = \frac{Z_1+Z_2}{Z_2}$$

b) NETWORK IDENTITY

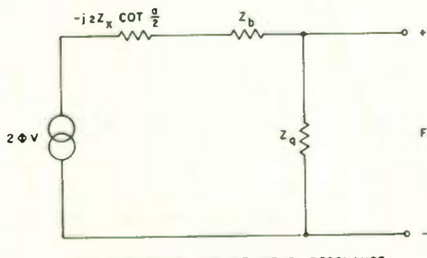


c) EQUIVALENT CIRCUIT FOR CRYSTAL LOADED ON BOTH FACES

Fig. 2. Equivalent circuits for thickness transducers loaded on both faces.



a) EQUIVALENT CIRCUIT NEGLECTING C_E



b) EQUIVALENT CIRCUIT NEAR RESONANCE

Fig. 3. Approximate equivalent circuit of crystal transducer loaded on both faces (driving circuit not included).

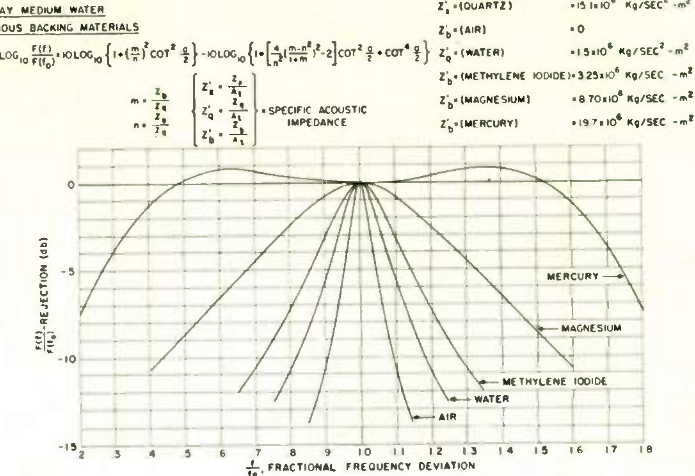
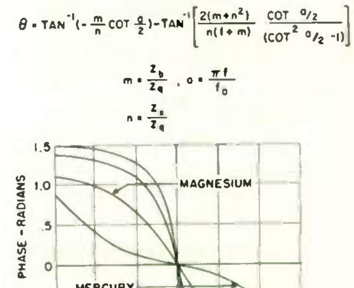


Fig. 4. Theoretical frequency response of x-cut quartz transducer loaded on both faces.

DELAY MEDIUM: WATER
VARIOUS BACKING MATERIALS



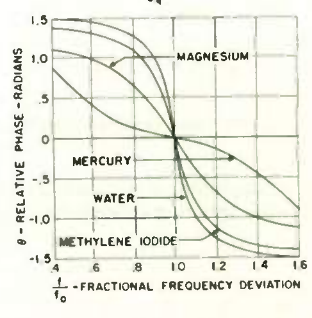


Fig. 5. Relative phase shift of backed quartz transducer.

DELAY MEDIUM: WATER
VARIOUS BACKING MATERIALS

Fig. 6. Delay as a function of frequency for backed transducer.

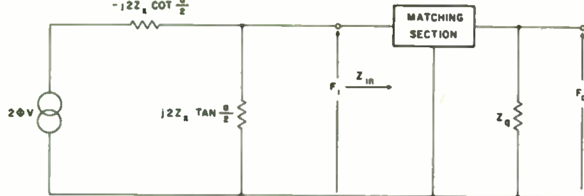


Fig. 7. Equivalent circuit of a backed transducer with quarter-wavelength matching section.

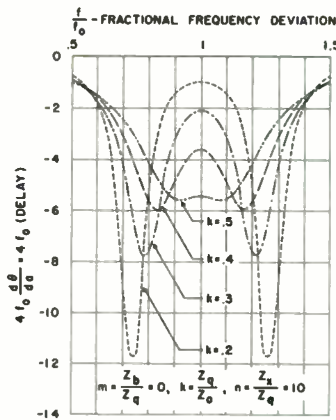


Fig. 10. Delay as a function of frequency for air-backed quartz transducer with various quarter-wavelength matching sections.

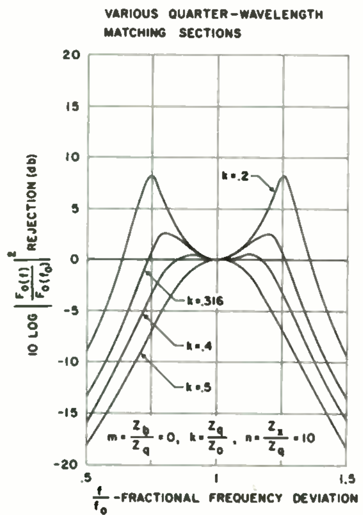


Fig. 8. Frequency response of air-backed quartz transducer with various quarter-wavelength matching sections.

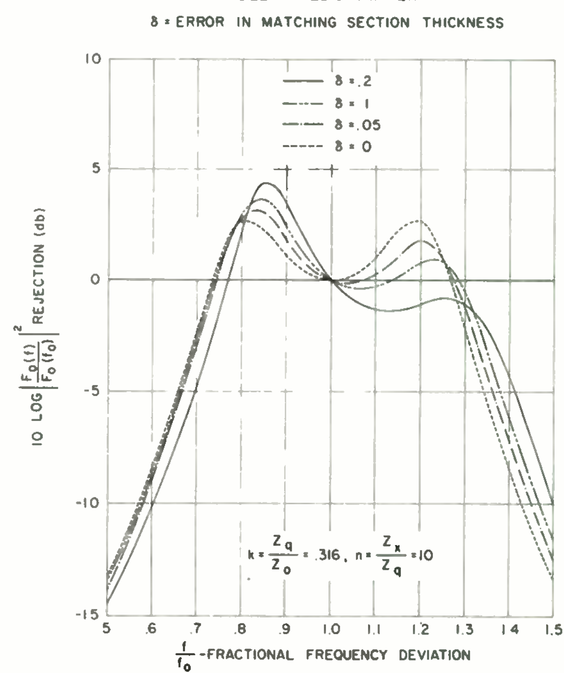


Fig. 11. Effect of error in thickness of quarter-wavelength matching section for k = .316 δ ≥0.

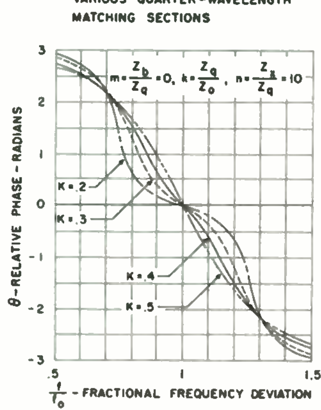


Fig. 9. Relative phase shift of air-backed quartz transducer with various quarter-wavelength matching sections.

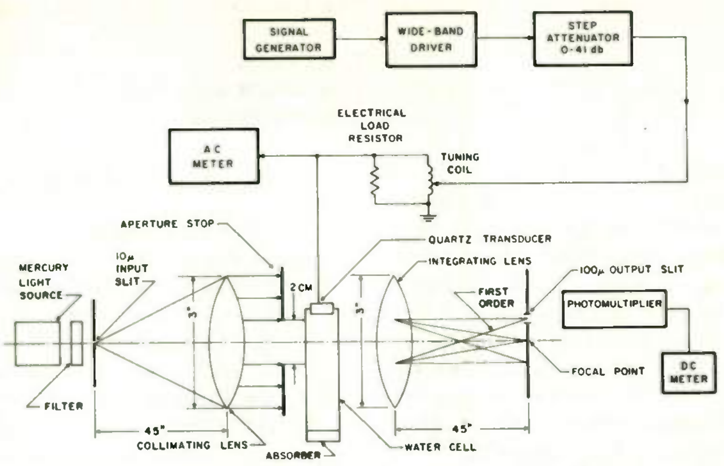


Fig. 12. Block diagram of system for measuring transducer frequency response.

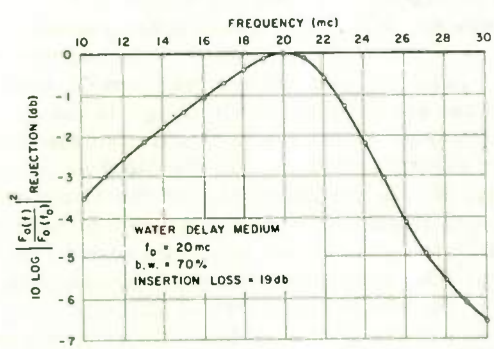


Fig. 13. Frequency response of 87 per cent tungsten—13 per cent epoxy backed transducer.

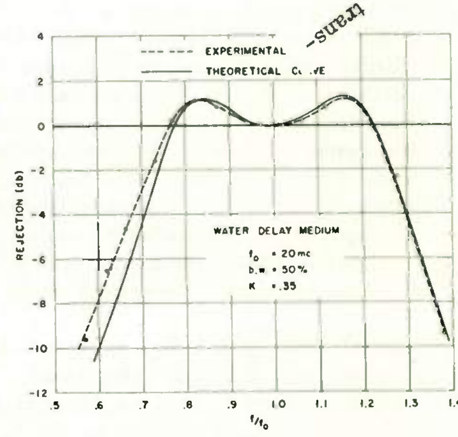


Fig. 14. Frequency response of transducer with Aluminum-epoxy matching section; $\delta = 0$.

THEORETICAL POWER LIMITS OF SONAR TRANSDUCERS

Ralph S. Woollett
U. S. Navy Underwater Sound Laboratory
New London, Connecticut

Summary. The power output of a sonar transducer may be considered to have an acoustical, an electrical, a mechanical, and a thermal limit. Many transducers encounter neither the acoustical, or cavitation, limit nor the thermal limit because of favorable operating conditions. The electrical and mechanical limits, on the other hand, are of practical importance in all high performance transducers. These limits are expressible in terms of fundamental transducer parameters, such as electromechanical coupling coefficient, mechanical Q, and maximum storable electric and elastic energy densities. The usefulness of various criteria for comparing power capability, such as watts/cm^2 and watts/lb, is discussed.

Introduction

The power output of a sonar transducer may be considered to have an acoustical, an electrical, a mechanical, and a thermal limit. The acoustical limit is reached when pronounced cavitation sets in, and it is determined by three factors: the compressive bias on the water due to depth, the dynamic tensile strength of the sea water, and the near-field pressure distribution of the radiated sound. While the cavitation limit is of prime importance for surface vessel sonars, many other sonars operate at depths sufficiently great so that this limit is not encountered.

The thermal limit is a function of the efficiency, duty cycle, and thermal conductivity designed into the transducer. Few generalizations about thermal design may be made, since transducers come in such a variety of configurations. In many sonars the transmitted signal has such a low duty cycle that the thermal limit is not encountered.

The power limits which will be considered in detail here are the electrical and the mechanical limits, since these are of practical importance in all high-performance transducers. The analysis will be for the common single-degree-of-freedom transducer, and we shall formulate equations for the electrical and mechanical power limits in terms of basic transducer parameters such as electromechanical coupling coefficient, mechanical Q, and the storable electric and elastic energies.

The Electrical Limit

The electrical limit will be treated first. Figure 1 shows the equivalent circuits which will be employed and shows relations of some of the basic parameters to circuit element values. The electromechanical coupling coefficient k is a function of the motional capacity C_Y and the blocked capacity C_b for the electric field transducer, or the motional inductance L_z and the blocked inductance L_b for the magnetic field transducer. The mechanical resonant frequency ω_r is determined by the motional inductance and capacity. The mechanical storage factor at resonance, Q_M, is a function of the motional capacity or inductance and the motional resistances. The power absorbed in the motional resistances R_Y and R_Z is partly useful radiated power and partly internally dissipated power.

In Fig. 2 expressions are given for the electric and magnetic energies stored in the blocked capacity or blocked inductance of the transducer. For most transducers the electric and magnetic fields are uniform throughout the active material and the second form of these equations then apply, in which ϵ_b is the blocked permittivity, $\mathcal{E}_p$ is the peak value of the alternating electric field, μ_b is the blocked permeability, H_p is the peak value of the alternating magnetic field, and V is the volume of material in which the electric or magnetic energy is stored. The next equations give the radiated power; the mechanoacoustical efficiency η_{ma} has been introduced to obtain this radiated portion from the total mechanical power flowing into the motional resistances. In these equations we shall eliminate the voltage E and current I in favor of the stored energies given by the equations above and shall next eliminate the various circuit elements by use of equations given in Fig. 1. The result is the general equation at the bottom of Fig. 2, which applies to all types of linear electroacoustical transducers.

The transducer's radiated power is limited because the storage capacity of its mechanically coupled electric or magnetic reservoir is limited. The limits on the stored energy U_e are imposed by such factors as insulation breakdown, depolarization of materials operating at remanence, distortion resulting from dielectric or ferromagnetic nonlinearities, and deteriora-

tion of efficiency resulting from a rise in the dissipation factor at high fields. Nonlinearities are thus included as a possible power limit, although the power equation was derived on the basis of linear circuit theory. Its usefulness, therefore, is restricted by the assumption that grossly nonlinear operation will be avoided.

The Mechanical Limit

Next, the mechanical limit, which depends on the transducer's capacity to store elastic energy, will be derived. The starting point is the power equation at the top of Fig. 3. By expressing the radiation resistance as a function of Q_M and eliminating the velocity in favor of the elastic energy, we obtain the general equation at the center of Fig. 3, which is free from explicit dependence on the equivalent circuit elements. The limit on the elastic energy U_m, which in turn limits the radiated power, is caused by factors such as fracture in the case of ceramic or crystal transducers, metal fatigue for transducers employing springs, diaphragms, or magnetostrictive laminations, gap closure for variable reluctance or electrostatic transducers, and excessive travel for moving coil transducers.

The transducer power must remain below whichever of the two limits included in Figs. 2 and 3 is lower. For an economical design, it is desirable to have the electrical and the mechanical power limits equal. When the optimum radiation loading to bring about this condition prevails, the mechanical Q will have the value given by the equation at the bottom of Fig. 3.

Maximum Storable Energy Densities

The two power-limit expressions which have been presented depend on the resonant frequency and the transducer size, inasmuch as the storable energies are a function of size. We shall now introduce some different parameters, defined in Fig. 4, which will make the results independent of resonant frequency and less dependent on transducer size. Energy densities rather than total energies will be used, and the radiated power will be divided by the radiating area to yield an average surface intensity. Other parameters required are the radiation resistance per unit area (averaged), the maximum value of the stress, T_{max}, which can be allowed without risking fracture or fatigue, and the density ρ and sound velocity c of the active material.

Before returning to the power limit equations we shall consider the energy densities which seem feasible for the common piezoelectric and piezomagnetic transducer materials (see Fig. 5). The first column lists the electromechanical coupling coefficient which is available when the material is used in its most favorable mode. The next column gives the estimated maximum electric or magnetic energy density. The third column gives the product of two of the factors which occur in the electrically limited power equation; it indicates the amount of electrical energy which has been converted to mechanical energy and is available for work. The last column pertains to mechanically limited conditions and gives the estimated maximum elastic energy density. All these estimates are extremely crude, and they are offered as targets for criticism rather than as definitive limits.

The Ring, Flexural Disk, and Tonpilz Transducers

Next, the two power-limit equations will be applied to three different types of transducers, but the results will be expressed as surface intensity rather than total acoustic power. In all cases radiation mass will be neglected. We consider first the piezoelectric or piezomagnetic ring resonator (see Fig. 6).

In its radial mode the resonant ring has uniform stress distribution as long as it is reasonably thin, and hence the elastic energy equation is simple. The ring can be constructed so as to utilize the highest coupling coefficient of the active material. The intensity limits are independent of resonant frequency, and size does not appear explicitly in their equations, though it must be such as to maintain the resonant condition. The mechanical Q equation assumes radiation from one cylindrical surface only, and it depends on the thickness-to-radius ratio b/a of the ring. The specific radiation resistance r_r is, of course, a function of the height of the ring, but in an array of transducers it is also a function of the spacing between transducers and hence is subject to considerable control.

Figure 7 gives similar equations for the edge-supported flexural disk transducer. The flexural disk is made of laminated ceramic, with the laminations on opposite sides of the neutral surface driven with opposite polarities so that flexural stresses are generated piezoelectrically. The intensity equations have the same form as that for the ring except for numerical factors, which account for the more complex velocity and stress distributions occurring in flexural vibrations. The mechanical Q for the disk depends on the square of the thickness-to-radius ratio, and the coupling

coefficient is a function of the planar coupling coefficient of the material.

Figure 8 gives the equations for a tonpilz, that is, a transducer consisting of a piezoelectric or piezomagnetic elastic member loaded on each end with masses. The transducer is assumed to radiate only from Mass 1. The tonpilz designer has a number of dimensional ratios at his disposal; accordingly, the equations must depend on these design variables as well as on the factors which were common to the equations for the two transducer types presented previously.

Intensity Limits of Barium Titanate Transducers

Next, the intensity limits for the three transducer types discussed above will be calculated in terms of the properties of a particular material. Barium titanate is chosen as the active material for all three transducer types; the maximum allowable stress is taken to be 3000 psi and the maximum electric driving field to be 2 KV/cm rms. For this illustrative example the specific radiation resistance r_r is assumed to remain constant at 1/2 the ρc of water, and the mechano-acoustical efficiency is given a constant value of 60%.

In Fig. 9 the intensity limits are plotted as a function of mechanical Q. The Q is varied by varying a dimensional ratio of the transducer over what seems like a practical range. The thickness-to-radius ratio is varied for the ring and the flexural disk transducers, and the ratio of active material area to radiating area is varied for the tonpilz transducer. In order to fix the design of the tonpilz it was necessary to choose values for the other two dimensional ratios appearing in the equations given in Fig. 8; the ratio of radiating mass to non-radiating mass was chosen to be 1 to 3, and the ratio of front mass length to active material length was chosen as 1 to 4. The condition that the specific radiation resistance remain constant as the dimensional ratios of the transducers are varied might be brought about reasonably well in an array by simultaneous variation of the spacing between transducers.

The results given in Fig. 9 are, of course, independent of the frequency of the design, though they do assume operation at resonance. Sonars normally operate in the lower decade of intensities covered on this graph. It is seen that the ring and the tonpilz appear capable of intensities extending into the second decade, while the flexural disk output falls near the bottom of the lower decade.

Watts/cm^2 versus Watts/lb as Criteria

Watts/cm^2 values, such as displayed here, are a useful indication of a transducer's capability, but they do not serve as a criterion of design excellence when different transducer types are compared. Thus, to evaluate these examples it would be necessary to give consideration to the amount of material in back of the radiating aperture required to produce the intensities shown. Since weight is usually of major importance in high-power sonar systems, watts/lb is a figure of interest. When a given design is scaled in frequency, its watts/lb figure will be directly proportional to frequency. If we choose a frequency of 5 kc and compute the watts/lb figures of the three transducer types at the value of mechanical Q where the electrical and mechanical limits are equal (or at the lower end of the plotted curves for the tonpilz), we obtain the following results: the ring ranks highest at 800 watts/lb, the flexural disk is next at 130 watts/lb, and the tonpilz is lowest at 60 watts/lb.

In conclusion, the examples which have been given are particular cases of a rather general analysis of power limits. The general method applies to variable reluctance, electrostatic, moving coil, and piezomagnetic transducers, as well as to the piezoelectric types illustrated.

WRH

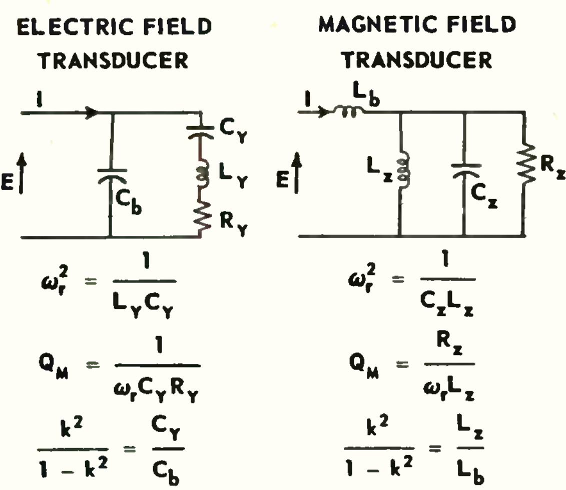


Fig. 1. Transducer equivalent circuits.

ELECTRIC FIELD | **MAGNETIC FIELD**

PEAK STORED ENERGY, TRANSDUCER BLOCKED:

Electric field: $U_e = \frac{1}{2}|E|^2 C_b = \frac{1}{2}\epsilon_b \mathcal{E}_p^2 V$

Magnetic field: $U_e = \frac{1}{2}|I|^2 L_b = \frac{1}{2}\mu_b H_p^2 V$

RADIATED POWER AT RESONANCE:

Electric field: $P_r = \left(\frac{1}{2}|E|^2/R_Y\right)\eta_{ma}$

Magnetic field: $P_r = \frac{1}{2}|I|^2 R_z \eta_{ma}$

$$\boxed{P_r = \eta_{ma}\,\omega_r \frac{k^2}{1-k^2} Q_M U_e}$$

η_{ma} = MECHANOACOUSTICAL EFFICIENCY

V = VOLUME OF ACTIVE MATERIAL

Fig. 2. The electrical limit equation.

RADIATED POWER: $P_r = \frac{1}{2}|V|^2 R_r$

PEAK ELASTIC ENERGY: $U_m = \frac{1}{2}\frac{|V|^2}{\omega^2 C_M}$

RADIATION RESISTANCE: $R_r = \frac{\eta_{ma}}{\omega_r C_M Q_M}$

V = PEAK VELOCITY

C_M = TRANSDUCER COMPLIANCE

$$\boxed{P_r = \frac{\eta_{ma}\,\omega_r U_m}{Q_M}}$$

OPTIMUM Q_M FOR POWER: $Q_M = \sqrt{\dfrac{(U_m)_{max}}{\frac{k^2}{1-k^2}(U_e)_{max}}}$

Fig. 3. The mechanical limit equation.

DEFINITIONS

$(u_e)_{max} = (U_e)_{max}/V$ = MAXIMUM ELECTRIC (OR MAGNETIC) ENERGY DENSITY.

$(u_m)_{max} = (U_m)_{max}/V$ = MAXIMUM ELASTIC ENERGY, AVERAGED OVER THE VOLUME.

T_{max} = MAXIMUM ALLOWABLE STRESS.

$I_s = P_r/A_r$ = SURFACE INTENSITY OF RADIATION.

$r_r = R_r/A_r$ = SPECIFIC RADIATION RESISTANCE.

A_r = RADIATING AREA.

ρ = DENSITY OF ACTIVE MATERIAL.

c = SOUND VELOCITY OF ACTIVE MATERIAL.

Fig. 4. Additional definitions.

MATERIALS	COUP. COEFF. $k = k_{33}$	ESTIMATED ENERGIES IN JOULES/m^3 $(u_e)_{max}$	$\frac{k^2}{1-k^2}(u_e)_{max}$	$T^2_{max}/2\rho c^2$ $(u_m)_{max}$
NICKEL	.30	200	20	4000
PERMENDUR POLARIZED	.29	400	37	4000
PERMENDUR REMANENCE	.20	90	4	4000
NICKEL FERRITE	.32	90	10	1500
ADP CRYSTAL	.28	200	17	4500
BARIUM TITANATE	.48	400	120	2000
LEAD TITANATE ZIRCONATE	.60	1200	670	3000

Fig. 5. Energy limits of transducer materials.

RING TRANSDUCER

a = MEAN RADIUS b = RING THICKNESS

MAX. ELASTIC ENERGY: $U_m = \frac{T^2_{max}}{2\rho c^2} V$

ELEC. LIMIT: $I_s = Q_M \frac{k^2}{1-k^2}\frac{r_r}{\rho}(u_e)_{max}$

MECH. LIMIT: $I_s = \frac{r_r}{2\rho^2 c^2} T^2_{max}$

STORAGE FACTOR: $Q_M = \frac{b}{a}\frac{\rho c}{r_r}\eta_{ma}$

MAX. COUPLING COEFF.: $k = k_{33}$

Fig. 6.

SUPPORTED-EDGE FLEXURAL DISK

a = DISK RADIUS h = DISK THICKNESS

MAX. ELASTIC ENERGY: $U_m = .15 \dfrac{T_{max}^2}{2\rho c^2} V$

ELEC. LIMIT: $I_s = .71\, Q_M^2 \dfrac{k^2}{1-k^2} \dfrac{r_r}{\rho} (u_e)_{max}$

MECH. LIMIT: $I_s = .10 \dfrac{r_r}{2\rho^2 c^2} T_{max}^2$

STORAGE FACTOR: $Q_M = 2.0 \left(\dfrac{h}{a}\right)^2 \dfrac{\rho c}{r_r} \eta_{ma}$

COUPLING COEFF.: $k = .77\, k_p \approx .5\, k_{33}$

Fig. 7.

TONPILZ

M_2 A_a ℓ_a M_1 ρ_1 A_r ℓ_1

$M_T = M_1 + M_2$, $V = A_a \ell_a$ $U_m = \dfrac{T_{max}^2}{2\rho c^2} V$

MAXIMUM $k \leq k_{33}$

ELEC. LIMIT: $I_s = \left(\dfrac{M_2}{M_T}\right)\left(\dfrac{\ell_a}{\ell_1}\right)\left(\dfrac{A_a}{A_r}\right) Q_M^2 \dfrac{k^2}{1-k^2} \dfrac{r_r}{\rho} (u_e)_{max}$

MECH. LIMIT: $I_s = \left(\dfrac{M_2}{M_T}\right)\left(\dfrac{\ell_a}{\ell_1}\right)\left(\dfrac{A_a}{A_r}\right) \dfrac{r_r}{2\rho^2 c^2} T_{max}^2$

STORAGE FACTOR: $Q_M = \left(\dfrac{M_T}{M_2}\right)^{3/2} \sqrt{\dfrac{\ell_1 A_a}{\ell_a A_r}} \dfrac{\sqrt{\rho \rho_1}\, c}{r_r} \eta_{ma}$

Fig. 8. Tonpilz transducer.

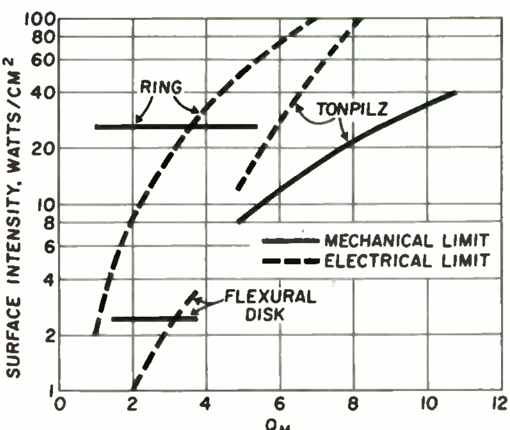


Fig. 9. Intensity limits of barium titanate transducers.

HIGH-FREQUENCY ELECTROSTATIC TRANSDUCERS FOR USE IN GASES+

Wayne M. Wright
Acoustics Research Laboratory
Harvard University
Cambridge 38, Massachusetts

Summary

Electrostatic pressure transducers of the type sometimes referred to as "solid dielectric condenser microphones" can be made to serve as useful sources and receivers of sound at frequencies as high as one megacycle per second. Microphones have been developed which have essentially uniform sensitivities up to half a megacycle and corresponding rise times of less than one microsecond for transient signals. The compliance which leads to the generation of a voltage proportional to the incident sound pressure is provided by the thin air film between a 0.00025" Mylar diaphragm and a smooth plane back electrode across which it is stretched; a light coating of aluminum on the exposed surface of the plastic provides the second electrode for the capacitor.

For theoretical analysis we assume the presence of a uniform air film whose stiffness, together with the diaphragm mass, determines the resonance frequency of the transducer. The theoretical receiving voltage sensitivity is constant below this resonance and falls off as the square of the frequency above resonance. On the other hand, when the transducer is operated as a sound source, the ratio of axial pressure to driving voltage would increase as the square of the frequency below resonance and remain constant above resonance if it were not for the attenuation in the air.

The narrow beam produced by a practical ultrasonic transmitter at frequencies above 100 kc, and the finite attenuation in gases at these frequencies, make it difficult to satisfy the conditions required by a reciprocity calibration technique. Hence the response characteristics of these transducers at ultrasonic frequencies are determined by measuring the voltage gain of transmitter-air-microphone systems.

Desirable features of some present units include excellent transient response, reasonably high sensitivity (of the order of -85 db re 1 volt/dyne cm^{-2}), and an accessible active surface (1/4" diameter) with no portion of the transducer housing protruding in front of the plane of this surface. Undesirable features include a dependence of both sensitivity and resonance frequency upon ambient air pressure and a physical size which is large in comparison with the wavelength of sound at ultrasonic frequencies.

Introduction

The condenser microphone for audio reception has been used for many years in applications for which its lack of sensitivity is more than compensated for by its flat frequency response below resonance. It typically comprises a tightly stretched metallic diaphragm which is exposed to the sound field and which is clamped along its edge at a small distance in front of a rigid metal back electrode. The resultant motion of this electrically grounded diaphragm changes the spacing and hence the capacitance between it and the back electrode; a corresponding voltage is induced if a fixed electrical charge is maintained on the condenser elements. The stiffness of the moving diaphragm, which influences the resonance frequency and the sensitivity at lower frequencies, is usually furnished either by the stretching or the elastic rigidity of the diaphragm.

The resonance frequency of a condenser microphone can be raised by decreasing the mass of the diaphragm or increasing the stiffness. The transducers developed in the present research utilize films of DuPont Mylar polyester, 0.00025" (6 microns) thick, on which has been evaporated about 1 micron of aluminum. This light diaphragm is actually supported by the microscopic and submicroscopic high spots on a nominally smooth back plate, and the motion of the plastic film between these points of support is cushioned by the film of trapped air, which has an average thickness of about 1 micron. Typical microphones constructed with Mylar film and smooth back plates have resonance frequencies between 400 and 700 kc and sensitivities below the resonance frequency of -85 to -95 db re 1 volt/dyne cm^{-2}.

Previous work upon which the present research is based was conducted at the Third Physical Institute of the University of Göttingen, Germany. In a paper which has become a standard reference,

+This work was supported in part by the Office of Naval Research.

Kuhl et al[1] presented details on the construction and performance of "solid dielectric" transducers having upper frequency limits of operation somewhat less than 200 kc. During the following ten years, a number of other research projects have been carried on at Göttingen which have both explored the operation of these devices further and used the transducers as research tools. Details of Japanese research on high-frequency electrostatic transducers are found in two papers by Matsuzawa.[2] Work in this country has included that of McCue,[3] who applied similar instruments to a study of the sonar used by flying bats.

Transducer Design

The two transducers shown in Figs. 1 and 2 illustrate the construction features of these devices. All of our experimental transducers have been designed to be physically interchangeable with Western Electric 640-AA condenser microphones and they are ordinarily operated with the same 200-volt bias. Two methods of securing the diaphragm to the brass housing are illustrated in these figures: mechanical clamping under a ring and screw-on cap, and bonding with an adhesive such as epoxy cement. The long shank of the transducer model at the right in the photograph and drawing is provided as an aid in mounting the unit; no portion of this housing extends in front of the plane of the diaphragm.

In addition to a thin, light-weight plastic film with stable physical properties, high dielectric strength, and a conducting coating on one surface, the essential components of these transducers are the metal housing and back electrode and an insulator to separate these two elements. The diaphragm must be rigidly attached to the housing, and some means must be provided to produce tension in this film and to hold it tightly against the back electrode. Electrical contact to the diaphragm coating is made through either a clamping ring or conducting silver paint.

Most of our experimental transducers have utilized brass back plates with their surfaces merely turned smooth in a lathe; in a few instances this surface has been roughened slightly by "sandblasting" with fine aluminum oxide powder. The back plate is advanced, by turning the screw which also serves as the electrical connection, to stretch the plastic film slightly. The sizes of the back plates in these experimental models were determined by consideration of radiation beam pattern, electrical loading of a transmitter on the power amplifier, and ease of assembly of the complete transducer.

Experimental variability in the mounting and stretching of the thin diaphragm over the smooth back plate has led to some difficulty in reproducing a given frequency response with this scheme. It is often necessary to assemble a transducer several times, each time with a new diaphragm, before the resulting sensitivity and resonance frequency are satisfactory. The plastic surface of the diaphragm must be clean and dry to avoid "charging" from the biased back plate, and the stretching over the back plate must be relatively uniform. The use of an adhesive rather than a mechanical clamp for attaching the diaphragm allows a reduction in the size and complexity of the transducer but makes diaphragm replacement more tedious.

As is the case with ordinary condenser microphones, these microphones require a relatively high-impedance termination as well as a biasing voltage. They are ordinarily placed physically close to the control grid pin of a cathode follower tube in such a way that the grid circuit is completely shielded.

Theoretical Results

The Mylar diaphragm in these transducers is actually supported at many points over the surface of the nominally smooth back plate; however, for theoretical analysis the resulting interrupted air film is replaced by an equivalent uniform air film with thickness d_2. The stiffness of this air film, together with the diaphragm mass, determines the resonance frequency. As might have been expected, we find that the calculated receiving voltage sensitivity is constant well below this resonance frequency, falls off as the square of the frequency above resonance, and is independent of the diaphragm area. On the other hand, the ratio of the axial pressure to the driving voltage on an ultrasonic transmitter would increase as the square of the frequency below resonance and remain constant above resonance if it were not for the increasing attenuation in the gaseous medium. This transmitting response is proportional to the area of the source, and the radiation becomes increasingly directive as the source area is enlarged.

Low- and high-frequency asymptotes for the theoretical response of these electrostatic transducers are shown as functions of effective air film thickness in the next two illustrations. The parameters of our 1/4 mil (0.00025") Mylar have been inserted into the theoretical equations, and we have assumed biases of 200 volts and an ambient pressure of one atmosphere.

In Fig. 3, which shows the asymptotes for a unit operated as a microphone, we see that the intersections of these asymptotes define a curve for resonance frequency vs low-frequency sensitivity which approaches straight lines at both frequency extremes. Thus we find that the maximum possible low-frequency sensitivity for these particular microphones and diaphragms is -74 db, and that the high-frequency sensitivity cannot be increased above that which occurs for an air film thickness of 0.1 micron.

Figure 4 gives the maximum pressure which would be induced at a distance of 6" from a transmitter having a 1/2" diameter back electrode. Unlike that of a microphone, the low-frequency response of a transmitter does not continue to

increase with air film thickness but is actually greatest with a thickness of about 2 microns.

We have made no attempt to calculate the damping of the diaphragm, which would be expected to control the response in the neighborhood of the resonance frequency, because the presence of a non-uniform air film in the actual transducers leads to a slight smearing of the resonance frequency and thus prevents the occurrence of a sharp resonance peak. Also, we shall not describe in this paper the severe effects of changes in ambient pressure on the transducer response characteristics.

Experimental Characteristics

There are no available standard microphones for the frequency range of interest in this research, and it is difficult to calibrate microphones by means of a reciprocity technique much above 100 kc. Hence our determination of the response characteristics of these transducers has been limited to measuring sensitivity and input capacitance at audio frequencies and comparing the responses of units which are operated in two-transducer pairs at ultrasonic frequencies. Gated signals have been used for the ultrasonic measurements in order to eliminate the effects of standing waves.

The shape of the theoretical response curve for any two-transducer system can be found by adding the effect of sound absorption in the air to data taken from Figs. 3 and 4. Thus we expect that a system with two transducers having a common resonance frequency will have a response which would first increase as the square of the frequency and then decrease inversely as the square of the frequency if it were not for the considerable attenuation of sound in air at frequencies of several hundred kilocycles.

Figure 5 shows the measured voltage gain between transmitter and microphone terminals for two transducers which were separated by 6" of room air. The behavior of the transducers themselves is shown with a corrected curve for which the effect of sound absorption in the air has been removed, and straight lines drawn on this latter curve indicate the 12 db slopes predicted by our theoretical analysis. The corrected response curve deviates from the straight lines by less than 1 db at most frequencies; this agreement is slightly better than would be the case for a typical pair of units.

The reproducibility of the frequency response of our experimental transducers is indicated by the curves of Fig. 6. Four nominally identical microphones have been used with a single transmitter, and the range of measured sensitivities is about 8 db. Similar variations in response are found when the diaphragm of a microphone is replaced. We thus conclude that our difficulty in reproducing a specific frequency characteristic is related more to the method by which the diaphragm is attached to the transducer housing and stressed than it is to the actual surface configuration of the back plate. More systematic control over the variables associated with the mounting and stressing of the diaphragm should lead to relatively consistent results.

We note that in these curves there is no marked resonance peak, and small-scale irregularities in the curves are limited to a couple of decibels. We thus expect that a high-frequency microphone will have a relatively smooth response up to the resonance frequency, and that there will be no ringing introduced by the microphone if it is subjected to an acoustic impulse. These conclusions have been verified, and the usefulness of the microphones demonstrated, with a study of the impulsive acoustic signal emitted by a small spark.

Figure 7 shows the output of a microphone at distances varying from 1 to 35 cm from a spark which was in room air and had a discharge energy of about 0.01 joule. The time scale is 5 microseconds per division, and the gain of the vertical amplifier was increased with distance as noted along the right-hand side of the oscillograms. A negative voltage corresponds to a positive variational pressure. The rise time of the trace at both ends of the N-wave is slightly less than 1 microsecond, which would be the expected rise time of a low-pass amplifier having a bandwidth of about 500 kc. There is no sign of ringing in these oscillograms.

The response of commercially available microphones to these impulsive acoustic signals is shown in Fig. 8. Here the output of one of our experimental high-frequency units (Mic B) is compared with that of a Massa M-213 ADP microphone, which has a nominal resonance frequency of 140 kc, and of a Western Electric 640-AA condenser microphone. The oscillograms within each set correspond to the same pressure waveform. Unlike that of the condenser microphones, the output of the M-213 has the same algebraic sign as the variational pressure. It is obvious that, at least without electrical or mechanical modification, these two commercial microphones cannot give a meaningful representation of an acoustic impulse of this type.

Desirable features available with our present transducers include excellent transient response, reasonably high sensitivity, and an accessible active surface with no portion of the transducer housing protruding in front of the plane of this surface. Undesirable features include a physical size which is large in comparison with the sound wavelength at ultrasonic frequencies and a difficulty in controlling the responses of these experimental units. The resonance frequency, sensitivity, and phase shift of one of these transducers will vary with time whenever the ambient pressure is changed. Variations will also occur if electrical charge is able to spread from the back plate over a major portion of the adjacent plastic surface of the diaphragm, but this problem can be avoided by careful fabrication of

the transducer. The driving voltage on a transmitter should be much less than the polarizing voltage, and the incident variational air pressure should be much less than ambient, if the introduction of harmonic distortion by these nonlinear devices is to be avoided.

Applications

Most of the sounds which occur in nature are within the frequency range of these microphones; they thus may be expected to yield a reliable characterization of sounds in the air under a wide range of conditions. Their usefulness for detecting transient signals has been demonstrated with the results of a study of the acoustic waveform generated by a small spark. Satisfactory adaptation of these microphones to detect a wide range of similar waveforms, such as occur in shock tubes and in blast waves, is not promising, unfortunately, because of the pressure restrictions mentioned above.

Operating as an ultrasonic transmitter, one of these electrostatic transducers will provide the wide pass band necessary to generate short bursts of high-frequency acoustic energy. One could presumably obtain a flat response even below resonance with an electrical network to compensate for the increase in low-frequency response with the square of the frequency. A relatively intense ultrasonic beam can be directed at a small object with a large-diameter electrostatic transmitter; the available energy will be determined by the amount of harmonic distortion which can be tolerated.

There are many interesting applications for these transducers operating in transmitter-microphone pairs, in addition to the usefulness of this scheme for the determination of transducer characteristics at ultrasonic frequencies. Sound velocity and absorption in gases can be measured by varying the separation between two transducers having a common normal axis. Information can also be derived concerning an acoustically opaque material which breaks this narrow sound beam. The scattering and reflection of sound from solid objects and sold or liquid surfaces can be studied for a continuous range of wavelengths in the neighborhood of one millimeter, and, by replacing the ultrasonic transmitter with a spark, one can also investigate the reflection of weak shock waves.

A major purpose of the present research program has been to provide a comprehensive treatment of the theory and design of these high-frequency electrostatic transducers and an investigation of their behavior in air under normal laboratory conditions. The results summarized in this paper are described at greater length in a forthcoming technical report.[4]

References

1. W. Kuhl, G. R. Schodder, and F. -K. Schröder, Acustica 4, 519-32 (1954).
2. K. Matsuzawa, J. Phys. Soc. Japan 13, 1533-43 (1958) and 15, 167-74 (1960).
3. J. J. G. McCue, I.R.E. International Convention Record 9, Part 6, 310-15 (1961).
4. W. M. Wright, Tech. Memo. No. 47, Acoustics Research Laboratory, Harvard University, April, 1962.

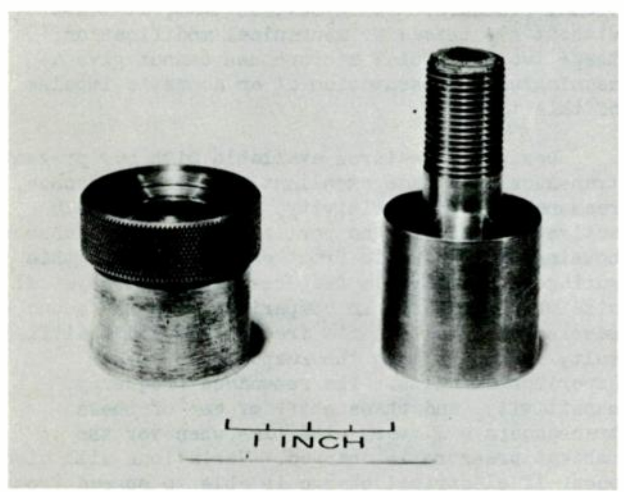


Fig. 1. Photograph of two high-frequency transducers.

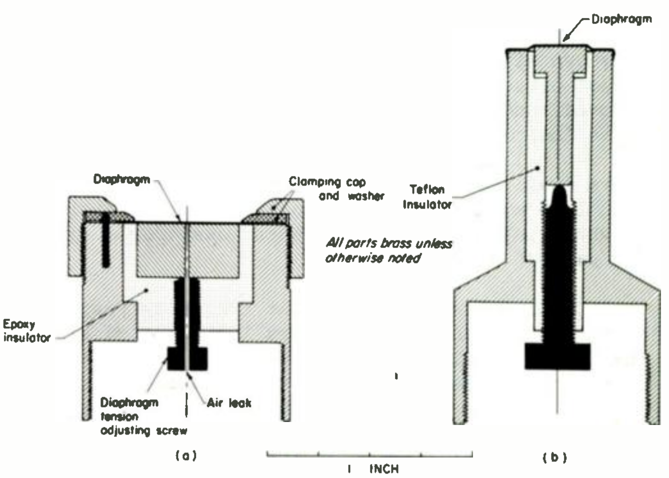


Fig. 2. Construction features of the transducers of Fig. 1.

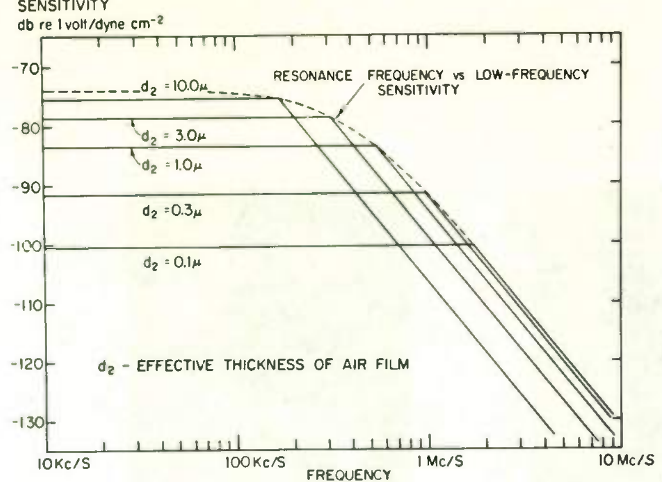


Fig. 3. High- and low-frequency asymptotes for the theoretical response of a microphone, as functions of effective air film thickness.

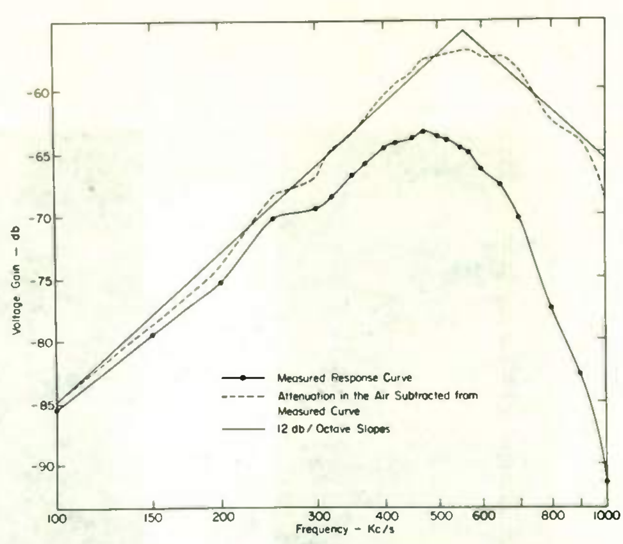


Fig. 5. Measured transducer voltage gain, with effect of sound absorption in the air.

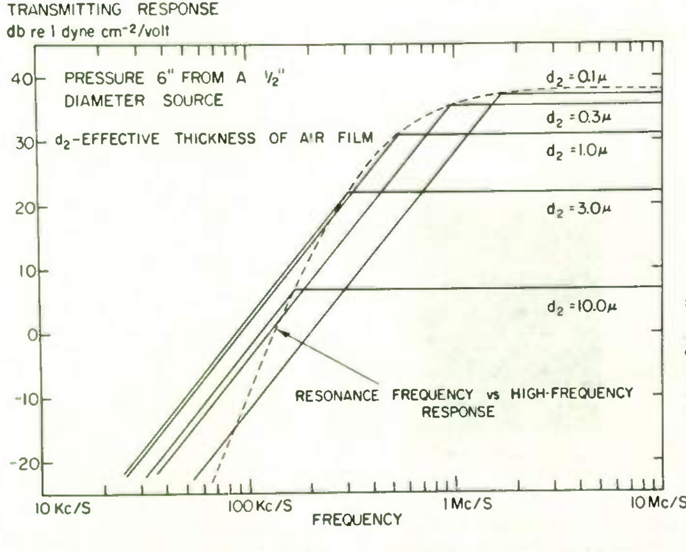


Fig. 4. Asymptotes for the theoretical response of an ultrasonic transmitter, as functions of effective air film thickness.

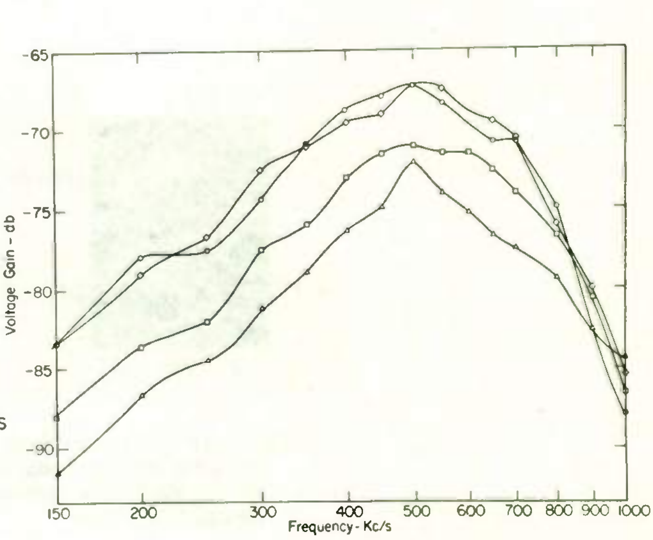


Fig. 6. Transducer voltage gain for a single transmitter and four similar microphones.

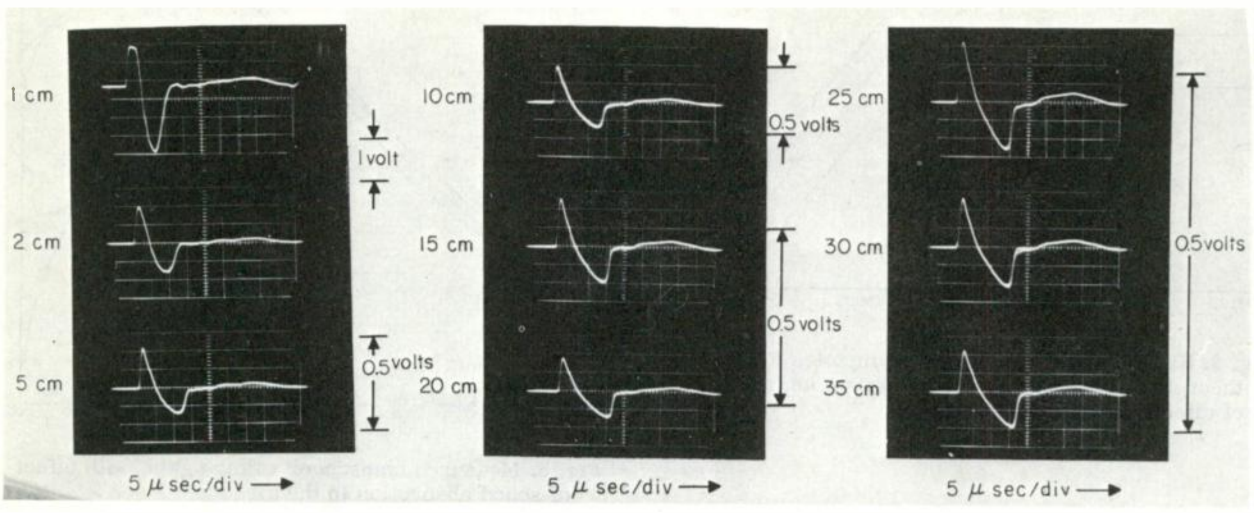


Fig. 7. Oscillograms showing the output of a high-frequency microphone at varying distance from a spark of length 0.5 cm; C=0.001μF, about 4 kv.

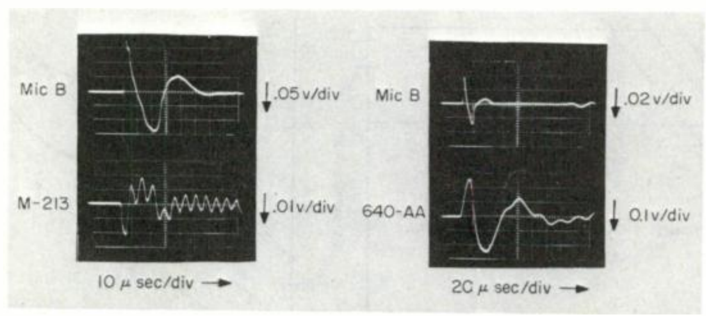


Fig. 8. Oscillograms comparing the response to the acoustic impulse from a spark of a high-frequency electrostatic microphone (Mic B) and two commercial microphones.

THIN-FILM CIRCUIT PACKAGING AND THE NUMBERS GAME

Arthur R. Meehan
Light Military Electronics Department
General Electric Company
Utica, New York

Summary

By neglecting to define their terms, some electronics manufacturers have played "the numbers game" in their claims to the customer. In discussing any form of micro systems electronics, claims of millions of parts per cubic foot have become commonplace. The rub is that it is seldom clear whether one is discussing the number of individual parts stacked on a shelf, the parts in one circuit, a module, or an actual operating equipment. Indeed, if complex assemblies were built using some of the parts density figures quoted, the equipment would become a molten mass in seconds.

With the thin-film technique as a frame of reference, the author indicates the kind of packaging densities which can be achieved realistically. He also identifies some of the parameters to be considered in using the thin film technique for digital equipment.

* * * * * *

Although some of the glamour has worn off micro electronics, the industry is still being flooded with claims that such and such a technique will give a 20, 50, or 100 to one improvement in size and weight over a conventional system. Claims of millions of parts per cubic foot have become almost commonplace. It is seldom clear, however, whether one is discussing a single circuit, a module, or an actual operating equipment.

If complex assemblies were actually fabricated using some of the parts density figures quoted, the equipment could become a molten mass in a matter of seconds. Before claims made for microcircuits can be realistically evaluated, one must first recognize that these circuit techniques can be used only for <u>low power dissipations</u> if the quoted packaging densities are even to be approached. Comparisons with "conventional" circuits are meaningful only if one knows what is meant by "conventional". Before examining what packaging densities can be realistically achieved using thin films and before discussing the parameters to be considered in constructing equipment using thin films, let us examine some typical densities of conventional assemblies.

G-E's Light Military Electronics Department's major interest in microsystems at present is in digital-computer applications. Because airborne digital-computer circuitry involves low power and is highly repetitive, it is felt that the first large-scale application of a microelectronics system should be in this area.

Many military digital computers use printed wiring boards (Figure 1) which are certainly a present-day technique. But in some of these applications, size and weight are not prime considerations. Therefore, the design is "loose", and considerable space could be considered wasted. If the microsystem is compared to this type of "loose" design, a large size reduction can be forecast; however, a redesign of the printed-board layout could also achieve a major size reduction.

Tight Design Lowers Ratio

Another conventional, present-day technique (Figure 2) makes use of small modules composed of densely packed components which are usually encapsulated. These modules are soldered or welded into a printed-wire board (mother board), or other interconnecting medium. This technique usually has a higher component density than conventional printed-wiring-board design. Obviously, any comparison of the microsystem with the modular assembly will give a much lower ratio than a comparison with the loose design.

With these cautions in mind, let us look at a typical missile computer and use it as a reference for a discussion of density. Using a neat, but loose, printed wiring board design, a packaging density of less than 4,000 parts per cubic foot would result. The actual production version of this unit, however, uses small modules soldered into mother boards, with approximately 14,000 parts per cubic foot.

The same computer has been redesigned to achieve the maximum component density practical with standard components, that is, one-eighth watt resistors and transistors in TO-18 cans. The packaging density of this design is 38,000 parts per cubic foot. It is the maximum possible density consistent with a feasible thermal design for this specific application.

Thus, by use of more critical and detailed design, component packing density has been increased about tenfold - from 4,000 parts per cubic foot to almost 40,000. And this has been done using conventional techniques. It is important to bear this in mind when discussing "im-

provements" using microsystem techniques.

A thin-film version of this computer will have a packaging density of 300,000 parts per cubic foot, which is about a 100-to-1 reduction over a loose printed board approach, but less than a ten-to-one improvement over an optimum conventional design.

Design Factors

Now that the density figures have been placed in proper perspective, let us turn our attention to the factors that must be analyzed to arrive at a satisfactory design. Among these factors are repairability, substrate size, substrate material, mounting required to resist shock and vibration, and manufacturability. Two of the more interesting ones--repairability and substrate size--will bear discussion in some detail.

What are the decisions that must be made to determine substrate size? Interconnection pattern, module throw-away value, the values of the deposited components, and thermal considerations all play an important part in the proper choice. The effect of the interconnection problem and the throw-away value on substrate size is illustrated by the following example.

For optimum connection reliability, it is desirable to vapor deposit as many connections within a substrate as possible. This practice yields a large substrate; however, the throw-away value and the predicted failure rate of the components must be carefully weighed to obtain the proper balance.

Unfortunately, no one has established a throw-away level which will meet all cases. But if we assume a 400-dollar limit as a reasonable value, a flip-flop which will require twelve input-output leads can be deposited on one substrate. If, however, the throw-away level were reduced so that four substrates had to be used to build the circuit, twenty-eight leads would be required. The interconnection problem, therefore, is considerably reduced by making the substrate as large as possible, provided that the factory yield is high and the throw-away value is acceptable.

Now let us examine circuit layout to see how placement of components can increase the overall density buildup. In the device shown in Figure 3 the semiconductor microdevices are all placed on one half of one surface of the substrate, and the deposited circuitry is placed on the other half. The portion of each substrate containing the microdevices is encapsulated. These split-level units can be nested so that the encapsulated area of one substrate rests against the deposited-component area of a second substrate. This arrangement works well if a flip-flop or double nor is the only circuit on a substrate.

Thermal Limitations

As more circuitry is added to the substrate, however, it becomes more difficult to arrange the active devices so that they "nest". Here again the mechanical packaging will depend on the throw-away and thermal limitations which are directly related to the amount and type of circuitry placed on any one substrate. The thermal problem can be alleviated by placing metal foil in the sandwich to improve heat flow.

The values of the deposited components and the thermal problem affect substrate size in that high-value resistors will require a large substrate area to permit adequate heat dissipation. We firmly believe that packing densities of several hundred thousand parts per cubic foot can be reasonably achieved while still maintaining a practical thermal design. We are somewhat skeptical, however, of the millions of parts claimed in some applications.

Many skeptics are quick to point out the thermal problem involved in reducing present equipment by several orders of magnitude. Part of what they say is true. Merely to shrink existing equipment by jamming parts closer together will most certainly result in serious thermal problems. But by redesigning circuits and by the generous use of lower power devices, we can overcome the thermal problems.

On the other hand, thermal-design problems cannot be solved unless the electrical and mechanical designers work as a closely integrated team. The smaller size and higher operating speed requirements of present-day military computers are increasing the heat problem and creating the need for other breakthroughs, particularly in thermoelectric cooling techniques. Unique mechanical designs and the increased use of new low-power devices enable the design team to build a microsystem equipment today and not exceed the thermal barrier.

One technique that can improve heat-transfer characteristics is the proper selection of a potting compound after a trade-off consideration is made between the thermal conductivity of the material and the weight penalty attached to the higher conductivity materials. For example, thermal conductivity can be increased as much as six times by the choice of a highly filled material rather than an unfilled material. As the thermal conductivity increases, however, the density will also increase. The exact effect of filling on density and thermal conductivity will depend, of course, on the resin system chosen.

Although in many applications an adequate supply of cooling air for the equip-

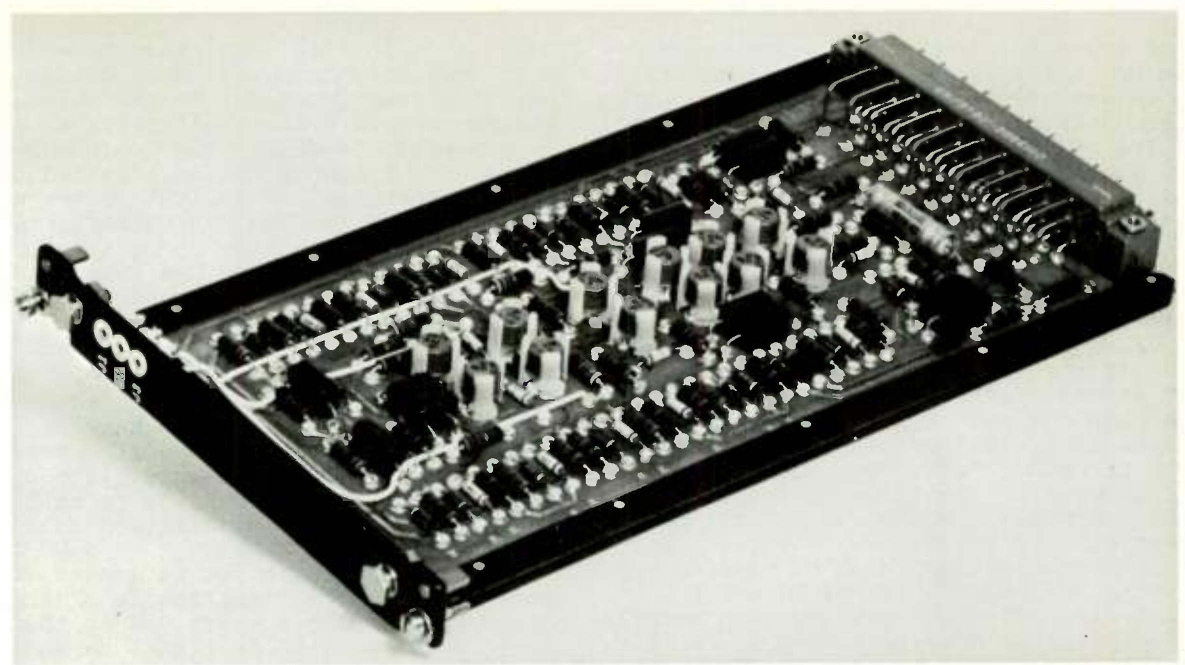


Fig. 1. Printed wiring board.

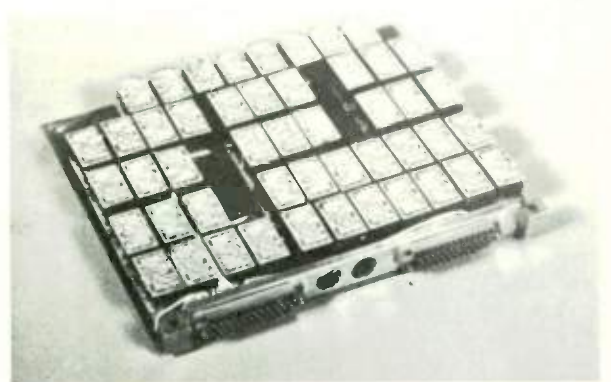


Fig. 2. Modular assembly.

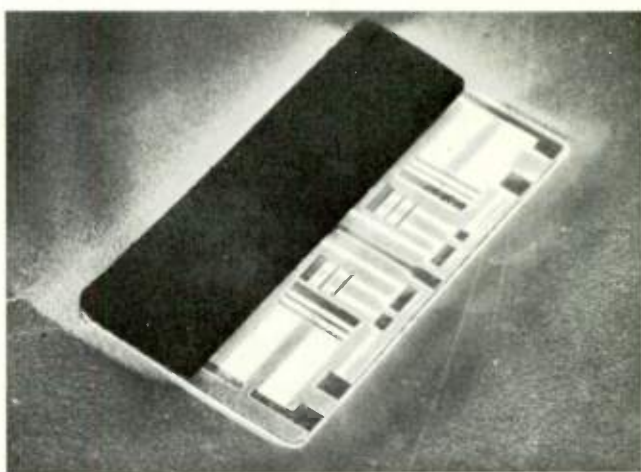


Fig. 3. Nested components.

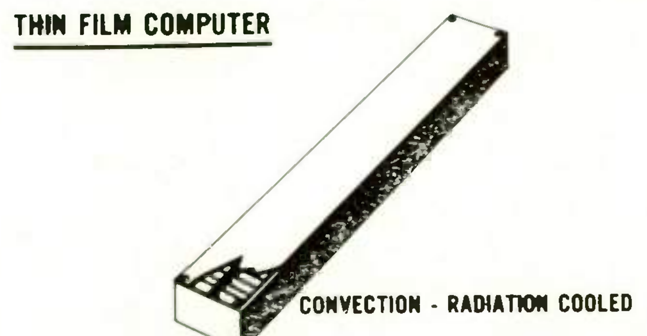


Fig. 4.

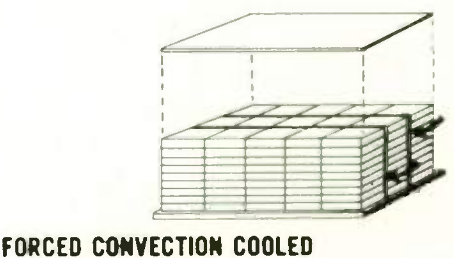


Fig. 5.

ment is available, more and more applications are depending entirely upon conduction or radiation for heat dissipation. Let us examine two approaches to packaging an equipment; the first without forced air cooling and the second with such cooling.

The first unit, shown in Figure 4, is similar to a 35-mm slide case. The substrate size is specified as 0.5 x 0.5 x 0.01 inch. After fabrication, the substrates are clipped into the box and held at two edges. Interconnections are made in the bottom of the equipment.

Heat to the case is transferred by conduction as well as by natural convection and radiation. The entire unit occupies slightly more than 19 cubic inches and has an overall density of 180,000 parts per cubic foot. The package is capable of dissipating 25 watts while maintaining a maximum component operating temperature of 104 C.

The temperature could be reduced by lowering the power level or by using forced-air or liquid cooling. For example, the density could be increased almost five times if the approach shown in Figure 5 were used.

In this approach, the circuit layout has been changed, and a substrate size of 0.5 x 0.34 x 0.1 inch has been used. After the substrates are processed and encapsulated, they are stacked and interconnected in three rows of five stacks each. The package size including the interconnection-matrix base plate is 2.6 x 1.1 x 0.94 inches. A spacing of 0.050 inch has been left between the inner row and the outer rows. Air at a rate of 0.66 ft^3/minute is blown down the two 0.75 x 0.050-inch channels created by the stacks, interconnection plate, and a cover plate. This air cools the middle row and part of the outer rows; the remaining heat is removed from the outer rows by radiation and natural convection. The maximum component temperature to be expected from this design is 93 C.

Repairability

Now let's take a brief look at repairability. For optimum densities the concept of nonrepairable substrates or of modules of several substrates is accepted. This does not, however, imply that a repairable package cannot be assembled. Figure 6 shows a three-bit parallel adder in a repairable form. The one hundred fourteen components are distributed over eight substrates. These substrates are placed on a thin sheet of tetrafluorethylene, which contains the interconnection pattern; connections are made to the substrate by soldering or thermal compression bonding. This assembly, with its micro connection at one end, is folded into a "W" and inserted into a flat can. No encapsulation is used. The micro semiconductor devices may be removed and replaced several times before the conductors break from repeated flexing. In repairable assemblies built in this fashion, the density is only about half that of a unit constructed in a nonrepairable form.

This discussion has attempted to shed some light on the significance of the component-density numbers and to indicate some of the factors that must be considered in achieving reasonable thin-film component densities. Although many problems are associated with the design of any microsystem equipment, they can usually be solved by a well-integrated design team of electrical and mechanical engineers, working closely with manufacturing personnel. Such teams are almost daily finding new ways to manufacture smaller and more reliable military equipment.

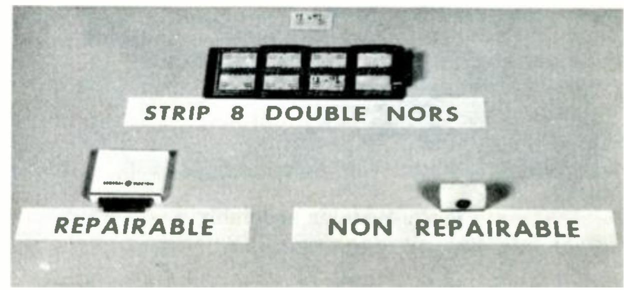


Fig. 6. Three-bit parallel adder.

THERMAL CHARACTERISTICS OF POTTED ELECTRONIC MODULES

J. I. Gonzales and C. E. Waugh
Orlando Aerospace Division
Martin Marietta Corporation
Orlando, Florida

Summary

This paper discusses the correlation and development of an analytical model for determining thermal transients in potted electronic modules.

The primary objective is to develop an analytical technique that can be used by the design engineer to predict thermal transients. By means of such an analysis it is hoped he can rapidly and efficiently (with a minimum of tests) design a thermally satisfactory and thus reliable potted electronic module.

The prototype module used in initial experiments had a high concentration of heat in a limited volume and thus indicated the need of efficient heat removal. Among the methods considered were direct forced air cooling, finned heat exchangers, liquid cooling and flat plate heat exchangers. Due to the size and shape of the module and space limitations, direct air cooling was used in conjunction with a flat plate heat exchanger (embedded heat sink).

Module heat loads and internal components were simulated during the experimental investigations because actual internal components were not available. These tests provided data for analytical correlation as well as background on different techniques for mounting the heat load simulators. Extensive tests were conducted which investigated the effects of air paths and variations in percentage of silica filler.

Nomenclature

Symbol	Description	Units
ρ	density	$\#/Ft^3$
C	specific heat	BTU/#-°F
V	volume	Ft^3
T	temperature	°F
θ	time	Hr
Δ	an increment	None
h	heat transfer coefficient	BTU/Hr-Ft^2-°F
A	area	Ft^2
K	thermal conductivity	BTU/Hr-Ft-°F
X	distance between node centroids	Ft
Q	heat transfer rate	BTU/Hr
WA	write amplifier	None
r	resistor	None
t	transistor	None

Subscripts

1, 2, 3, ...	integers for nodes
a	ambient air
s	heat sink
pc	potting compound
BS	back side of module
RS	right side of module
LS	left side of module
B	bottom of module
F	front of module
TS	top side of module
i	internal component
R	radiation

Introduction

As the trend in electronic packaging continues toward miniaturization, component temperature problems become more prevalent. This trend is quite apparent when one considers the fact that the packaging density is rapidly increasing. There are two factors working towards generating these component temperature problems:

1. Decreased volume or size of component with its own internal heat dissipation.

2. Increased power dissipation of additional components.

To assure component reliability and satisfactory operation, there is a definite need for thermal analytical prediction techniques of microminiaturized components. This need for analytical prediction techniques can be readily appreciated if one considers a basic rule-of-thumb used by electronic packaging engineers[1]. This rule-of-thumb states that semiconductor failure rates are halved for each 18°F reduction in junction operating temperatures below the maximum. The capability of evaluating the effects of such characteristics or parameters as noted below is needed:

Number of nodes required in analytical model.
One, two and three dimensional heat transfer effects.
Potting compound thermophysical properties.
Ambient air temperature.
Heat transfer coefficient.
Internal heat dissipation.

The method of analysis as discussed herein concerns itself with the prediction of thermal transients in potted electronic modules. A technique is presented and results obtained which are compared with experimental data. The effects of varying such parameters as noted above are investigated and discussed.

Discussion

Background - Problem Statement

With all electronic equipment, there is associated a certain amount of energy which evidences itself as heat. During early aerospace applications, heat dissipation did not result in excessive component temperatures since packaging densities used were lower than those presently employed. This larger volume generally allowed the transfer of heat by natural means; i.e., convection, conduction and/or radiation.

Now, however, the advent of microminiaturization has brought about a marked increase in packaging density. This reduction in volume, together with its associated heat dissipation, brings rise to the apparent problem of excessive component temperatures. Reference 2, for example, discusses some of the thermal limitations imposed by power dissipation and packaging density.

A knowledge of when, where and how this excessive component temperature occurs is very necessary if one is to adequately describe the thermal characteristics of a component and assure its operational reliability. Unless the thermal characteristics of a component are known, then the magnitude of the problem is undefined. Having defined these thermal characteristics, the designer can then evaluate the magnitude of the problem and make recommendations for alleviating same while the design is still in the conceptual stages.

For the case of potted electronic modules, the analytical technique should be capable of investigating variations and/or effects of the following:

Number of nodes required in analytical model.
One, two and three dimensional heat transfer effects.
Potting compound thermophysical properties.
Ambient air temperature.
Heat transfer coefficient.
Internal heat dissipation.

The study as outlined herein, discusses the steps taken in evolving an analytical technique that is capable of considering the effect of some of the above variables of parameters.

Analysis and Correlation

The analytical approach used was basically that of dividing the module into discrete nodes and writing a heat balance relation for each node. These relations were then solved as a function of time and correlated with test data.

Nodal Breakdown. The first problem encountered in the analysis is that of defining the nodal breakdown for the mathematical model. It is apparent that the desired objective is to achieve a minimum number of nodes for analytical simplicity. However, since the variation in temperature within a node should theoretically be zero (i.e. node should be isothermal) care must be exercised in defining both the size and location of each node with respect to internal components.

Early studies indicated that the module designs being considered would require forced cooling. This is evident when one examines the fact that the wattage density of the typical module was 3.75 watts/in^3. Thus, the nodal breakdowns favored were those which provided the most simplicity and correlated best with forced convection conditions.

Background data, not shown herein, indicated that the gradients resulting along the heat sink were negligible (due to the high thermal conductivity of aluminum). Thus, each heat sink was considered as an individual node in its entirety. Additional studies also showed that nodal breakdowns exceeding 12 in number did not result in any appreciable improvement in design accuracy.

Figure 1 shows the variation in temperature (analytical and experimental) for typical nodes between a 10 and 12 nodal breakdown. The configuration details of these breakdowns (size, location of nodes and internal components) appear in Figure 2. Note that the data of Figure 1 are for forced cooling (h = 4) and with module ambient air varying from 130 to 90°F. The data for initial time zero correspond to steady state conditions resulting from free convection (h = 1). A comparison of Figures 1 (a) and (b) shows that the 10 nodal breakdown analytical data correlates best with experimental results. This comparison was based on typical nodes whose locations were closest to one another, i.e. nodes 1, 2 and 4 of the 10 nodal breakdown compare with nodes 1', 3' and 6' of the 12 nodal breakdown. The maximum variation in the 10 node breakdown results in the heat sink node 1 and is on the order of 10%. The data emphasize a basic point, that being, that the criteria for a nodal breakdown is not necessarily that of increasing the number of nodes in the analytical model.

The symmetry (location of resistors and transistor) of the Write Amplifier module allowed both the 10 and 12 node breakdowns to be analyzed with 6 and 8 nodes, respectively. This symmetry occurs on either side of the centerlines (℄) shown in Figure 2. Thus, by means of this symmetry an adiabatic condition is assumed to exist to the right of the centerline.

To avoid confusion, further reference in the data that follow is made to the nodal callouts and locations of Figure 1 (a).

One, Two and Three Dimensional Heat Transfer Effects. Consider the top view of Figure 2 as containing the X and Y dimension while the elevation and side views contain the Z and X and Z and Y dimensions, respectively.

Early studies showed that for the typical module being investigated there were no appreciable temperature gradients along the Z (vertical)

axis. This resulted because the heat dissipating components, resistors and transistors, extend practically the entire height of the module. Thus, for this type module and arrangement of heat generating components, the transfer of heat by conduction in this direction was not present.

However, the transfer of heat by conduction in the X and Y directions is of enough magnitude that it must be considered. Evidence of this can be seen in the thermal transient plots of Figures 1, 3, 4, 5, 8, 9, and 10. Also, the transfer of heat by radiation (externally) in all dimensions for the conditions investigated, constituted a small part (8% and less) of the total and was neglected in final analysis. The major mode of transfer externally was convection. Studies omitting the convection mode were found to be in significant error.

Thus, the analytical model used in this study considered the following dimensional transfer effects:

a) three dimensional convection.
b) two dimensional conduction.
c) no radiation.

Heat Balance Relations. Heat balance relations for the analytical model were written in as general a form as possible. A typical relation for node 1 of the 10 node breakdown of Figure 2 is given below:

$$\left[\rho_1 C_{s_1} V_{s_1}\right]\frac{\Delta T_1}{\Delta\theta} = \frac{K_{pc_2}}{X_{1-2}} A_{BS_2}\left[T_2 - T_1\right] + \frac{K_{pc_3}}{X_{1-3}} A_{BS_3}\left[T_3 - T_1\right] - h_{LS}A_{LS_1}\left[T_1 - T_a\right] - h_{BS}A_{BS_1}\left[T_1 - T_a\right] - h_T A_{T_1}\left[T_1 - T_a\right] - h_B A_{B_1}\left[T_1 - T_a\right] - Q_R$$

Relations similar to the above were written for each node in the module and varied depending upon the module configuration and type of nodal breakdown. The general capability of the above and similar relations is quite flexible. For example, the capability of considering a heat sink and its varying thermophysical properties can be accomplished since the relations include this term. Also, there is the capability of varying the magnitude of the internal heat dissipation, potting compound thermophysical properties, external heat transfer coefficient, ambient air temperature, areas and volumes. Thus, in this manner, a general parametric investigation can be made with comparative ease by varying or omitting any of the above parameters as desired or necessary. Examining each of the terms in the above relation in detail will point out the flexibility of the analytical model. The bracketed term on the left hand side of the above relation is the heat storage in node 1. Refer to Figure 2 for layout of the typical module and node breakdown. Node 1 in this case is the aluminum heat sink. The heat storage term for the heat sink is then multiplied by the ratio of incremental temperature and incremental time. The first term on the right hand side is the transfer by conduction from node 2 to node 1. The second is the transfer by conduction from node 3. Keeping in mind the symmetrical breakdown of the module allows the assumption of an adiabatic condition at the centerline. The next four terms define the transfer by convection from the left, back, top and bottom sides of node 1. The final term is used to account for radiation transfer to or from the node (in this study this was found to be negligible). The remaining relations for the 10 node model are similar to the node 1 relation varying only slightly in the conduction and/or convection paths considered as well as the inclusion of heat dissipation when it occurs. Relations similar to those given above were used to generate the analytical data presented herein. Detailed discussions of the numerical procedures used to solve these type relations appear in References 3, 4, and 5.

Effect of Per Cent Filler in Potting Compound. As the per cent filler (silica powder) in the epoxy potting compound was varied from 0 to 50%, the basic thermophysical properties changed as follows:

Thermophysical Properties		Per Cent Filler 0	Per Cent Filler 50
Thermal conductivity	- BTU/Hr-Ft-°F	0.109	0.273
Specific heat	- BTU/#-°F	0.27	0.20
Density	- #/Ft^3	70.5	108

Figure 3 shows the difference in thermal performance between the use of 0 and 50% filler, respectively, for typical nodes. It is apparent from a review of these analytical results that the epoxy with the 50% filler is more efficient thermally. A 50% filler concentration was the practical limit from a viscosity or flow property aspect. The thermal efficiency is based on the fact that the heat generated internally is transferred more effectively from the inner nodes, where the heat is dissipated, to the external surfaces and finally to the heat sink (s) and also results in smaller spreads in temperatures.

For a detailed discussion of potting compounds and their properties see References 6 and 7.

Effect of Ambient Air Temperature and Heat Transfer Coefficient. The net effect of varying the ambient air temperature while keeping the heat transfer coefficient (h) constant at typical values of 4 and 50 is essentially that of displacing the nodal temperature distribution. Comparing the h = 4 values of Figure 4 (a) with those of Figure 4 (b) will show that the only difference is their displacement. The data of Figure 4 (a) start at an ambient air temperature of -30°F and reach a maximum of 198°F, giving a maximum increase of 228°F. The data of Figure 4 (b) start at 130°F and reach a maximum of 358°F, also giving a maximum increase of 228°F.

The éffects of varying the heat transfer coefficient are readily apparent in Figures 4 (a) and 4 (b). As the external heat transfer coefficient is increased, the thermal resistance between the external module surface and ambient air sink is reduced. This reduction in external thermal resistance has a twofold effect:

a) The response time of each nodal temperature is reduced. Note the short time required to stabilize with an h of 50 as opposed to that with an h of 4. This reduction in response is about 75%.

b) The net rise in nodal temperature is decreased with an increase in h. See Figure 4 (b) and compare the rise in temperature for node 4 with an h of 4 and 50. The net rise in temperature in this case has been reduced by 80%.

The heat transfer coefficients used in the analysis were evaluated as per Reference 8. For the case where free convection exists, the value of h was calculated to be 0.88 BTU/Hr-Ft^2-°F. Basic data used for this evaluation required a knowledge of the airgap between columns (↑) and rows (→) of modules which was 0.4 and 0.1 inch respectively.

In the case of forced convection, the calculated value of an h = 4 resulted with an airflow volume rate of about 4 cfm.

Effect of Internal Heat Dissipation. The typical module investigated was a Write Amplifier (WA) which is an integral part of a memory unit in a data computer. The heat dissipating resistors, r, (see Figure 2) were attached to fuse clips which were in turn welded to the aluminum heat sink. The transistor, t, in the typical module was attached to a circular transistor clip (IERC type) which was also welded to the aluminum heat sink.

The heat dissipation of each resistor was 1 watt and that of the transistor 0.9 watt. Thus, the total heat dissipated by the typical module was 3.9 watts.

Figure 5 shows the effect of variation in internal heat generation or dissipation in free convection. The thermal transients shown as solid lines are for a module whose total heat dissipation was 3.9 watts while those with broken lines applied to a module dissipating 7.8 watts. Thus, increasing the module heat dissipation by 100% increased the nodal thermal transients by 30 to 52%. The nodes showing the largest increase in temperature (46 to 52%) were the internal ones which contained heat dissipating components.

Experimental Investigations

Experimental Setup. Tests were conducted on an actual size model of a typical memory tray. Resistors were potted within blocks of epoxy (hysol #6020) which were equal in size to the actual production modules being investigated. The resistor values were chosen to dissipate the same amount of heat as the actual circuit. Separate resistors were used to simulate the heat dissipation of the transistor, emitter resistor, collector resistor and heat from other components. These blocks or module simulators were then located on a tray (see Figure 6) in a prescribed order as required by the circuitry.

Details of the test setup appear in Figure 7. The insert of Figure 7 is a schematic showing internal details of the module tray. Instrumentation included thermocouples (28 gage iron constantan) which were mounted inside the modules, on the heat sinks, in the air paths between the modules and in the tray air inlet and outlet. Module ambient air temperature was obtained by averaging four pickups along the air flow path. Temperature readings were recorded for tests with and without forced air cooling, for filled and pure epoxy and for various types of heat sink arrangements.

The module which was considered to be the most desirable for detailed testing and was used as the analytical model was the Write Amplifier (WA). The WA module dimensions are 1.0" x 0.8" x 1.3" and it dissipated 3.9 watts which resulted in a wattage density of 3.75 watts/in^3.

Experimental Results. Early tests utilizing modules made of pure unfilled epoxy indicated poor thermal conductivity. This was such that the effective removal of the heat dissipated by the internal components could not be accomplished even with forced convection resulting from an airflow volume rate of 4 cfm. During free convection tests, the maximum internal temperature measured in the WA module reached 308°F. This temperature was sufficient to soften unfilled epoxy and destroy the embedded transistor (silicon).

The next step was that of evaluating different types of fillers. This was done to obtain an improvement in the over-all thermal performance of the module. As a result, a 50% concentration of silica filler was chosen. Although an increase in filler concentration would further improve the thermal conductivity of the compound, excessive viscosity becomes a practical limitation.

Tests utilizing modules made of 50% filled epoxy indicated an increase in the thermal conductivity over the unfilled epoxy by a factor of 2.5. The data of Figure 8 compare the thermal performance between an unfilled and a 50% filled epoxy module. Note the closer grouping of temperatures and generally cooler performance of the module. This is particularly true with forced convection. Further improvements in heat transfer characteristics were obtained with the use of embedded aluminum heat sinks. Figure 9 compares typical thermal transients for a module with one and two heat sinks. Note that the module with two heat sinks tends to minimize, as one might expect, the spread in temperatures. The basic advantage of the dual sinks is to provide a low

thermal resistance for internally dissipated heat between its source and the forced convection air-flow path. This is particularly desirable in areas of high heat dissipation.

Additional techniques were investigated for improving the transfer of dissipated heat from the resistors to the external module surfaces and heat sinks. These techniques consisted of wrapping the resistors in aluminum foil as well as mounting them on fuse clips which were riveted (in the final production module these were welded) to the heat sinks.

Figure 10 shows the effect of wrapping the heat dissipating resistors in aluminum foil. The solid lines show thermal transients for typical nodes in a module wherein the resistors were wrapped in aluminum foil. The broken lines show thermal transients for the same nodes in a module where resistors were not wrapped in foil. The effect of reducing thermal resistance and increasing the heat dissipation area of the resistor is evidenced by the lower thermal transients.

Conclusions

Every attempt has been made to generalize conclusions as much as possible. It must be emphasized, however, that in the majority the conclusions given herein apply primarily to the WA module discussed in the body of the paper. Thus, depending on the configuration and function of other potted electronic modules the conclusions could agree or vary markedly as the case may be.

a) A total nodal breakdown of 10 gives adequate results for engineering design purposes (accuracies of the order of 10%). Every effort should be made to achieve symmetry in the location of heat dissipating components. In the case of the 10 node system, the analytical model reduces to 6 nodes.

b) The nodal breakdown should favor that part of module which contains heat dissipating components.

c) Due to the length of heat dissipating components as compared to module height, the analytical model was simplified by only accounting for two dimensional conduction effects.

d) Three dimensional radiation effects were found to be negligible - particularly in the case of 50% filler.

e) An epoxy with 50% silica powder filler best satisfied both thermal and viscosity effects.

f) A variation in ambient air temperature will result in the same net displacement of temperature transients.

g) Increasing the external heat transfer coefficient by a factor of 10 reduces response time by 75% and limits rise in nodal temperatures by about 80%.

h) Increasing module heat dissipation by 100% results in a maximum increase of nodal thermal transients of only 50%.

i) Wrapping resistors in aluminum foil reduces maximum internal temperatures by about 6%.

References

1. "Semiconductor Cooling - Theory and Practice," T. D. Coe, Wakefield Engineering, Incorporated, 1961.

2. "Circuit Considerations Relating to Microelectronics," J. J. Suran, Proceedings of the IRE, February 1961.

3. "Numerical Methods for Transient Heat Flow," G. M. Dusinberre, Trans. ASME, Vol. 67, p. 703, November 1945.

4. "The Solution of Transient Heat Conduction Problems by Finite Differences," G. A. Hawkins and J. T. Agnew, Purdue University Engineering Bulletin, Research Series 98, March 1947.

5. "Transient Heat Flow," R. P. Benedict, Electro-Technology, pp. 94-112, December 1961.

6. "Thermally Conductive Cast-Resin Compounds for Heat Dissipation," C. A. Harper, Electro-Technology, pp. 148-152, April 1961.

7. "Potting, Embedment and Encapsulation," C. G. Clark, Space/Aeronautics, pp. 25-30, December 1961.

8. Introduction to Heat Transfer, A. I. Brown and S. M. Marco, McGraw-Hill Book Company, Incorporated, 1951.

The authors would like to gratefully acknowledge the assistance given them by personnel in the Orlando Aerospace Division Applied Mathematics Section. The analytical programming effort of Mr. W. M. Howell, Jr. is particularly appreciated.

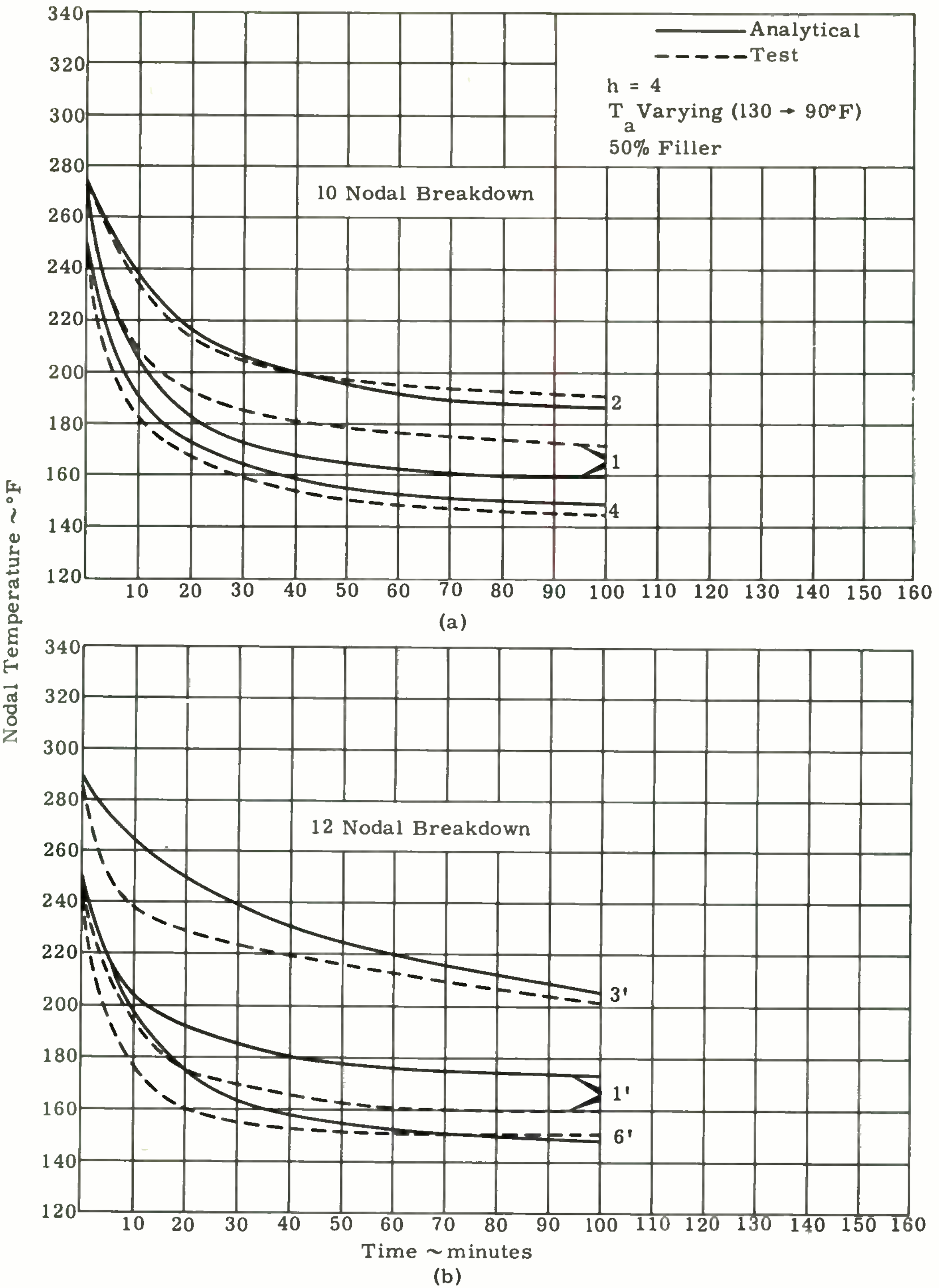
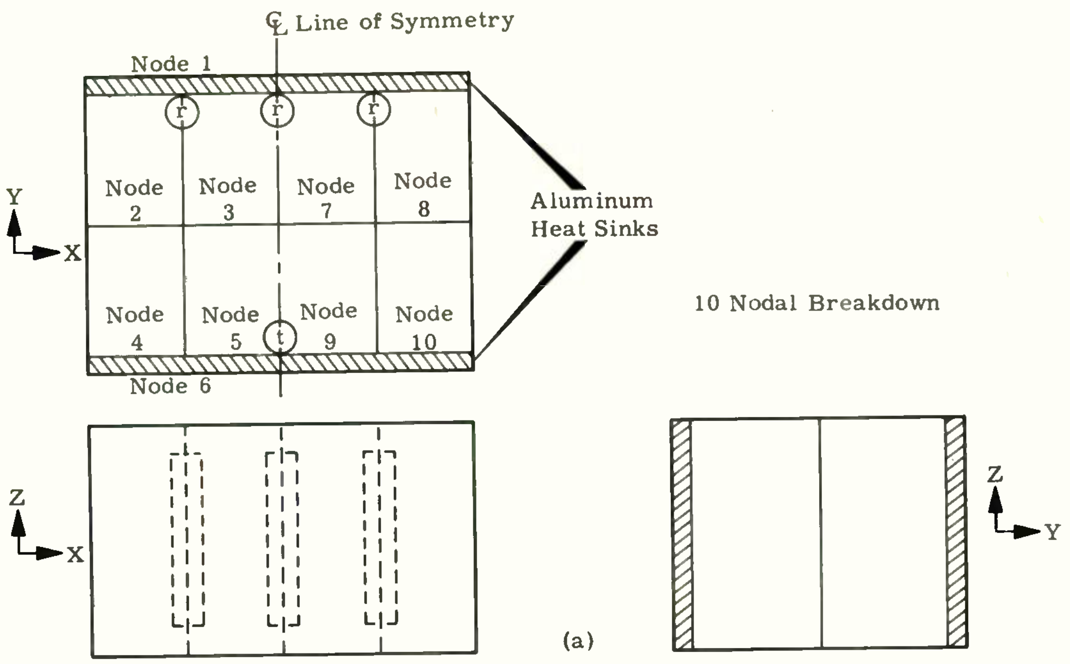
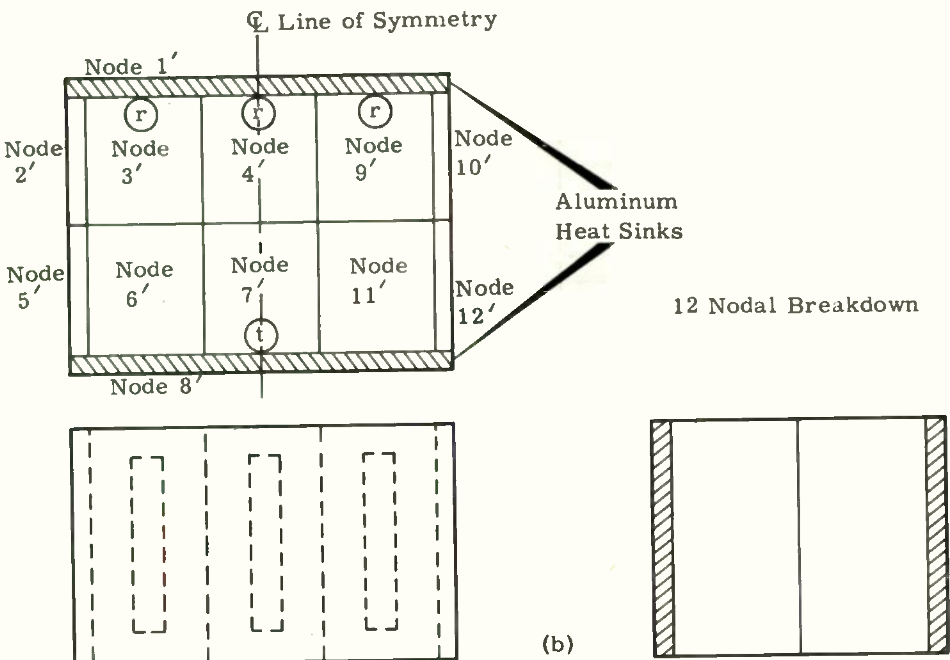


Figure 1. Effect of Nodal Breakdown
(See Figure 2. for Nodal Locations)



(a)



(b)

Figure 2. Nodal Breakdown Results

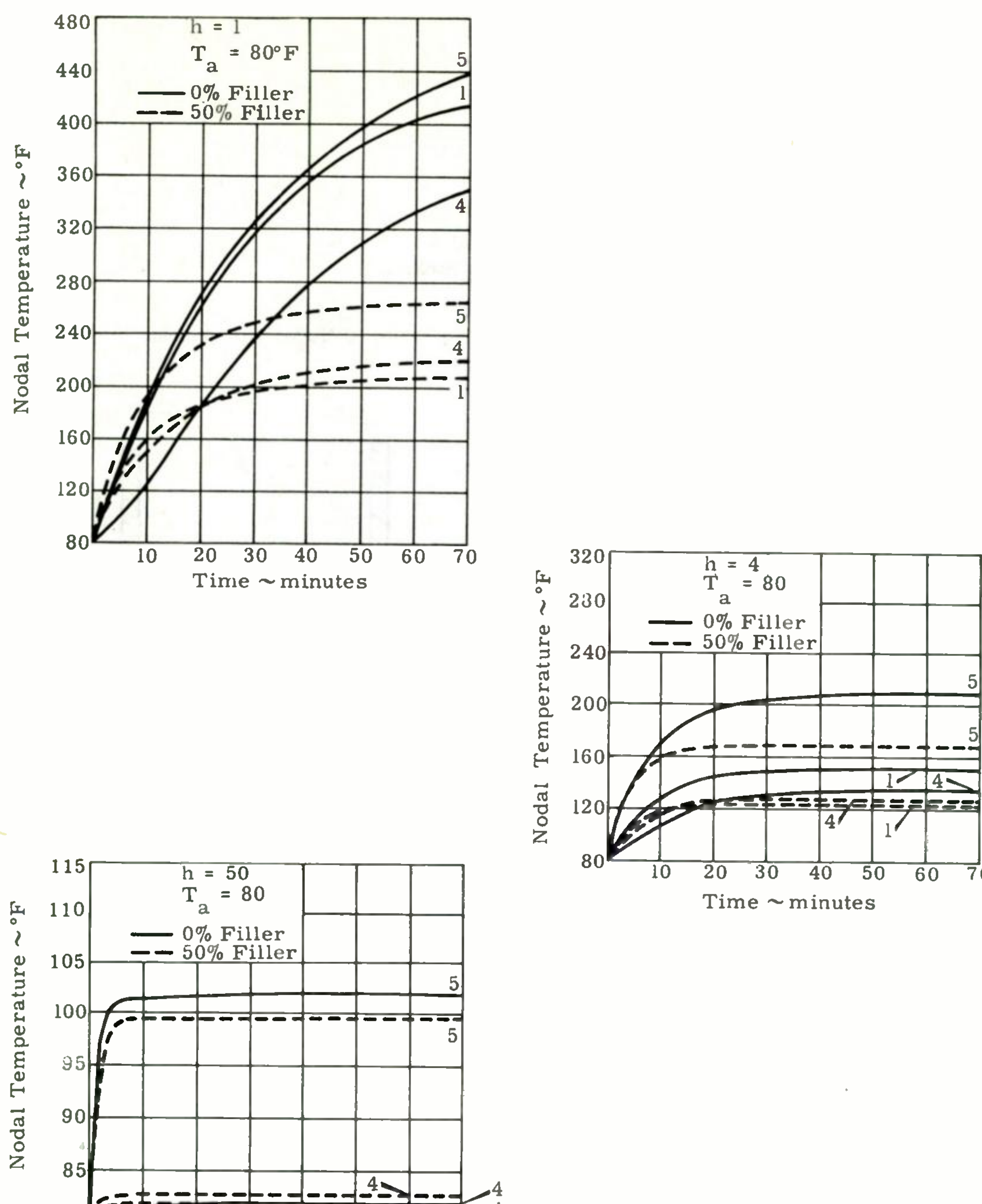


Figure 3. Effect of Percent Filler in Potting Compound

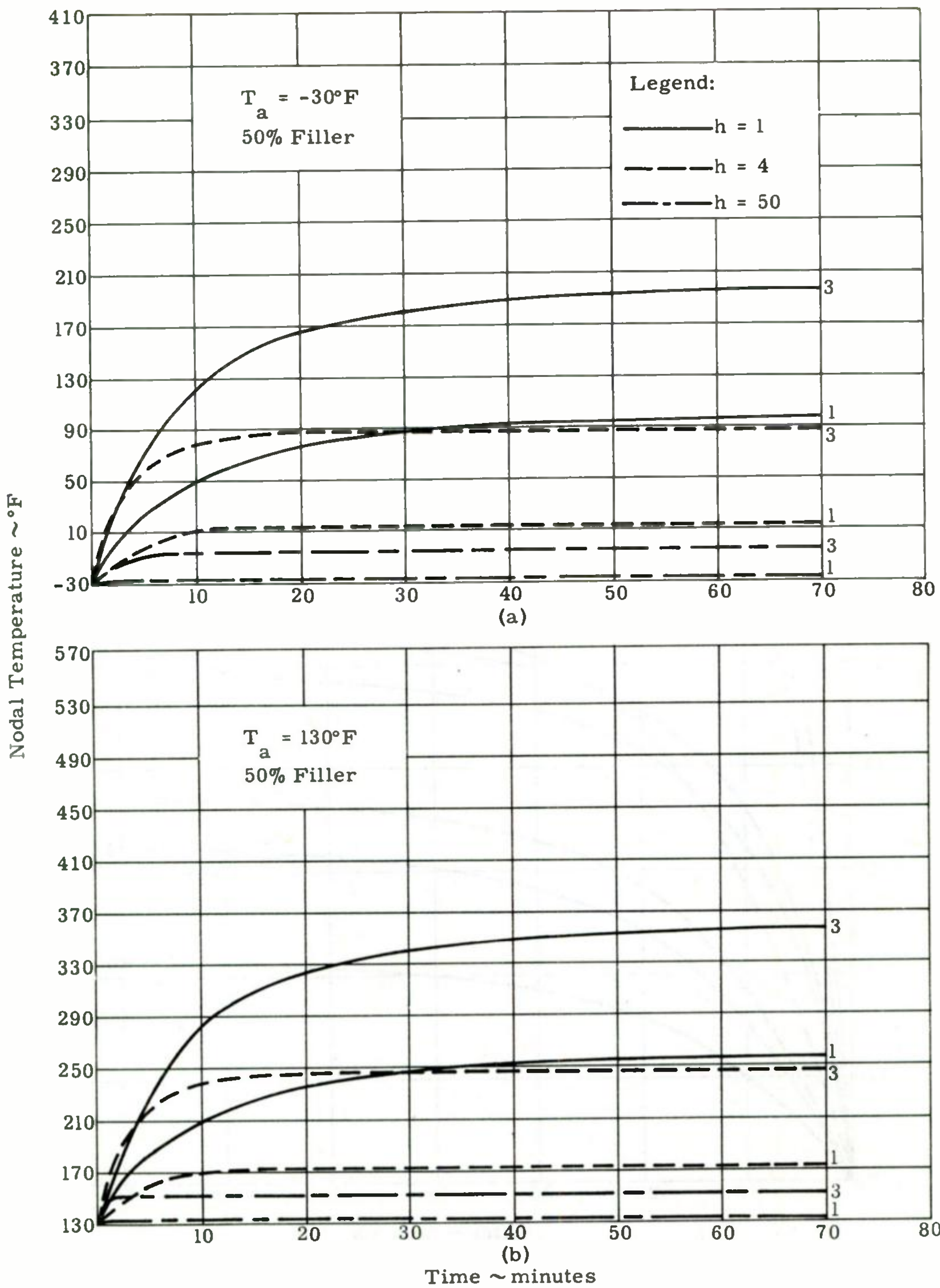


Figure 4. Effect of Ambient Air Temperature and Heat Transfer Coefficient

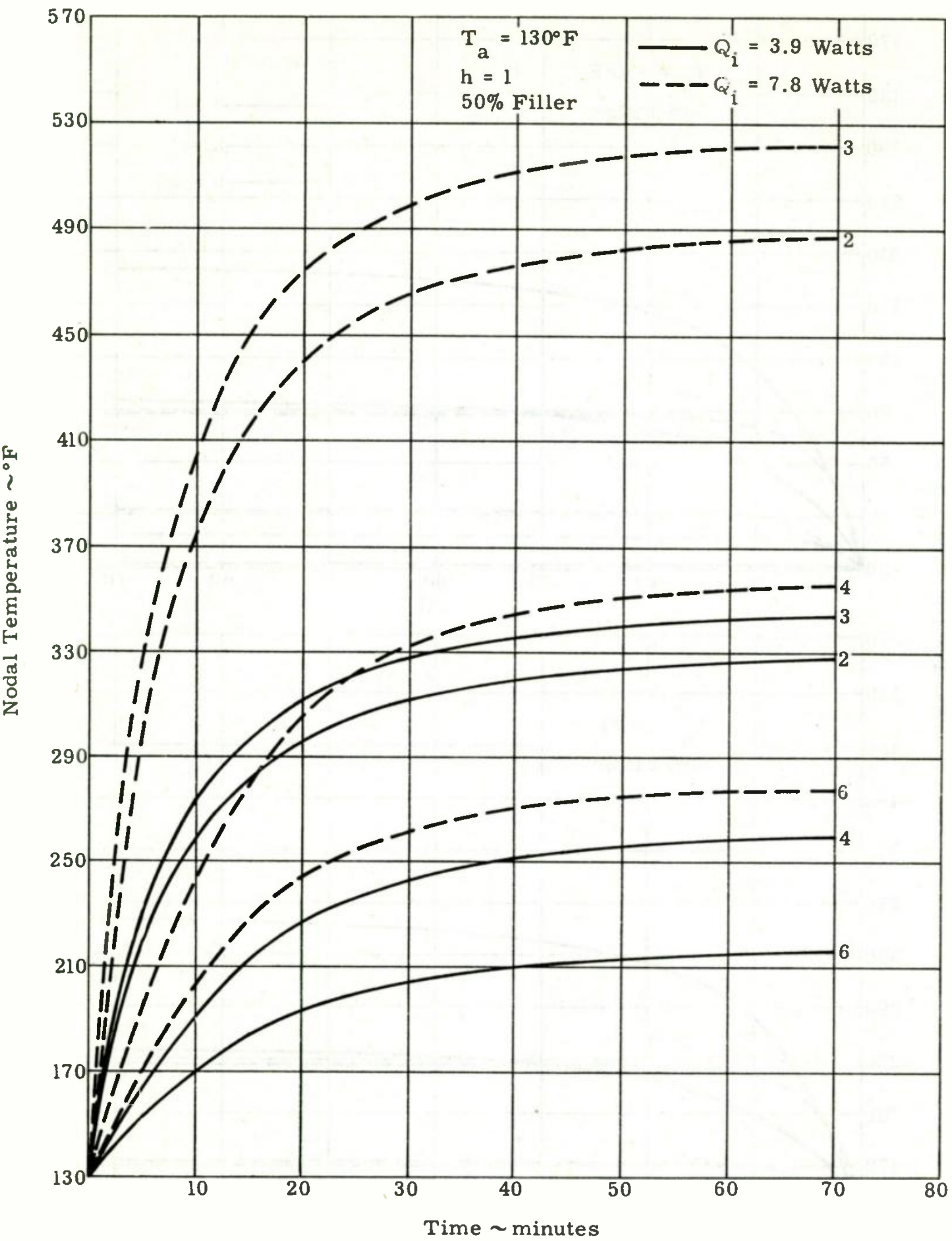


Figure 5. Effect of Internal Heat Dissipation

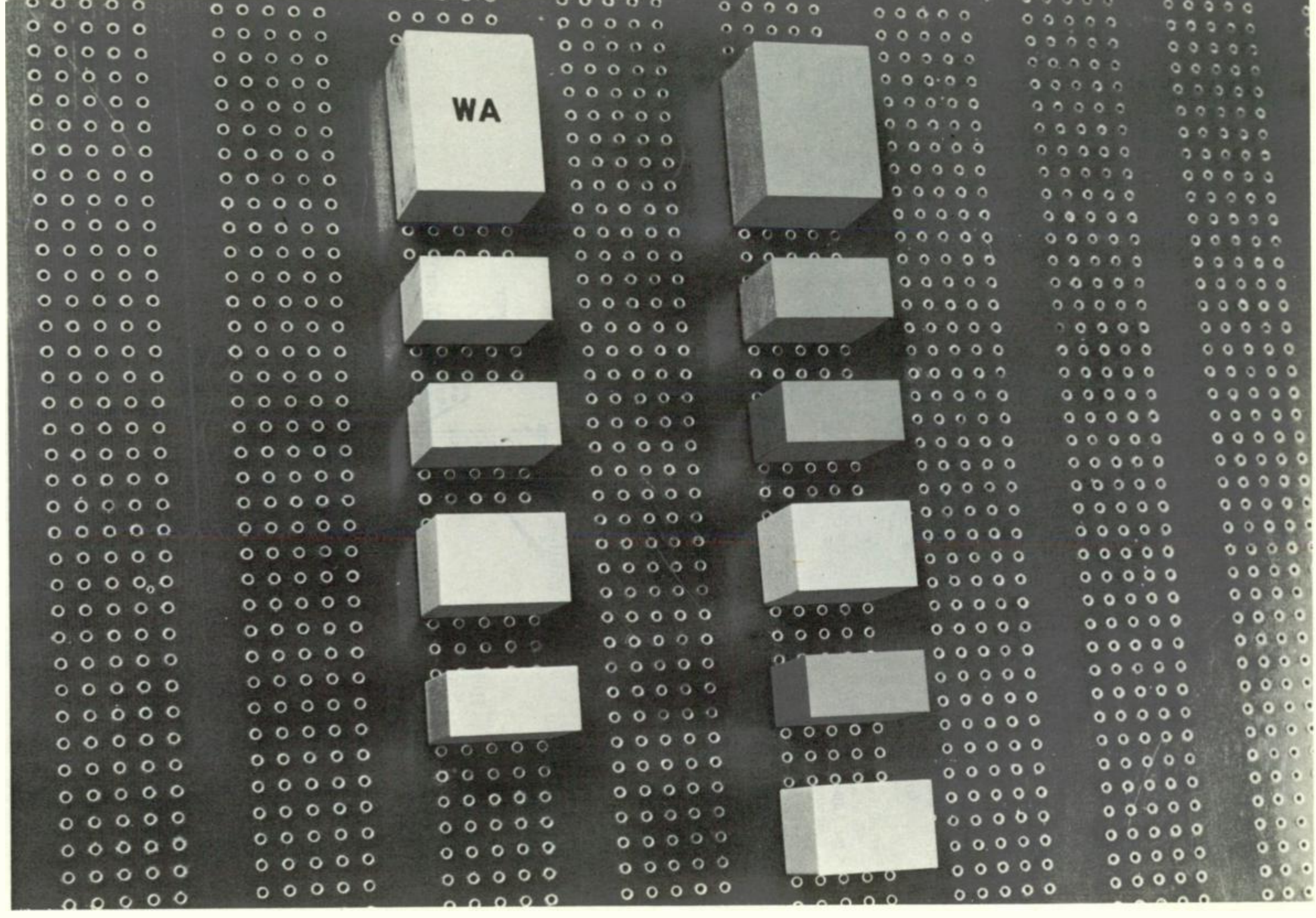
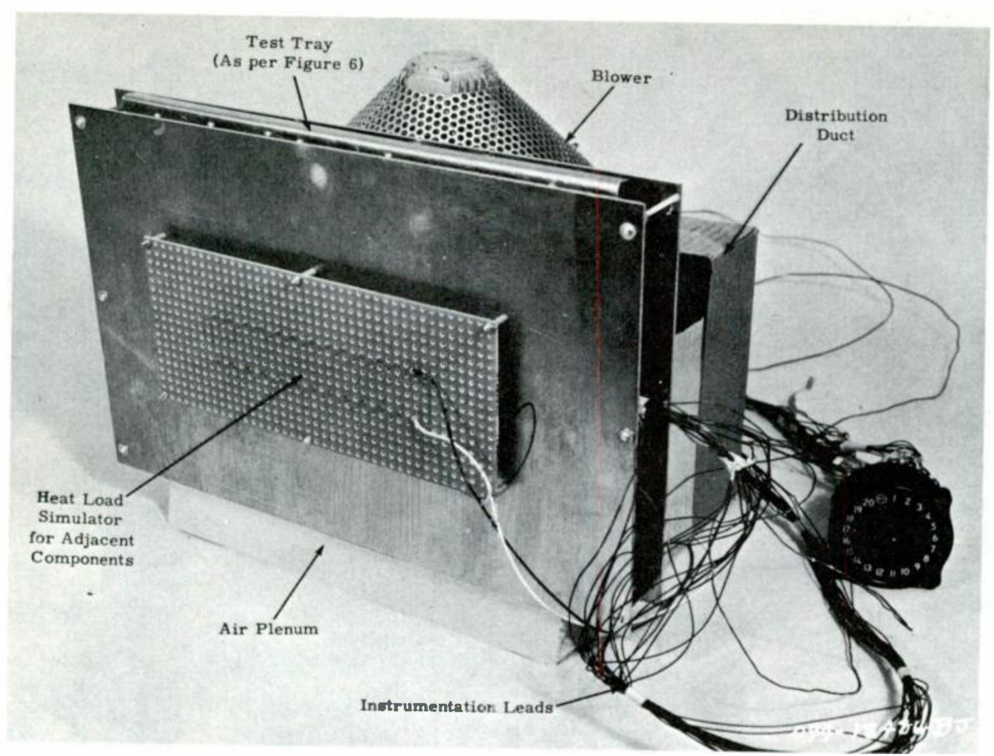


Fig. 6. Photo of module tray.



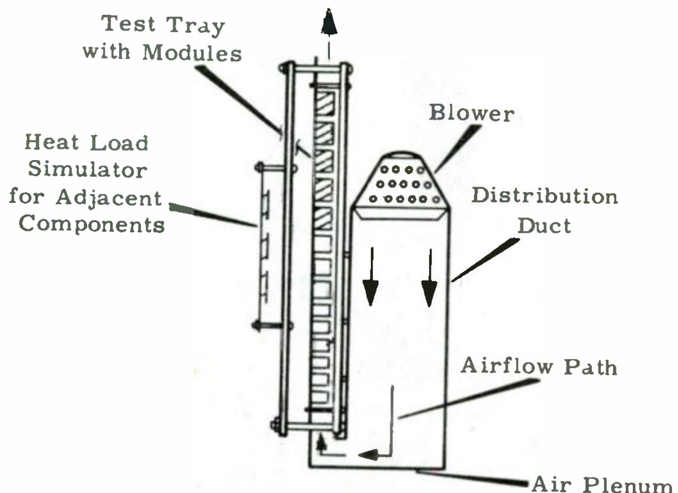
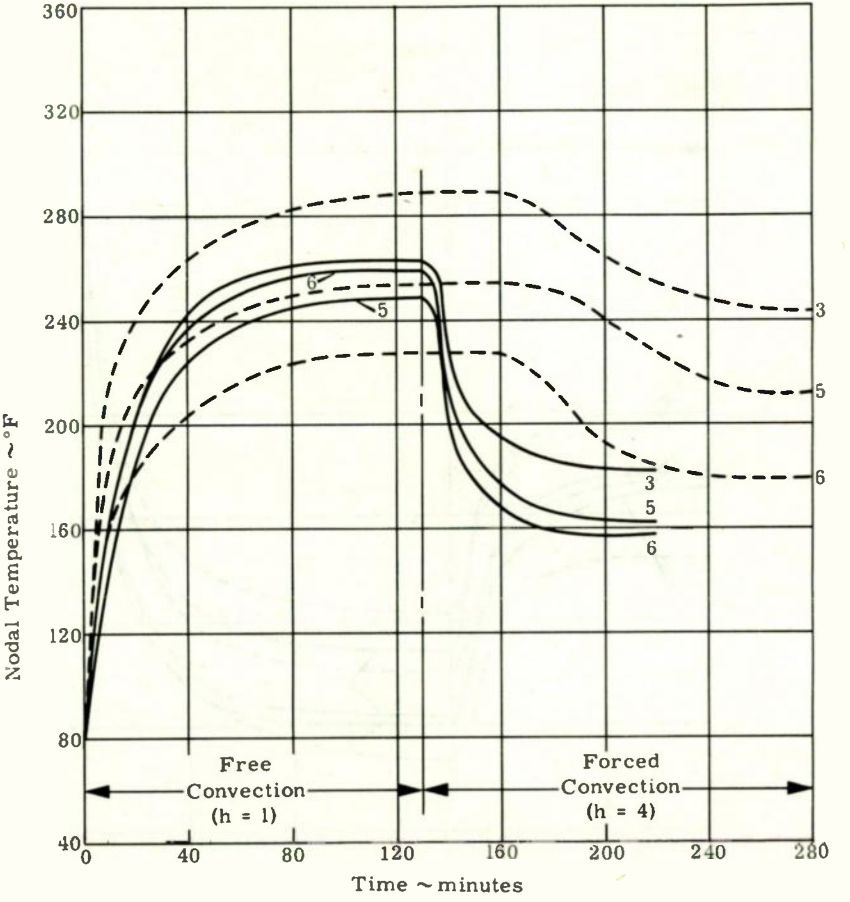


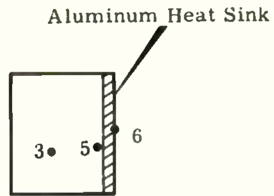
Figure 7. Details of Test Setup



- - - - Unfilled

——— 50% Filled

Aluminum Heat Sink



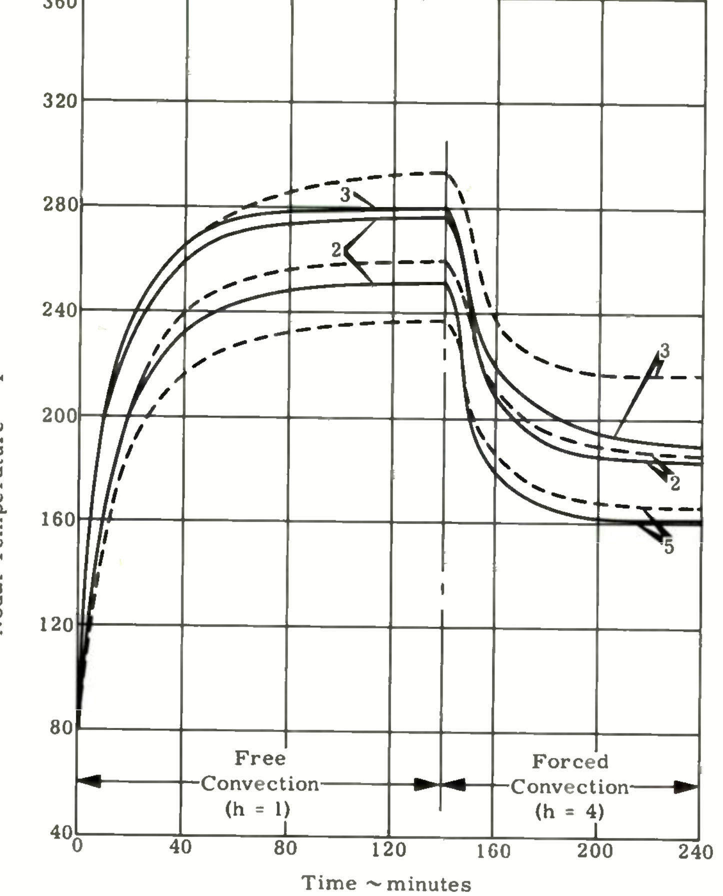
Plan View of Module

• ~ thermocouple locations

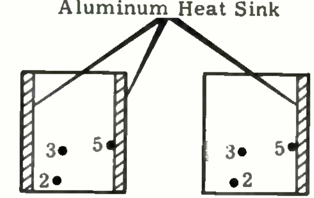
Node Locations and Numbers Refer to Figure 2 (a)

Q_i = 3.9 Watts

Figure 8. Effect of Unfilled versus 50 Percent Filled Epoxy (Experimental Data)



- - - - -1 Heat Sink
———2 Heat Sinks



Plan View of Modules

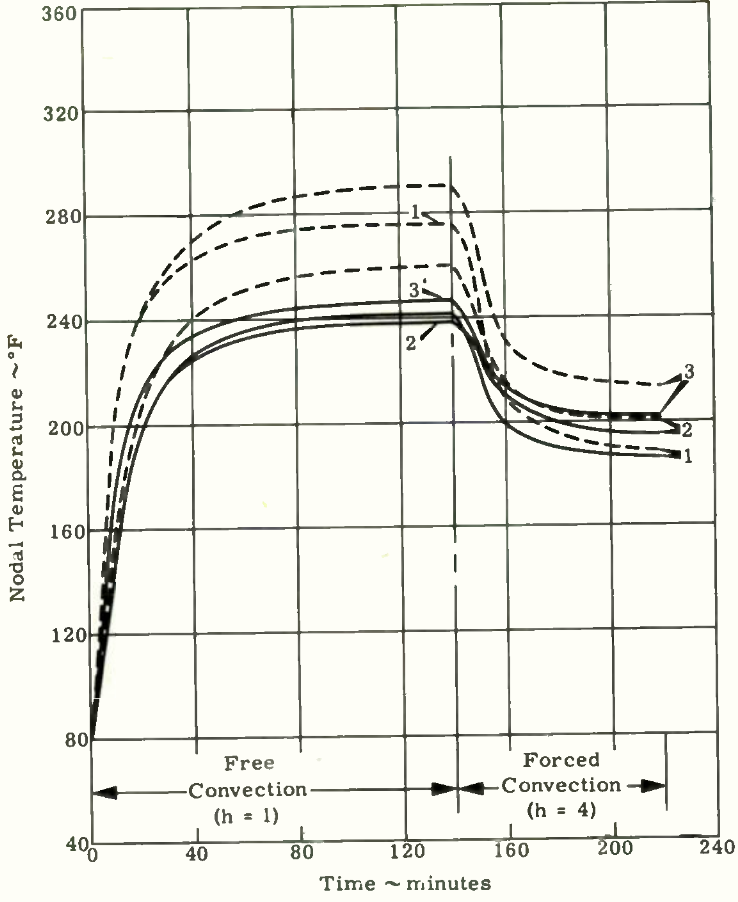
• ~ thermocouple locations

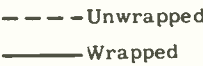
Node Locations and Numbers Refer to Figure 2 (a)

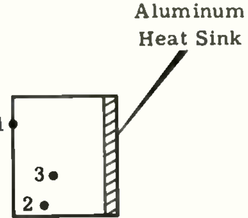
Q_i = 3.9 Watts

50% Filler

Figure 9. Effect of One versus Two Heat Sinks (Experimental Data)







Plan View of Module

• ~ thermocouple locations

Node Locations and Numbers Refer to Figure 2 (a)

Q_i = 3.9 Watts

50% Filler

Figure 10. Effect of Aluminum Foil Wrapped Resistors versus Unwrapped Resistors (Experimental Data)

HIGH-DENSITY-PACKAGE JOINING TECHNIQUES

Charles W. Johnson
Steven G. Konsowski, Jr.
Sperry Electronic Tube Division
Sperry Rand Corporation
Great Neck, L.I., New York

Introduction

Reliable high-density packaging has been high priced. This paper discloses an economical packaging and joining concept that retains the desirable features of previous high-density-packaging systems. These well-known advantages include:

1. High reliability
2. Size and weight reduction
3. Excellent environmental performance capabilities
4. Minimum system assembly costs

Several years of experience with high-density packaging have directed our engineers in the establishment of a set of guiding design concepts. The obvious considerations of high reliability and reduction of weight and size are accepted requirements and will not be pursued in this paper. Some of the less publicized but equally important considerations warrent further consideration:

- o Manufacturability - the capacity of the system to be manufactured with short lead time and low labor costs while remaining flexible to design changes.
- o Hybridization - the ability to incorporate into a system the various technological advancements in simple and complex components, e.g., subminiature standard components, solid circuits, and thin films.

These guiding design concepts have aided in the selection of a joining and packaging technique that incorporates a controlled repeatable process, fulfills temperature-exposure requirements while keeping component leads short, and provides environmental performance capabilities.

Soldering, for example, does not provide a controlled repeatable process, because variations may occur. The length of time and temperature to which the joint is subjected may very, and there may be incomplete flux removal problems and cold solder joints. Dip soldering has an advantage over point-to-point joining with respect to the speed with which a circuit can be completely joined, but there are clean-up and repair problems associated with this technique which, to some degree, minimize the advantage of multiple simultaneous joining.

Welding is a precise, controllable technique because all of the parameters (such as welding energy and electrode material and pressure) are held constant. Welds that exhibit high tensile breaking strength and bend and peel resistance are commonly made with a high degree of repeatability.

The use of short component leads, with resultant higher density packing, is possible with welding as contrasted to soldering. To avoid thermal degradation, the heat generated during the fabrication of an ordinary solder joint must be conducted away from heat-sensitive components via longer leads and/or heat sinks. But with capacitor-discharge welding, the temperature rise of the component during a weld is exceedingly small - not above 50°C - because of the short current pulse duration of one millisecond.

Complete encapsulation of a welded unit assures a high degree of environmental resistance.

After considerable exploration, the Sperry Honeycomb[1] packaging concept was developed, offering the following advantages:

1. Easy assembly
2. Easy welding
3. Accurate parts positioning
4. Minimal dielectric change
5. Minimal hot-spot defects
6. Lower curing exotherm
7. Component identification during welding
8. Minimal lead interference during welding
9. Simplified lead crossovers and feed-throughs
10. Component case-to-case shorts eliminated
11. Color coding reduces assembly error
12. Economical process due to rubber-mold tooling

This system is presently being employed to fabricate modules used in systems for an NASA Satellite experiment, the Polaris Mk III Navdac Console, the APN118 Flight Director System, and

the NASA X15 stable platform. A present production rate of 2000 modules per month has exposed us to the many development and manufacturing problems connected with high-density packaging, in addition to giving us valuable insight into their solutions.

The applications of welded circuits at present demands the highest reliability obtainable. The intraconnection welding of components represents the primary source of catastrophic failure, and must be precisely controlled. The well-known statistical method of developing a weld schedule (Fig. 1), reported by Space Technology Laboratories,[2] is the method used to determine the parameters for reproducible, reliable joints. In order to verify the statistical choice, it is further necessary to examine samples of welds by metallographic sections (Fig. 2). The fabrication, testing, and qualification of welds requires such production control procedures as the making of sample welds at the beginning of each day's work and after every fiftieth weld or each time the weld schedule is changed. These sample welds are subjected to a destructive tensile test (Fig. 3), and the breaking strength is recorded on a chart that is kept for each welding machine. Acceptance limits are set up on this chart, indicating any drift of machine setting well in advance of trouble. The welding-energy meter, the electrode-pressure gage, the tensile-testing machine, and welding-circuit-continuity meter must be routinely calibrated.

It was found necessary to screen the assembly and welding operators to assure efficient operations. However, recent fixturization and methodization have broadened the type of personnel acceptable for this type of work, while at the same time increasing the reliability through the elimination of subjective operator influences. To illustrate the economics of fixturization (exclusive of considerations of increased reliability over the standard process), a typical example follows:

> 8600 modules, each containing 10 ribbon-tabbed resistors requiring 86,000 welds at 3 minutes per weld and wire trimming, were fixturized to enable these resistors to be semi-automatically welded and trimmed at the rate of one-half minute per weld. A 215,000 minute or 1 1/2 man-year savings was realized as a result of the use of this fixture.

Tinned copper component leads, such as those of transformers, resistors, and capacitors, because of their nonuniform welding surface, must be reprocessed by chemical stripping and gold plating to assure reliable welding.[3]

A 10X binocular microscope is employed in the visual inspection of all welds in the uncapsulated module (Fig. 4). Dynamic electrical testing, both before and after encapsulation, qualifies the electrical performance.

The obvious large labor savings resulting from the fixturization of the resistor tab operation made apparent the tremendous potential of a highly reliable technique employing the simultaneous multiple joining of the components within a module.

A review of possible assembly and joining techniques consistent with the guiding principles outlined above indicated the desirability of a planar configuration of the components and their required joints. If the subminiature standard components are arrayed with their leads intersecting at right angles with each other or the hookup wire, a multiheaded joining device may be used to effect simultaneous connections.

Ultrasonic welding, thermal compression bonding, and cold welding were evaluated as possible simultaneous multiple joining techniques, all lending themselves to a simplified multiheaded tool.

To illustrate the process, a typical circuit will be traced through the design and manufacturing steps. After components compatible with high-density-packaging requirements have been selected (Fig. 5) and mutually agreed upon by the customer and design group, ten-to-one scaled cardboard cutouts are employed for the rapid attainment of a high-density configuration (Fig. 6). A sample unit is then fabricated using spot welding techniques for electrical evaluation (Fig. 7). This unit is employed as a master for the manufacture of a vacuum-formed or cast form-fitting tray which will be employed to hold subsequent parts during the joining operation. To provide access to the joining sites, the trays are drilled or punched. The trays may be marked for component identification as an aid to assembly. The fabrication of a module consists of the following:

1. Tray is loaded with components.

2. Loaded tray is aligned on base plate of joining machine.

3. Joining machine is activated to effect bond.

4. Complete assembly is removed from tray and excess leads are trimmed (trimming may be accomplished simultaneously with the joining operation by providing trimming blades on the multiheaded joining fixture).

5. The encapsulated module is tested mechanically and electrically.

6. Tested circuit is placed in a preformed shell provided with slots for the input and output leads. It is then potted with appropriate resin.

7. Mechanical inspection is performed, and the units retested electrically.

It can be seen from the above process that hand operations have been minimized, with further economies being possible through the use of automatic loading equipment and indexing trays.

Final assembly to higher order system configurations may be accomplished by stacking these wafer-like modules with or without metallic heat sinks and interconnecting the individual modules by spot welding (Fig. 7).

Replacement may be easily effected by cutting the defective wafer free from the interconnecting wiring and removing it. A sandblasting unit such as the S.S. White dental abrasive unit may be used to expose the damaged components in the wafer for replacement. After replacement of the components, sealing is accomplished by the addition of potting material.

Summary

This paper has presented a Sperry-developed packaging concept that offers the following advantages of joining and assembly:

1. Rapid, economical assembly
2. Adaptable to semi-automatic handling
3. High reliability
4. Small size, light weight
5. Low-cost flexible tooling
6. Efficient heat transfer (large surface-to-volume ratio)
7. Hybridization capabilities.

This concept has evolved from the activities of the packaging facility of the Sperry Electronic Tube Division and is offered as a contribution to advance the state of the art of high-density packaging.

Bibliography

1. C.W. Johnson, "High-Density Electronic Packaging via Welded Honeycomb Structures," Fifth National Convention of the IRE Professional Group on Product Engineering and Production, Philadelphia, June 15, 1961.

2. D. Armstrong, "Adoption of Welded Electronic Circuits for Missile and Space Vehicles," Fourth Symposium on Welded Electronic Packaging, Melville, N.Y., March 24, 1961.

3. C.W. Johnson and S.G. Konsowski, Jr., Unpublished Notes.

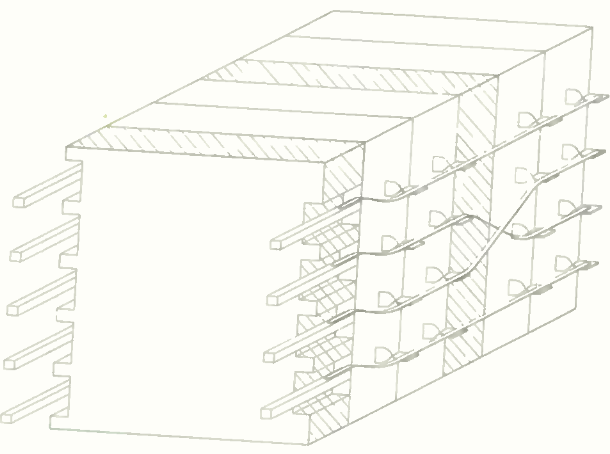


Fig. 8. Stacked interconnected planar modules with interleaved heat sinks.

WRH

THE MATHEMATICAL BASIS OF THE AUTONETICS ETCHED INTERCONNECTION DESIGN PROGRAM

John Weissman, Senior Engineer - Research
Department 3341-51, Autonetics, Anaheim, California

Summary - This discussion will demonstrate by example how the mathematics of the Autonetics etched interconnection design program achieves the intended technical result and an unintended abstract result. The technical result is a procedure that designs a so-called multilayered etched interconnection board with the minimum number of layers. The abstract result is a procedure that determines a coloring scheme for a mathematical map, which uses the minimum number of colors. The mathematical proofs of the categorical statements in this discussion are contained in another paper being prepared for publication.

Introduction

Autonetics, as well as other computer manufacturers, is interested in reducing the time from inception to production of its computers. It was felt that a significant reduction in design time could be realized if a large-scale computer could be used to design multilayered etched interconnection boards. Consequently, a study was initiated to formulate a mathematical procedure upon which such a computer program could be based. Certain requirements necessitated that the design use the minimum or a near minimum number of layers of etched panels. This paper describes a mathematical procedure that satisfies these requirements and contains a brief discussion of a programing problem and its solution. The mathematical procedure was found to have a surprising application to mathematical maps. Given a map, a coloring scheme that uses the minimum number of colors is produced.

The Map Coloring Application

To apply the foregoing procedure to a mathematical map, the map must have a finite number of regions and all pairs of adjacent regions (those pairs that must be colored differently) must be determinable. The procedure is independent of the topological space containing the map. The procedure apparently does not shed light on the well-known and unresolved four-color map problem of mathematics. (No map has ever been drawn on a plane or on the surface of a sphere that requires more than four colors; however, a proof that five colors are sufficient is all that exists. For a thorough and elementary discussion, see Courant and Robbins, What is Mathematics?, Oxford University Press.)

Fig. 1 contains a sample mathematical map with its nine regions arbitrarily labeled with the letters a through i. From it is derived the top expression of the Boolean algebra below by the following means. For example, the fourth factor (e+bd) results from the observation that only regions b and d precede e in alphabetical ordering and are adjacent to e. The next to the last factor (h+abd) results from the observation that only regions a, b, and d precede h in alphabetical ordering and are adjacent to h. If for a particular region there is no preceding region adjacent to it, then there would be no factor starting with that particular region's letter. This is always true for the region chosen to be labeled by the letter a.

(b+a)(c+b)(d+bc)(e+bd)(f+de)(g+df)(h+abd)(i+dh)
= (b+ac)(d+bc)(e+bd)(f+de)(g+df)(h+abd)(i+dh)
= (bd+bc+acd)(e+bd)(f+de)(g+df)(h+abd)(i+dh)
= (bd+bce+acde)(f+de)(g+df)(h+abd)(i+dh)
= (bdf+bde+bcef+acde)(g+df)(h+abd)(i+dh)
= (bdf+bdeg+bcefg+acdeg+acdef)(h+abd)(i+dh)
= (bdfh+abdf+bdegh+abdeg+bcefgh+acdegh+acdefh)(i+dh)
= bdfh+abdfi+bdegh+abdegi+bcefghi+acdegh+acdefh.

The seven lines following the top expression above are indicative of the methods used in the Autonetics program to convert by Boolean algebra the top expression into the bottom expression, a standard form of a nonredundant sum of products. Each of the seven products in the bottom expression above corresponds to the set of all regions not present in the product as follows.

bdfh → A = {a, c, e, g, i}
abdfi → B = {c, e, g, h}
bdegh → C = {a, c, f, i}
abdegi → D = {c, f, h}
bcefghi → E = {a, d}
acdegh → F = {b, f, i}
acdefh → G = {b, g, i}

Each of these sets of regions, A through G, has the following properties:

1. All of its member regions may be colored with the same color.

2. Every region not in the set is adjacent to at least one region in the set.

Furthermore, sets A through G are the only sets with these properties. The processing of sets A through G by another Boolean multiplication completes the procedure. Each of the factors in the top expression of the Boolean equation below corresponds to a region. The first factor (A+C+E) results from the fact that sets A, C, and E contain region a, the second factor (F+G) results from the fact that sets F and G contain region b, etc. The bottom expression below is the result of the conversion of the top expression into the same standard form of a sum of products and by the same methods as used in the prior multiplication. The intermediate steps are omitted.

$$(A+C+E)(F+G)(A+B+C+D)E(A+B)(C+D+F)(A+B+G)(B+D)(A+C+F+G)$$
$$= ADEF+BEF+ADEG+BCEG+BDEG.$$

BEF is the product of the bottom expression of this equation that has the least number of factors. Therefore, the map may be colored with a minimum of three colors. BEF represents this coloring scheme as illustrated in Fig. 2 and this coloring scheme is the only way to color the map with only three colors (barring permutations of the colors). The other products in the bottom expression of this equation also represent coloring schemes. For instance, the term ADEF represents the 16 coloring schemes using 4 colors as indicated in Fig. 3.

The same mathematical procedure, with a slight generalization in the second multiplication, will be used now to solve the following problem in etched electrical interconnections.

The Interconnection Application

Imagine pins sticking out of a plane surface of a box of electronic components. A certain list of pairs of these pins is available. Each pair of pins in this list is to be connected electrically. For each pair of pins there is a list of etched paths that connect the pins, one of which must be used. These paths are to be contained in a board through which the pins protrude. Usually one such board cannot contain all of the paths necessary to connect all the pairs of pins. So, many boards containing etched paths are used, stacked one upon the other, with the pins protruding through every board. Finally, complete information is available or computable on whether two paths, if placed on the same board, would interfere with one another. By definition, two paths interfere if they establish an unwanted electrical connection (the definition may be extended to include problems of capacitance, manufacturing tolerances, etc.). The problem is to choose one path for each pair of pins and assign it to a board so that no two paths on the same board interfere and so that the number of boards used is the MINIMUM possible.

Fig. 4 illustrates the geometric configuration of such an interconnection problem with 10 pins. The pin configuration is repeated to demonstrate the following information:

Pins 1 and 8 are to be connected by either path a or path b.

Pins 7 and 8 are to be connected by either path c or path d.

Pins 2 and 9 are to be connected by either path e, path f, or path g.

Pins 4 and 6 are to be connected by either path h or path i.

Pins 6 and 10 are to be connected by path j.

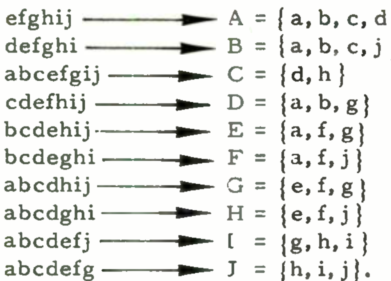
Fig. 4 also contains a table indicating by an "x" when two paths interfere, if placed on the same board. Notice that paths b and d interfere in the geometric sense, but no "x" is entered for them, since in the electrical sense the use of both of them on the same board establishes an efficient, wanted connection between three pins. Also of interest but less obvious is the fact that if paths connecting the same pair of pins (a and b, for example) were considered as interfering with one another, the final result would be unaffected.

Paths a, b, c, and d precede e in alphabetical order and would interfere with e, if placed on a common board with e; hence, the first factor in the top expression of the equation below. Paths b, c, and d precede f and interfere with f; hence, the second factor, etc.

$$(e+abcd)(f+bcd)(g+cd)(h+abcef)(i+abcdef)(j+dg)$$
$$= efghij+defghi+abcefgij+cdefhij+bcdehij$$
$$+bcdeghi+abcdhij+abcdghi+abcdefj+abcdefg.$$

The bottom expression is the same standard form as seen before and, in the same manner as before, leads to the sets of paths, A through

J, listed below:



Sets A through J are the only sets with the following two properties:

1. The paths of each set may be contained on one board with no interference.

2. No other path may be placed on that one board without interference.

The top expression of the next equation below results from the following information. Only sets A, B, D, E, and F contain at least one path connecting pin pair 1 and 8. Only sets A, B, and C contain at least one path connecting pin pair 7 and 8, etc.

$$(A+B+D+E+F)(A+B+C)(D+E+F+G+H+I)(C+I+J)(B+F+H+J)$$
$$= ADJ+AEJ+AFJ+AGJ+ACH+AHJ+AFI+AHI+AIJ+BDJ+BEJ+BFJ+BCG+BCJ+BCH+BHJ+BI+BCD+CDH+CDJ+CEH+CEJ+CF.$$

BI and CF are the terms of the bottom line of the last equation that contain the least number of factors. Hence, the desired interconnections using the available paths are possible on a minimum of two boards. BI and CF each indicate such a design as illustrated in Fig. 5.

A Programing Problem and Solution

As of the date of this writing, an IBM 7090 program incorporating the above methods to design a multilayered etched interconnection board is under study. A major programing problem was how to store the computer representation of an intermediate step or the final sum of products in the first multiplication. For a worthwhile number of pairs of pins to be connected, this information is thought to exceed the 7090 memory space even if only a few paths are considered for each pin pair. The use of high speed tape units apparently would result in too long an execution time. The device that overcame this problem, at the cost of some increase in the number of boards in the final design, is the particular way used here to set up the first multiplication as a product of sums. To illustrate, the left-hand factor of the fourth line of the first multiplication in the map procedure (bd+bce+acde) contains sufficient information to get the best result for a sub-map consisting of regions a, b, c, d, and e alone. So the computer program processes pin pairs in order from the top of the pin configuration to the bottom by entering the equivalent of the left-hand factors of the first multiplication in memory. As a new factor is computed, the old one is erased. At some particular point the program realizes that it cannot store the next factor. Then the best answer is obtained for those paths considered up to that point. Then the program considers the next set of pin pairs that the computer can handle. Eventually a design like that illustrated in Fig. 6 is produced.

Other serious programing problems, in addition to the one described above, exist and are under study. However, this much can be said: the design of many types of etched interconnections can be and has been automated; this automation can and does have certain optimal features; and the development of underlying mathematical theories and programing techniques foretells much optimism about this area of automation.

A Shortcoming

It is believed that the procedures used and the results ascribed to them are mathematically foolproof. However, the ground rule of a predetermined list of pair of pins to be connected is faulty. For example, in Fig. 4 why is it better to connect pin 1 to pin 8 and pin 7 to pin 8, instead of connecting pin 1 to pin 7 and pin 8 to either pin 1 or pin 7? A substantially improved theory would be one that could handle these possibilities as well as the possibilities of alternate paths on alternate boards. The Autonetics program determines the set of pairs of pins to be connected so that the total length of all etched paths used is close to the minimum. This low total length is important in its own right, but seems independent of the more important problem of having the minimum number of layers in a multilayered etched interconnection board.

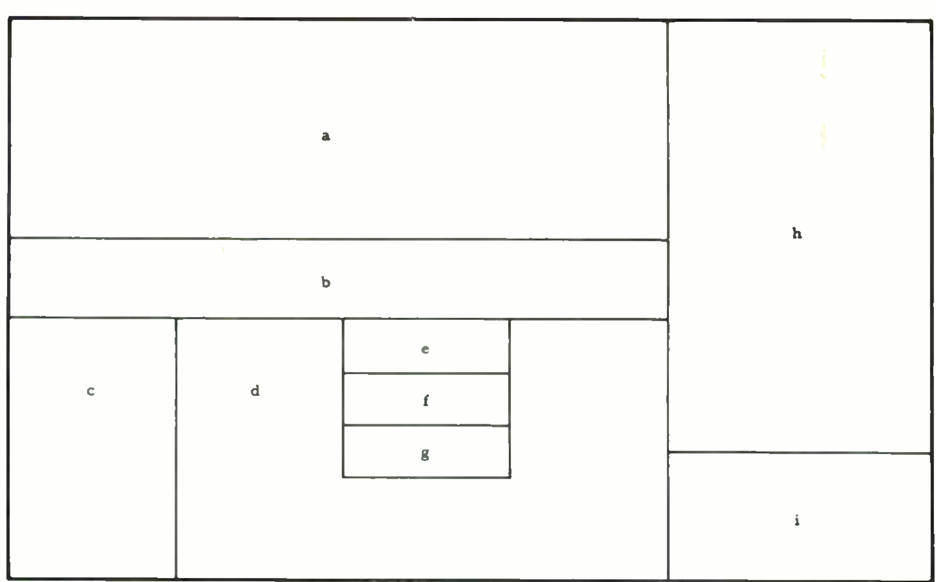
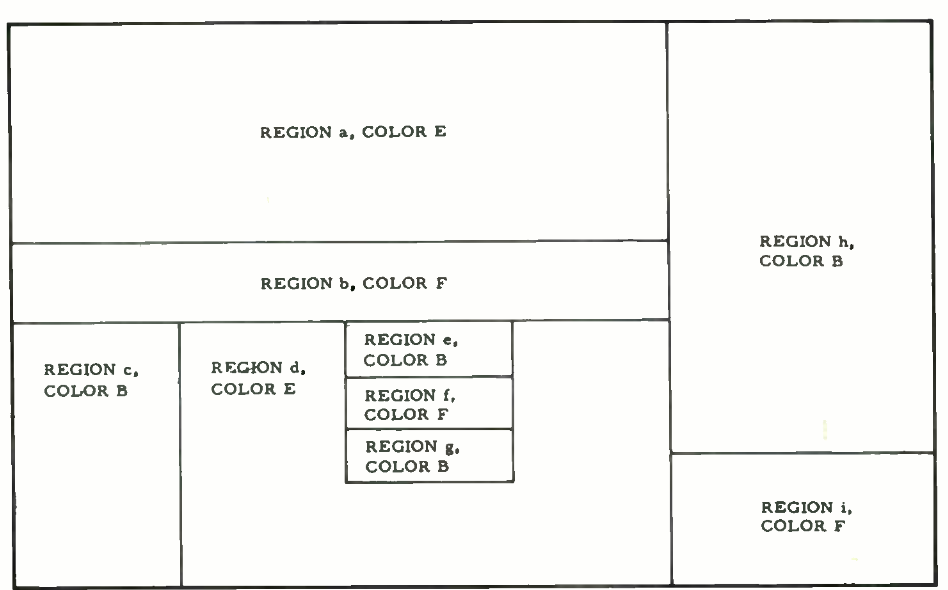
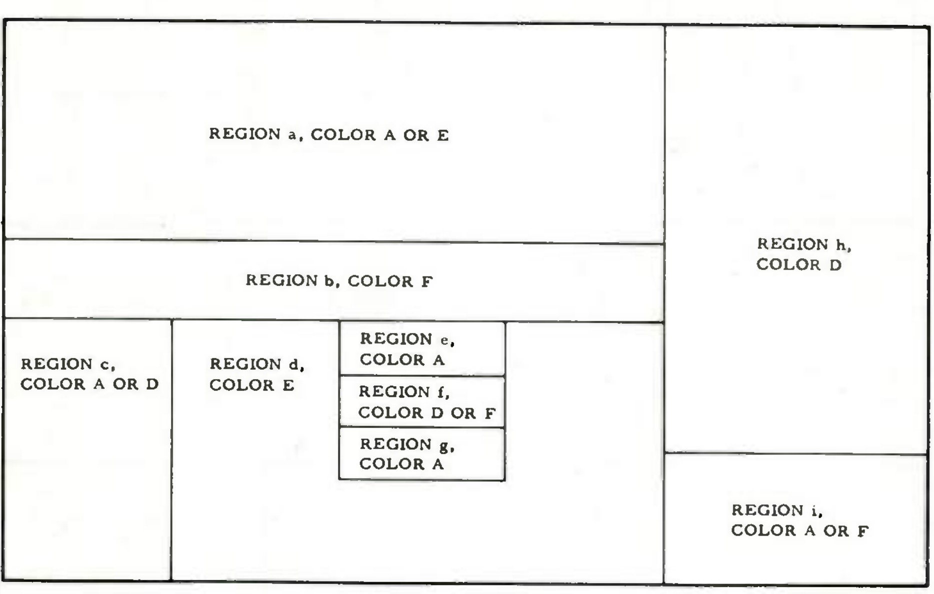


Figure 1. Sample Map



B = {c, e, g, h}
E = {a, d}
F = {b, f, i}

Figure 2. Minimal Map Coloring Solution



$A = \{a, c, e, g, i\}$
$D = \{c, f, h\}$
$E = \{a, d\}$
$F = \{b, f, i\}$

Figure 3. Four Color Solution

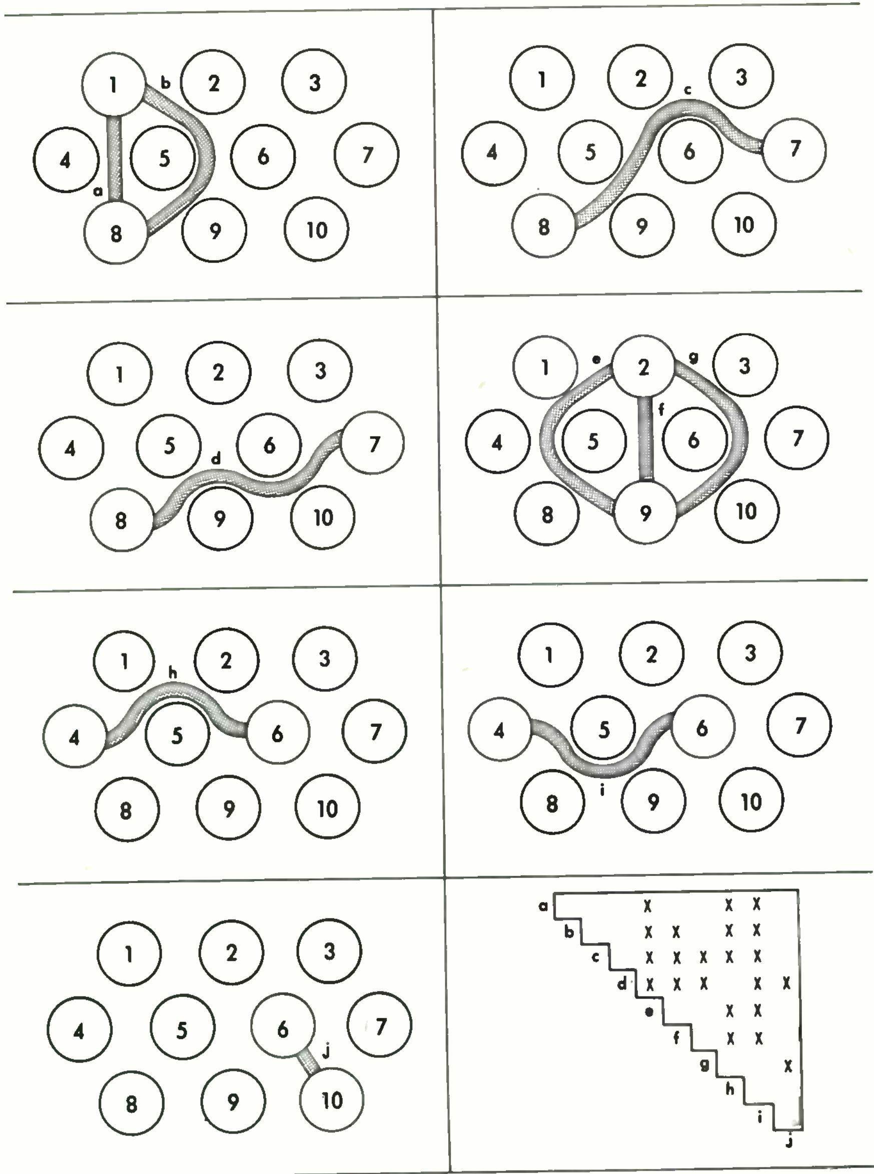
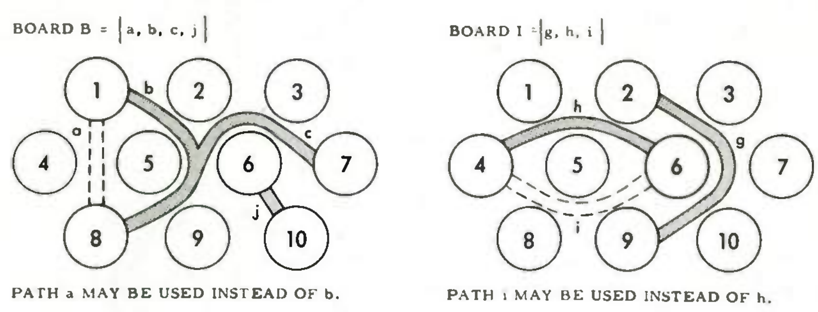


Figure 4. Sample Interconnection Problem

SOLUTION BI



SOLUTION CF

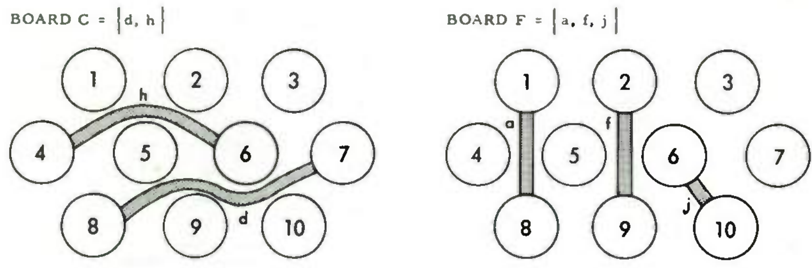


Figure 5. Minimal Interconnection Solutions

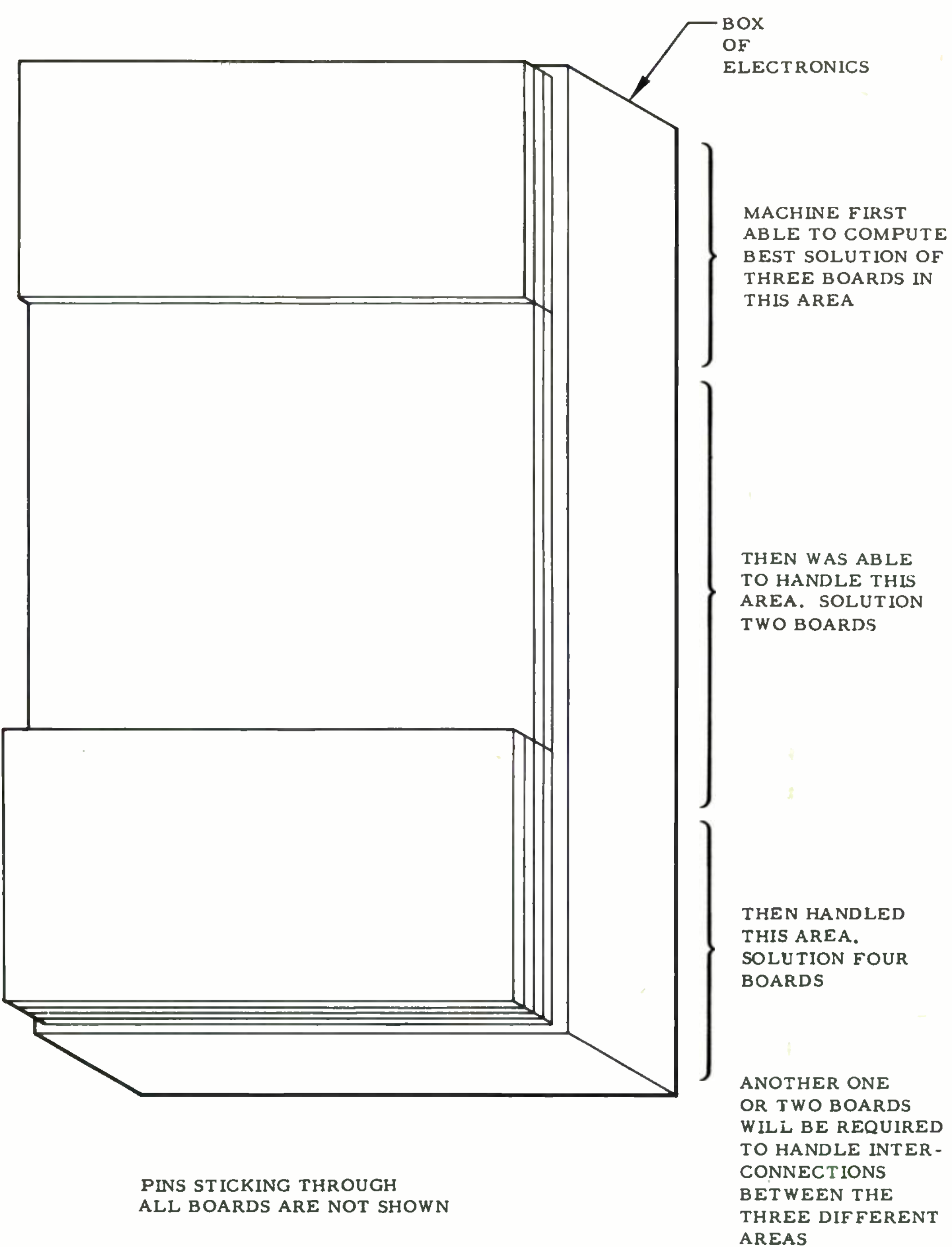


Figure 6. Example of Computer Results of Method

NEEDED: WIRE DESIGN DOCUMENTATION FOR MORE EFFICIENT PRODUCTION

John T. LaForte
Light Military Electronics Department
General Electric Company
Utica, New York

Summary

One of the most critical phases of equipment design is the wiring. Since even minor changes in wiring can affect equipment performance adversely, the details of equipment wiring are fully as much a part of the equipment design as the mechanical and electrical portions. Detailed wiring information, therefore, should be recorded on formal manufacturing drawings for every project.

Defense budgets clearly indicate that participation in military electronics programs requires large commitments to development efforts. Clearly, it is industry's responsibility to meet the challenge which the Secretary of Defense put to the service secretaries when he asked them to cut development costs, shorten delivery schedules, and improve reliability. Development programs, unfortunately, are fraught with changes which can obscure comparative evaluations unless these changes are controlled and documented.

To improve reliability and to shorten delivery time, General Electric's Light Military Electronics Department has integrated mechanical and wiring design. The result is that the location of wiring is planned at the same time as the location of parts, insuring circuit integrity and resistance to adverse environment.

* * * *

One means of insuring reproducible equipment is to include wiring details in a formal documentation system. Complete wiring information is recorded for every modification or development so that when the design is "fixed" production can be started with the assurance that the production equipment will be equal to or better than the approved prototype.

To shorten delivery time, much of the planning and design have been integrated at the layout stage. When the drawings are released to Manufacturing, wire layout and process information are complete enough so that production can begin even before the longer lead-time parts are procured and assembled.

To cut costs and improve efficiency, the manual rearrangement of basic wiring information has been eliminated, and process instructions are now generated by automatic data processing procedures.

With the old system, two drawings--a connection diagram and a connection list--were used to detail wiring. The connection diagram was simply a spur diagram which identified and oriented parts and terminals. Although it did not detail harness branch location; it did specify which wires were connected to what terminals.

The connection list identified wire size and type and the "from-to" connection points. It also specified, when applicable, methods of termination, markers, sleeving, and terminals.

When these two drawings were released to Manufacturing, the wire design group began procuring parts. These parts together with mock-ups of long lead items were then assembled, and the wires were routed, connected, and bundled within the available space. Ordinarily, this less-than-optimum harness was then removed and laid flat on a board. A tracing was next made for future assembly and a wire sequence list was drafted.

Wire and strip lengths were then determined and recorded as a wire reference list. Production orders required that this list be sorted into a wire processing list, which grouped all similar wire types, sizes, colors, cut lengths, and strip lengths for the most economical set up of wire-processing equipment. Because the equipments contained hundreds of wires, the manual rearrangement of information was a time-consuming chore. Still another job of manual sorting and rewriting was an assembly sequence list which combined mechanical assembly instructions with wire assembly instructions. As it turned out, the wire reference list, wire sequence list, cutting list, and assembly sequence list incorporated essentially the same information as the original connection list. The primary difference was that their individual arrangement was peculiar to a particular process.

No small consideration was the draftsman's effort in lettering thousands of entries legibly on the connection list to suit the exacting requirements of microfilming. When such a list was to be included in instruction books it was also typewritten so that the same data was processed manually four or five times.

A search for a new method was begun with the basic premise that Engineering bears complete responsibility for the design and performance of equipment and for detailing requirements on final drawings. Engineering must therefore develop and

record simultaneously the interrelationship of parts and wiring.

It was determined that in most cases the connection diagram could be eliminated by including wiring information on the assembly drawing. Figure 1 is part of a typical assembly drawing, and Figure 2 is a segment of a connection diagram; both of these documents were standard prior to the study. Close examination of Figure 1 will show that the only delineation lacking for wiring is specific terminal identification and harness location. Figure 3 is part of an assembly drawing with the wiring details added.

For a number of reasons, this kind of assembly drawing is preferable. In the first place, it provides a complete picture of the compatibility of wire and component parts. It is also a means of determining the order in which assembly operations should be sequenced for optimum combination of assembly and interconnection. Secondly, it requires the designer to consult with Manufacturing early in his efforts, to discover assembly details; eg., whether wire harnesses should be made up as subassemblies for later installation or whether the unit should be wired point-to-point with wires tied and bundled. This early integration results in a much better design.

When a pre-assembled wire harness is indicated, the harness must be designed and detailed. An outline of the harness is shown on the assembly drawing and ordered as a specific part. A dimensionless and single-scale detail drawing of the harness assembly is next made. A print of this drawing can be laid out on a board and used as a manufacturing tool as shown in Figure 4. Of course, fixtures and nails must be added to hold the wires and connectors in place.

The real key to successful harness design is to derive distances from the center line of the harness branch to the appropriate terminal. Additional length should be allowed for service loops and strip lengths. This requires that purchase part drawings are complete in detailing terminal sizes and locations. To facilitate manufacture, the harness should also be reduced to a single plane whenever possible.

Each wire is identified by a wire number corresponding to its identity on the connection list. Harness branches are specified by a progressive two-letter designation which identifies the paths in which wires are routed. The drawing format is zoned to assist in locating branches and wire ends.

The connection list (Figure 5) is the support document that ties the harness assembly drawing to the equipment assembly drawing. This list itemizes each individual wire or hung-in component and provides directions for connecting them between termination points. Each "from-to" connection (Figure 5A) is listed only once. The wire or component column (Figure 5B) either lists an abbreviated size and color description of conductors or includes the reference designation of the hung-in components. All "item no." columns (Figure 5C) refer to the parts list item number for ordering material. The "wire end zone" (Figure 5D) refers to the zone on the harness assembly drawing in which the "from" and "to" ends of the wire can be found. The "harness route" column (Figure 5E) identifies the paths along which the wire is routed. (When routing is unmistakable because of the absence of closed loops, this information is unnecessary.) "Notes" and "remarks" (Figure 5F) are used for special instructions for terminating wire ends and assembly.

The "wire length" column (Figure 5G) and "strip length" column (Figure 5H) are included to minimize the paper work when harness assembly is accomplished with precut and stripped wires. These columns enable the wire developers to record the appropriate lengths in making wire-processing sheets.

The connection list is unique in that it has been designed to use data processing equipment. Basic to the process is data processing cards, two of which contain all of the information for a single wire or line entry.

The use of data processing equipment permits a much faster tabulation of a running list than is possible with hand lettering. Punched cards also supply a retrieval method that permits variations and alterations in the list with a minimum amount of repetitive manual labor. The cards themselves are a means for sorting and tabulating the various forms of paper work required in both Engineering and Manufacturing. For example, when harnesses are built with pre-cut and pre-stripped wires, the sorting of the wire processing list is mechanized, and wire sizes, colors, and types are specified in a standard format.

It is also possible to sort in the "from-to" column to arrange all of the wires and components that terminate on a particular component or in a particular location. The result is the most efficient assembly sequence list or terminal check list possible. A typical example would be a connector having 20 or 30 wires terminating on the pins. A wire developer might sort down the "from-to" column for the connector J101 and tabulate a list which would collect all wires terminating on J101. These wires would then comprise a sub-assembly of the major harness.

Sorting in the harness-route column, besides supplying all of the wires which

are bundled in the same branch, permits an analysis of the harness sub-assembly and calculations of the harness-branch diameter. In addition where a computer program can be derived to develop the shortest wire paths between terminations, path assignments can be made according to harness routes.

By sorting and tabulating in the wire length column, it is possible to tally the total lengths of the various types of wire. This information is useful in material and inventory control as well as in purchasing. A wire type code is assigned in column 2 of the data processing cards to permit sorting of wire types. The reason for this sorting is that not all wires can be cut and stripped on the same type of equipment. Therefore, the following code, which is punched into the cards by a key punch operator but omitted from the final running list, is written in by the designer on the rough draft:

KEY	WIRE TYPE
A	Ground wires attached to shields
1	Insulated single conductor
2	Shielded wire
3	Coax cables
4	Bare wires
5	Multi-conductors

The color presentation is coded on the wires with standard letters used to designate respective colors:

CODE	COLOR
K	black
N	brown
R	red
O	orange
Y	yellow
G	green
B	blue
V	violet
A	gray
W	white
T	natural

For example, to describe a wire which has a white base color with a red broad tracer and a blue narrow tracer, the combination of letters WRB would be used.

Columns 79 and 80 on the cards contain the strip length information provided in the "from" and "to" column of the running list under the heading "strip length". The logic for this arrangement is that by retaining the same information on two cards, it is possible to keep them together during the sorting process. This, of course, is only important when sorting strip lengths and wire lengths to arrive at a listing of wires with progressively longer strip lengths. The strip length column contains three individual columns to provide for wires with a jacket, a shield, or a combination of both over the individual conductors. The first column lists the primary conductor insulation strip length; the second column lists the shield strip length, and the third column lists the jacket strip length.

This kind of wiring documentation has been highly effective at G-E's Light Military Electronics Department, guaranteeing the reproduction of critical designs with original performance. The loss or destruction of Manufacturing tooling or drawings is no longer a concern, for the design is documented in the Department's archives.

Harness assembly drawings also permit the designer to arrive at a quantitative "make or buy" decision. Harness assembly drawings have helped materially in spare parts provisioning even to the point of expanding markets for spare harnesses. And most important of all, this documentation system has literally shaved months from delivery schedules. The manufacture of harnesses now begins as early as three months before the longest lead-time item can be delivered to the factory. When all of the parts are available, they are assembled, and the harnessing can be in place with a minimum of additional wire development and planning.

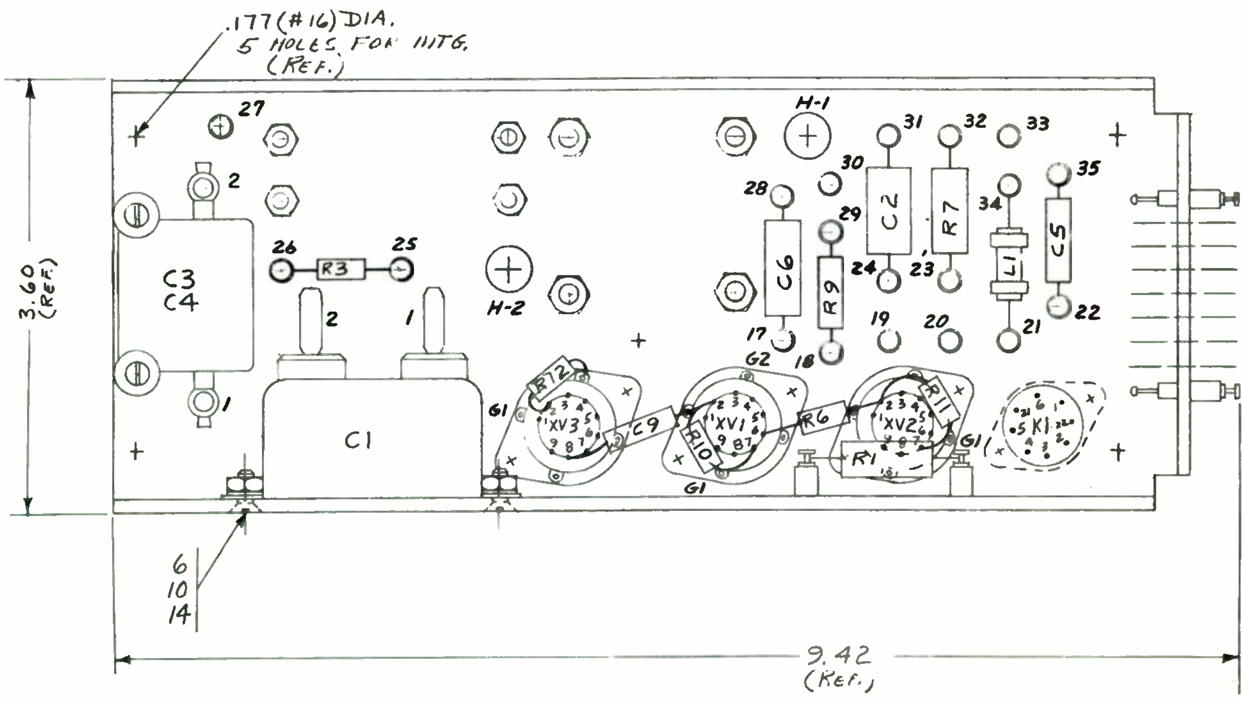


Fig. 1. Portion of typical assembly drawing.

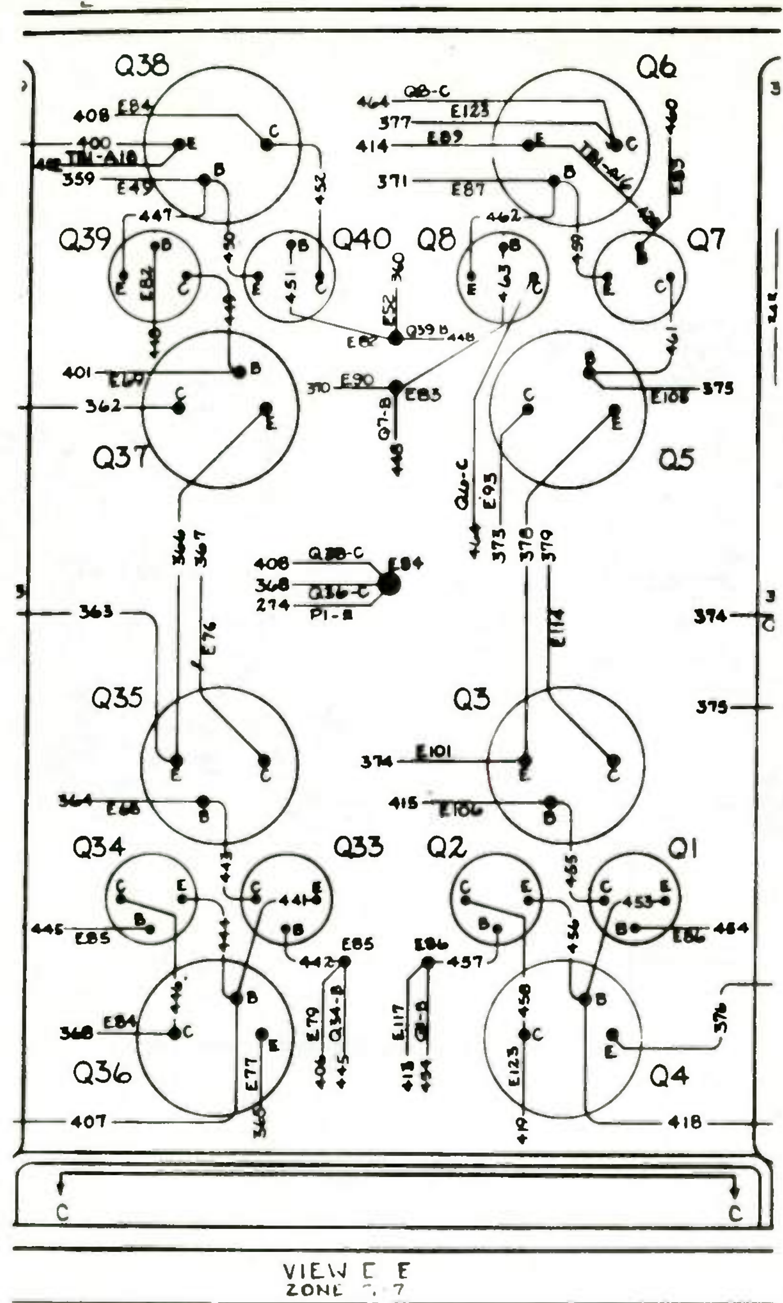


Fig. 2. Part of a connection diagram.

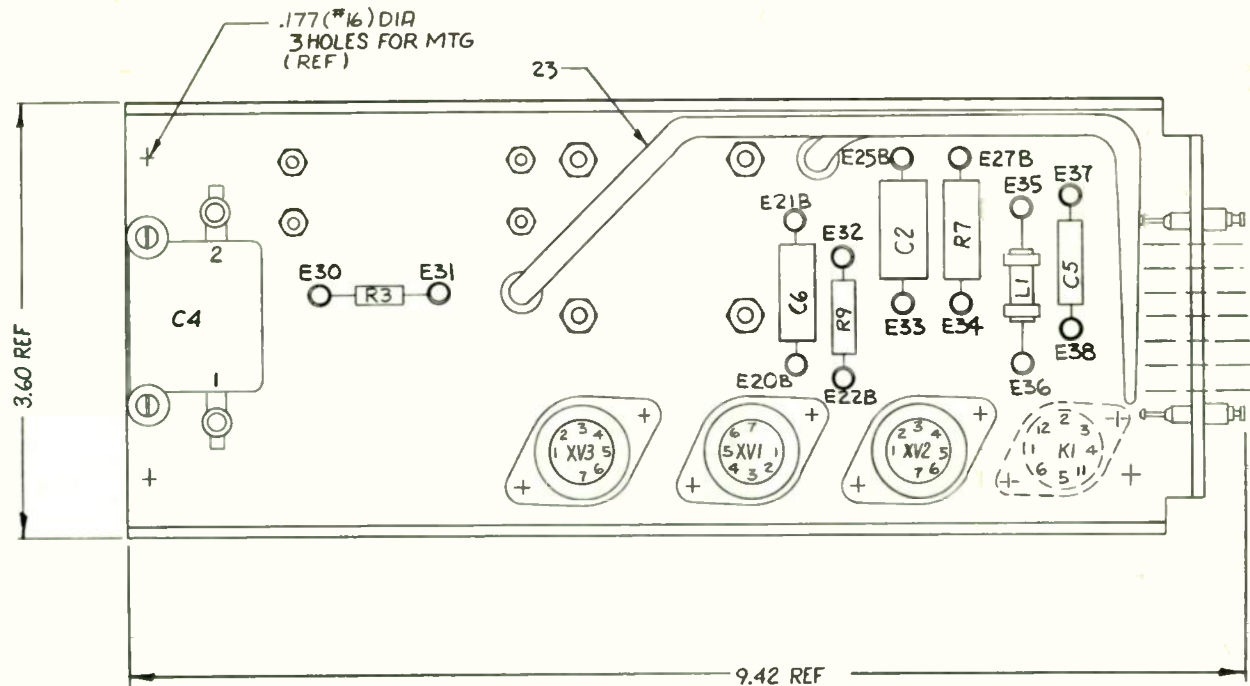


Fig. 3. New assembly drawing incorporating wiring details.

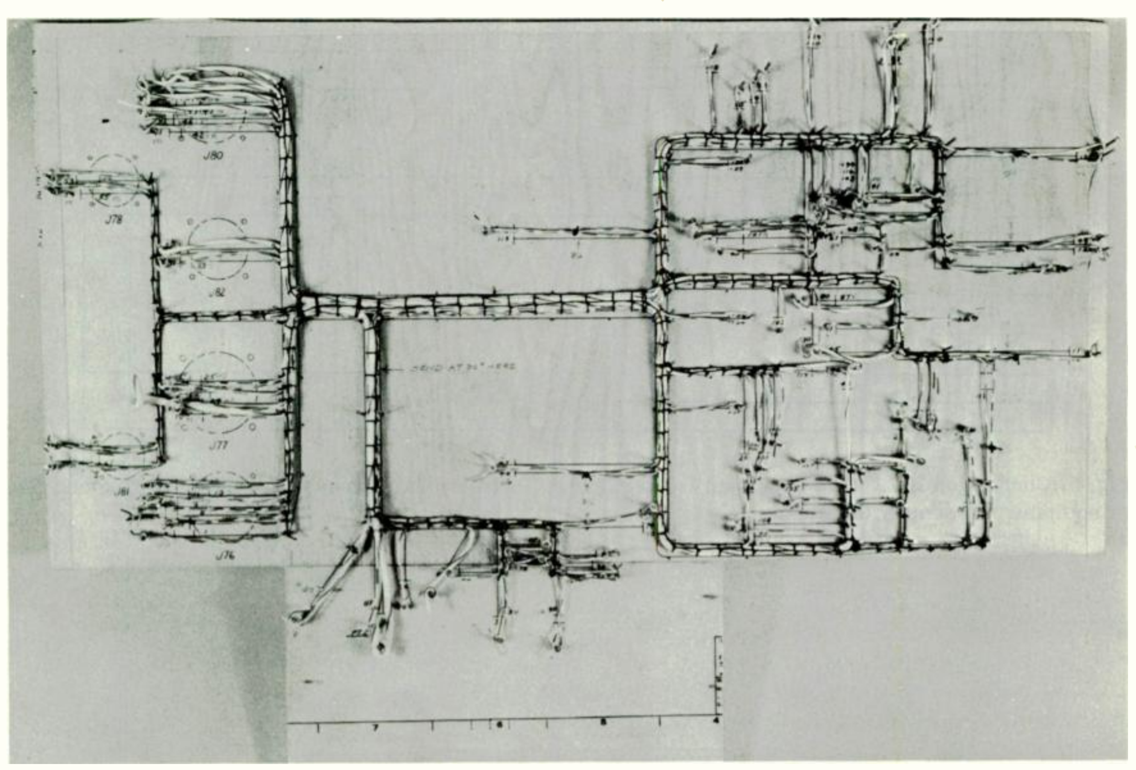
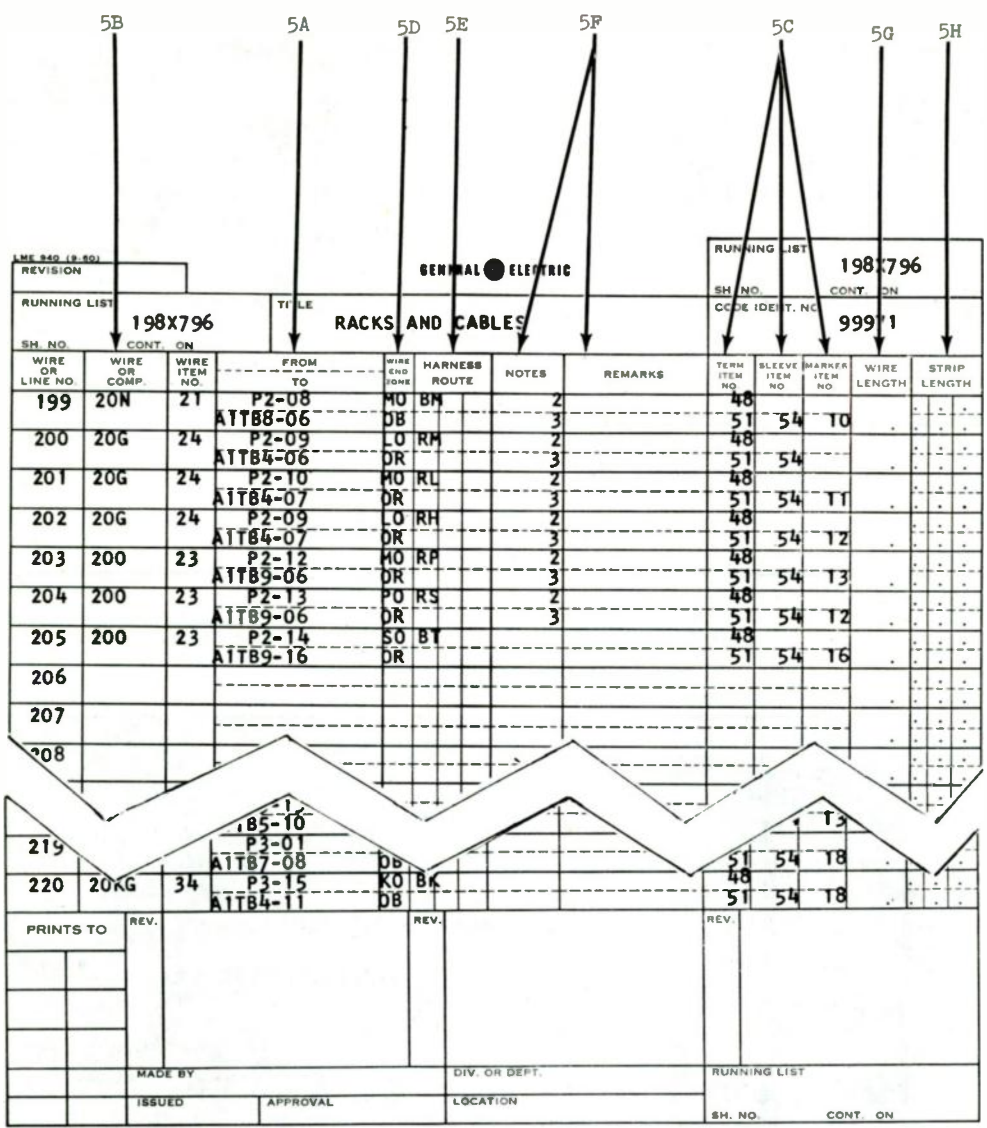


Fig. 4. Pre-assembled harness assembly; note detail drawing in background.



LME 940 (9-60)

REVISION

GENERAL ELECTRIC

RUNNING LIST 198X796 SH. NO. CONT. ON

TITLE RACKS AND CABLES

RUNNING LIST 198X796 SH. NO. CONT. ON

CODE IDENT. NO 99971

WIRE OR LINE NO.	WIRE OR COMP.	WIRE ITEM NO.	FROM / TO	WIRE END ZONE	HARNESS ROUTE	NOTES	REMARKS	TERM ITEM NO	SLEEVE ITEM NO	MARKER ITEM NO	WIRE LENGTH	STRIP LENGTH
199	20N	21	P2-08 A1TB8-06	MO DB	BN	2 3		48 51	54	T0		
200	20G	24	P2-09 A1TB4-06	LO DR	RM	2 3		48 51	54			
201	20G	24	P2-10 A1TB4-07	MO DR	RL	2 3		48 51	54	T1		
202	20G	24	P2-09 A1TB4-07	LO DR	RH	2 3		48 51	54	T2		
203	200	23	P2-12 A1TB9-06	MO DR	RP	2 3		48 51	54	T3		
204	200	23	P2-13 A1TB9-06	PO DR	RS	2 3		48 51	54	T2		
205	200	23	P2-14 A1TB9-16	SO DR	BT			48 51	54	T6		
206												
207												
208												
			B5-10							T3		
219			P3-01 A1TB7-08	DB				51	54	T8		
220	20KG	34	P3-15 A1TB4-11	KO DB	BK			48 51	54	T8		

PRINTS TO | REV. | REV. | REV.

MADE BY | DIV. OR DEPT. | RUNNING LIST

ISSUED | APPROVAL | LOCATION | SH. NO. CONT. ON

Fig. 5. Connection list. This document is the "bridge" between the harness assembly drawing and the equipment assembly drawing.

SOME MYTHS ABOUT INDUSTRIAL ELECTRONICS

Sidney Feldman, Editor
"Shepherd's Electronics Marketing"
140 West 42d Street, New York,NY.

Summary: Engineers may be surprised or even disappointed to learn that electronics manufacturers pursuing the industrial electronics market may be as successful as their marketing--as well as their engineering. However, there are many mistaken notions about industrial electronics where buyers are not interested in engineering virtuosity or sophistication but are interested in solving their problems or cutting costs to raise profits.

Electronics engineers have a professional and economic stake in the broad scope of industrial electronics markets. At the outset, this is a "tent term" which broadly comprises: computers and industrial controls; measuring and test equipment; sea-air-ground navigation aids; two-way radio; microwave communications; nuclear instrumentation; medical and therapeutic equipment; commercial audio and industry-closed circuit television.

Yet, so far, most business seems to be in the communications and computer fields; the others are still developing or are fairly well developed and growing modestly. Many electronic products have yet to come out of their defense electronics cocoon by being re-engineered for industrial use.

In round figures, industrial products (at factory sales) are: $1.6 billions in 1959; $1.75 billions in 1960. Estimates are: $1.9 billions for industrial products sales in 1961 and $2.1 billions for 1962.[1/] Yet, it is my impression that these statistics fluctuate along with the definitions of industrial products. In part, the problem is part-electronics, part-semantics. Thus,statistics change with definitions.

As a result, industrial electronics represents a subject which many electronics people talk about --but which seems to be many things to many men, including engineers. For this reason, industrial electronics has become part-market, part-myth. A fair rule of thumb here seems to be: if you subtract all electronic products which are for defense, aerospace or consumer use--whatever then remains therefor must be: industrial electronics. Consequently, it seems it's not so much what engineers and management people don't know, but what they think they know that isn't so, to paraphrase humorist Josh Billings.

Here, for your consideration, are a dozen myths about industrial electronics. Let us acknowledge at the outset that industry's goods are sold to business and institutional buyers to run their enterprises. (The rest of industry output may be sold to individual ultimate consumers for home or personal use). All goods and services, except for defense and aerospace use, serve individuals as consumers--directly or indirectly--in the long run.

Myth #1.

Electronics engineers should keep their noses to the grindstone, concentrating on transforming electronic technology into hardware--rather than worry about the industrial customers' particular needs and applications.

Not so. Unlike the military and aerospace buyers, most industrial electronics customers do not quite know what they want. Only the big private companies have staff engineers who can write specifications, then work with electronics engineers--provided they both speak the same technical language. Electronics engineers may have to learn to work with middlemen who are consultants hired by private companies. Or, electronics engineers may have to determine private industry's needs, then write specifications, then proceed to develop electronic hardware. And, all along this process, electronics engineers may have to learn how to be "marketing men, diplomats or public relations representatives" for their firms in dealing with certain industrial electronics customers.

Myth # 2.

Industrial electronics serves "industry" only.

Not quite so. This is the big semantics myth. Just how do statisticians define "industrial electronics?" Here is a typical case of misnomer, or mistaken identification. I recently received an announcement from one of the country's top communications firms. The announcement described an "industrial data transmission system." But then it described this "industrial" system which linked the user's "missile plant" with its "computer center." To my way of thinking, ballistic missiles are made in defense plants--not in industrial factories.

May I here suggest two prime distinctions? First, defense is an industry, but it does not directly create commercially useful products. Second, although this electronics communications firm did build an "industrial data transmission system," this system seems to be entirely used for defense purposes and was paid for by funds from the Department of Defense.

This situation parallels that of the employee who was told he was working in a factory making baby carriages during World War II. Since his wife was expecting a baby, this employee volunteered to take home spare parts which he rationalized as "fringe benefits." But, no matter how he combined these supposed baby carriage parts he always assembled a machine gun!

Or, let us briefly consider the electronics components distribution business. For years, these distributors sold spare/repair parts

to electronics servicemen to fix radios, then TV sets. Distributors also sold parts to amateur radio "hams," then to commercial sound systems men, and to high fidelity audio fans who bought component-equipment sooner than component-parts. In recent years, electronics engineers have been placing small-to-medium-size orders (one-sies to ten-sies and more) with electronic parts distributors at the same prices charged direct from the factories making these parts. But these engineers are primarily engaged in R&D and related work--chiefly on defense and aerospace contracts. But, such orders generally are classified by parts distributors as "industrial electronics" accounts. To wit, the rule of thumb here tends to be: if it is not used by a service technician, a ham or an audio hi fi fan--it must be used for "industrial electronics."

Parenthetically, it may be noteworthy here[2/] to recall that only about 24% of the nation's electronics engineers are engaged in commercially supported activities. But about 76% of all engineers and scientists employed by the electronics industries in the U.S. are supported by Government funds, almost entirely for defense/aerospace work.

Myth # 3.

Industrial electronics serves " big business."

So far, yes. So far, mostly big businesses have been able to afford industrial electronics, so they have been able to benefit from industrial electronics. Such big industries are ranked here by the size of their capital outlays in 1960: Chemicals, petroleum, electric/gas utilities, iron and steel, paper and pulp, food and beverages, textiles, stone and glass and clay, non-ferrous metals and rubber products. Please note that there is not a single corner drugstore or bakery or blacksmith's shop on this list.

Most of the big businesses mentioned here are flow-line, process control industries which readily lend themselves to electronics and other automatic controls of big operations, either close-by or in remote areas. However, here too, much experimentation remains to be done. One big three-initial company even has a special group which handles "risk" enterprises, where technical feasibility (let alone marketability) remain to be achieved. Only after feasibility has been demonstrated, then a regular product division is assigned to carry on the operation commercially. "Commercially" here means that the electronics firm anticipates making profits by engineering and making (or assembling) this system, even as the industrial electronics users look forward to making profits from utilizing such a system. To repeat: industrial electronics customers want commercially beneficial results; sheer electronics engineering virtuosity does not thrill them at all.

Moreover, the United States of America is a big country of millions of small businesses, but a few thousand big businesses. Obviously, the bulk of industrial electronics customers initially comes from these big businesses. Yet, it will take more electronics marketing (and that mainly means sales engineering) to innovate compact, relatively cheap, reliable electronic gear to serve the preponderant number of relatively small-to-medium-size businesses in the U.S.

Statistically: in 1959, for example, some 4.8 million firms in the U.S. hired only from 1 to 19 persons. In other words, more than 95% of the businesses in the U.S. hire fewer than 20 persons each. To consider manufacturing only: about 20,000 firms are in manufacturing, but they turn out about 75% of all our manufactures. [3/] More, there is a trend to increasing services as part of our national output. That is, greater numbers of people are being hired in service industries. And here is an area being explored by marketing men and electronics engineers.

Myth # 4.

"Reliability" is to be striven towards only for defense/aerospace electronics hardware. And, even then, this is to be done only in proportion to financing. But customers for industrial electronics will have to manage with whatever reliability can be obtained.

Nothing is further from the truth. The Department of Defense is accustomed to buying new, untried, sophisticated hardware which, at times, does not work. But practically all U.S. business firms, down to a comptroller, may be somewhat untutored in electronics but-they-do-insist-on-getting-their-money's-worth-down-to-a-penny. Comptrollers in private industry are just not interested in subsidizing experimentation in electronics. And, private industry, accustomed to the reliability of automobiles and refrigerators, generally tends to take it for granted that most other products are reliable, too--or should be. Then why pay extra for reliability?

Myth # 5.

Industrial electronics is only a market for hardware.

Not quite. On a percentage basis of dollars spent, it may turn out that more money will have been allocated for "software" than for "hardware" in the long run. In the computer field, more money and more attention are being directed to computer programming , known as "software." A heavy percentage of costs for industrial electronics ultimately will be accounted for by cost accountants as going for: consultation, engineering, installation, testing, maintenance, personnel training, manual preparation, etc.

Myth # 6.

Industrial electronics is a quick and easy contracting job.

Actually, it turns out to be a somewhat

more involved engineering-consulting operation, as suggested in Myth #5. Mark closely that defense/aerospace electronics serves one huge customer: the Department of Defense, which seeks to spend about $50 billions for Fiscal Year 1963, of which about 1/10th (estimated $5.6 billions) may go to electronics firms. Consumer electronics, which runs an estimated $2.1 billions for 1962, can be determined by counting eyes (or ears) for radios, TV sets, tape recorders, etc, then dividing by two to determine the market. But industrial electronics requires a much subtler and elusive blend of economics and electronics.

Industrial customers must be ascertained, then educated, then convinced, then sold, then financed, then kept sold. Electronics firms just do not have to do all this when they sell transistor radios. And the defense/aerospace customer is often a few shades more sophisticated than the hardware he is specifying.

Myth # 7.
Industrial electronics offers a "safe and sure" U.S. electronics market domestically.

Maybe not. So far, U.S. electronics firms are making inroads into foreign markets, bringing to them systems and equipment (especially in electronic data processing and controls). But certain U.S. electronics firms may do technical consulting and engineering work domestically, but cut electronic hardware costs by importing foreign-made components, equipment or whole systems made in foreign countries--to U.S. specifications.

This practice may gain momentum in future particularly since most foreign countries (outside of mutual defense participation by NATO countries) often are kept out of U.S. markets by "Buy American" defense procurement regulations. Yet these same foreign countries have good capabilities in electronics engineering and production. Such countries are saturating their consumer electronics needs, hence seek to build up their industrial electronics markets in their own countries and abroad.

Myth # 8.
Industrial electronics is an " All Electronics" show.

Not always. You get no guarantee. In the process control industries, for example, users long have used pneumatic, hydraulic, and mechanical and electrical and electro-mechanical products and systems which have performed most satisfactorily for a long time.

In some cases, technological mixed marriages are evolved. One example is the "happy marriage between electronic controller and pneumatic control valve" advertised by one company.4/ It hails the "advantages of the electronic controller and superior power and smooth throttling action of pneumatic control valve, etc."

In other cases, electronic controls may be unsafe because of dangerous dust or inflammable raw materials--unless such electronic controls or devices are heavily or expensively shielded. Or, costly heavy cables may be required to connect electronic equipment with devices to be controlled. And, in still other cases, industrial processors are not only satisfied with present non-electronic systems, but they will not change to "new-fangled" electronics equipment.

Myth # 9.
Industrial electronics engineers can handle all the engineering.

Not quite so. At best, virtually all industrial planning which involves use of electronic equipment and/or systems necessitates close cooperation between electronics firm and a non-electronics firm. Each is a specialist in its own right, yet each must cooperate and communicate and integrate two or more technologies and disciplines to help solve industrial customer needs. Simple or trite as this may sound, it can range from merely difficult to nearly impossible.

Here are two examples:

a) In the machine tool industry, machine tool manufacturers try to compete with electronics manufacturers by making their own electronic numerical control units. However, machine tool buyers (customers) prevailed by specifying both the machine tool (made by machine tool makers) and separate numerical control units (made by electronics equipment makers). Thus, in the technically oriented machine tool field, the industrial customer's engineers decided which machine tool to buy and which numerical control to go with it--as an integrated package. But the user-buyer, not the maker-seller, controlled this situation.

b) In bio-medical electronics, there is often a wide communications gap. The biologists and the medical men have yet to determine their needs, then set specifications, then finance purchases. Yet these technical-scientific men also must learn how to communicate their needs and specifications for equipment to electronic engineers to enable them to design useful products and systems to serve bio-medical needs.

Myth # 10.
Industrial electronics offers a good market for pieces of equipment and replacement parts.

Well, ultimately every technical product may require replacement parts. But industrial electronics involves variations on a theme:

a) It is primarily a systems market, not a hardware market. (However, as noted, an "electroler" that is, an electronic controller may be "mated to a pneumatic control valve" but the offspring here represents part of a system, not a complete electronics system.)

b) Industrial electronic systems may be specified by various sources: 1) Outside consultant, serving as middlemen between electronics firms and client customers; 2) The customer, do-in his own systems engineering, may call in an electronics firm as contractor or sub-contractor; 3) Electronics firm may do a "brokering" job, working between client company and a group of sub-contractors.

One big electronic controls firms prides itself as being "the first for which a single manufacturing source supplied all basic elements as well as virtually all components." However, while this was true in this case where a public electric utility was the industrial customer, this is not always the case with all customers--particularly in the chemicals industry.

In the chemicals industry, firms often prefer to have their own engineers design a system, then specify and buy and assemble components. The reason is not "do-it-yourself" economy. The reason is some chemical companies fear having integrated electronics systems suppliers (or any other firms) learn their non-patented "proprietary" or "trade secrets."

Myth # 11.

Industrial electronics is a fairly stable business, unlike the ups and downs of the defense/aerospace business where contracts may be canceled overnight, or projects tightened up, or stretched out.

Nothing, but nothing, is stable in this dynamic world. Although industry may be relatively more sure of itself in capital goods expansion periods when the economy is on the upswing, our economy also is known to have its recessions and depressions.

Industrial electronics managements do have to keep up with several fronts at the same time: a) Industry customer activities, b) Both industry customer competitors and electronics company competitors, c) The U.S. Government and its actions and policies, d) The national economy and its trends.

Myth # 12.

Electronics engineering capability is the foremost prerequisite for prospering in the industrial electronics business.

Not quite so. In the broad sense, management will play the deciding role. Management here includes engineering capability as a base on which to add financing and marketing.

Financing is important in many ways here, particularly in: a) Underwriting a joint venture between an electronics equipment maker and a non-electronics industrial firm, such as a paper and pulp maker, to explore an "art" and possibly transform it into a "technology."

b) In the case where the electronics firm provided both the engineering and virtually all the componenets of the electronic system its engineers designed for a public utility--three of the six industrial products group divisions had been acquired companies. Such corporate planning required foresight and financing.

On the other hand, astute electronics engineering firms may establish themselves as consultants, brokers, or team managers to handle the overall system design, then manage the team of individual corporate electronics sub-contractors participating in an industrial systems installation. Incidentally, this idea which is used in electronics defense/aerospace team-bidding, team-contracting, was adapated from a private industry approach.

In short, corporate engineering-and-management-and-financing-and-marketing systems precede the design and creation of industrial electronic systems.

# # #

Notes: 1/ Electronics Industries Association statistics.

2/ Survey by Department of Defense and Electronics Industries Association.

3/ National Industrial Conference Board "Economic Almanac" and Fortune.

4/ Mason-Neilan advertisement, June 12, 1961, Chemical Engineering.

# # #

"VWP; A NEW APPROACH TO SMALL, LIGHT, EFFICIENT, HIGH POWER REGULATED POWER SUPPLIES"

Victor Wouk
Electronic Energy Conversion Corporation
New York 17, New York

Summary

A regulated power supply has been developed, employing no mains power transformer, with very low power dissipating control elements, and resulting high efficiency. The size and weight are substantially less than those of conventional designs for comparable power ratings.

The mains power is converted directly into dc, and silicon controlled rectifiers are utilized to generate high frequency variable width pulses (VWP) of voltage, for output control, and for line and load regulation. Pulse width modulation is used to reduce line frequency ripple. The system is advantageous for load power ratings above 2 kw.

Introduction

Reduction of size and weight is an increasingly important objective in the development of modern electronic equipment. Where possible, higher operating efficiency is also desirable. The regulated power supply system described herein effects a substantial saving in size and weight over conventional designs, by eliminating the main power frequency transformer, usually a 50/60 cycle transformer, normally employed to raise or lower the line voltage.

The system consists of an input rectifier that immediately converts the main ac power to dc, (without employing a transformer), a chopping circuit that converts the dc to high frequency bi-directional pulses, and a high frequency transformer for stepping the output voltage up or down as desired. In order to increase efficiency, variable width pulses (VWP) are developed to control and regulate the output voltage, rather than utilizing power dissipative techniques such as in series or shunt regulating circuits. The variable width pulses are available either as an inverter for ac output, or, when applied to a rectifier and an LC averaging circuit, as dc output. The over-all system is illustrated in Fig. 1, in block form, for dc output.

Basic Variable Width Pulse Switching Technique

It has been pointed out by Morgan[1] and others, that very efficient regulation of dc power supplies can be achieved by switching techniques. This also applies to ac power supplies; but since the concepts are more readily comprehended when discussed in relationship to dc input, the balance of this paper is devoted to dc supplies. The switching technique, indicated in Fig. 2, does not dissipate energy in a series or shunt impedance, as is done in conventional regulated power supplies. The input dc is switched on and off at a rapid rate by switch "s", developing rectangular pulses of voltage, spaced "T" apart, that are applied to the choke input filter. With proper inductance design so that

$$fL \gg R_l \qquad (1)$$

where $f = \frac{1}{T}$

the voltage Eav developed across the resistor R_l will equal the average of the rectangular wave, which is determined by:

$$Eav = E_i \times \frac{t}{T} = k \times E_i, \text{ where } k = \frac{t}{T} \qquad (2)$$

If "k" is referred to as the duty cycle, the dc output voltage equals the input voltage times the duty cycle. This well established concept is not discussed further herein.

It is important to note that a rectifier CR1 connected as illustrated in Fig. 2, must be employed, for this circuit to operate properly. CR1 is commonly referred to as a "free-wheeling" rectifier[2], or "flyback" diode.

This circuit can be made to regulate the output voltage against input voltage changes and load current changes, by varying the width of the pulse, or the duty cycle, i. e., by varying the time "t". This is done by allowing the switch "s" to remain open or closed for greater or shorter periods of time. If the switching rate is rapid (over 1 kc per second), the pulse width can be modulated to compensate for input dc ripple voltage, when the system is operated from rectified ac. Hence, this system of output voltage control is referred to as "Variable Width Pulse", (VWP), as the average width is varied to give the average output voltage desired, and rapid width variations, or pulse width modulation, compensates for perturbations of the output by various causes.

Although this concept of variable width pulse control of dc and ac power has been known for over half a century, it has been made practical only by the development of modern high speed switching

devices having very low voltage drops when carrying large amounts of current, and negligible conduction when "off". These switches may be power transistors, or silicon controlled rectifiers (SCRs). For high power levels, above 2 kilowatts, SCRs are superior to power transistors, due to, among other reasons, the smaller driving power required for SCRs, and the lower voltage drops.

The balance of this paper is devoted to analyses of circuitry and problems peculiar to the use of SCRs in the variable width pulse (VWP) mode.

Dual Polarity Switching

In the switching technique described for Fig. 2, and in all dc "choppers" described in other literature, the output voltage is almost invariably lower than the input voltage, and single polarity pulses are generated for subsequent filtering and smoothing to the desired output voltage level. Where output voltage greater than the input voltage are desired, for both dc power supply and inverter applications, it is necessary to employ a transformer to step the voltage up to the desired level. Also, for very low ratios of output dc to input dc voltage, a step down transformer is usually employed, otherwise very short duration pulses in simple chopper circuits are mandatory, resulting in extremely high peak currents in the switching device, and consequent uneconomical design.

Dual polarity pulses must be developed to drive a transformer, as unipolarity pulses will cause the transformer to saturate. If this transformer is operated at a frequency higher than line frequency, the transformer can be much smaller for a given power rating, than the equivalent 60 cycle transformer.

> NOTE: It is feasible to obtain output voltages higher than input voltages without a transformer, by using the well known voltage multiplier circuit of Greinacher[3]. All these circuits have the disadvantage of providing no isolation between the primary and secondary circuits. Hence, all the discussions hereafter will be limited to transformer input.

Typical of the circuits developed for variable width pulse operation is that shown in Fig. 3, from the paper by McMurray and Shattuck[4]. This circuit employs 4 SCRs in a bridge circuit, to utilize fully the ratings of the SCRs. The improved commutation, effected by the diodes, and the general performance of the circuit, are described in the paper.

There are commutation losses of power in L and C of Fig. 3, associated with the turning on and off of each pulse. Because of this, an inverter whose efficiency might be 96% at 120 cps, may be down to 67% at 1 kc, with the output power rating dropping more than 50% from 120 cps to 1 kc. For maximum efficiency, particularly in dc power supplies, where it is desired to operate at a high switching frequency to reduce the size and weight of the filter components, a circuit to eliminate this commutating power loss is desirable.

Such a circuit, and several variations thereof, have been developed by Wouk and Poss[5], a typical one being illustrated in Fig. 4. In this circuit, it is important to emphasize, <u>no power at all is inherently lost in the commutating process</u>, (except internal SCR and rectifier drops, lead losses, etc.) All of the commutating capacitor energy passes through the load, so that the losses in the commutating capacitor are only those incidental to dielectric losses. Hence, this is inherently a very efficient circuit, and is referred to hereafter as "lossless" commutation.

The operation of the circuit of Fig. 4 is as follows:

1: Assume that capacitor C is charged to the polarity and voltage shown, i.e., equal to the input supply voltage "E". For simplicity of analysis, it will be assumed that the output transformer is a 1:1 voltage transformer, i.e., twice as many total primary turns, center-tapped, as secondary, or , as illustrated, $N_1 = N_2$. It is further assumed that the system is operating under steady state conditions, and a constant load current, I, is flowing in the output filter choke.

2: SCRs 1 and 2 are fired, and e_{ad} and e_{dg} are zero, to energize what is arbitrarily called the positive power pulse. This is illustrated in Fig. 5, at the start of the period "t_1", as are the fact that the source current, i5, and the SCR currents i and i_2, equal the load current, I (due to N1/N2 =1).

3: After a time t1, SCR3 is fired to deenergize the power pulse. This places point "e" at ground potential, and drops point "d" to a voltage -E, below ground. This applies a reverse voltage to SCR1, which ceases conducting in less than 5 microseconds. The load current which had been flowing through SCRs 1 and 2 now flows through SCRs 2 and 3. This is also illustrated in Fig. 5, as i_3 is identical to i_2 during this interval.

4: In a rectifier application, with inductance input, it can be assumed that the load current is a constant through the inductor as a function of time, particularly if equation (1) applies. Therefore, capacitor C will discharge linearly, and the volt-

age across the transformer input will decrease linearly with time to zero. This is also shown in Fig. 5, curve e_c, which is also e_{de}.

5: When the voltage of point "d" reaches that of the input voltage, i. e., "E" volts, SCR2 will cease to conduct, as will SCR3, and the power pulse output will be zero. This is illustrated in Fig. 5, at the end of period t2.

A very important fact is that at the end of t1, when SCR3 fires to turn off SCR1, point "d" is at -E voltage, and, since SCR2 is still conducting, the voltage at point "a" is also essentially -E. However, at the input to the transformer, point "b", the dc voltage is +E. Therefore, there appears across the conducting portion of the primary of the transformer a voltage step of 2E, shown in Fig. 5, in curve e_{xy}, at the beginning of period t2. This is an extremely important phenomenon, as it is the heart of the concept of the zero commutation loss circuitry, and, although it aids in the efficiency, it introduces several design problems. For example, the voltage across SCR4 is momentarily sent to a forward voltage of 3E, and hence a much higher voltage rating is required for this SCR.

After a quiescent period, in order to start the negative pulse, SCRs 3 and 4 are fired. Then, to turn the negative pulse off, SCR1 is fired. Action similar to that described for SCRs 1 and 2, and 3, now obtains for SCRs 3 and 4, and 1, as capacitor C had been charged, during the turnoff of the positive pulse, to the correct polarity for the turnoff of the negative pulse.

In the curves of Fig. 5, ideal commutating conditions are assumed, such as no transformer leakage inductance, no circuit "ringing", etc. These effects exist in any practical circuit, and can be corrected for, where necessary, by standard techniques.

Since voltages and currents in the SCR3 and SCR4 branch of the circuit are identical to corresponding values in the SCR1-2 branch, shifted 180° they are not illustrated in Fig. 5 .

Note that in both cases of pulse turnoff, the energy stored in the commutating capacitor C is returned to the load, through the mechanism of the twice-input-voltage triangular pulse shown at the turnoff of each power pulse.

Further, since the capacitor is charged through the load circuit, the $1/2\ CE^2$ loss normally associated with charging a capacitor from a dc source[6] actually is dissipated in the load. It is this $1/2\ CE^2$ loss that accounts for the low efficiency of conventional "parallel inverter" circuits at high frequency. This loss is NOT PRESENT in the VWP circuit described herein.

The operation of the rectifier and pulse averaging portion of this circuit is hereafter analyzed at length, mathematically, because of the fact that the output voltage is a function of the output resistance, and therefore the regulation problem is somewhat more complex than in non-energy-conserving regulating systems.

Waveform And Circuit Analysis

In Fig. 6 the secondary of the output transformer of Fig. 4 is shown with a full wave bridge rectifier. The waveform input to the filter is shown now as unidirectional pulses, the average of which will appear at the output. This waveform and some of the important parameters, are shown in Fig. 7.

Two of the important parameters of the system are the rectangular section duty cycle, k1, and the triangular section duty cycle, k2.

Obviously,

$$Eav = \frac{Et_1 + \frac{2\,Et2}{2}}{T} = E\,\frac{(t1 + t2)}{T} = E(k1 + k2) \quad (4)$$

or,

$$\frac{Eav}{E} \equiv k3 = k1 + k2; \text{ as a limitation, } k1 + k2 \leq 1 \quad (5)$$

This set of parameters is plotted in Fig. 8, with further parameter limitations discussed later.

In practice, t2 cannot be controlled directly, but t1 can. With a given input voltage E, the output voltage is controlled by varying t1. However, this is not the only factor determining the output voltage Eav. It is shown in Appendix I that:

$$k3 = \frac{k1}{2} + \sqrt{\left(\frac{k1}{2}\right)^2 + k4} \quad (6)$$

$$\text{where } k\,4 = \frac{2RC}{T} \quad (7)$$

$$\text{or, } \frac{Eav}{E} = \frac{t1}{2T} + \sqrt{\left(\frac{t1}{2T}\right)^2 + \frac{2RC}{T}} \quad (8)$$

where "C" is the commutating capacitance of Fig. 4, and R is the load resistance of Fig. 6, transformed to the primary as indicated in Fig. 4.

Equation (8) shows that the ratio of the average voltage to the input voltage depends not only on the rectangular portion of the "on" part of the SCR chopping wave, but also depends upon the load resistance, the commutating capacitance, and the

repetition period of the pulses.

Since, in any practical device, the commutating capacitance is normally fixed, the output average voltage can be controlled to some degree by varying the period T. But, since all factors entering into the equation for k3 are inverse functions of the period, it is best, from an analytical point of view, to maintain the period fixed, and vary only the nominal pulse on time t1. With T fixed, equation (6) can be plotted as a series of curves for fixed k1, with k4 varying independently, or vice versa. If T is varied, this may not be done.

The variation of the output voltage as a function of the rectangular portion of the wave, and the ratio of the resistance-capacitance product to the period, i.e., equations (6) and (8), is indicated in the upper part of Fig. 9. It is seen from the upper half of this drawing, that for a small rectangular duty cycle, k1, the output voltage can vary over a very wide range, depending upon the load resistance. Or, for a fixed, narrow duty cycle k1, the inherent regulation of a dissipationless VWP dc power supply is very poor. However, the system is still very efficient. Further, the load resistance R may also vary over a very wide range for a small k1.

As the duty cycle, k1, of the rectangular portion is increased, the output voltage increases, and the permissible variation of k3 decreases. This corresponds to higher output power, and a narrower permissible range of R, at lower resistance values.

This is a limiting factor on the performance of this device, namely, operation over a comparatively narrow range of output voltages and load currents, without some type of modification of a major parameter.

The triangular period t2, or duty cycle, k2, is of great practical importance. As stated in equation (5), $t1 + t2 \leq 1$, otherwise the inverter pulses will overlap, and both arms of the inverter will conduct simultaneously, causing a short circuit on the dc input. So, t2 must not be too large.

The turnoff period of the SCRs, i.e., the time that the voltage across the SCR being turned off by the negative voltage from the capacitor C (e.g., SCR1 or 3 in Fig. 4), is negative, must be greater than a minimum value, depending on the type of SCR, among other factors. This period can be as low as 10 µs, reliably, under certain favorable conditions, or as high as 30 µs. From curve e_{dg} of Fig. 5, it can be seen that this period is $\frac{t2}{2}$. Therefore, t2, or k2, cannot be allowed to be too low in value.

In Appendix I it is shown that the triangular pulse "duty cycle" k2, is a function of the input voltage and load current as follows:

$$t2 = \frac{2EC}{I} \tag{9}$$

In Appendix II it is shown that k2 depends upon k1 and k4, so that:

$$k2 = \frac{-k1}{2} + \sqrt{\left(\frac{k1}{2}\right)^2 + k4} \tag{10}$$

This relationship is shown in the lower portion of Fig. 9. From this it again is seen that for a given k1, k2 decreases as the load resistance decreases, for the load current increases, and hence the slope of the triangular section is much steeper.

It will be noted that the curves of Fig. 9 are drawn for a minimum limit of k1 = 0.05, and maximum k1 = 0.95. Although in theory both variables k1 and k2 could extend to 0, and k1 and k2 could each equal 1, the practical aspects of SCRs require a minimum rectangular pulse width, a minimum triangular pulse width, and a minimum time between pulse turnoff of one polarity, and pulse firing of the next polarity.

Due to the buildup time of gate current of most SCRs with simple gate pulsing circuits, plus the requirement of time for current to build up in the SCR, it is impractical to have t1 reliably much smaller than 20µs. Similarly, it is unreliable to have an off time between pulses much less than 20µs. Finally, t2 must be at least 20µs, as previously mentioned, otherwise the SCRs may not turn off, and there may be simultaneous firing of SCRs in different arms of the inverter, which will short the dc supply. A k1 of 0.05 at 20µs would correspond to a rectified pulse repetition rate of 2.5 kc, or an inversion frequency of 1250 cycles per second. For a dc power supply, this operating frequency will yield substantial reduction in size and weight of the power transformer, and of the output filter components.

In order to have smaller values of k1, the inversion frequency must be lower, causing a consequent increase of size of transformer, choke and capacitor.

If the operating frequency is higher, then the range of variation of k1 and k2, will be smaller, with an even smaller allowable range of load resistance R1.

The curves of the lower half of Fig. 9 further indicate that for a wide rectangular pulse, or large duty cycle, k1, the allowable variation of load resistance ratio is not great, as for too low a resistance there will be too low turn off time, k2,

and for too high a resistance the pulse turn off time will be too great, causing pulses to overlap.

For example, with a rectangular duty cycle of 0.45, the allowable resistance ratio, determined by k2 max = 0.5, and k2 min = 0.05, (from the lower curve of k1 = 0.45 of Fig. 9) is from k4 = 0.45 to k4 = 0.025, or a resistance ratio of approximately 20:1. However, the voltage will drop from a ratio of k3 = approximately 0.95, to k3 = approximately 0.50, almost 2:1. The current swing will be greater than the voltage range, but less than the resistance range, a ratio of

$$0.48/0.025 \Big/ 0.95/0.5 \cong 10:1$$

The above data are all for an UNREGULATED system, with fixed k1.

These curves will also indicate the pulse width variation necessary for regulation against load current changes. It is apparent that the closer the desired output voltage to the input voltage, i.e., the greater k3, the wider the range of output current variation permissible. An example of the use of these curves is given in Appendix III.

The curves of Figs. 8 and 9 could be called the "Characteristic Curves" of the VWP, "lossless" commutating circuit. Fig. 8 has the "Output Voltage Characteristics", and Fig. 9 the "Regulation Characteristics" (upper curves) and the "Turnoff Characteristics" (lower curves).

Fig. 10 shows an experimental run to verify equation (6). The data are plotted as a function of the dimensionless parameters k1, k3 and k4. The correlation between theory and experiment is excellent. It is seen that the inherent regulation, for fixed k1, is very poor.

The curves of Fig. 11 show regulation in a more conventional form, i.e., output voltage versus output current. It is seen that the inherent regulation is much better for wider pulses, as the no-load voltage for all pulse widths tends to the same value, corresponding to k3 = 1, per Fig. 9.

The regulation of the power supply can be made as good as desired, by standard techniques, illustrated in block form in Fig. 1. Automatic regulation circuits are not discussed herein.

Filtering

The harmonic content of the wave form, of the VWP "lossless" commutation power supply, for various rectangular portion and triangular portion duty cycles, has been calculated in Appendix IV, and the first 4 components are indicated in the drawings therein, Figs. 13 - 16. The amplitudes of the harmonics are plotted as a percentage of the rectangular pulse height.

It is seen, for example, that the maximum fundamental harmonic shown in Fig. 13, in case of the dc wave, is 0.84, occuring with k1 = 0.15, at k2 = 0.52. As is to be expected, as k1 increases beyond 0.5, the fundamental drops substantially, and becomes even smaller as k2 increases. Thus, if a design can be such that wide rectangular pulses are used throughout, the filter problem is reduced substantially.

The second harmonics, illustrated in Fig. 14, show much greater variation as a function of k1 and k2 than does the fundamental. Further, after k = 0.5, the curves repeat for k1 intervals of 0.5. For the larger k1, since k1 + k2 ≤ 1, on any curve the terminal point is indicated.

As is to be expected, as the harmonic order increases, the convolutions of the curves increase in frequency. Fig. 15 shows the 3rd harmonic, and Fig. 16 the 4th harmonic.

Inverter

Since the VWP dc power supply circuit is essentially an inverter to begin with, it can be used as an inverter, with the understood limitations of waveform impurity. Here, the operating frequency can be much lower than in a dc output power supply, and still obtain advantages, namely, of very high efficiency. Under these circumstances, in most practical cases, it can be assumed that the current is constant during the commutation period, in which case all the formulas previously evolved can be applied. If it is assumed that the output waveform consists of rectangles and triangles, but not rectified, as shown in Fig. 5, curve e_{xy}, then the harmonic content can be analyzed to determine the necessary filtering for purity of ac output waveform. This is done in Appendix V, where the equations for the harmonics are derived, the amplitudes as a function of k1 and k2 are plotted in Figs. 17-20, for the fundamental, 3rd, 5th and 7th harmonics.

Practical Design

A dc power supply incorporating the VWP principle and the elimination of the 60 cycle transformer has been built and operated successfully, in accordance with the block diagram of Fig. 1. The anticipated advantages of small size, light weight and high efficiency were achieved. Because of the cost of the solid state rectifiers and switches, the system does not become economically competitive, using SCRs, below an output of 2 kilowatts. There-

for, a 2 kw design was the first one attempted.

Fig. 21 is a photograph of a 2 kw unit, 7" high (compared with 15" or more for conventional designs), weight 60 pounds, (compared with 200 pounds or more for conventional designs). The overall conversion efficiency ac to dc, for 1/2% regulation is 83%, compared with 70% for conventional series impedance designs. Ripple and regulation are less than 1/2%, though they can be made lower without undue engineering effort. The unregulated dc input has 10% ripple at 120 cycles and this is reduced to below the specified value by pulse width modulation.

This system has the further advantage of operating from a wide range of input frequencies. The lower limit of input frequency is determined by the amount of output voltage control that may be sacrificed due to the requirement of additional pulse width modulation for derippling. The upper limit of input frequency is determined by the input rectifiers' characteristics. Operation ranges from 25 cps to 2000 cps are anticipated.

The one limiting performance factor of the basic circuit shown in Fig. 1 is transient response with respect to load changes, due to the output filter choke. At 1 kc operating frequency, the transient recovery time is of the order of tens of milliseconds. However, this can be improved by auxiliary circuits to control transient energy.

As an ac-ac inverter, the overall efficiency is 90%. As a dc-ac inverter, operating from 150 volts dc, the efficiency is close to 95%.

Conclusion

A new approach to dc and ac power supplies, regulated and unregulated, resulting in smaller size, lighter weight, and higher efficiency than conventional designs, has been developed. Analytical approaches, and experimental verification of theory, have been described in this paper. As the costs of solid state devices continue to drop, this technique will prove to be economically competitive with existing devices, and hence more advantageous.

Appendix I

Calculation of $\frac{Eav}{E} \equiv k3$

Referring to Fig. 4, when SCR3 is fired and SCR1 extinguished, a process that takes but a few microseconds, thereafter the circuit can be considered that of Fig. 12. The following assumptions are made in reference to both Fig. 4 and Fig. 12:

(1) Transformer T has a 1:1 ratio of input and output, as employed. (2:1 ratio of total input and output turns).

(2) T is lossless, as are all connecting leads.

(3) All SCRs and rectifiers are lossless.

(4) Choke L is lossless.

(5) The current I that flows during commutation is constant.

(6) The change in capacitor C voltage during turn-off of SCR1 is negligible.

Then, from equation (5),

$$k3 = \frac{Eav}{E} = k1 + k2 = \frac{t1}{T} + \frac{t2}{T}$$

To determine t2, referring to Fig. 7 and Fig. 12, we use the fact that:

$$\frac{dE}{dt} = \frac{i}{c} = \frac{\Delta E}{\Delta t}$$, since all changes are linear.

Applying this to the triangular portion of the wave of Fig. 7,

$$\Delta E = 2E$$

$$\Delta t = t_2$$

$$i = I$$

so $t2 = \frac{2EC}{I}$, which is equation (9).

but $I = \frac{Eav}{R}$

so $t_2 = \frac{2ECR}{Eav} = \frac{2CR}{k3}$

Putting the above into (4) yields:

$$k3 = k1 + \frac{2CR}{k3\,T}$$

Let $k4 \equiv \frac{2CR}{T}$

Then $k3 = k1 + \frac{k4}{k3}$

$$k_3^2 - k_1 k_3 - k_4 = 0$$

or $k_3 = \frac{k1}{2} \pm \sqrt{\left(\frac{k1}{2}\right)^2 + \frac{2RC}{T}}$

only the + solution is physically acceptable, yielding equations (6) and (8). This is plotted in the upper half of Fig. 9.

Appendix II

Calculation of $k_2 \equiv \frac{t_2}{T}$

From equation (4),

$$k_3 = k_1 + k_2$$

From equation (6),

$$k_3 = \frac{k1}{2} + \sqrt{\left(\frac{k1}{2}\right)^2 + k4}$$

Subtracting the above two equations, (6) from (4), yields:

$$k2 + \frac{k1}{2} - \sqrt{\left(\frac{k1}{2}\right)^2 + k4} = 0$$

or $k2 = -\frac{k1}{2} + \sqrt{\left(\frac{k1}{2}\right)^2 + k4}$, which is equation (10).

This is plotted in the lower half of Fig. 9.

Appendix III

Determination of Pulse Widths Required in VWP "Lossless" Commutation Circuit, for Regulation Against Load Changes

Referring to Fig. 9, assume k3 = 0.85, or $\frac{Eav}{E}$ = 0.85, a desired ratio of output dc to input dc (normalized for unity transformer ratio).

Then, entering horizontally from k3 = 0.85, we have a range of k1 = 0.85 to 0.05. This corresponds to k4 = 0 and k4 = 0.67. The latter value is acceptable, the former is not, as it requires R=0, and k2 =0.

Since k2, as discussed in the main text of this paper, must be greater than a minimum amount, say 0.05, we must seek a maximum k1, whose value of k4, for k3 $\doteq$ 0.85, will yield k2 $\geq$ 0.05.

Trying the k1 =0.80 curve (interpolated), on the horizontal line k3 = 0.85 we find k4 = 0.04. Descending vertically on the k4 = 0.04 line, we find this line intersets the LOWER k1 =0.80 characteristic line very close to k2 = 0.05.

So, the range of k1 = 0.80 to 0.05. If T were 1,000 µs, this would correspond to a rectangular pulse width of 50 to 800 µs. The output voltage would be constant for this set of parameters (assuming the input voltage constant) over a range of k4 = 0.04 to 0.66, a range of current of approximately 16.5/1.

At a lower ratio of $\frac{Eav}{E}$, corresponding to a practical power supply design in which it is desirable to vary the output voltage over a reasonable range, say 50-100%, the range of acceptable resistance changes is much less at the lower extreme of output voltage. In the above case, 50% x .85 = 0.425 = k3.

At k3 = 0.425, going horizontally to the upper curve of k1 = 0.05, yields k4 = 0.165 maximum. As above, we must determine k1 maximum such that k2 ≥ 0.05. Trying k1 = 0.40 (interpolated), yields k4 = 0.02 and, on the k4 =0.02 line, the intersection with the LOWER k1=0.40 curve yields k2 = 0.04, too low. So, trying k1 = 0.35 yields k4 = 0.03, and k2 = 0.055, satisfactory.

Hence, the current ratio permissible at the lower output voltage is 0.165/0.03 = 5.5/1, for regulation of the output voltage.

Appendix IV

Calculations of Harmonics of Rectified DC Output Waveforms

Referring to Fig. 7, the waveform e (t) can be expressed in a Fourier series as

$$e(t) = \frac{Ao}{2} + \sum_{n=1}^{\infty}\left\{An \cos\left(\frac{2\pi n\, t}{T}\right) + Bn \sin\left(\frac{2\pi n\, t}{T}\right)\right\} \quad (11)$$

$$\text{where } An = \frac{2}{T}\int_0^T e(t) \cos\left(\frac{2\pi n\, t}{T}\right) dt \quad (12)$$

$$\text{and } Bn = \frac{2}{T}\int_0^T e(t) \sin\left(\frac{2\pi n\, t}{T}\right) dt, \quad (13)$$

$$\text{or } e(t) = \sum_{n=o}^{\infty} Cn \sin\left(\frac{2\pi n\, t}{T} + \phi n\right) \quad (14)$$

where $Co = E\,(k1 + k2)$ (15)

$\phi o = \pi/2$

$Cn = \left[An^2 + Bn^2\right]^{1/2}$ (16)

$\phi n = \tan^{-1} \frac{An}{Bn}$ (17)

Following a technique of Parzen (7) to simplify the mathematical analysis, in Fig. 7 have t = 0 at the END of the period t_2.

$$\left.\begin{array}{lll} \text{Then } e(t) = E & ; -T(k1+k2)<t<-k_2 & (a) \\ = -\frac{2E}{Tk2}\,t & ; -T\,k2<t<0 & (b) \\ = 0 & ; 0<t<T\,(1-k_1-k_2) & (c) \end{array}\right\} \quad (18)$$

Putting equations (18) into (12) and (13), and integrating from t = -T (k1 +k2) to t =0 (the period 18c can be neglected, as e(t) =0 therein), and evaluating the intergrals yields:

$$An = \frac{1}{n\pi}\left[\sin\left\{2\pi n(k_1 + k_2)\right\} + \sin(2\pi nk_2)\right] - \frac{1}{n^2\pi^2 k_2}\left[1 - \cos(2\pi nk_2)\right] \quad (19)$$

$$Bn = \frac{1}{n\pi}\left[\cos\left\{2\pi n(k_1 + k_2)\right\} + \cos(2\pi nk_2)\right] - \frac{1}{n^2\pi^2 k_2} \sin(2\pi nk_2) \quad (20)$$

$$\phi n = \tan^{-1}\frac{An}{Bn}$$

The values of (19) and (20), (16) and (17), have been calculated on a computer, at intervals of 0.05, for $0.05 \leq k1 \leq 0.9$, and $0.05 \leq k2 \leq 0.9$. *The coefficient of Cn, i.e., equation (16), are plotted in Figs. 13 - 16 herewith.

Appendix V

Calculations of Harmonics of Inverter Output Waveforms

Referring to Fig. 5, waveform e_{xy}, the waveform e(t) can be expressed in a Fourier series as:

$$e(t) = \sum_{n=o}^{\infty}\left[A_{2n+1}\cos\left\{\frac{2\pi(2n+1)}{2T}t\right\} + B_{2n+1}\sin\left\{\frac{2\pi(2n+1)}{2T}t\right\}\right] \quad (21)$$

In comparing equation (21) to equation (11), it is seen that there is no dc component, and there are no even harmonics, due to the symmetry around the zero axis.

Equation (21) may be rewritten as:

$$e(t) = \sum_{n=o}^{\infty} C_{2n+1}\sin\left[\left\{\frac{(2n+1)\pi t}{T}\right\} + \phi 2n+1\right] \quad (22)$$

$$\text{where } C_{2n+1} = \left[A^2_{2n+1} + B^2_{2n+1}\right]^{1/2} \quad (23)$$

$$\phi 2n+1 = \tan^{-1}\frac{A2n+1}{B2n+1} \quad (24)$$

Using the same technique of Parzen (7), to simplify the mathematical analysis, we set t=0 at the end of the first period t2, in Fig. 5, waveform e_{xy}, and from $t = -T(k_1 + k_2)$ to $t = T(1-k_1-k_2)$, the conditions of equation (18) apply.

Hence,

$$A_{2n+1} = \frac{2}{T}\int_{-T(k1+k2)}^{0} e(t)\cos\left\{\frac{2\pi(2n+1)}{2T}t\right\} dt \quad (25)$$

$$B_{2n+1} = \frac{2}{T}\int_{-T(k1+k2)}^{0} e(t)\sin\left\{\frac{2\pi(2n+1)t}{2T}\right\} dt \quad (26)$$

Putting equations (18) into (25) and (26), and integrating over the limits as shown in (25) and (26) and evaluating the integrals, yields:

$$A(2n+1) = \frac{2}{(2n+1)\pi}\left[\sin\left\{(2n+1)\pi(k_1 + k_2)\right\} + \sin\left\{(2n+1)\pi k2\right\}\right] - \frac{4\left[1 - \cos\left\{(2n+1)\pi k_2\right\}\right]}{(2n+1)^2\pi^2 k_2} \quad (27)$$

$$B(2n+1) = \frac{2}{(2n+1)\pi}\left[\cos\left\{(2n+1)\right\}\pi(k1 + k2) + \cos\left\{(2n+1)\pi k2\right\}\right] - \frac{4}{(2n+1)^2\pi^2 k_2}\sin\left\{(2n+1)\pi k2\right\} \quad (28)$$

The values of (27), (28), (23) and (24), have been calculated and tabulated at intervals for 0.05, for $0.05 \leq k_1 \leq 0.9$, and $0.05 \leq k_2 \leq 0.9$*. The coefficients C2n+1 (equation (23)) are plotted in Figs. 17-20 herewith.

---

*These are available from Electronic Energy Conversion Corp.

Acknowledgements

The author wishes to acknowledge the valuable consultation of Mr. Eliasz Poss on circuit and commutation concepts, and to Computech, Inc., of New York City, for assistance in computing the coefficients plotted in curves 13 - 20.

References

1. Morgan, R. E.: "A New Power Amplifier Using A Single Controlled Rectifier And A Saturable Transformer." AIEE, 60-410 (1960).
2. Gutzwiller, F. W., et al: "Silicon Controlled Rectifier Manual." General Electric Co., (1961).
3. Greinacher: "Uber Eine Methode Wechselstrom Mettels Elecktrischer Ventile undKondensatorem in Hochgespannten Gleichstrom Umzerivandeln," Zeitschrift Fur Physics. vol. 4, p. 195 (1921).
4. McMurray, W., Shattuck, D. P.: "A Silicon Controlled Rectifier Inverter With Improved Commutation." AIEE, 61-718 (1961).
5. Wouk, V.: "Energy Wasted In Charging A Condenser." Communications (April 1944).
6. Poss, E., Wouk, V.: Patent Applied for.
7. Parzen, P., Plainview, N.Y., private communication.

Addendum

The author believes that in the VWP circuit, the phrase "SCR turnoff" is more accurate than "commutation". However, common usage has been followed.

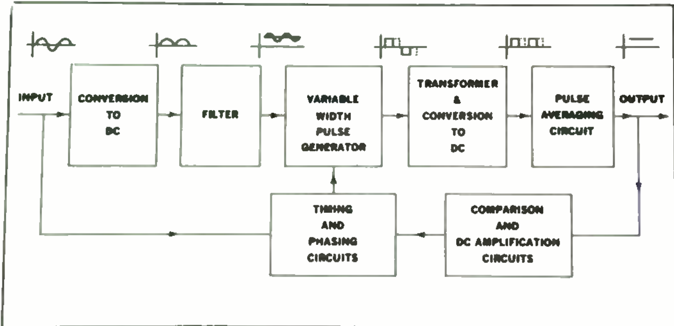


Fig. 1. Block diagram of system employing no main power transformer, and using SCRs for variable width pulse generation.

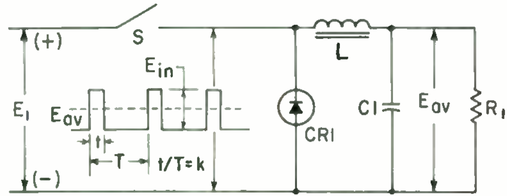


Fig. 2. Basic dc switching control and regulating circuit.

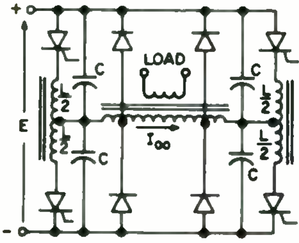


Fig. 3. Typical VWP circuit, employing SCRs, with inherent commutation losses.

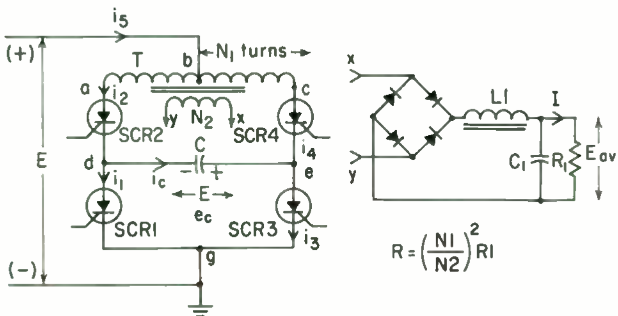


Fig. 4. SCR circuit for VWP generation, with no inherent losses.

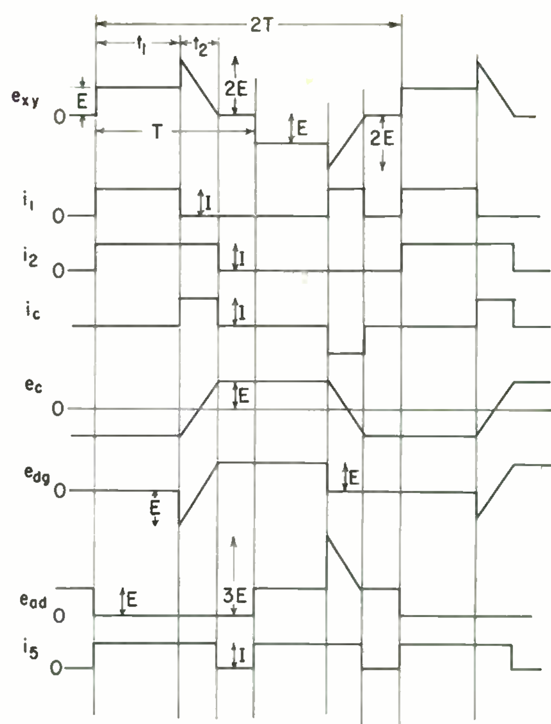


Fig. 5. Waveforms of voltages and currents for Fig. 4.

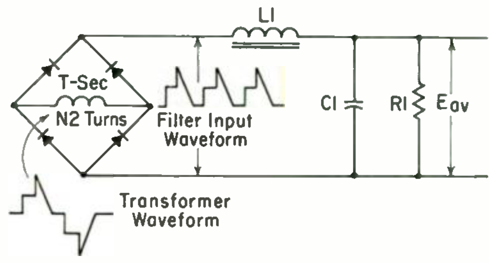


Fig. 6. AC waveform rectifying and filtering circuits, for Fig. 4.

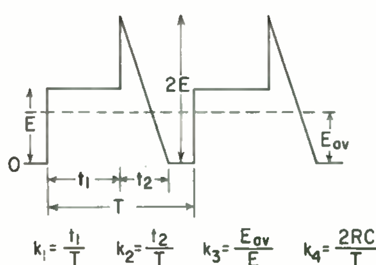


Fig. 7. Details of waveform into filtering circuit, for dc output.

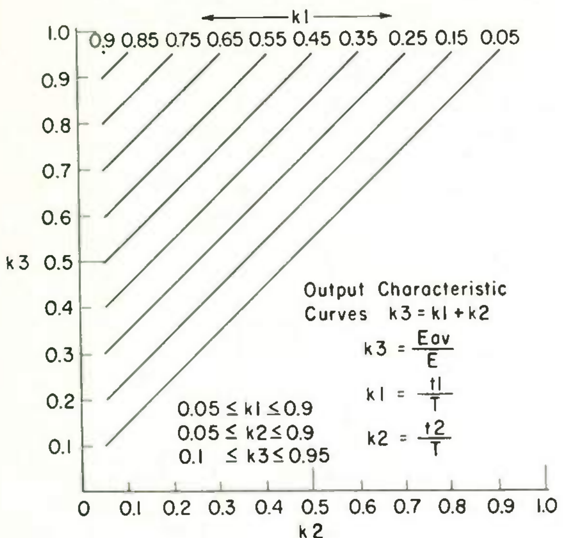


Fig. 8. Output voltage "characteristic curves" for VWP "lossless" commutation circuit.

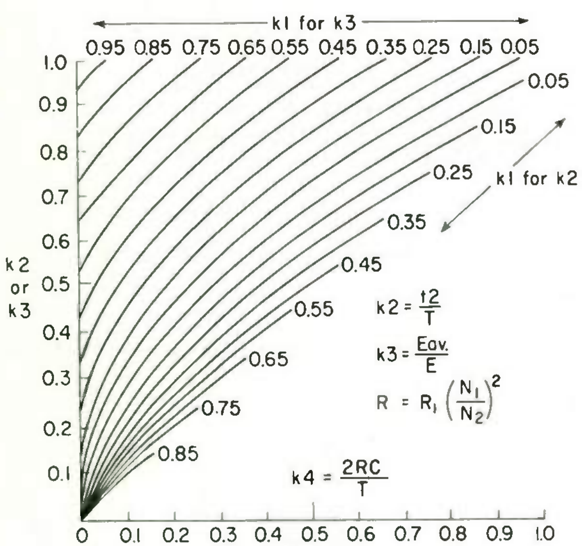


Fig. 9. Regulation and turnoff "characteristic curves" for VWP "lossless" commutation circuit.

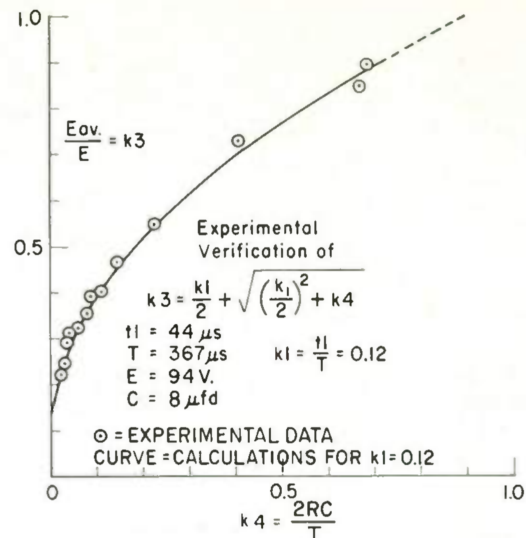


Fig. 10. Experimental verification of regulation curve for VWP "lossless" commutation circuit.

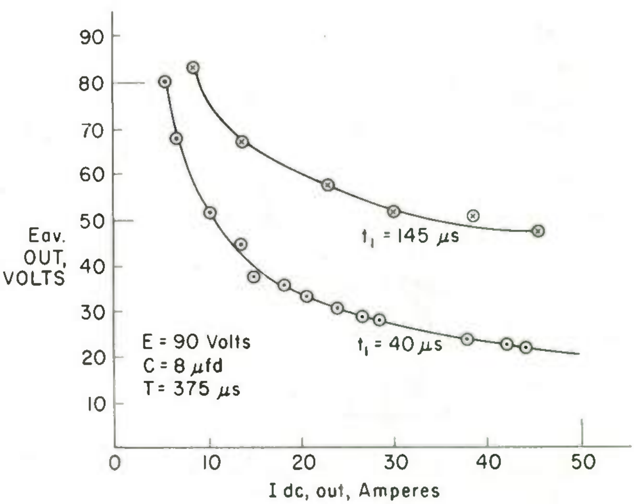


Fig. 11. Typical regulation curves of VWP "lossless" commutation circuit, for various rectangular pulse duty cycles.

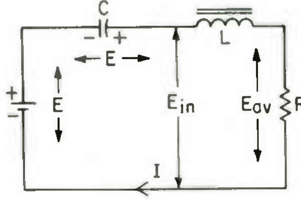


Fig. 12. Simplified circuit during power pulse turnoff.

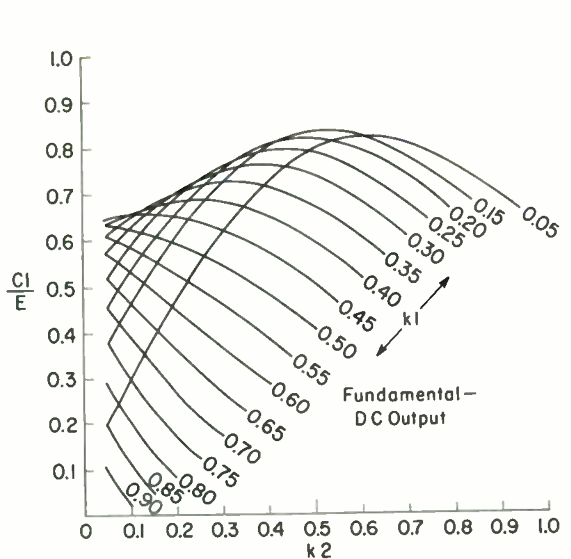


Fig. 13. Fundamental of harmonic content of waveform of Fig. 7, for dc output.

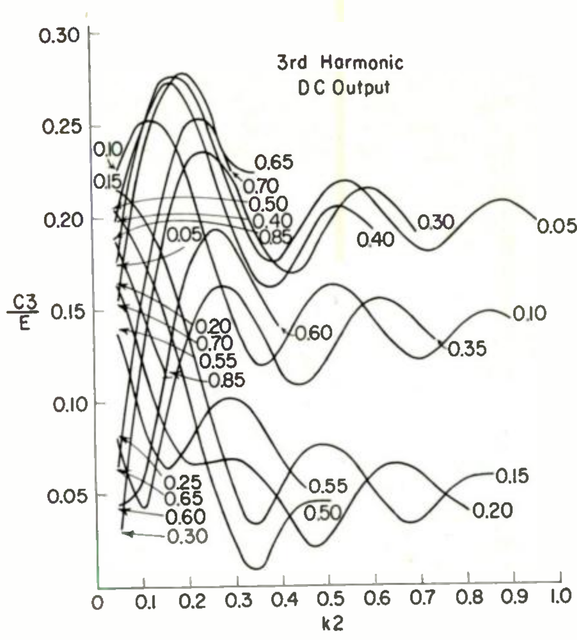


Fig. 15. 3rd Harmonic of waveform of Fig. 7.

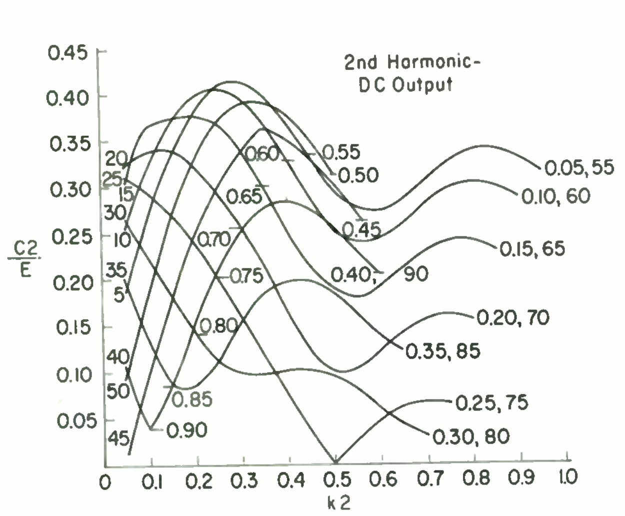


Fig. 14. 2nd Harmonic of waveform of Fig. 7.

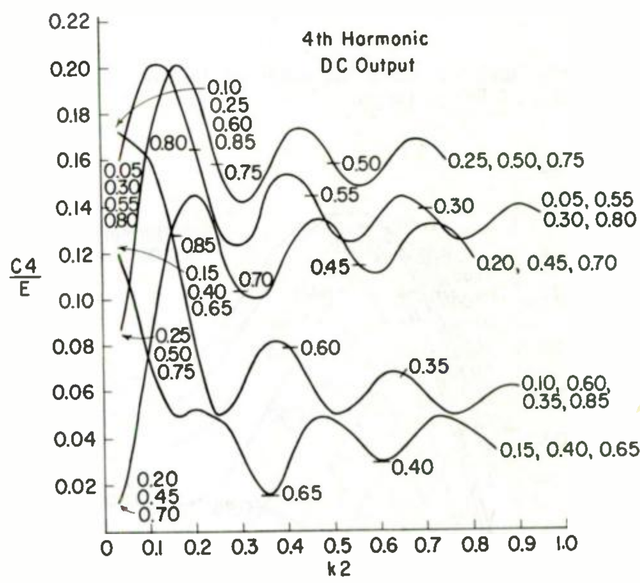


Fig. 16. 4th Harmonic of waveform of Fig. 7.

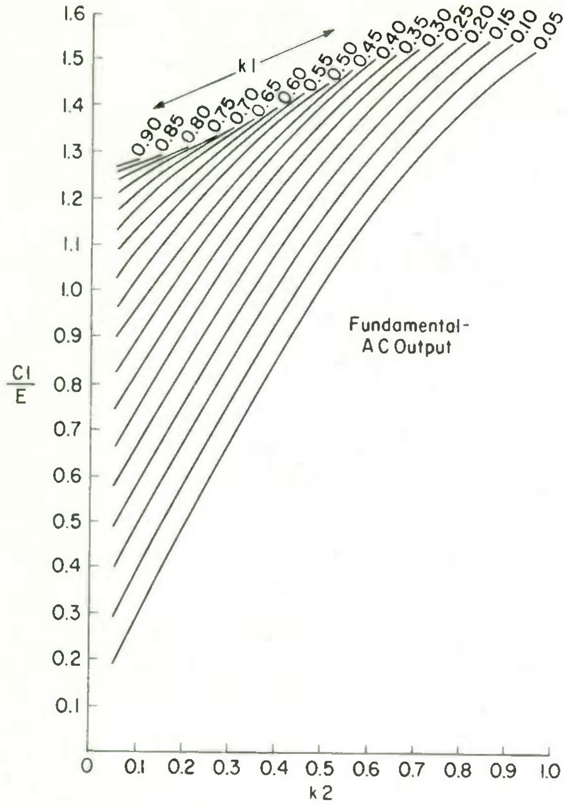


Fig. 17. Fundamental of harmonic content of waveform of Fig. 5 for ac output.

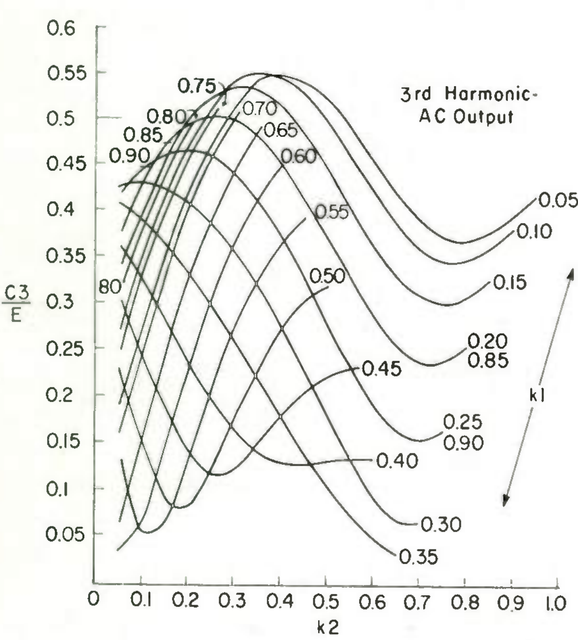


Fig. 18. 3rd Harmonic, ac output.

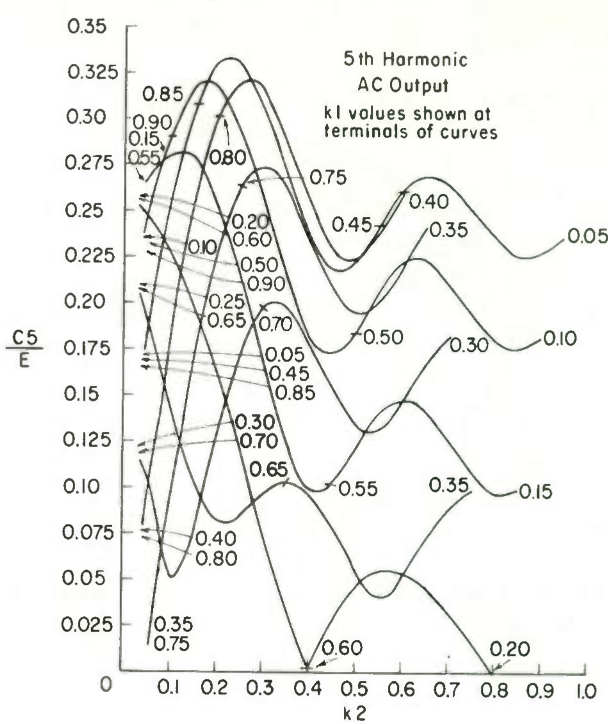


Fig. 19. 5th Harmonic, ac output.

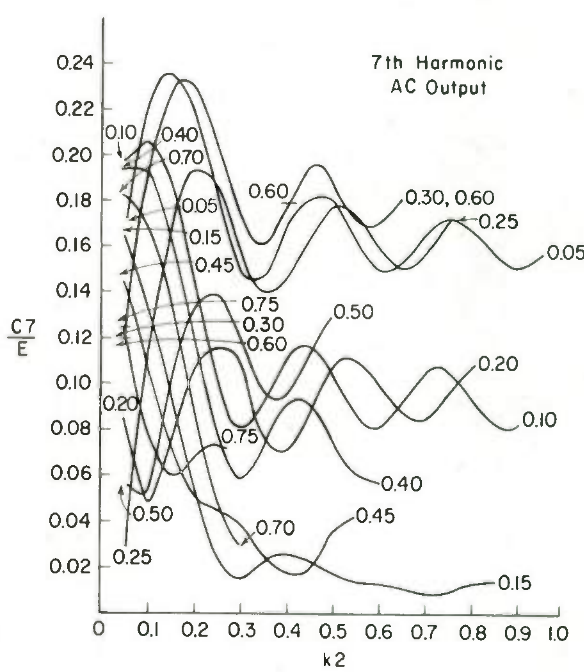


Fig. 20. 7th Harmonic, ac output.

Fig. 21. Photograph of 2 kw, dc output, regulated VWP power supply.

A LOW COST INDUSTRIAL TELEVISION SYSTEM

Harold R. Walker
Consulting Engineer
Metuchen, New Jersey

Summary

A new type of television camera utilizing an electrostatic camera tube of the vidicon type is described. The camera also contains some new and novel circuitry in its dual tetrode video amplifiers and its built-in intercarrier 4.5 mc. sound system.

* * *

Since the earliest days of commercial television broadcasting, engineers have dreamed of a low cost television camera for industrial and consumer use. Many attempts have been made over the past ten years to produce such an item. Generally, these cameras have centered around the 1" electromagnetic vidicon as exemplified by the 6198 and its successors. The camera to be described is the first to depart from this tube and there is every reason to suspect that in the future there will be further departures in attempts to lower cost and complexity.

The "Pixicam" represents the first attempt by any manufacturer to package a complete audio-visual closed circuit transmitter in a single low-cost unit. Several design firsts will become apparent as the design features are explained. The camera utilizes a 2" electrostatic vidicon of a unique design and a 2" f 1.9 lens interchangeable with any Leica lens or accessory. These are obvious departures from the 1" tube and lens of the past. Cameras of this type have tended to copy the RCA T.V. eye video amplifier circuits. The departure here is again complete in that this camera utilizes the Sylvania 10 pin dual tetrode 6C9 in a series peaked circuit. Because of the different circuit requirements for electrostatic scanning, the scanning circuits are also totally different from other cameras of this class. The modulator circuit is unique in that only two tubes perform the functions of audio preamplifier, reactance tube, 4.5 mc. FM sound oscillator and R.F. oscillator. The complete unit has 7 tubes plus the 2" "Pixicon"; weighs 11 pounds and consumes 40 watts.

In designing and producing the camera system, first consideration was given to price and second consideration to performance. The end result was a camera that actually matched the performance of its competitors, so that little or nothing was sacrificed. Unfortunately, the hoped for production economies have not as yet been realized and the camera has little price advantage.

Figure #1 shows an early developmental 2" electrostatic vidicon with a frit glass seal at the face. Production tubes utilize the gun shown and an indium pressure seal at the face. Note that a split deflection plate system is used to give a common deflection center. The electron optics are such that a .0015" defining aperture is magnified at the target surface 1/1. Resolution of the tube has been measured in a high definition system as 650 lines at the center and 350 lines at the corners. In this camera it is limited by video bandpass to 400 lines. The tube is unique in that it uses no mesh or wall screen and it is also unique among camera tubes in that the bulb is made of G 12 glass with a lime glass faceplate. The absence of a mesh almost completely eliminates beam flutter, a common problem with electromagnetic vidicons.

Being electrostatic, the tube requires no yoke or focus coil. It is wrapped in a high mu 80 shield and clamped into place using motor block clamps, a high production item in the fastener industry, which was used to reduce costs.

The lens was especially designed for the camera. It is a 7 element anastigmatic, coated, color corrected lens. To provide interchangeability, a compromise was made on Leica threads. However, the flange may be changed to accomodate the Pentacon type thread as well. This makes the camera useable with Leica accessories such as the microscope adapter.

The camera was originally designed and laid out for the more conventional video amplifier and with a more compact form factor. This design had to be abandoned because adequate video gain required that the video amplifier tubes work close to maximum allowable dissipation ratings and the magnetic coupling between the power transformer and camera tube was almost impossible to eliminate. The present video amplifier circuit is quite stable, operates the tubes well below their ratings and gives a 1.5 volt composite video output.

The long case was made necessary by the transformer which is located on the tube axis to prevent magnetic interaction. Since mass production was planned, printed circuits are used throughout. The complete camera is shown in Figure #2. The schematic diagram is given in Figures 3 and 4.

The video amplifiers are series compensated tetrodes. Contrary to expectations, there is no perceptible front end noise. The first stage has no plate circuit peaking since it was found the overall gain was quite high and peaking tended to cause a regenerative white noise. Because of the high output capacity of the "Pixicon" tube, the correction required is about 3 times that of a 6198 output. Under these conditions an LR peaker would have caused a loss of gain compared to the RC peaker used. Aperture correction is provided by the RC combination in the cathode of the fourth stage. Biasing is provided by various means with the first three stages operating with about 2.0 volts of bias while the fourth stage operates at about 4.0 volts. The average stage gain is about 12.

An anode follower output stage is used for coax line driving, D.C. restoration and sync mixing. The plate of the tube is at ground potential so that no video coupling condenser is required for the 75 ohm coax output. Sync pulses are mixed in the plate load resistor. The grid/cathode combination serves as a D.C. restorer, clamping on the blanking signal from the "Pixicon."

Various 60 cycle line drive schemes were tried and finally abandoned in favor of the more expensive blocking oscillator which gave better blanking and sync signals and was less responsive to power line fluctuations. The output of the vertical blocking oscillator is direct coupled to a phase inverter which drives the deflection plates.

Horizontal deflection is provided by a ringing coil stabilized multivibrator and a direct coupled phase inverter.

Various sync and blanking schemes were considered and tried, but the one finally used proved to be the most satisfactory even though it did require the use of an additional tube section which most low cost cameras avoid. Positive pulses are taken from the horizontal and vertical oscillators and applied through RC networks to the grid of the blanking pentode. This tube is biased so that it is cut off between pulses and the pulse peaks saturate the tube below the knee of the pentode curve. The blanking and sync pulses are therefore squared, of uniform height and of more than adequate amplitude to blank the "Pixicon."

It will be noted the "Pixicon" is blanked in a somewhat unusual way, and the beam control is also quite different from the usual practice. The "Pixicon" grid is always at 0 volts with respect to the cathode, except during blanking, so that the grid/cathode diode effect can be used to provide D.C. restoration to the blanking pulses. Beam current is varied by varying the voltage on G_2. This is analogous to varying the plate current in a triode by varying plate voltage. There are two reasons for this. The first is that it permits the grid to be used as a D.C. restorer. The second is that by lowering the voltage on G_2 to the lowest possible limit, the effects of secondary emission from the defining aperture are reduced. When G_2 is high, a white "ion" spot sometimes appears at the center of the picture. This could be eliminated by using a wall screen or mesh in the tube, but this would greatly increase tube cost.

This camera is the first known camera to have intercarrier sound built in. Because of this, the modulator is more complex than the simple R.F. circuit used in other cameras.

A microphone input for a microphone having approximately -55 db output is provided. There is no volume control on the camera but one could be incorporated in the microphone. A starved screen grid amplifier is used to raise the microphone input level to about 8 - 10 volts peak to peak. This signal is applied by means of a cathode follower to a triode reactance tube circuit. The reactance tube varies the 4.5 mc. oscillator frequency over a suitable range for a good audio signal at the receiver.

The video and audio signals are mixed at the R.F. output jack in such a way that there is little or no perceptible audio effect on the video signal. The R.F. oscillator tunes through channels 2 - 6 but it has strong harmonic outputs which can give acceptable signals on channels 7 - 13. A bandpass filter is needed if the R.F. is to be mixed into a master antenna system.

The camera has two controls not generally found on cameras. An astigmatism control is located internally to correct for gun astigmatism and a collimation control is provided on the back to correct for beam landing on the target surface. All other controls are normal.

The engineers who worked on this camera and its tube had high hopes that it would bring about a new era in closed circuit television because of its simplicity and expected low price. However, it is now apparent that this is not the ultimate camera, for even simpler lower cost cameras are now in the prototype stage. Although its time before the public may be brief, it does mark the start of a trend toward new ideas in cameras and camera tubes made necessary by the cost of cameras which is still too high for mass acceptance.

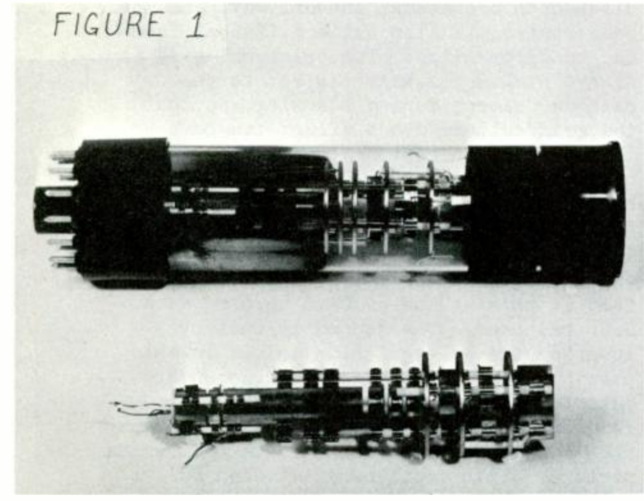


FIGURE 1

FIGURE 2

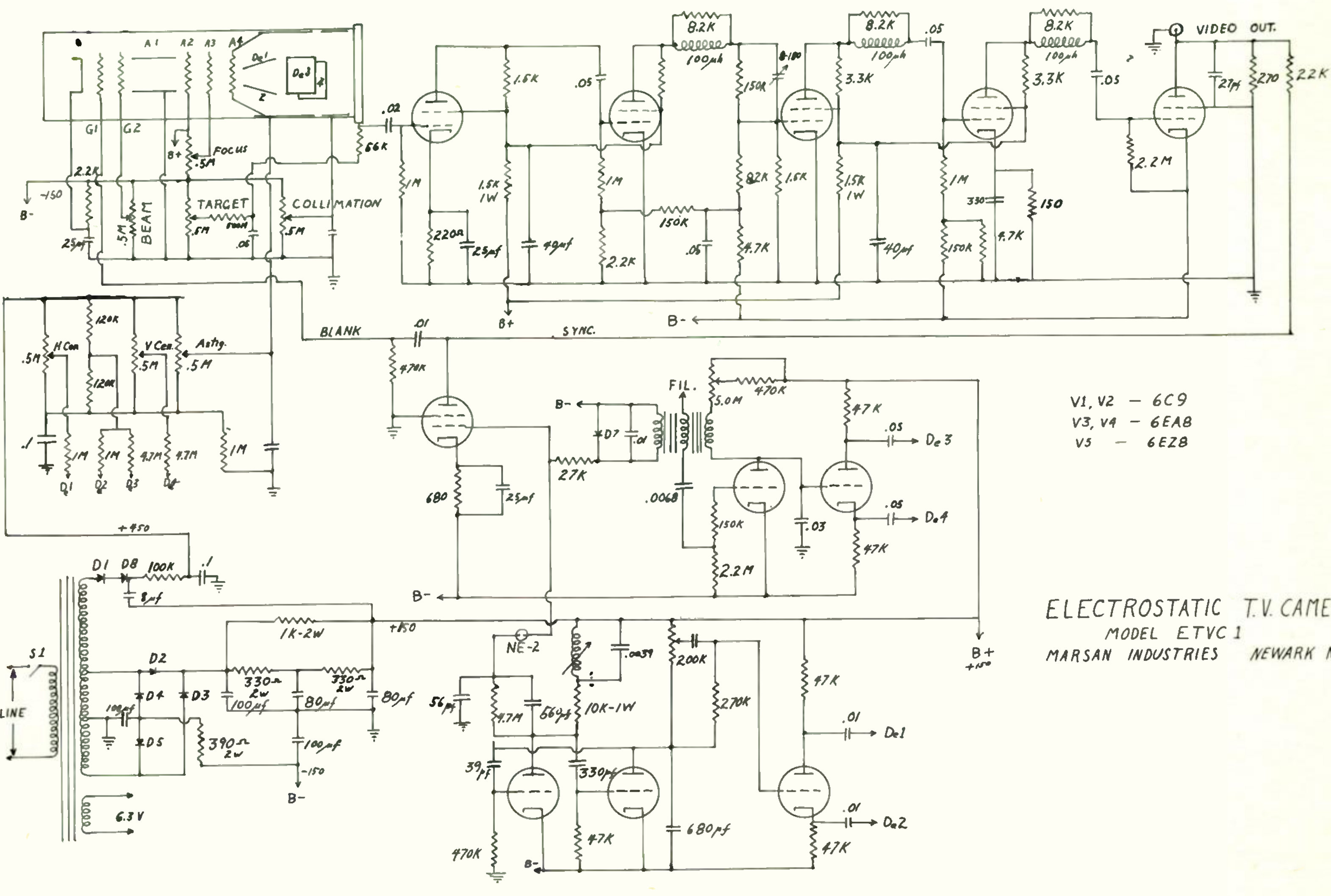


Fig. 3.

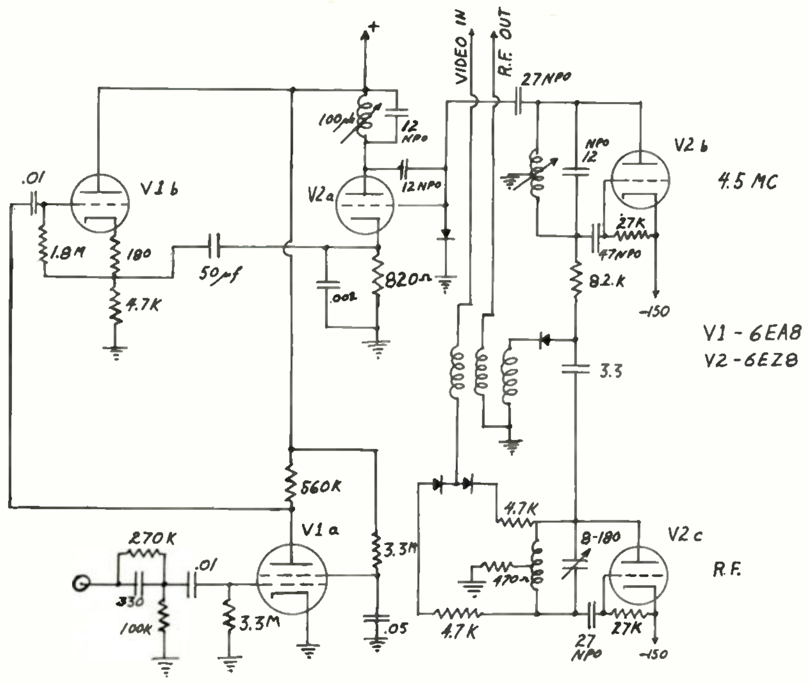


Fig. 4.

APPLICATION OF AUDIOFREQUENCY FOR ANESTHESIA OF ANIMALS

John E. Thompson
Reliable Packing Company
Chicago, Illinois

Summary

In seeking an improvement in the existing methods of preslaughter anesthesia of meat animals, a program of research was undertaken in which sine wave audiofrequency currents were applied to the animal's heads in the region of the skull overlying the frontal portion of the brain. The frequency, amplitude and duration of application were varied. The ability of this type of current to render an animal unconscious was found to be frequency dependent, as well as requiring a minimum value of amplitude and duration. The most suitable frequencies appear to lie in a band bounded by 200 cps and 2000 cps. Amplitude varies with the size and species of animal. Duration should be greater than 0.25 seconds for consistent results. Rats, rabbits, dogs, hogs, sheep and cattle were tested.

Introduction

It is entirely possible that Adam and Eve saw an animal killed by lightening in the midst of a storm. The observation that electricity could kill must have come almost this early. Not many generations could have passed before someone came into accidental contact with an electric eel and, thus, became aware of the shocking power of electrical currents. Our ancient forebearer might have had the good fortune to avoid the traumatic experience personally. He might have noticed the action of the electric eel on an animal. He would have noticed that the electrical discharge did not always kill, but rendered the animal unconscious and paralyzed. Thus nature has provided electrical anesthesia long before the advent of the electronic sciences.

The experimental study of electrical currents to render dogs unconscious is recorded by Leduc[1] in 1902. Many other experimenters in the classic period of physiology also experimented with electrical currents.

Following the astute observation that epilepsy victims were seldom bothered by schizophrenic symptoms following a seizure, various schemes for giving mental patients convulsions were tried. At a Swiss meeting in 1937, Cerletti's assistant, Bini, first mentioned the possibility of using electricity. It was first demonstrated by Cerletti and Bini in 1938.[2] Bini's method used the 50-60 cycle electric light circuits. This is still widely used. Kalinowsky and Hoch[3] describe a modern clinical procedure as comprising 70 volts for 0.2 second which results in a current flow of about 200 ma. This is reported to produce immediate unconsciousness and the patient has no recollection of the application.

The commercial use of both direct and alternating currents to stun meat animals prior to slaughter has been in use in Europe for a considerable period of time. The subject was investigated quite thoroughly in this country under the auspices of the Institute of American Meat Packers in 1932 by Ivy.[4]

A number of European countries have laws requiring the stunning of meat animals prior to handling for killing. In 1958, a similar law was enacted in the United States, Public Law 85-765, requiring most meat packers to render the animals unconscious and insensitive to pain before any hoisting operation. Chemical, mechanical percussion and electric means have come into use.

Most of the electric preslaughter stunning systems have a serious disadvantage, severe convulsions, as previously reported in connection with human electroconvulsive therapy. The sudden tightening of muscles when the electricity is applied is so strong as to rupture blood vessels and even to break bones. The resulting injury leads to economic loss by condemnation of the damaged parts. A frequent observation in the course of our early experimental work was hemorrhage from the lungs. None of these factors is in any way painful to the animal because it is killed while still unconscious.

The objective of any preslaughter anesthesia system is primarily to render the animal unconscious, immobile and insensitive to pain. It differs from human or veterinary surgical anesthesia in that the animal must remain quiet and anesthesized for a period after the application is removed. It is in this period that the butcher must handle and dispatch the animal. A satisfactory system must also leave no after-effect that will depreciate the value of the meat carcass. In order not to interfere with the smooth flow of production, the time and labor to give the treatment must be minimized. The net profit of the meat packing industry last year was such that the return on the average hog carcass was in the order of 20¢, so it is obvious that the cost of the treatment should be small.

The Present Studies

Work on electric anesthesia in our laboratory began before 1947 when I made a personal survey of the European meat packing industry and brought back a crude piece of Danish apparatus. From this we began our own development program. At first our main aim was to provide a saving in meat yield by avoidance of the bruises and other injuries to

WRH

the animals incident to struggling during handling, and to provide safer working conditions for the handlers who were occasional victims of attack by frightened, infuriated animals. Later, it became evident that some sort of "humane" pre-slaughter treatment was going to become mandatory.

The program was divided between work with direct current and work with alternating current. The d.c. work was the first to meet with success and a patent was recently issued, Morse #3,012,271. Our a.c. program was continued because the interrupted d.c. of the proper characteristics requires costly and cumbersome apparatus that would not be within the means of the smaller packers to acquire and operate. We hoped that an a.c. apparatus could be developed that would not cost appreciably more to build or operate than a high power public address amplifier.

The Experimental Apparatus

The laboratory apparatus set-up is diagramed in Figure 1. This apparatus can be conveniently described as consisting of a variable frequency audio oscillator, an audio frequency power amplifier, metering and monitoring apparatus, an electronic timer and the applicator. All but the applicator are mounted, for convenience, in a standard relay rack cabinet.

The oscillator is a Hewlett-Packard model 200-B which covers the range of 20 to 20,000 cps. The output of the oscillator is fed to a McIntosh model MI-200B 200 watt power amplifier which is provided with a multi-tap output transformer.

The monitoring panel is an assembled unit incorporating an Atcotronic model 5231A26A03XX electronic timer, a 40 watt incadescent lamp, a variable power resistor to provide a dummy load, and standard voltage and current meters. An oscilloscope was occasionally used to check the applied treatment waveform, frequency and amplitude.

The electronic timer contains a double throw switch controlling the output of the amplifier. During the treatment cycle, current flows via the applicator through the animal's head; at other times, the current is diverted to a dummy load consisting of a bank of three Ohmite model L variable resistors rated at 150 watts each. These were set to the approximate resistance of the animal's head. Thus, the monitoring meters were not required to move substantially between stand-by and treatment indications, so errors in readings due to wide meter swings were minimized.

The applicator for small laboratory animals is simply a pair of test prods separated by a piece of lucite and fastened securely to it. This provides two contact points spaced to apply to the animal's head just above and slightly lateral to the eyes. A SPST switch at the applicator handle is provided to actuate the timer. This type of applicator was used in the treatment of rats, rabbits, cats and dogs.

The hog applicator was developed in our plant by the operating people who quickly became disgusted with the awkward handling of the lab-rigged applicator when they were asked to conduct in-plant trials of our electrical anesthesia apparatus at production rates exceeding 360 animals per hour. Patent 2,977,627 has been issued to Morse and DiPasquale of our company for this appliance. With either type of applicator, it was unnecessary to use conductive pastes, brine or any other preparation to facilitate good transfer of electrical current.

The Experimental Procedure

Preliminary work was done with adult, white albino rats of 370 grams average body weight. The animals were taken at random from a colony cage and the treatment applied to a given animal was also randomized by shuffling cards on which the treatments were marked. The same group of animals was used repeatedly, but not on the same day.

Treatments were made at 75, 100, 150 and 200 volts RMS. Various frequencies of sine waves were applied at each of these voltages in the range 20 to 20,000 cps. The time of treatment was varied from 1/32 to 4 seconds in eight steps: 1/32, 1/16, 1/8, 1/4, 1/2, 1, 2 and 4.

The effectiveness of a given treatment was judged by the production of unconsciousness from which the animal did not recover for one minute or longer. Simple tests for insensitivity were made by lightly touching the cornea of the eye and by pinching the tender skin between the toes with a small hemostat.

Similar treatments were applied to a group of eight young adult albino rabbits ranging in body weight from 2900 to 3700 grams. These animals were treated daily each working day for three weeks. The voltages employed were 75, 100 and 150. The frequency range was the same. The time of treatment was limited to three steps: 0.1, 0.2 and 0.4 seconds.

One cat, an adult male Siamese weighing about 3 kilos, and one dog, an adult male mongrel weighing about 20 kilos, were treated with 100 volts, 600 cps for 0.20 second.

The apparatus was transported to the cattle dressing floor of a small Chicago packer where it was used successfully on ten heavy bullocks weighing approximately 1300 lbs. each. Only the highest available voltage (200) and a one second long treatment was tested at the insistance of the plant superintendent. The frequency employed was 600 cps. Further, the unconscious cattle were immediately shot with a captive bolt pistol, so there was no opportunity to evaluate the length of the period of unconsciousness.

Hog experiments were conducted on a group of animals ranging in body weight from 50 to 165 kilos. Each animal was used repeatedly, one treatment following another within minutes, allow-

ing only sufficient time between treatments to permit the animal to recover from the obvious effects.

The voltages employed were 75, 150 or 200 volts RMS. The frequencies were in the 20 to 20,000 cps range. The time ranged from 1/8 to 2 seconds, with a few treatments for longer times up to 10 seconds. The particular treatment conditions used in a series were selected at random, using a table of random numbers or by drawing billets from a container. Some specific treatments were omitted, however, when they were judged to be redundant.

Experimental Results

The results of the rat experiments are summarized in Table 1. The higher voltage treatments are more effective in producing unconsciousness in a given period of time than is the lowest voltage. The highest and the lowest frequencies are somewhat less effective in producing unconsciousness, particularly when the time if short and the voltage low. A band of frequencies in the range 200 to 2,000 cps is able to produce unconsciousness in less time and at lower voltage than either higher or lower frequencies.

The rabbit experiment results are summarized in Table 2. The pattern of results is the same as for rats, and the most effective band of frequencies is the middle range.

The one cat and the one dog treated were effectively anesthesized by the only treatment applied to them. This treatment was selected from the treatment region that was expected to be effective.

The cattle were rendered unconscious by the treatment given. The experimental conditions were the strongest available from this apparatus and supplied only qualitative information to the effect that such treatments are capable of producing anesthesia in large meat animals, as well as small stock and laboratory animals

The hog experiment results are summarized in Table 3. The pattern here is also similar to that observed in the rat and rabbit experiments.

Discussion

At least four types of responses to electric shock can be clearly identified:

1. Lethal shocks. These may kill immediately or indirectly by stunning the rupture centers of the brain, resulting in death by asphyxia. Fortunately, the electric characteristics of such a shock, even when administered directly across the brain, are well beyond the strength required to give effective anesthesia. Accidental, premature deaths are virtually non-existant. In all our work, one rat and one dog died as a result of the experimental treatment. These might have been prevented by artificial respiration.

2. Effective anesthetic shock. These make the animals unconscious immediately and apparently insensitive to pain. Such shocks are most often accompanied by a strong tonic seizure, followed by a period of quiet relaxation.

3. Sub-effective shocks. These produce immediate unconsciousness from which the animals recover almost at once after the electric current is interrupted. This sort of shock is most often accompanied by violent, clonic thrashing and frequently with snorting and loud squealing, as if the animal was in severe pain. There is very little opportunity to test the animal's sensitivity to painful stimuli following sub-effective shocks.

4. Ineffective shocks. These produce pain rather than protect against it. The typical response is struggle and fright.

Both treatments 3 and 4 above are ineffective in producing anesthesia, and appear to be painful and frightening to the animal. Animals that have been so treated appear to remember the painful experience and resist handling for further treatments. By contrast, animals receiving a type 2 treatment do not appear to have any recollection of the treatment at all. They exhibit no more than normal anxiety when handled for a later treatment.

I have personally conducted experiments on myself, using gradually increasing voltage applied to electrodes held against my temples. I did not experiment with sufficiently strong treatments to produce any level of anesthesia or unconsciousness. My only subjective impression was of pain--a sharp, burning pain at the skin surface under the electrodes. Treatments intense enough to almost, but not quite, produce unconsciousness must be extremely painful, indeed. This is an experience I would not care to impose on either man or animal.

It was especially interesting to discover that 50 cycle and 60 cycle currents were not particularly effective in producing anesthesia. This is worthy of note because it eliminates a simple transformer configuration from consideration in commercial applications where compliance with Public Law 85-765 is important. This law clearly requires that a minimum effective treatment be used. Fifty cycle or 60 cycle currents require substantially stronger treatments than frequencies in the more effective 200 cycle to 2,000 cycle band.

According to the computations of Offner,[5] modified square waves or interrupted d.c. pulses should provide the basis for effective treatments at lower average current flow than with sine waves. We have undertaken to make a comparative study of the effectiveness, for preslaughter anesthesia of meat animals, of sine waves versus shaped wave and pulses. In a preliminary way, it appears true that smaller average currents can produce unconsciousness when the waves are

interrupted d.c. pulses or are shaped to give high peak amplitudes of short duration. It may develop, however, that such treatments must be more prolonged in order to produce the desired post-treatment period of anesthesia. Recall that this post-treatment period is the time in which the butcher must handle and dispatch the animal.

Conclusions

Comparatively simple and inexpensive electronic apparatus can be used to effectively anesthesize meat animals prior to handling for abbatoir operations. The middle range of audio frequencies were found to be more effective than either lower or higher frequencies. By use of audio frequency preslaughter anesthesia, the handling of meat animals in the abbatoir can be made more humane and safer for the men who must work with the animals. Valuable meat can be saved by the avoidance of ante-mortem bruises to the animals.

References

1. Leduc, S. Production du sommeil et de l'anesthesie generale et locale par le courant electrique. Compt. rend. Acad. de sc. 135:199, (1902); 135:878, (1902).

2. Bini, L. Experimental researches on epileptic attacks induced by electric current. American Journal of Psychiatry. 94(Supp.):172, (1938).

3. Kalinowsky, L. B. and Hoch, P. H. Shock Treatments, Grune and Stratton, New York, (1952) p.97.

4. Ivy, A. C. and Barry, F. S. Studies on the electric stunning of dogs. American Journal of Physiology. 99:298, (1932).

5. Offner, Franklin. Stimulation with minimum power. Journal of Neurophysiology. 9:387, (1946).

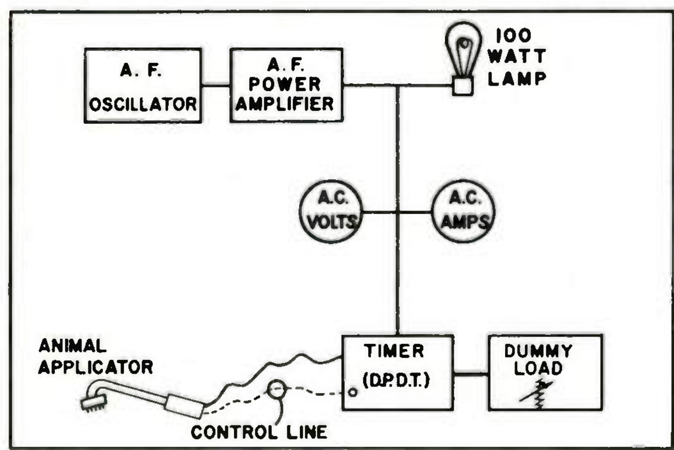


Fig. 1.

Table 1. The Response of Adult Albino Rats to Sine Wave Electrical Treatment

Volts RMS	Time Seconds	20	25	30	40	50	60	80	100	150	200	250	300	400	500	600	800	1K	1.5K	2K	2.5K	3K	4K	5K	6K	8K	10K	15K
	1/32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0
	1/4	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	0	0	+	+	0	0	0	0	0	0	0	0
75	1/2	0	0	0	0	+	0	0	0	+	+	0	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0
	1	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	+	0	0
	2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	+	0	0
	4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	0	0
	1/32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/16	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/8	0	0	0	0	0	0	0	0	0	0	+	+	+	+	0	+	+	0	+	0	0	0	0	0	0	0	0
	1/4	0	0	0	+	0	+	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0
100	1/2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
	1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
	2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	+
	4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+
	1/32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/16	0	0	0	0	0	0	0	0	0	+	+	+	0	+	+	+	0	0	+	0	0	0	0	0	0	0	0
	1/8	0	0	+	+	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0
	1/4	+	+	+	+	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
150	1/2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0
	1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	1/32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1/16	0	0	+	0	+	0	0	+	0	0	+	+	+	0	+	+	+	+	0	0	+	0	+	0	0	0	0
	1/8	0	0	+	0	+	0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0
	1/4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
200	1/2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	0
	1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0
	2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0
	4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 2. Response of Young Adult Rabbits to Sine Wave Electrical Treatment.

Volts RMS	Time Seconds	30	60	120	300	600	1.2K	3K	6K	12K
		Sine Wave Cycles Per Second								
75	0.1	0	0	0	0	+	0	0	0	0
	0.2	0	0	0	0	+	0	0	0	0
	0.4	0	+	+	+	+	+	0	0	0
100	0.1	0	0	0	+	+	+	0	0	0
	0.2	0	+	+	+	+	+	+	0	0
	0.4	+	+	+	+	+	+	+	+	0
150	0.1	+	0	+	+	+	+	+	0	0
	0.2	+	+	+	+	+	+	+	+	0
	0.4	+	+	+	+	+	+	+	+	+

+ Favorable response (anesthesized) on two or more test treatments out of three, on separate animals.

0 Not unconscious in two or more test treatments on separate animals.

Table 3. Response of Hogs to Sine Wave Electrical Treatment.

Volts RMS	Time Seconds	20	40	80	200	400	800	2K	4K	8K
		Sine Wave Cycles Per Second								
75	1/8	-	-	-	-	0	-	-	-	-
	1/4	-	-	-	0	+	0	0	0	0
	1/2	0	0	0	+	+	+	0	0	0
	1	0	0	0	++	++	++	+	0	0
	2	0	0	++	++	++	++	+	0	0
150	1/8	-	-	0	0	+	+	0	-	-
	1/4	-	-	0	0	++	++	0	-	0
	1/2	0	0	0	++	++	++	0	0	0
	1	0	0	++	++	-	-	+	0	0
	2	+	+	++	-	-	-	++	+	0
200	1/8	0	0	+	++	++	+	0	0	-
	1/4	+	+	++	++	++	++	+	0	-
	1/2	+	+	++	++	++	++	+	0	0
	1	+	++	-	-	-	-	++	+	0
	2	++	++	-	-	-	-	-	++	0

0 No unconsciousness produced.
+ Short duration unconsciousness; petit mal seizure.
++ Fully unconscious.
- Combination not tried experimentally.

WRH

SOLID-STATE LINEAR AND SINUSOIDAL SYNCHROS

T. W. Parsons
Res. & Adv. Development Dept.
Precision Components Division
Kearfott Division, G.P.I.

D. R. Simon*
Physics & Space Science Dept.
Vitro Laboratories
West Orange, New Jersey

Summary

The recent development of new permanent magnet ferrites having considerable coercive force, coupled with the development of extremely linear Hall generators, permits new design concepts in electrical transmission systems. The overall characteristics of a solid-state synchro depends in part on parameters which have not hereto fore been considered in standard wire-wound units.

The major areas which had to undergo analysis are: "misalignment voltages" of the sensing crystals, position of the Hall generators, effects of gap size and placement of the flux distribution, magnetic properties of retentivity of the magnetic return paths. The "misalignment voltage" is the result of a resistive unbalance and can seriously affect the overall accuracy of the unit. This voltage can be compensated for by an external network; however, if the field magnitude is varied by changing the position of the shaft, the misalignment voltage changes. These changes can be minimized and the resistive unbalance can be tuned out.

The position of the major air gaps in the units are the controlling parameters in the error analysis. Equations are derived relating the flux in the air gap to the magnetic characteristics of the ferrite rotor. The equations are complicated by the fact that the air gap is not constant, and, therefore, the equations must be broken into several terms which are summed over $\Delta\theta$. The limits of the summations are constantly changing. The equations are not easily solved by analytical methods, but lend themselves to digital computer methods very nicely. Several different return paths are considered and the predicted results are compared favorably to the test results. The configuration showing the most promise is then analyzed for hysteresis or retentivity. An interesting outcome of the study indicated a reduction of the theoretical error with smaller size units. This was born out in comparative tests of a size 4 and size 11 frame.

The frequency response of the units are calculated using network theorems and compared to experimental results. The result here indicates again better characteristics with smaller size.

It is apparent from the analytical and experimental results that a whole family of resolvers can be built with the largest size being .4 inches in diameter and .6 inches in length.

Technical Analysis

The overall characteristics of solid-state linear and sinusoidal signal generators depends on the interaction of several phenomena which have not manifested themselves in wirewound units. This is in part due to the time independent field contributed by the rotor and in part due to the different characteristic inherent in "Hall effect" devices.

To achieve a sinusoidal output from a linear field sensor, the original field itself must be sinusoidal. Since one of the objects of the program was to eliminate windings and brushes and another to achieve an extremely small unit, the only possible solution was a permanent magnet rotor of some sort. The overall size of the synchro, dictating the small volume of the rotor, eliminated all but the high coercive force magnets such as ferrites. If we begin with a cylincrical rotor we can intuitively expect something similar to a sinusoidal distribution of flux, if the direction of magnetization is perpendicular to the axis of rotation, since the coercive force of a magnet is proportional to the length of it, and this length varies sinusoidally for a cylinder. The flux density emanating from the surface of the rotor can be approximated by the sum of fluxes emanating from "n" parallel segments comprising the cylinder. This is illustrated in Figure 1.

To compute the flux density generated by a segment, we must make three assumptions:

a. The fringing effects are negligible.

b. The direction of magnetization is not changed in the rotor after removal from the magnetizing fixture.

c. The rotor ferrite is uniform, homogeneous, and isotropic.

We can now resort to the B-H curve of the material and determine the operating characteristic curve for PM-4 (lead ferrite). The load line from standard permanent magnet analysis has a slope:

$$L_m A_g / L_g A_m$$

where Lm is the length of the magnet, Ag is the area of the gap, and Am is the area of the pole face of the magnet. We then have the following two equations:

$$B_m = 1700 - H_m \quad (1)$$

$$B_m = L_m A_g (L_g A_m)^{-1} H_m \quad (2)$$

which can be combined and rewritten as follows:

*Formerly of Kearfott Division, G.P.I.

**Solid State Sinusoidal Synchro work performed under Buweps Contract number NOW-60-0013

$$Hm = 1700 \left(1 + \frac{Lm\ Ag}{Lg\ Am}\right)^{-1} \qquad (3)$$

$$Bm = 1700 \left[1-\left(1 + \frac{Lm\ Ag}{Lg\ Am}\right)^{-1} \right] \qquad (4)$$

since $Bm = Bg\ Ag\ Am^{-1}$ where Bg is the flux in the air gap, we can write the equation (4) as

$$Bg = 1700 \left[1 - \left(1 + \frac{Lm\ Ag}{Lg\ Am}\right)^{-1}\right] \frac{Am}{Ag} \qquad (5)$$

Using the relationship (5) we can determine the flux contribution of each of the "n" segments as a function of the rotor angle related to the sensing element.

To consider the specific case of the size 11 unit, we can let the diameter of the rotor equal 1.265 cm, the length equal 1.27 cm, and $\Delta\theta = 10°$, the following values are then obtained:

Table I

Segment Number	Average Length (cm)	Cross Sectional Area (cm 2)
1	1.260	.2790
2	1.222	.2708
3	1.156	.2737
4	1.036	.2294
5	.8945	.1980
6	.7256	.1606
7	.5346	.1183
8	.3274	.0725
9	.1103	.0244

To compute Bg for each segment we can evaluate (5) for the following parameters:

Ag = .4283 sq. cm
Lg = .127 cm

The following values are obtained from equation (5)

Table II

Segment #	Calculated Bg	Reference Values where B max = 544.5
85° (1)	541.6	541.6
75° (2)	520.1	525.9
65° (3)	487.3	494.5
55° (4)	441.3	446.0
45° (5)	380.4	385.0
35° (6)	311.8	312.2
25° (7)	229.6	230.1
15° (8)	140.7	140.9
5° (9)	47.4	47.5

(Which approximates a sine function within 2%)

Return Path

Since a preliminary analysis indicates the presence of a sinusoidal flux distribution within tolerable limits, our task is now to examine the return path in detail with respect to the method of crystal placement.

Our knowledge of Fourier series allows us to represent our output function as a fundamental sine function with higher order sinusoids superimposed as the error signal. It appears obvious then that we can reduce the magnitude of the error signal by integration of the flux distribution, since the nth frequency will be reduced by a factor of 1/n.

A return path which will perform the integration process is illustrated in Figure 3. The one illustrated is built in four quadrants A_1, A_2, A_3, A_4; however, it can have A_n quadrants if necessary, as long as the unit has symmetry about any diagonal plane bisecting it and parallel to the rotor axis. The pole faces P_1, P_2 ... P_n (P_1 and P_3 shown) cover the active area of the crystals which are inserted into slots S_1, S_2 ... S_n.

The operation of the integrating unit can be shown in the following manner. One can see that if the rotor is turned through an angle θ_s, the total flux emanating from the reference point to the position directly under the crystal must pass in a perpendicular manner through the crystal. The crystal then sees the integral over the surface of the rotor from the zero position to θ_s. Since there are four gaps in order to obtain both the (sinθ) and (cosθ) functions careful observation indicates that the reluctance of the return path will not be constant but a function of the angular position and the particular segment under observation. For angles less than 45° the segments being integrated see a two slot reluctance path; however at angles greater than 45° some segments will see only one gap. The effect of this characteristic is not immediately clear to the casual observer and if a true understanding of the unit is to be had it should be analyzed in detail by the reader. (Figure 3A)

To obtain the new Bg as f(θ) the condition of one and two gaps must be computed separately for each segment and then a numerical integration must be performed choosing from each case as the situation dictates. Equation (5) must be written as

$$Bg=1700\left[1-\left(1+\frac{Lm\ Ag}{(Lg_1+Lg_2+2L_{ag})Am}\right)^{-1}\right]\frac{Am}{Ag} \qquad (6)$$

where Lg_1, is the length of slot S_1, and Lg_2 is the length of slot S_2. The term $2L_{ag}$ arises because of the small air gap between the rotor and the stator proper. For the values $Lg_1 = Lg_2 = .061$ cm, $L_{ag} = .00254$ cm, and Ag = .4283 sq. cm, we obtain the results tabulated in Table III. The numerical summation resulting from these figures appears in table IV. A plot of the error vs shaft angle appears in Figure 4. The actual data taken on a size four function error experiment is also plotted on the same axis. The agreement of the two curves indicates the analysis made above is at least on a solid foundation.

Table III

Segment #	Bg for one slot	Bg for two slots
1	544.4	541.6
2	528.3	520.1
3	495.0	487.3
4	447.0	441.3
5	386.3	380.4
6	313.4	311.8
7	230.8	229.6
8	141.4	140.7
9	46.7	47.4

Table IV

Computer	Actual value on the basis of K sin 90° = 3132.7	Error%
0° = 0	0° = 0	0
10° = 541.6	10° = 543.9	.1
20° = 1061.7	20° = 1071.4	.32
30° = 1549.0	30° = 1566.0	.53
40° = 1990.3	40° = 2013.6	.75
50° = 2370.8	50° = 2399.8	.95
60° = 2695.7	60° = 2712.9	.55
70° = 2933.0	70° = 2943.7	.31
80° = 3083.2	80° = 3085.1	.07
90° = 3132.7	90° = 3132.7	0

Misalignment Voltage of Hall Generator

In the construction of solid state function generators, we would like to reduce the misalignment of the Hall leads (Figure 5) to a minimal value. When the crystal is excited, the misalignment of the Hall leads gives rise to a voltage which is independent of the magnetic flux density and therefore a direct source of "functional error" and "Proportional Voltage Bridge error."

The theoretical effects of the probe misalignment can be derived from the following equation which expresses the Hall output as a function of the crystal parameters.

$$V_H = R_H\ I\ B\ d^{-1} + I\ r_m \qquad (7)$$

The first term on the right-hand side represents the normal Hall effect voltage, and the second term represents the misalignment voltage drop. Substituting the following definitions into equation (7) $R_H = \rho\mu$, $I = \int JdA$, $r_m = \rho\ l_m A^{-1}$ we get:

$$V_H = \rho\mu B d^{-1} \int JdA + \rho\ l_m A^{-1} \int JdA \qquad (8)$$

since

$$\rho = (n\ q\mu)^{-1}\ , \text{ and } \mu = \frac{V}{E_e} \text{ we can say}$$

$$\rho = E_c\ (n\ q\ v)^{-1}$$

and equation (8) becomes

$$V_H = \left[B(n\ q\quad d)^{-1} + E_c l_m\ (n\ q\ v\ A)^{-1} \right] \int JdA \quad (9)$$

since $J = n\ q\ v$ we can rewrite (9) and

$$V_H = \left[B(n\ q\ d)^{-1} + E_c l_m\ (n\ q\ v\ A)^{-1} \right] n\ q\ v\ A \quad (10)$$

which reduces to

$$V_H = E_c \left[\mu\ Bw + l_m \right] \qquad (11)$$

It can be seen from this equation that the errors introduced by the misalignment voltage will be directly proportional to the physical misalignment of the probes.

Figure 7 is a summary of the functional errors we might expect to be introduced for given values of misalignment voltage and maximum coupling voltages. One can see from Figure 8 that the error introduced in the P.V. Bridge test is much greater for a given misalignment voltage and can seriously impair the operational value of the unit. It is interesting to look at some numbers to get an idea of the quality control problems that might be encountered in production. For example, if a potential difference of 3v is applied across a uniform crystal of L = 1 cm, a displacement δ L of .1 mm gives rise to a misalignment voltage of 30 m v or 10^4 times as large as we desire for an overall angular accuracy of ± 5 minutes.

Some of the methods of reducing misalignments at the present are:

a. Purchase the crystal in cruciform shape (Figure 6) within .0005 in dimensional tolerances.

b. Abrasive etching at A'A or B'B while monitoring the voltage on the Hall terminals.

c. Passivation of the surfaces.

These techniques have resulted in misalignment voltages on the order of 1 volt at an excitation voltage of 7 volts.

From the engineering standpoint it may be more practical to eliminate misalignment voltages within the buffer amplifiers of an integrated system as will be discussed in a following section.

Null Voltage Due To Inductive Coupling

A Hall generator can be simulated by a two loop inductively coupled network in order to determine the magnitude of the error signal introduced. If we simulate the input and output loops with the network in Figure 9 we can solve the loop equations as follows:

$$V_C = R_C\ I_1 + jwL_1 - jwMI_2 \qquad (12)$$

$$0 = -jwMI_1 + jwL_2 I_2 + R_h I_2 + R_3\ (1 + jwR_3 C_L)^{-1} I_2 \qquad (13)$$

$$V\text{null} = I_2\ \cdot R_L\ (1 + jwR_L C)^{-1} \qquad (14)$$

Rearranging equation (12)

$$I_1 = (V_C + sMI_2)(R_C + sL_1)^{-1} \text{ where } s = jw \qquad (15)$$

Substituting for I_1 in equation (13)

$$0 = -sM\,(V_C + sMI_2)(R_C + sL_1)^{-1} + sL_2 I_2 + R_H I_2 + R_L\,(1 + sR_L C_L)^{-1}\,I_2 \quad (16)$$

$$I_2 = sMV_C \left[(R_C + sL_1) \left(\frac{(sM)^2}{R_1 + L_1} + sL_2 + R_H + \frac{R_L}{1 + sR_L C_L} \right) \right]^{-1} \quad (16a)$$

To determine the circuit parameter L, we can set

$$L = 4\pi r \left(\ln \frac{8}{a} - \frac{7}{4}\right) \times 10^{-7} \text{ henrys} \quad (17)$$

where a = wire radius in meters, r = radius of current loop in meters, setting R = .01 meters and a = .0001 meters.

$$L = (12.5)\ (.01)\ (\ln 800 - 1.75) \times 10^{-7} \text{ henrys}$$

$$L = .616 \times 10^{-7} \text{ henrys} \quad (17a)$$

The circuit parameter C_h can be approximated by

$$C = \frac{27.7\ Ke}{\ln(2rla)} \times 10^{-12} \text{ farads} \quad (18)$$

where Ke = dielectric constant 1, a = wire radius in meters, 2 = distance between centers of two parallel wires in meters, setting R = .0020 meters and a = .0001 meters.

$$C = \frac{27.7}{\ln \left(\frac{.004}{.001}\right)} \times 10^{-12} \text{ farads} \quad (18a)$$

$$C = 2 \times 10^{-11} \text{ farads}$$

The proportionality constant M called the mutual inductance can be determined by

$$M = K\sqrt{L_1 L_2} \quad (19)$$

where K = the coupling factor

$$L_1 = L_2 = L = .616 \times 10^{-7} \text{ henrys}$$

For a circuit such as we have, K is between the limits of zero and 1. A reasonable value seems to be K = .1

$$M \cong .6 \times 10^{-8} \quad (19a)$$

The parameter R_H, R_C, R_L, and V_C can be chosen to fit the actual experimental values of a germanium crystal and a resolver bridge test circuit.

$$R_C = 250\Omega,\ R_H = 60\Omega,\ R_L = 10^4\Omega,\ V_C = 10 \text{ volts} \quad (20)$$

Combining equations (16a, 17a, 18a, 19a, 20) we get

$$I_2 = s.6 \times 10^{-7} \left[(250 + s.6 \times 10^{-7}) \left(\frac{.36 \times 10^{-16}}{250 + 6 \times 10^{-7}} + .6s10^{-7} + 60 + \frac{10^4}{1 + (10^4)\,2 \times 10^{-11}} \right) \right]^{-1} \quad (21)$$

Eliminating all terms of extremely small magnitude, we get

$$I_2 = s.6 \times 10^{-7} \left[(60 + 10^4)(250 \times s.6 \times 10^{-7}) \right]^{-1} \quad (22)$$

$$I_2 \cong s.6 \times 10^{-7}\ (s.6 \times 10^{-3} + 2.5 \times 10^6)^{-1} \quad (22a)$$

To find the zero field voltage, we can substitute equation (11a) in equation (3)

$$V_{null} = \frac{s.6 \times 10^{-7}}{s.6 \times 10^{-3} + 2.5 \times 10^6} \left[\frac{10^4}{1 + s10^4 \times 2 \times 10^{-11}} \right]$$

Table V

Tabulated Results

Frequency (cps)	Magnitude (volts)	Phase
400	$.6 \times 10^{-6}$	89° 58'
4,000	$.6 \times 10^{-5}$	89° 4'
40,000	$.6 \times 10^{-4}$	87° 9'
400,000	$.6 \times 10^{-3}$	63° 26'
4,000,000	1.5×10^{-3}	11° 19'

The effects of this coupling characteristic are indistinguishable from those of the misalignment voltage and therefore represent a frequency limitation of the input signal. The tabulated results use figures representing a size 11 unit. It should be obvious that the smaller crystal used in a size 4 unit will move the frequency limitation to approximately 6 megacycles.

Hysteresis

In considering errors related to time invarient field sources, such as permanent magnets, the retentivity of the return path plays an important role. Obviously, the narrower the hysteresis loop of the material making up the rotor, the smaller the "dead band" that will exist when the direction of rotation is reversed. The problem was entirely solved in the size 4 unit by using 4750 material which was machined .002 inches oversize, superannealed and then finish machines. The final assembly was then

finish machined. The final assembly was then thermally shocked at 100°C before testing the unit. This process reduces the error introduced by the retentive qualities of magnetic materials to a minimal value.

Solid-State Synchro-Resolver

We are now in a position to calculate the the transfer function of a resolver transmitter, and a synchro control transformer. One can begin by defining the electrical angle as the angle (a) displaced in a positive (counter-clockwise) direction from synchro zero which satisfies the relative magnitudes and polarities of the Hall generator output voltages of a synchro transmitter in accordance with the following equation:

$$E_{RS} = AE^{mn} \sin \left[a + \frac{360^\circ}{B} (R-1) \right] \quad (24)$$

A resolver transmitter is a special case of the synchro transmitter. It is a unit where B = 4 and the equation reduces to:

$$E_{RS} = AE^{mn} \sin \left[a + 90^\circ (R-1) \right] \quad (25)$$

Illustrative case: take equation (25)

$$E_{RS} = AE^{mn} \sin \left[a + 90^\circ (R-1) \right]$$

for two input voltages and four output voltages we get:

$$E_{15} = AE^{12} \sin (a + 0^\circ) \quad (25a)$$

$$E_{26} = AE^{12} \sin (a + 90^\circ) \quad (25b)$$

$$E_{37} = AE^{34} \sin (a + 180^\circ) \quad (25c)$$

$$E_{48} = AE^{34} \sin (a + 270^\circ) \quad (25d)$$

Determination of A

The flow of an electric current through a Hall generator such as the control current may be regarded as a result of an average drift of electrons. Consider a conduction electron which moves with a drift velocity, $\vec{V}$, in a crystal. Due to the magnetic flux density, $\vec{B}$, the electron will be acted on by a force, $\vec{F}$, in the direction given by:

$$\vec{F}_L = q (\vec{V} \times \vec{B}) \quad (26)$$

This force, called the "Lorentz force," can be expressed as a field gradient

$$\vec{F}_L = q\vec{E}_H \quad (27)$$

Equating equations (26) and (27) we get

$$\vec{E}_H = (\vec{V} \times \vec{B}) \quad (28)$$

The drift velocity, V, may be related to the current density, J, by the simple expression:

$$\vec{J} = q \vec{V} n \quad (28a)$$

This may be seen from the fact that (qn) represents the charge density in the solid. We may eliminate $\vec{V}$ from equations (28) and, (28a) and, hence, write

$$\vec{E}_H = \frac{\vec{J}\vec{B}}{qn} \quad (29)$$

The Hall voltage across the crystal may be obtained by integrating the electric field with respect to distance.

$$\vec{V}_H = \int_0^w \vec{E}_H \, ds \quad (30)$$

or

$$V_H = E_H w \quad (31)$$

since

$$\vec{J} = \frac{\vec{I}}{wd} \quad (32)$$

we can say

$$\vec{V}_H = \frac{\vec{I}\vec{B}}{dqn} \quad (33)$$

$$I = \frac{Vin}{R} \quad (34)$$

$$R = \rho \frac{L}{A} \quad (35)$$

$$I = \frac{(Vin)\ wd}{\rho\ L} \quad (36)$$

$$p = q \mu_H n \quad (37)$$

$$p = q \frac{3\pi}{8} \mu n \quad (38)$$

$$I = \frac{8Vinwd}{3\pi q \mu nL} \quad (39)$$

we can then set

$$V_H = Vinput \frac{8}{3\pi} \frac{w \vec{B}}{(qn)^2 \mu L} \quad (40)$$

$$\frac{V_H}{Vinput} = \frac{8}{3\pi} \frac{\vec{B}\ w}{(qn)^2 \mu\ L} \quad (41)$$

therefore

$$A = \frac{8}{3\pi} \frac{\vec{B}max\ w}{(qn)^2 \mu\ L}$$

and equation (25) becomes

$$E_{RS} = \frac{.848\ Bm\ w}{(qn)^2\ L}\ E^{mn} \sin \left[a + \frac{360°}{p} (R-1) \right]$$

This equation is true only where the load is at least 8 times the output impedance of the Hall generator.

The equation for a synchro control transformer can be developed in a similar manner. A synchro control transformer is a unit where the Hall voltage out is proportional to the sine of the difference between the electrical input angle and the permanent-magnet rotor angle.

If the output voltage is:

$$E_{RS} = AE^{mn} \sin \left[a + \frac{360°}{B} (R-1) \right] \quad (42)$$

where

$$E^{mn} = A'^{m'n'} \sin \left[B + \frac{360°}{B} (R-1) \right] \quad (43)$$

then

$$E_{RS} = AA'\ E^{m'n'} \sin \left[s + \frac{360°}{B} (R-1) \right] \sin \left[B + \frac{360°}{B} (R-1) \right] \quad (44)$$

Illustrative case:

a. Two crystal, 2 phase CT

$$E_{13} = AA'\ E^{(24)'} \sin (a+0°) \sin (B+90°) \quad (45)$$
$$= KE^{(24)'} \sin (a) \cos (B)$$

$$E_{24} = AA'E^{(24)'} \sin (a + 90°) \sin (B + 0°)$$
$$= KE^{(24)'} \cos (a) \sin (B) \quad (46)$$

Summing the two crystal outputs

$$E_{13} \pm E_{24} = KE^{(24)'} \sin (a) \cos (B) \pm KE^{(24)'} \cos (a) \sin (B) = KE^{(24)'} \sin (a \pm B) \quad (47)$$

which is the desired out put from a control transformer.

Hall-Effect Linear Synchros

The linear synchro differs from the type described above in the shape of the output function obtained. Where the synchros described above produce an output that is a sine or cosine function of the shaft angle, the linear synchro puts out an approximately triangular function. Just how closely the function approaches a triangle depends on a parameter which we shall call the linear range, which is illustrated in Figure 10.

(There is no standard name for this type of device. The wire-wound prototype is generally called a linear synchro at Kearfott; it is also known as an "induction potentiometer." The present device has been informally called a "Hall-effect potentiometer." For taxonomic reasons, the term, "linear synchro," preceded by the modifier "Hall-effect" (or "solid-state"), seems preferable although there are objections to calling a single-phase device a synchro. In this paper, however, we shall feel free to say "potentiometer" or "linear synchro" indiscriminately.)

In the design of linear synchros, additional problems appear, concerned mainly with getting the proper flux distribution. It has been shown above that the flux through the Hall generator, as a function of the shaft angle, is the integral of the angular flux-density function on the inner surface of the return path. This is a result of the shape of the return path. The return path for the linear synchro is the same as that for the sinusoidal type, except that there are only two radial slots, instead of four. It follows that, except for the reduced slot effects, the potentiometer return path has the same integrating effect. Therefore, if we wish to get a triangular output function, we must have a square-wave flux-density function on the inner surface of the return path. This is shown in Figure 10.

The flux-density wave is particularly useful for two reason; first, as a derivative, it naturally tends to emphasize the irregularities of the output. After a certain amount of linearity has been achieved, all output curves look pretty much the same; their failings are immediately made clear by deriving the flux-density curve. In practice this is done by plotting the increments between successive output voltage readings. The resulting curve contains some noise resulting from errors of measurement, but this is slight and does not obscure the information contained in the derivative itself. The second reason for the usefulness of the flux-density function is the fact that it indicates what the physical conditions inside the air-gaps are like, and frequently suggests the direction which further improvements must take. For these reasons, the $B(\theta)$ curve is new as a criterion of device performance, instead of the error curve used for sinusoidal synchros.

In attempting to get the desire square wave, two points of departure were immediately apparent. The first is an imitation of the conventional wire-wound potentiometer in which a lumped winding is used to generate a square wave of flux density. (Figure 11A.) In the solid-state version, the winding is replaced by a slab of ferrite, sandwiched between two semi-cylindrical pole-pieces. (Figure 11B.) The second approach is to use a solid ferrite rotor, specially shaped to give the required flux-density function. (Figure 11C.) The operation of this type of rotor is not quite as straight forward as that of the sandwich type rotor, but depends on an interaction of the shapes of the ferrite and

the air gap.

Each type has its advantages and disadvantages. The sandwich rotor is capable of a higher output, because of the minimal reluctance of the magnetic circuit. On the other hand, the ferrite slab must have a certain minimum thickness, which decreases its linear range. In a rotor 1/4 inch in diameter, for example, a ferrite slab 1/16 inch thick subtends an angle of almost 15 degrees on each side; hence the linear region covers only approximately 165 . For many applications, this is much more than needed, and in those cases range can be traded off for output.

The solid-ferrite rotor operates with a high-reluctance magnetic circuit. This is necessary because of the importance of the air-gap in shaping the flux distribution, as will be shown. As a result, the peak flux through the Hall generator is less than it would be with a sandwich rotor using a comparable amount of ferrite. On the other hand, its linear range is very great, and approaches 180° on each side. For applications in which this is important, the shaped ferrite rotor may be important enough that its relatively lower output is not a problem. As the size of the linear range approaches 180°, the sandwich rotor loses ground rapidly anyway. A little calculation will show that the ferrite slab for a 1/4-inch rotor can be only about ten mils thick if it is to subtend an angle of less than 5 degrees. Somewhere around this point, the available mmf from the ferrite becomes too small to produce a respectable output voltage from the Hall generator.

The remarkable linear range of the shaped ferrite rotor, together with the way in which the shapes of the ferrite and the air gap work together to produce the desired output, make this type of rotor particularly interesting, so that most of the research done, was devoted to perfecting this rotor type. To show how the shape was arrived at, we will start by considering a square rotor, which has been magnetized diagonally from corner to corner. This rotor can be imagined as fitting inside the return path rather like a square peg in a round hole. The output of this rotor is shown in Figure 13. Now the interesting thing here is the splendid sharp peaks on the output function. These seem like almost more than we have a right to expect, starting with such a simple shape. Just under these peaks, the flux-density function looks a little like a square wave, in the abrupt way in which it crosses the axis.

Now, this flux-density wave can be displayed as the product of two other functions, one a triangular MMF wave, the other a peaked "performance wave." This is done in Figure 12. The triangular MMF wave reflects the shape of the rotor. The validity of this rests on the assumptions that the demagnitization curve of the ferrite is linear, and that the lines of flux, inside the ferrite, are parallel. The peaked curve is the permeability function which we must have to get the flux density wave of Figure 12 out of the MMF wave of Figure 14. The large peaks are finite. The permeability function is not symmetrical about the 45 and 135 degree abscissae, as might be expected, because the flux in the air gap is not symmetrical. It appears that all that is necessary is to grind off the unwanted corners of the rotor, to remove the peaks from the flux density function and give us the desired triangular $V(\theta)$ function.

Unfortunately, beyond this point, the design of the rotor shape must be largely empirical. The MMF and permeance waves cannot be plotted accurately for a given rotor shape, first because we don't know quite how the actual MMF function deviates from that predicted from the shape of the rotor; and secondly, because to plot the permeance wave, we must know the actual flux paths in the air gap. The trouble is that the boundary conditions at the ferrite-air interface are not presently well understood. Investigation of these boundary conditions, along with analysis of the air-gap flux distribution, has only begun at this writing.

As it happens, the empirical approach has been fairly successful. In testing various rotor shapes, the flux-density function has proved a powerful and useful tool. So far, we have achieved less than 1% functional error over a range of about 90° in a size 4 unit. The output and flux-density curves for this unit are shown in Figure 15. Figure 14 represents a guess at the MMF and permeance functions for this configuration. The effect of the discontinuities (A,A,A,A) has been extremely slight. They are barely discernible in the flux-density function of Figure 16. Up until now, they were masked by much larger errors stemming from defects in the ferrite itself, and in fact the shoulders (A - B and A' - B') may also be in part due to the same cause.

Non-linearities in this type linear synchro are mostly traceable either to defects in the rotor or to disturbances in the air-gap. The importance of the air-gap in shaping the flux distribution renders the device somewhat sensitive to external magnetic fields. If the synchro is used in a d-c system (i.e., with d-c on the Hall generators), then the effect of external a-c fields is nullified. We have had more trouble with rotor defects. These have been mostly inhomogeneities in the ferrite. Even in non-oriented ferrites, anisotropies have been observed which appear to have resulted from uneven pressure, or directional pressures, introduced in the preparation and heat-treatment of the ferrite. These can wreck a flux distribution, and we have not yet discovered a way to detect unsatisfactory ferrites before machining. If solid-state synchros eventually become as important as we believe they will, it may prove profitable for ferrite producers to undertake development programmes aimed toward improving this situation.

Amplification

Until there is a genuine breakthrough in the manufacture of Hall generators, the low-output voltage of these synchros is going to present problems. They obviously cannot be used iteratively, because of the insertion loss in each unit, presently about 30 db. Hence, it was

thought worthwhile to consider the problem of amplification of the output signal. Three questions are of particular interest: (1) how to deal with the absence of a common ground point between the input and output; (2) how to achieve satisfactory d-c amplification (where desired) economically; (3) whether there is a way of nulling out the misalignment voltage by means of the amplifier circuitry.

The use of a differential amplifier is the natural solution to these problems. The output of a Hall generator is a differential signal; with proper handling a differential amplifier can be made to put out a single-ended signal capable of driving another Hall generator.

Since the operation of differential amplifiers is fairly widely understood, there seems no reason to belabor it here. We will similarly pass over questions of drift stability as being beside the point, although of course important. On the other hand, it is of considerable interest to see whether the effect of misalignment voltage can be nulled out by manipulation of amplifier circuitry.

As a starting-point, we will use equivalent circuits for the Hall generator and amplifier shown in Figures 16 and 17 respectively. In the Hall generator, A and B are the input terminals and C and D are the output terminals. The Hall-effect itself is indicated by the two voltage sources, Ki_cB, and the misalignment voltage is shown by the resistive drop across r_{ml}. We are assuming linear operation here, which is justified at moderate power and flux levels. In our analyses, r_c and r_d will be neglected, since they are very small compared to the input impedance of the amplifier (typically $r_c = 50\Omega$, $Z_{in} = 5000\Omega$).

The amplifier has been kept simple; if we can null out the misalignment voltage with this circuit, we can also do so with more elaborate circuits. It is assumed that the emitter current for this amplifier comes from the collector of another transistor; since this is a high-impedance source, we have used a Norton equivalent circuit consisting of I and G_1. We will simplify things further by assuming:

$$I_{co} = 0$$

$$h_{FE} = \infty$$

$$V_{be} = 0$$

$$h_{oe} = 0$$

Using this circuit, the single-ended gain of this amplifier can be shown to be:

$$V_4 = V_{cc} - R_5 \frac{IR_2 + (V_2 - V_1) + V_2 R_2 G_1}{R_2 + R_3 - R_2R_3G_1} \qquad (48)$$

$$V_3 = V_{cc} - R_4 \frac{IR_3 + (V_1 - V_2) + V_1R_3G_1}{R_2 + R_2 - R_2R_3G_1} \qquad (49)$$

The differential gain is:

$$V_4 - V_3 = \frac{\left[I(R_2R_5 - R_3R_4) + V_2(R_4 + R_4 + G_1R_2R_5) - V_1(R_4 + R_5 + G_1 R_3 R_4) \right]}{R_2 + R_3 + R_2R_3G_1} \qquad (50)$$

Typical Circuits

We will consider two circuits, to show how the misalignment signal may be nulled-out. In the first, the amplifier has a differential input and a single-ended output. We will assume the Hall generator has a d-c bias which ensures that the output voltage is positive under all signal conditions (so the transistors Q_1 and Q_2 will not be cut off). Since the device may be an element in a synchro chain, we must assume that there are 2 information inputs: shaft-angle information from previous synchros enters as an input voltage superimposed upon the bias, and shaft angle information from the present synchro appears as a variation in the flux-density in the Hall generator.

Assume that the input signal is an a-c voltage superimposed on the d-c bias. Then $i_c = I_b + I_s \sin wt$. Then the output voltages from the generator are

$$V_c = (I_b + I_s \sin wt)\,(r_b - KB) \qquad (51)$$

$$V_d = (I_b + I_s \sin wt)\,(r_b + r_m + KB) \qquad (52)$$

Now if these are made to be V_2 and V_1, respectively, of the amplifier, it can be shown that the a-c component of the output voltage is

$$V_4(a\text{-}c) = I_sR_s \frac{rm + 2KB - (r_b - KB)\, R_2G_1}{R_2 + R_3 - R_2R_3G_1} \sin wt \qquad (53)$$

Now if the misalignment-voltage signal is bucked-out, then $V_4\,(a\text{-}c)$ must be zero when B = 0. This requires that

$$r_m = r_b R_2 G_1,$$

or

$$R_2 G_1 = \frac{r_m}{r_b}$$

Now a null-d-c system has the advantage that the effect of external a-c magnetic fields can be minimized. In such a system, the input information is a variable d-c signal superimposed on the fixed d-c bias. Now the input signals from the Hall generator are

$$V_c = (I_b + I_s)\,(r_b - KB) = V_2$$

$$V_d = (I_b + I_s)\,(r_b + r_m + KB) = V_1$$

and it can be shown that

$$V_4 = V_{cc} - R_5 \left[\frac{IR_2 - (I_b + I_s)(r_m + 2KB + (KB - V_b) R_2G_1)}{R_2 + R_3 - R_2R_3G_1} \right] \quad (54)$$

Now presumably when B = 0, V_4 has some quiescent value, V_4 (0). Then if the misalignment component is bucked-out, this should remain constant when I_s varies. This requires that

$$\left. \frac{\partial V_4}{\partial I_s} \right|_{B = 0} = 0$$

It turns out that this is satisfied if $R_2G_1 = r_m/r_b$, which is just the same as in the a-c case.

Now it may be that a second differential stage is operated from the output of the first. So before we leave this topic, it may be worthwhile to establish that the misalignment-voltage signal can be eliminated from the differential output of the first stage. If we take the d-c case, the differential output is

$$V_4 - V_5 = \frac{I(R_2R_5 - R_3R_4)}{R_2 + R_3 + R_2R_3G_1} + (I_b + I_s) \quad (55)$$

$$\frac{(r_b-KB)\ (R_4 + R_5+G_1R_2R_5)-(r_b+r_m+KB)(R_4+R_5+G_1R_3R_4)}{R_2 + R_3 + R_2R_3G_1}$$

We require that $V_4 - V_5 = 0$ when B=0. Obviously, in the a-c case, I will have to be zero, which in practice saves one transistor. Then the remaining requirement is that

$$r_bG_1R_2R_5 = r_bG_1R_3R_4 + r_m\ (R_4 + R_5 + G_1R_3R_4)$$

$$R_3 = \frac{r_bR_2R_5 - V_m\ (R_4 + R_5)/G_1}{R_5(r_m + r_2)} \quad (56)$$

In practice, this expression comes out positive. It looks difficult on paper, but in the actual circuit, use of a resistance with a variable tap for R_2 and R_3 solves the problem. A similar result is obtained in the a-c case, except that I does not have to be zero.

The important point in each of these circuits is that the misalignment voltage is bucked out, actually, by using a little of the common-mode signal resulting from the drop across r_b. If the Hall generator is properly connected, the r_m and r_b voltages can be made to cancel each other out, and in this way the misalignment voltage problem is eliminated. It should be observed, however, that r_m and r_2 are not necessarily constant over variations in power level and temperature. Nevertheless, it seems reasonable to suppose that an optimum arrangement would ordinarily be arrived at.

Example of an Iterative Unit

We may now consider an iterative unit, consisting of synchro and associated amplifier capable of being used in a chain. In the interest of simplicity, it has been assumed that a single differential input, single-ended output amplifier will have sufficient gain to drive the next Hall generator. (This implies Hall generators with somewhat higher output than are presently available.)

Figure 18 is a block-diagram of 2 iterative units in a chain. We assume that the information is transmitted on a a-c carrier. The first synchro in the chain starts with an unmodulated carrier. A d-c bias is added, greater than the peak amplitude of the a-c signal. The output of the synchro is a differential a-c signal with various d-c and common-mode and misalignment signals present. The output of the amplifier is a single-ended modulated carrier, capable of driving the next synchro.

Figure 19 shows one way in which this might be done. R_x and R_y supply the bias, which along with the carrier input, drives the Hall generator through a common-collector stage, Q_3. The amplifier is similar to the ones described above: G_1 and R_2 are adjusted to cancel the common-mode and misalignment signals. The current-source, I, is the collector of Q_4. The output from R_5 passes through a buffer stage, Q_5 and R_6, and is ready to be applied to the next link in the chain. With present day techniques, this circuit could probably be made to fit in a size 11 synchro package, along with a size 4 synchro.

Conclusion

The above technical discussion describes the characteristics of a new family of electrical transmission components especially adapted to the requirements of high reliability, long-term operation, and small size characteristic of the technical level of our time. Figure 20 illustrates size 4 sinusoidal synchro, and Figure 21 illustrates the change in rotor design characterizing the linear synchro. These units will be the largest size of a new generation of components whose ultimate accuracy will probably be in the order of .05%.

Acknowledgment

The authors would like to express their appreciation to Gary Bauerlien for the many stimulating discussions and helpful suggestions toward the development of these units.

List of Symbols

V_H = Hall voltage

R_H = Hall Constant

I = Control current

B = Magnetic field

d = Crystal thickness

r_m = Misalignment resistance

ρ = Resistivity

μ = Mobility

J = Current density

A = Crossectional area, control current direction

l_m = Misalignment distance

n = Electron density

q = Electron charge

v = drift velocity

$\vec{E}_c$ = Electric field in control direction

w = Crystal width

E_{RS} = Output voltage for Hall effect generator

A = Constant

E^{MN} = Input voltage for Hall effect generator

a = Electrical angle

B = Number of phases

R = Constant

$\vec{F}_L$ = Lorentz force

$\vec{V}$ = Drift velocity

$\vec{B}$ = Magnetic field vector

$\vec{E}_H$ = Hall field

n = Charge density

w = Width in Hall direction

L = Length of crystal

μ_h = Mobility in Hall direction

μ = Mobility in control current direction

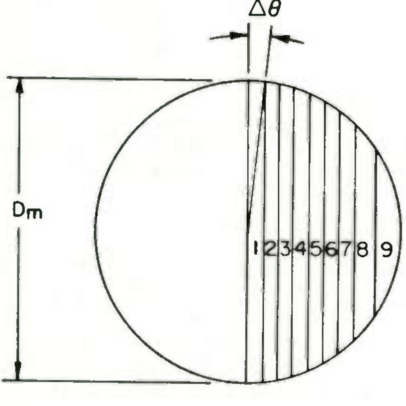


FIG. 1

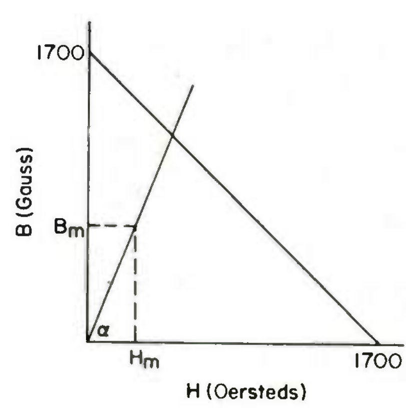
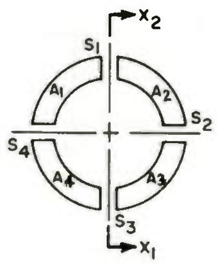


FIG. 2



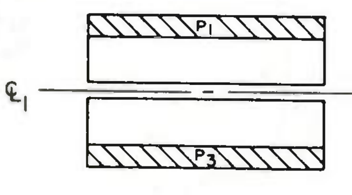


FIG. 3

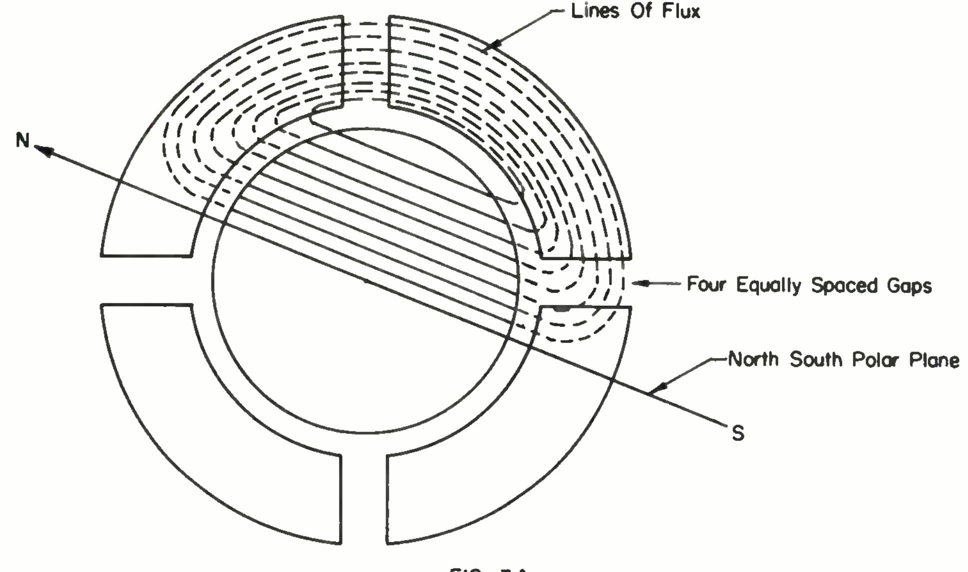


FIG. 3A

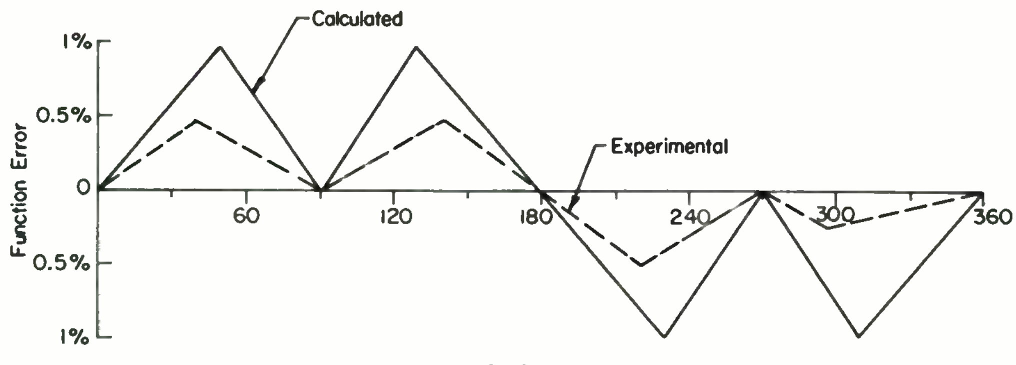


FIG. 4

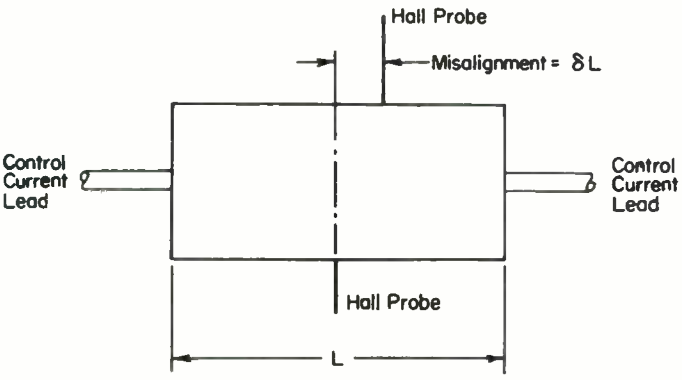


FIG. 5

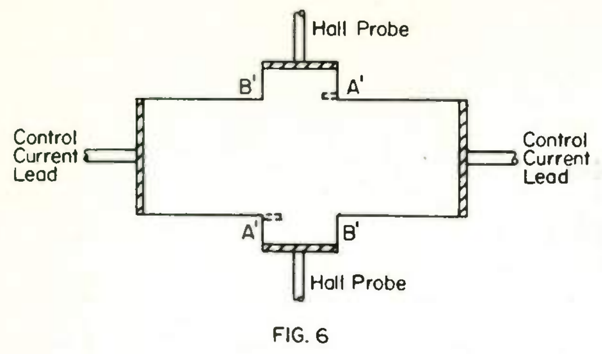


FIG. 6

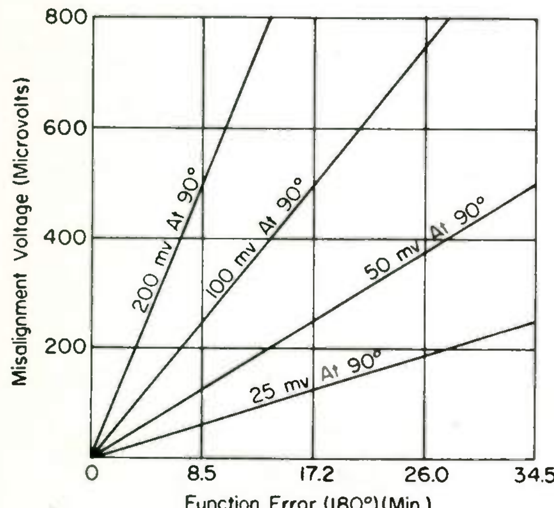


FIG. 7

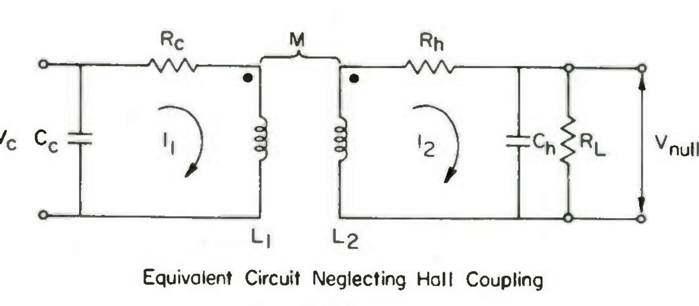


FIG. 8

Equivalent Circuit Neglecting Hall Coupling

FIG. 9

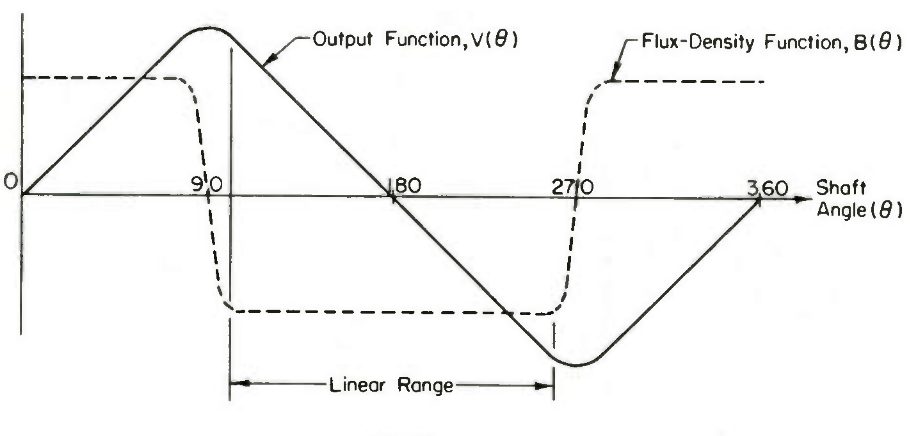


FIG. 10

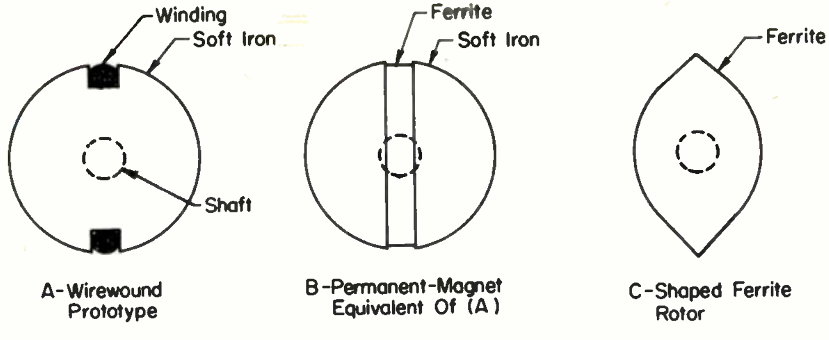


FIG. 11

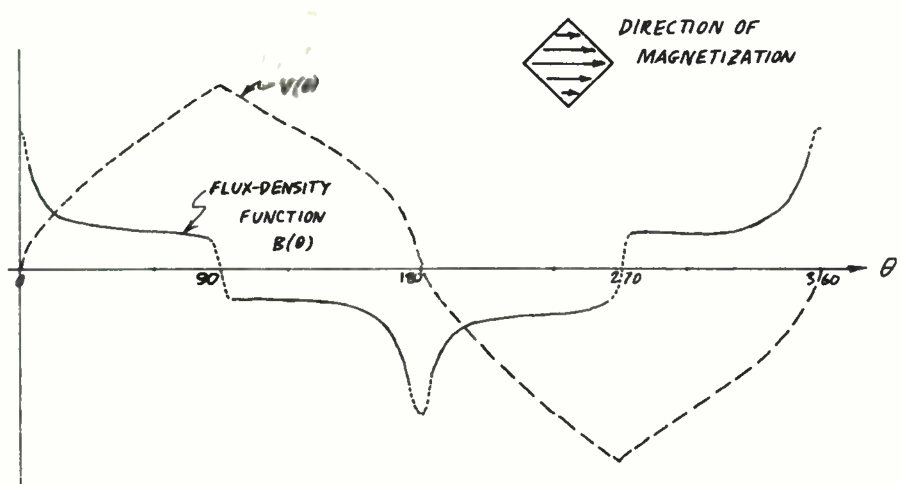


FIG. 12

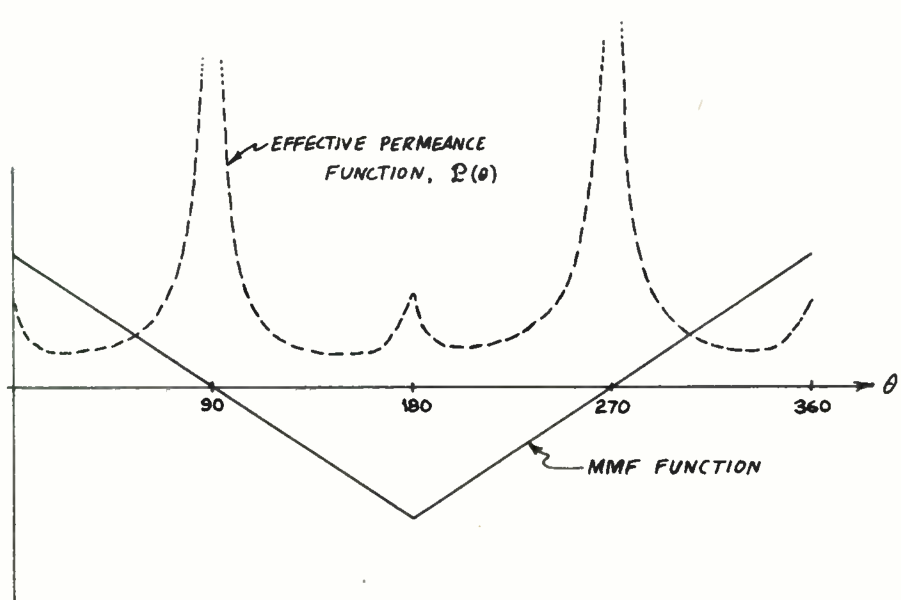


FIG. 13

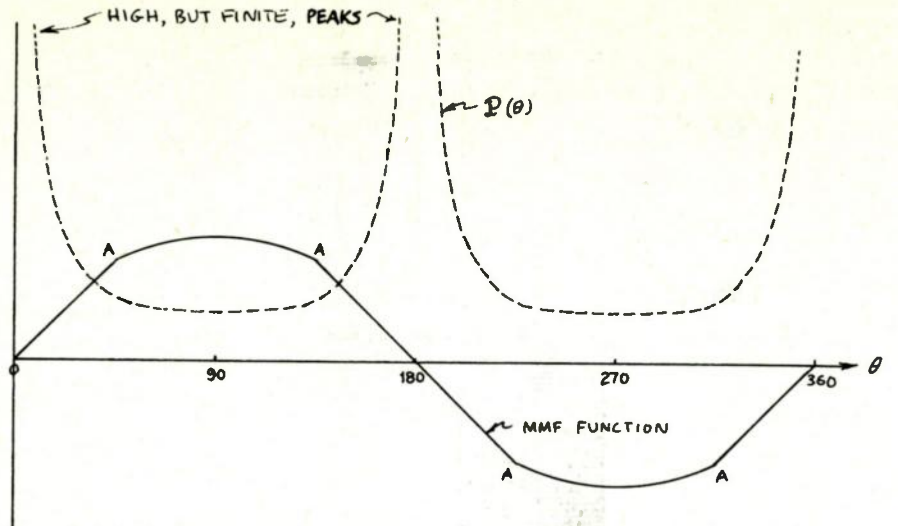


FIG. 14

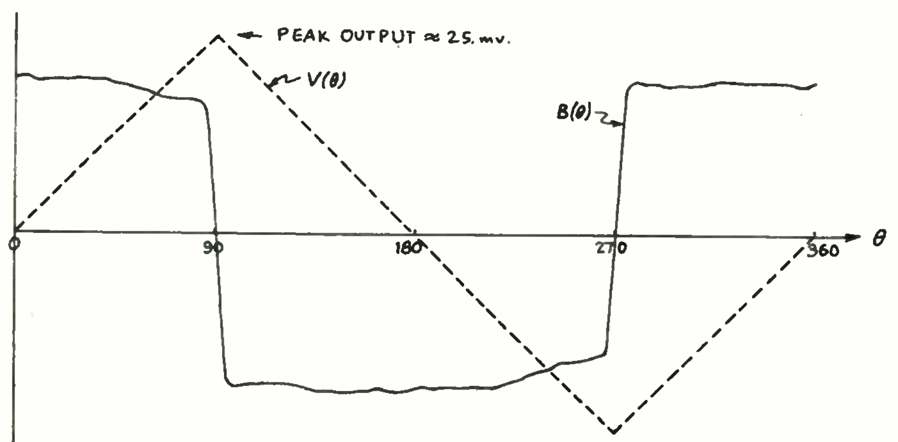


FIG. 15

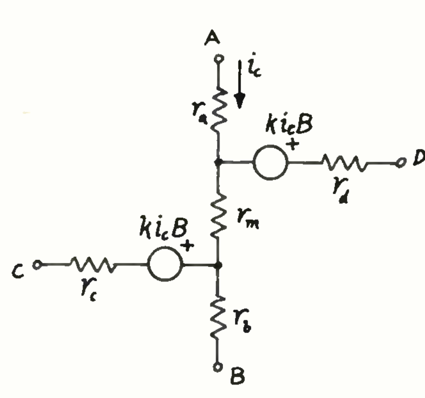


FIG. 16

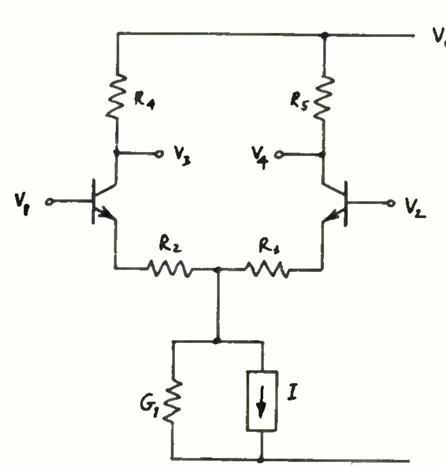


FIG. 17

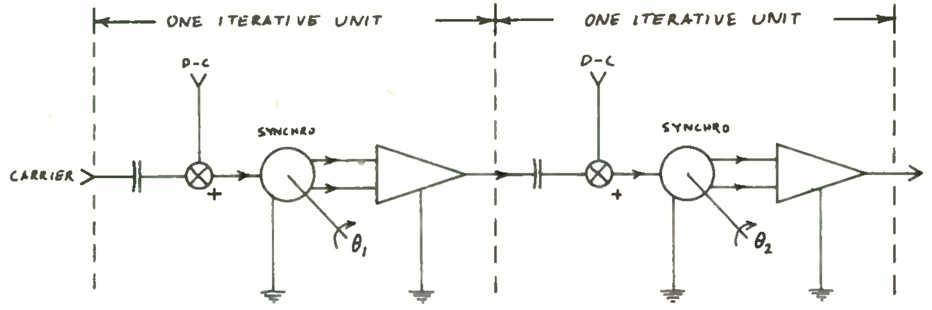


FIG. 18

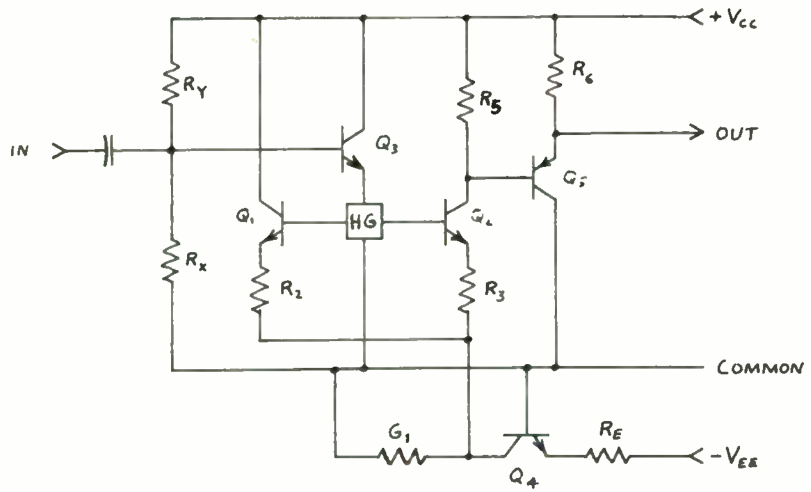


FIG. 19

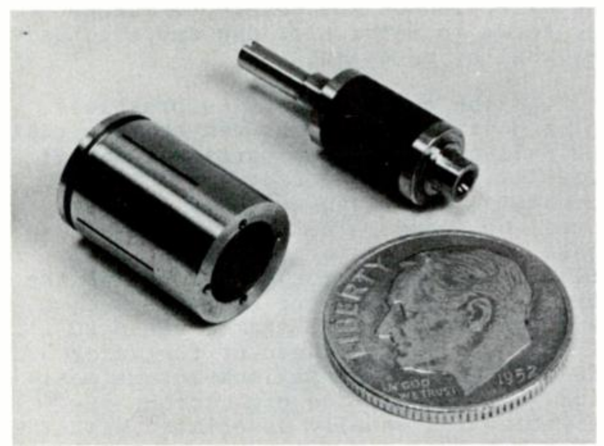


Fig. 20.

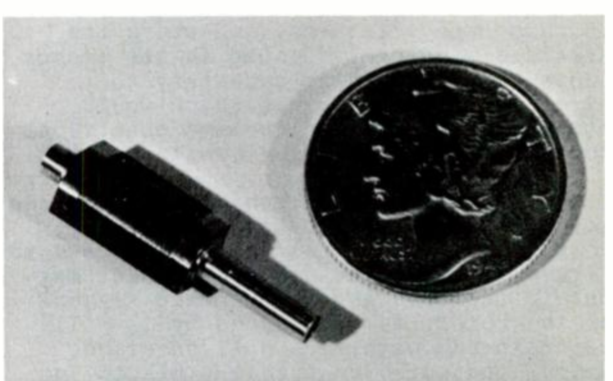


Fig. 21.

A BRUSHLESS DC MOTOR WITH SOLID-STATE COMMUTATION

Gerhard Bauerlein
Kearfott Division
General Precision, Inc.
Little Falls, New Jersey

DC Motors have long been recognized as efficient and easily controllable machines. The drawbacks associated with their commutator, however, have limited their use in many applications. Attempts have been made to eliminate these drawbacks, but these have usually led to the elimination of some of the desirable dc motor characteristics as well.

A brushless dc motor has been developed which promises these desirable dc motor characteristics:

a. high torque-to-weight ratio because the torque angle is maintained at 90° regardless of loading, and

b. a linear speed-torque characteristic which is adjustable by varying a low-power signal input.

In addition, these motors, because they are brushless, should have longer maintenance-free life, high reliability, and less friction.

General Description

The approximate placement of the subassemblies of the subject motor are shown in Figure 1. The construction is that of an inverted dc motor wherein the field structure - a permanent magnet in this case - rotates, and where the armature winding is placed in the stator slots. The latter is necessary to eliminate the slip rings which would otherwise be required to make electrical connection to a rotating armature.

A schematic representation embodying the brushless dc motor design concept, using Hall effect generators, is shown in Figure 2. Here the inner permanent magnet ①, representing the forward segment of the rotor, is surrounded by 12 Hall generators ③ with each Hall generator output connected to a coil ④ placed in the appropriate slots in the stator. These coils are so located that they produce a field which is displaced 90 electrical degrees from its associated Hall generator. This is represented by showing the outer rotor, whose field reacts with the armature winding to produce torque, displaced 90° from the inner rotor.

The excitation of the armature coils by the Hall generators produces a stator field which leads the outer rotor poles 90° in space. The outer rotor attempts to line up with stator field and, in so doing, rotates. As it rotates, the two Hall generators under the trailing pole tips experience a change of field. This field reversal causes the Hall generator output to reverse also, switching the polarity of the two coils nearest the advancing rotor pole face. The two Hall generators under the leading pole tip also experience a field reversal and their output polarity is thereby reversed. This switches the two coils farthest from the advancing rotor pole tips. It can now be seen that with this arrangement as the rotor tries to follow the stator field it initiates action to cause the stator field to rotate away from it, producing continuous rotation.

If the control current to the Hall generators is reversed, the polarity of the output of every Hall generator in the array is reversed, which changes all the stator north poles to south poles and vice versa. This reverses the direction of the developed torque and leads to motor slowdown and eventual reversal of rotation.

In the operation of the brushless motor just described, it can be seen that the Hall generator and permanent magnet directly replace the functions of the commutator and brush assembly of the normal dc motor. These functions are to sense the position of the rotor relative to the stator and to switch the current in the armature coils in such a manner as to keep the stator field 90 electrical degrees ahead of the rotor field. This is the maximum torque position and leads to the high torque-to-weight ratios usually associated with the dc motor.

The Hall Generator As a Rotor Position Sensor

The Hall generator possesses characteristics which make it a good rotor position sensor. It is small in size (1/8" x 1/16") and thin (.025") and can therefore be conveniently placed in the air gap of a magnetic circuit, as shown in Figure 1. It senses the magnitude of the flux and not its rate of change with respect to time. This provides two advantages: there is an output produced at speeds near and at standstill, and the output does not become excessively large as the speed increases. Figure (3a) shows the output voltage produced by one Hall generator at various rotor positions, accurately reproducing the flux distribution of the rotor. The maximum voltage amplitude of this wave is small in comparison to that of inductive pickoffs - on the order of 100 millivolts.

The Hall Generator As a Switching Mechanism

As an armature current switch the Hall generator has a major drawback: its low efficiency of power transfer. A limitation exists on the maximum efficiency a Hall generator can achieve since the Hall generator is basically a resistive device. Efficiency has been variously estimated at between 10 and 20% in a 10 kilogauss field and using high mobility materials (such as In As and In Sb). With a practical magnet-Hall commutator design, a maximum field of four kilogauss can be realized. With a typical In As Hall generator operating into a matched load, this yields an efficiency of:

$$\frac{P_{out}}{P_{in}} = \frac{(V_{ho})^2/4R_h}{I_c^2\, R_c} = 1.86\%$$

where

control current (I_c) = .4 amps

control resistance (R_c) = 1.4 ohms

open circuit Hall voltage (V_{ho}) = .152v/4 kilogauss,

if all resistance (R_h) = 1.4 ohms.

This low efficiency is clearly too great a penalty to pay for the switching function.

An experimental motor which was built using the above design indeed did not rotate. However, a rotating field was generated by the armature winding when the shaft was mechanically driven.

The motor was then modified by having three of the 12 Hall generators work into three one-transistor amplifiers. This provided the needed extra power which enabled the motor to maintain rotation.

Variations Utilizing Semiconductors

Transistors and SCR's may be used as switching devices in order to improve efficiency. Here the Hall generators are used only as position sensors and handle only low-power signal currents. Thus their low efficiencies cause only a small total power loss. The need to dissipate less power also allows the use of smaller Hall generators.

The associated electronic circuitry can be divided into two modes of operation by function: amplifying or switching. The application determines which mode will be used. If speed control with a low power control signal or a smooth torque vs. position characteristic is required, the amplifying mode of operation should be used. If high efficiency is the most important consideration, the switching mode of operation is preferable. This mode is analogous to the brush-type dc motor.

In the switching mode greater efficiency can be obtained because the transistors (or SCR's in larger units) operate either full-off or full-on, reducing the time the transistors are conducting partially to the actual switching transient and thereby minimizing the power losses in the transistors. This in turn allows the use of smaller transistors and heat sinks. Speed can be controlled by reducing the dc supply voltage to the amplifier. This is equivalent to armature speed control in an ordinary dc motor. If a permanent magnet is used for the rotor, field control of speed is not possible.

When amplifiers are used, their output voltage magnitude is determined by the magnitude and polarity of both the commutator field and the control current, in addition to the amplifier gain. This is shown in Figure 3b where the curve labeled I_c represents the output wave shape at maximum control current and $I_c/2$ represents the output at one half this value.

The use of amplifiers allows a reduction in the complexity of the Hall commutator. In a multipole machine the winding currents are of the same magnitude and reverse under every pole. Since

this means most of the Hall generators supply redundant information about rotor position, (n-1)/n Hall generators can be eliminated in an n-pole machine. Figure 4 shows a schematic representation of such a machine using four separate electrical circuits whose current flow will be controlled by the four Hall generators shown distributed under one pole arc. For clarity only one of the series of coils is shown connected to the amplifier.

Breadboard Motor Design

A machine based on the foregoing principles was designed, built, and tested. The brushless dc motor was to drive an inertia wheel to control the attitude of a satellite in space. This would be done by utilizing the principle of the conservation of momentum within the satellite. The motor was designed to operate in a closed loop where its speed is controlled by a command signal which represents the difference between the desired and actual satellite attitudes. Since accurate speed control and smooth torque were requirements of the design the amplifying mode of operation was required. A dc amplifier with a differential input was used in order to minimize the number of components while still maintaining amplifier drift at a low value. Since the motor was to be used in a closed loop the remaining drift would be integrated out by the system.

The design of an amplifier for this purpose presented some problems since the input circuit of a Hall generator is resistively coupled to the output circuit; because of the nature of this coupling only one side of the input or one side of the output can be grounded. Figure 5 shows one of the four amplifiers used in conjunction with the Hall generators. Resistors R_1 and R_2 were adjusted to balance out the non-symmetry in the division of the resistance of the Hall generator in the control direction (R_c). The amplifiers were required to operate over a frequency range of dc to about 1 kcps. They had a voltage gain of 120 and a power gain of 250,000.

Figure 6 is a photograph of breadboard machine. Observation of the amplifier output on an oscilloscope showed the waveform to be generally trapezoidal in shape as seen in Figure 3b. This type of wave shape is advantageous since it can be summed with others of the same shape but time-displaced to reflect a constant current draw (back to the dc source). The four separate circuits were approximately similar in wave shape and spaced 45 electrical degrees apart in time. This time displacement of current together with the space displacement of the windings caused the armature winding to produce a rotating four-phase field. Thus the operating motor can be considered to be a synchronous machine whose rotating field is maintained at such a speed that it is always 90° removed from the rotor field axis.

Test results (Figure 7) show the speed-torque characteristic to be linear for a given control current. However the increases in the individual speed torque lines do not follow linearly the increases in the control current. This is in part due to the fact that the amplifier outputs were not all balanced properly and some clipped earlier than others. All amplifiers were clipping at I_c = .165 amperes as can be seen by the fact that an increase in control signal to .222a does not cause an increase in motor output.

The efficiency of the whole motor is a maximum of about 3% (Figure 8). This compares to about 30% or more in the normal ac instrument servo motors. I believe that with the improvement of Hall generators to yield higher output and the refinement of the amplifier circuitry, marked improvements in the efficiency of the motor can be achieved.

Future Trends

The brushless dc motor utilizes devices which are currently the subject of intensive research. The Hall generators are being made smaller and with higher voltage outputs. The high mobility semiconductors are also under investigation with a view of obtaining more-efficient Hall generators. Transistors in matched pairs using the same substrate material are coming into the market. These transistors used in differential amplifier stages should track each other's variations more closely and therefore reduce the drift associated with dc amplifiers. Advances in any of these fields will improve the brushless dc motor performance.

Acknowledgments

The author wishes to thank Messrs. D. R. Simon and T. W. Parsons, Project Engineers at Kearfott Division of G.P.I., who assisted in the construction of the breadboard motor and Mr. Z. R. S. Rataj-ski, Chief of Research and Advanced Development Department at Kearfott whose encouragement and advice assisted in the preparation of this paper.

Bibliography

1. Hall Effect Devices, by W. J. Grubbs, Bell System Technical Journal, May 1959.

2. The Hall Effect Compass, by I. M. Ross, E. W. Saker, and N. Thompson, Journal of the Scientific Instruments, Vol. 34, Dec., 1957.

3. Designing Solid-State Synchros with Hall-Effect by Z. R. S. Ratajski, Electronics, Sept. 3, 1961.

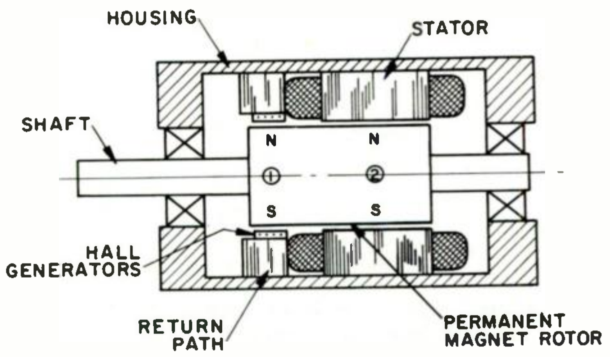


Fig. 1. Components of a brushless dc motor.

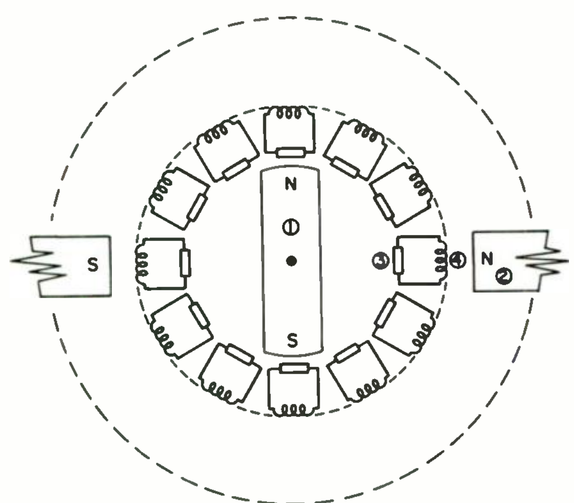


Fig. 2. Schematic diagram of brushless dc motor.

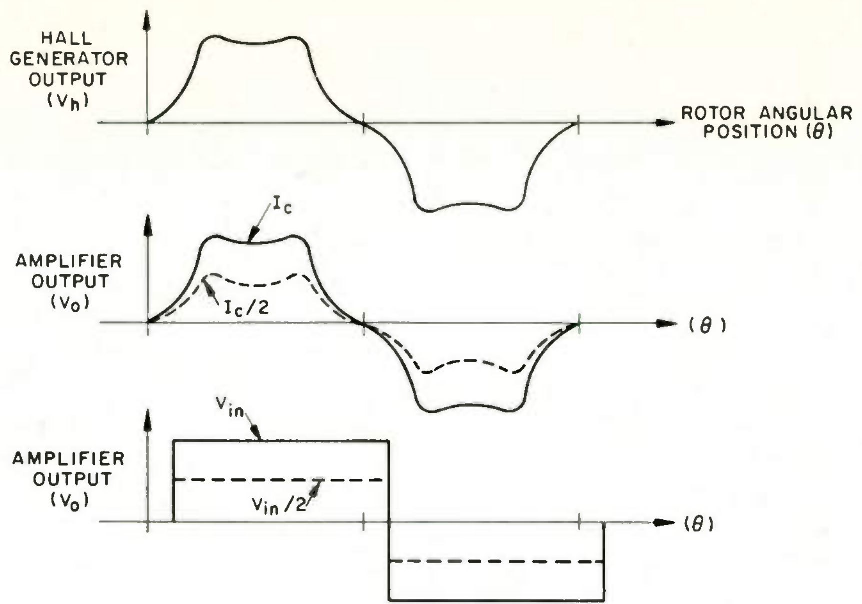


Fig. 3. a) Output of position sensor. b) Input to stator winding in amplifying mode. c) Input to stator winding in switching mode.

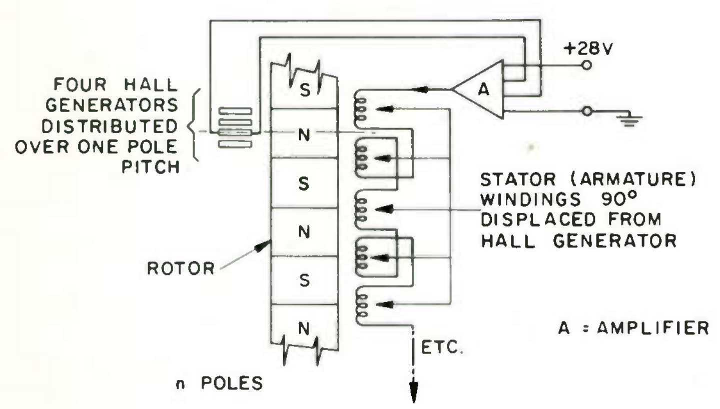


Fig. 4. N-pole brushless motor using four separate electrical circuits whose current flow is controlled by the four Hall generators shown distributed under one pole arc.

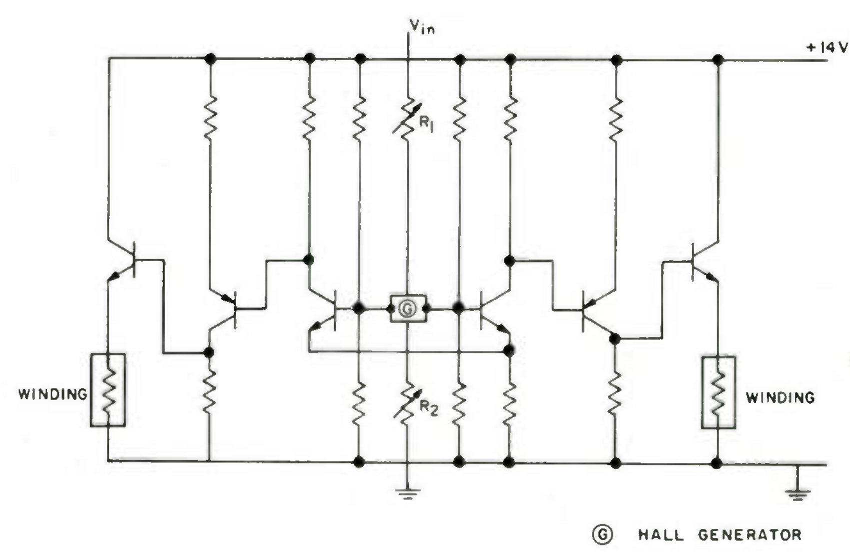


Fig. 5. The electronics associated with one Hall generator and winding group.

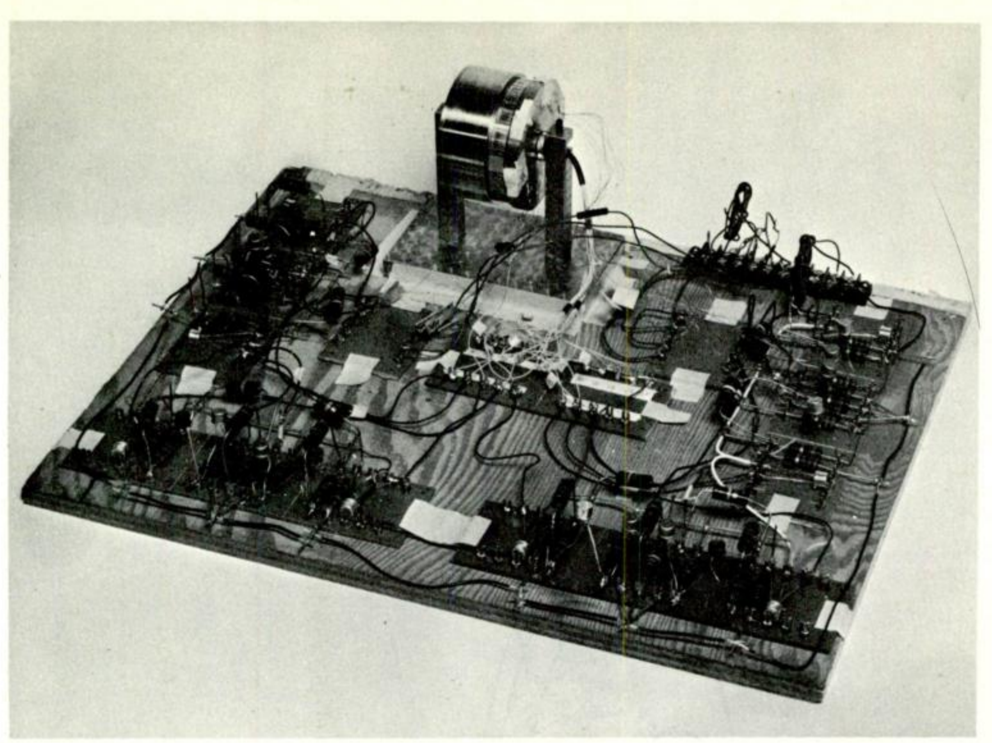


Fig. 6. Breadboard of brushless dc motor.

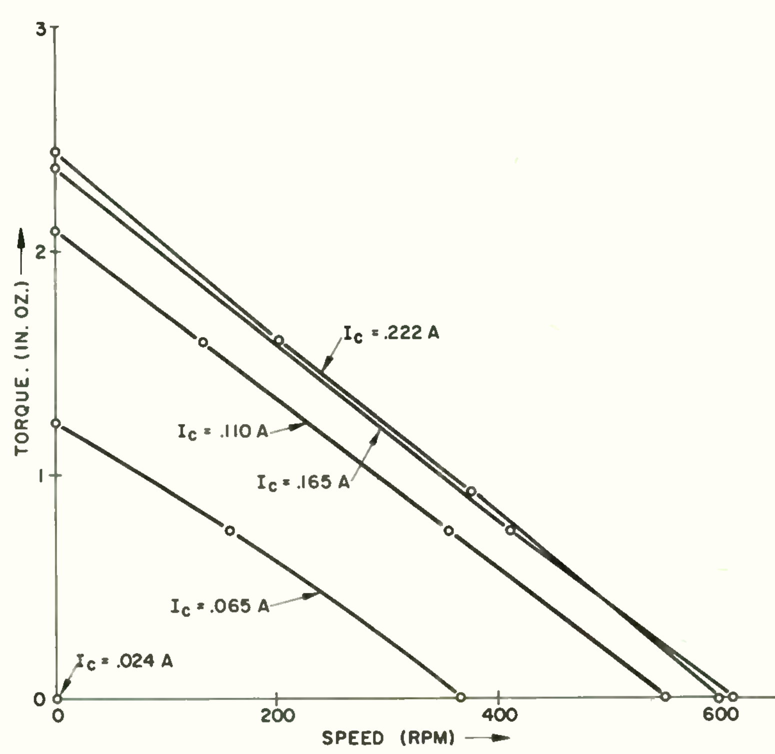


Fig. 7. Test results show the speedtorque characteristic to be linear for a given control current.

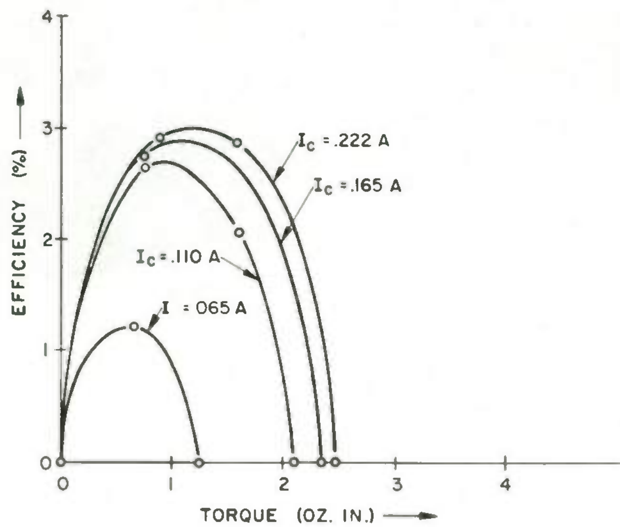


Fig. 8. Graph of efficiency vs torque.

THE SYSTEM CONCEPT OF PRECISION POTENTIOMETER SPECIFICATION

David C. Hoos
Spectral Electronics Corp.
San Gabriel, Calif.

Abstract

The system concept of potentiometer specification is a streamlined method of determining which potentiometer parameters need be specified. The method which is presented frees the design engineer from much work by permitting him to specify potentiometers in terms of system requirements. The paper further shows that not only does the streamlined method of potentiometer specification save time for the design engineer, but will also result in lower cost, and potentiometers that are more nearly suited to the application. A specification check list is also included.

SPACE RADIATION RESISTOR EVALUATION

I. Doshay, Space-General Corporation, El Monte, California

Abstract

In this paper we describe an evaluation of 324 Sprague resistors of types RN65B* and RN65C.** It involves a statistically designed experiment of functioning and non-functioning parts exposed to a combined vacuum-radiation environment. The paper includes observations of resistors under load during exposure to a maximum of 7.2×10^{7} Roentgens of gamma radiation (in a vacuum environment of 10^{-5} to 10^{-3} mm Hg) from a 10 Kilo Curie Cobalt-60 source.

The Space Component Evaluation Program

In the next ten years man will probe into the far reaching regions of space. The missions will involve covering hundreds of thousands and possibly multimillion miles of expanse in between planets and traversing heretofore unconquerable barriers of environment. The spacecraft that will be designed to satisfy these mission requirements must therefore be fabricated with components known to have a high probability of surviving their environment applications.

Van Allen belt radiation will affect the lunar and planetary probes as well as numerous other spacecraft which orbit at greater than several hundred miles. Radiation trapped in the Van Allen belts includes dose rates of up to approximately 10^{2} Roentgens per hour; mostly of electrons from .02 to 2.5 Mev and a small fraction of high energy protons. We shall discuss this in more detail later on. However, it may also be noted that periodic solar flare activity increases the radiation intensity.[5, 12]

This paper is a preliminary report and the first in a series to describe the results of exposing essential mechanical and electronic parts to a combined radiation-vacuum environment and establishing their suitability to space vehicle applications currently being evaluated at Space-General Corporation laboratories. This report covers, specifically, Sprague resistor type RN65B (±1% deposited carbon) values of 1K ohms, and 1M ohms, and type RN65C (±1% metal film) values of 1K ohms and 750K ohms. In this space simulation test the 318 non-functioning resistors were placed in small capsules that were sealed with a special leak-proof gasket. These non-instrumented containers were then pumped down to the required vacuum via a tubulation and then pinch sealed and soldered. An appendage-type ion pump gage and valve were permanently attached to the instrumented container, containing 6 resistors. These capsules were then introduced to the radiation environment, consisting of a Kilo-Curie Cobalt-60 gamma radiation source.

*Sprague 406E - **Sprague 419E

Although there is a wealth of reports on nuclear radiation effects for general classes of components and materials, there is still lacking data on specific component qualification in space radiation environment. Due to lack of knowledge concerning the operational characteristics of specific electronic components in a space radiation environment, a program has commenced which will result in the exposure of certain electronic components to suitable radiation levels and in the performance of functional, non-destructive testing and analysis of the components prior to, during, and subsequent to exposure. The resultant knowledge obtained will allow preparation of standards on "Space Radiation Components" which will include reliable performance levels in exposure to Van Allen belt radiation and in systems exposed to limited radiation from nuclear reactors, a power source for space engines.

Reliability in the Space Environment

The effects of radiation upon materials has been the subject of investigation for many years. However, electronic components represent complex arrangements of many different materials performing a variety of electrical and mechanical functions. The arrangement of the materials and the functions performed may improve, deteriorate or not affect the component performance when radiation changes occur.[8] It will therefore require individual assessment of the effects of radiation on each class or style of component in its intended application. For example, a disc of Teflon may be used as a seal or an insulator in a tantalum capacitor. Subjecting the part to a space radiation environment would

cause it to swell and it also will exhibit some deterioration in insulating properties. In addition, it releases Flourine gas, which may be the most significant factor. Under vacuum conditions the part may burst as a result of excessive internal to external pressure differential. All other factors may not exceed performance criteria. Polyethylene also swells during radiation and then shrinks subsequently accompanied by an improvement of insulation properties. Discovery of this condition recently in a routine research project led to the formation of a company that intentionally radiates polyethelene and other materials for insulation and other applications. The combination of effects of the numerous materials within a component are thus extremely difficult if not impossible to analytically anticipate. However, the resultant effects can be obtained and exhibited for analysis by placing the component in a suitable simulated environment. To quote from the conclusion of Dr. Wiser's paper "The various and intense nuclear radiation environments in which electronic systems must perform interact with the systems providing undesirable complex effects. Optimistically, I believe that continuing and expanding work will produce component radiation specification ratings, deratings and preferred circuits for application to particular radiation environments".[21]

These effects can be measured and expressed in terms of various reliability indexes in accordance with military specifications and standards. It is the intention of this program to provide necessary space radiation evaluation criteria, as well as data, to enable a common basis for component rating and selection to be established.

Basis for Proper Simulation Criteria

Reasons for use of the Cobalt-60 source for purposes of space radiation simulation of effects on components whether composed of one or more organic materials or of covalently bonded metals is sometimes questioned. As a prelude to our further discussion we quote from reference 1, "It has been repeatedly confirmed that the major reactions in polymers, whether produced by fast electrons, x-rays or gamma rays, or mixed radiation including neutrons from atomic reactors, depend primarily on the total energy absorbed and sometimes on the radiation intensity, but rarely, if ever, on the type of radiation or its source. Although exceptions may eventually be found, there is little doubt that this generalization is of considerable value, as it enables the results obtained with one source of radiation to be immediately compared with those obtained under very different radiation conditions." and from reference 3 "In gases, liquids, and covalently bonded solids the chemical effects of ionizing radiations can be ascribed almost entirely to ionization, excitation and dissociation of molecules."

Of particular concern is the change in physical properties of the materials involved. Let us first understand that all materials are affected to some degree by radiation. Metals are the most radiation resistant of materials, ceramics and other inorganics intermediate while organics are most affected. It may be noted, however, that many organics can absorb more than 10^7 rads without significant deterioration. In contrast, however, 500 to 1000 rads is sufficient to cause death of most mammals, including humans. For organic materials the dose (quantity of absorbed radiation) has been determined to be the principle damaging effect with the energy absorbed from different types of radiation approximately additive. The principal mechanism of organic material degradation is the ionization induced by radiation which alters the chemical structure. In metals a mass particle displacement effect becomes significant when collisions occur in the atomic structure. Through mutation within the nucleus the atomic weight of the metal increases along with the metal becoming radioactive and itself a source of gamma radiation.[2, 4] Changes in the crystal structure will also occur due to atomic displacement, however, this is usually accompanied by a temperature annealing process. As a result, only minor changes are known to occur at room temperatures in tensile strength, impact strength, hardness and volume resistivity.[13, 14] It is also noted that a small amount of metallic shielding will reduce to almost insignificance, most proton or neutron particle radiation effects on components.[12] Although all types of radiation cause ionization, displacements come only from the more energetic types. Electrons and gamma rays, for example, usually produce only a small number of primary displacements. With neutrons and heavy charged particles enough energy is delivered in atomic collisions that primary displacements cause secondary displacements, and in fact thousands of cascade displacements can occur. It is important to realize, however, that ionization and displacement effects cannot always be clearly separated. In both organic and inorganic insulations ionization effects can persist so

that it is hard to draw a clear distinction between permanent and transient effects. In alkali halides ionization produced by x-ray or gamma rays results in a persistent coloration caused by electrons trapped in negative ion vacancies.[20]

Radiation Units and Comparative Effects

The basic unit of radiation is the Roentgen which is defined by the ionizing effect in air. One Roentgen of exposure will cause 83.3 ergs to be absorbed per gram of air. One rad is equivalent to 100 ergs/gm of gamma radiation. One Roentgen/hour of gamma radiation exposure will result in approximately one rad/hour of absorbed energy. On the other hand 10 Roentgens/hour of fast neutrons are required to effect 1 rad/hour of absorbed radiation energy.

The energy level of fast neutrons is greater than 1 Mev, therefore, it is on a par with the Cobalt-60 gamma source which radiates at an energy level of approximately 1.2 Mev. The higher absorbed energy effect of gamma radiation may be attributed to its lack of mass. In comparison, a 1-3/4" lead shield will reduce the energy level of gamma radiation by 90% and fast neutrons by almost 100%, while 3" of wax will not measurably shield from the gamma radiation but will reduce fast neutron radiation energy by 64%.[4] For most insulators it requires approximately 10^8 fast neutrons/cm^2 to produce 1 rad of absorbed energy, and about 100 times as many slower neutrons/cm^2 (thermal and epithermal).[8] It thus develops that consideration of employing a live reactor versus radioactive gamma source produces no apparent advantages for composite organic materials. In fact a recent paper describing tests on capacitors using a nuclear reactor defines the effects in terms of gamma radiation.[17] The principal effect on metals, of the live reactor, is that of causing them to become radioactive. However, this generally does not result in a detriment for its structural applications.[13] In the case of semiconductor materials, a neutron source has peculiar usefulness since gamma radiation creates the dynamic effects of change in leakage currents while neutron irradiation changes its front to back ratio by decreasing the minority carrier lifetime.[18] Since gamma radiation is electromagnetic in nature, it will produce dynamic effects in materials, particularly, those that are used to produce electromagnetism. Thus the gamma source is suitable for inducing high levels of radioactive energy in components of composite metals and organics to cause dynamic and passive degradations.

Van Allen Radiation Belts

For correlation of simulation techniques to that environment present in the Van Allen Radiation belts, it would be well to understand a bit more about them. (Figure 1) Let us first note that most of the data on what these belts are consistent of is still poorly substantiated, particularly, as far as the energy spectra is concerned. It is known that the radiation consists of high energy charged particles trapped in a dipole like geomagnetic field. The motion is properly called cyclotron motion. The energy consists of two components, an electron component and a proton component. The knowledge of the electron component is stated in the literature to be highly tentative, with major areas of uncertainty and controversy indicated. Reviewing a wide variety of reports from numerous space flights with measurements made by energy level scintillation counters (sodium iodide crystal detectors) one may approximate that a flux rate of about 10^8 electrons/cm^2 sec above 20 kev is consistent with most experiments.[5, 10, 11, 12] However, as pointed out by Hess (1960), "Making a quantitative deduction from a phenomenon caused by an unknown number of unidentified particles of uncertain energies is a questionable procedure."[5] The proton flux values are much better known and high intensity belt region measurements are recorded to be approximately 3 x 10^4 protons/cm^2 sec for energy levels of about 40 Mev.[5] Thus, in terms of establishing a proper simulation technique, it is well to consider and synthesize the foregoing radiation technology. We may then state the following as our major determinents:

1. The principal mechanism of radiation deterioration is ionization which is proportional to the total quantity of all absorbed radiation energy, particularly when organic materials are involved.

2. A gamma radiation source can effect a maximum of absorbed energy in a material within a minimum time period when compared to other radiation sources of like energy levels.

3. It takes about 10^8 fast neutrons/cm^2 (energy level about 1 Mev) to effect an equivalent absorbed dosage of 1 rad of energy.[8]

4. High intensity Van Allen electron

radiation belt flux is about 10^8 electrons/cm^2 sec at an energy level of about 20 Kev and proton high intensity flux of 3×10^4 protons/cm^2 sec at an energy level about 40 Mev.[5]

5. The exact nature and intensity distribution of the Van Allen radiation belt must yet be established.

From 1 and 2 we may conclude that a gamma source would be most desirable as our radiation medium. From 3, 4, and 5 that we may establish conservative equivalents in lieu of exact simulation.

Consider that the energy absorption may be calculated from the equation:[16]

$$E_a = (\phi_\gamma \bar{C}_\gamma + \phi_n \bar{C}_n)\, t$$

where

E_a is the energy absorbed in ergs/gm or rads

ϕ is the radiation flux in gammas or neutrons/cm^2 - sec

C is the energy absorption coefficient in $\frac{\text{ergs/gm}}{\text{gammas or neutrons/cm}^2\text{-sec}}$

t is the time in seconds

Thus, if we conservatively consider that energy from fast neutron radiation has as great or greater ionization effect as Van Allen belt electrons, we conclude that it effects an absorbed dose of about 60 rads/hour in the high intensity region. Considering the higher energy level of the proton radiation belt and assigning a factor of 40 times the fast neutron in its ionization effect (on the basis of a linear relationship) we arrive, at a dose rate of about 40 rads/hour. Thus, our total absorbed dose rate should be approximately 100 rads per hour on even more conservatively, an exposure rate of 100 Roentgens/hour of 1. 2 Mev gamma radiation. We have thus arrived at a most, most conservative basis for simulating the ionization deterioration effect of an equivalent high intensity Van Allen radiation belt dose rate. We note in comparison that statements in the literature quote 10 to 200 rads/hour for high intensity Van Allen belt unshielded dosage.[12, 19]

Facilities and Staff

The Kilo-Curie Cobalt-60 Irradiation Facility permits simultaneous exposure of hundreds of components. (Figure 2) Variable sample positioning permits exposure to intensities of up to 6×10^5 Roentgens per hour. (Figure 3) The ionization energy of the Cobalt-60 source duplicates in hours the equivalent ionization energy of years in the Van Allen belts. There is no residual radioactivity in the components after exposure to Cobalt-60. Exposed samples may be handled, tested, and shipped without fear of radiation contamination. Sample containers over a right circular volume of up to 3. 5 inches diameter and 18 inches long may be evacuated down to 10^{-5} mm Hg pressure with provision for electrical leads. (Figure 4) Suitable measuring and recording equipment is used for functional parameter measurement before, during, and after radiation exposure. (Figure 5)

The technical staff involved in this program includes personnel experienced in the fields of radiation, electronics, reliability, mathematical analyses and space environmental studies. The work reflected herein was done at Aerojet-General, Azusa.

Metal Film Resistor Dose Effects

Stability of Resistor types RN65C, Metal Film, 1K and 750K ±1% and RN65B, Deposited Carbon, 1K ±1%.

As can be seen in Figures 6, 8, and 9, changes in resistance values after initial measurement of the above resistor types appear to be of a random nature, not significantly affected by either total vacuum-radiation exposure or exposure rate. The random nature of the variations indicate that the values vary within the accuracy of the techniques and equipment used. In no case did the measured resistance value deviate (after exposure) as much as 0. 1 percent from initial value nor as much as . 3 percent in weight, for any group of resistors.

Deposited Carbon Resistor Dose Effects

Resistance changes for Resistor type RN65B, Deposited Carbon, 1M ±1% were substantially greater than all others. Figure 7 shows the resistance increased in value with an increase in radiation exposure until the samples were exposed to a total radiation of 3.36×10^7 Roentgens. The average change in resistance was noted to be nearly a linear function of total radiation. How-

ever, the resistors appeared to stabilize so that no further increase in resistance was noted subsequently. The last sample was removed from the radiation environment after 7.2×10^7 Roentgens of radiation exposure.

Figure 10 presents histograms of the 26 resistors of the above type which were exposed to the maximum radiation (7.2×10^7 Roentgens). The shift in mean value of the resistances is shown as well as the distribution of resistance values before and after radiation. Values of standard deviation were calculated based on the assumption of a normal distribution of resistance values and the 3s limits are shown in Figure 10. For a normal distribution of values, 99.73 percent of the samples will fall within the $\pm 3s$ limits.

Figure 11 is a histogram showing the distribution of mean value and resistance changes with the 3s limits shown, based on a normal distribution of values.

Changes in weight for this type of resistor were as negligible as for the other types, averaging about 1/1,000 of a gram loss, which is at the accuracy limit of the weighing equipment.

Dynamic/Transient Response of Resistance Due to Insertion in a Vacuum-Radiation Environment

Six resistors were monitored using a bridge network and a six channel Offner Recorder which would detect a change in resistance of as low as 0.2 percent. There was no detectable change in resistance of any of the monitored resistors at either the time of insertion in the radiation field or at any other time during the 135 hours of continuous monitoring. The radiation exposure intensity during this portion of the test was approximately 102 Roentgens/hour or the estimated equivalent of the maximum radiation intensity in the inner or outer Van Allen radiation belt.[12, 19] Subsequent to exposure, the vacuum level of the container was checked and found to be at 5×10^{-4} mm Hg.

Method of Analysis

Resistor values obtained before and after radiation exposure were first separated into groups according to resistor type, exposure rate and total exposure. Changes in resistance and weight were then calculated for individual resistors. A comparison of changes within each group was then made to determine if the values were in the same population, i.e., were all of the same order of magnitude. Three recorded values of weight (noted in the data) were found to differ by an order of magnitude from previous and subsequent measurements, were assumed to be in error and therefore, not included in the averages of sample effects.

Average values and percentages of resistance and weight changes were calculated and charted for all groups of resistors. Only one type of resistor (RN65B, 1M $\pm 1\%$) appeared to be significantly affected by radiation exposure.

Further analysis was made of the RN65B, 1M $\pm 1\%$ resistor group which received the maximum radiation exposure of 7.2×10^7 Roentgens. Histograms were plotted of the resistor values before and after exposure and of the calculated resistance change. Standard statistical methods were used to calculate the mean values and standard deviations of the sample based on normal distribution of value and to determine the 3s limits.

The analysis of the monitored instrumented capsule of resistors consisted of examination of over 100 feet of paper. This disclosed no significant pen movement in the recorder which would indicate changes in resistance values at the time of insertion of the resistors into the radiation field, nor at any subsequent time during the test.

References

1. J. J. Harwood, et al, Effects of Radiation on Materials, P. 262, Reinhold Publishing, 1958

2. R. D. Evans, The Atomic Nucleus, P. 232-233, McGraw-Hill, 1955

3. Friedlander and Kennedy, Nuclear and Radio Chemistry, P. 212, John Wiley, 1956

4. S. Glasstone, Principles of Nuclear Reactor Engineering, D. Van Nostrand, 1956

5. F. S. Johnson, Satellite Environment Handbook, Stanford University Press, 1961

6. E. R. Pfaff, The Effects of Nuclear Radiation on Electronic Components, WADC Technical Report 57-261, Vol. IV, 1960

7. Hamman, et al, The Effect of Nuclear Radiation on Electronic Components, REIC Report No. 18, 1961

8. D. S. Toomb, Selecting Insulators and Elastomers for a Radioactive Environment, Aerojet-General Nucleonics, I. S. A. Conference, 1960, Preprint No. 40-SF60

9. Irradiation Effects of 22 and 240 Mev Protons on Several Transistors and Solar Cells, NASA TN D-118

10. J. A. Van Allan, J. Geophysics Research, 1959, 64, 1683

11. E. Helde Kallmon, Space Research, Proceedings of the First Int'l Space Science Symposium, Nice, North-Holland Publishing Co., 1960

12. N. K. Ganguly & J. T. Lence, Shielding Manned Space Vehicles from Space Radiations, Journal of British Interplanetary Society, May, 1961, pp. 111, 112, 113

13. A. Boltax, Effects of Radiation Environment on Structural Materials, Electrical Manufacturing, June, 1958

14. Billington and Crawford, Radiation Damage in Solids, Princeton University Press, 1961

15. J. J. Taylor, Application of Gamma Ray Buildup Data to Shield Design, WAPD-RM-217-1954

16. R. R. Paddock, A Reliability Analysis of the Effects of Nuclear Radiation on the Electrical Properties of Capacitors, General Dynamics Corporation, Fort Worth, Texas, IRE Transactions, PGRQC, April, 1958

17. Wicklein & Dickhart (Boeing-Seattle), Transient Radiation Effects in Capacitors & Dielectric Materials, 1961, WESCON 20/3.

18. Gardner and Kaufman (Litton-Woodland Hills, California) Semiconductors in a Hyper-Nuclear Environment, 1961, WESCON, 20/2.

19. H. S. Appleman (MATS Air Weather Service, USAF), Radiation Effects on Manned Space Flights, Journal of Environmental Sciences, December, 1961

20. J. W. Clark and T. D. Hansrome (Hughes Aircraft), Radiation Pulse, Nucleonics, September, 1960

21. H. L. Wiser (Hughes Aircraft), Radiation Problems with Electric Components, AIEE Aero-Space Conference, Philadelphia, Pa., June, 1961

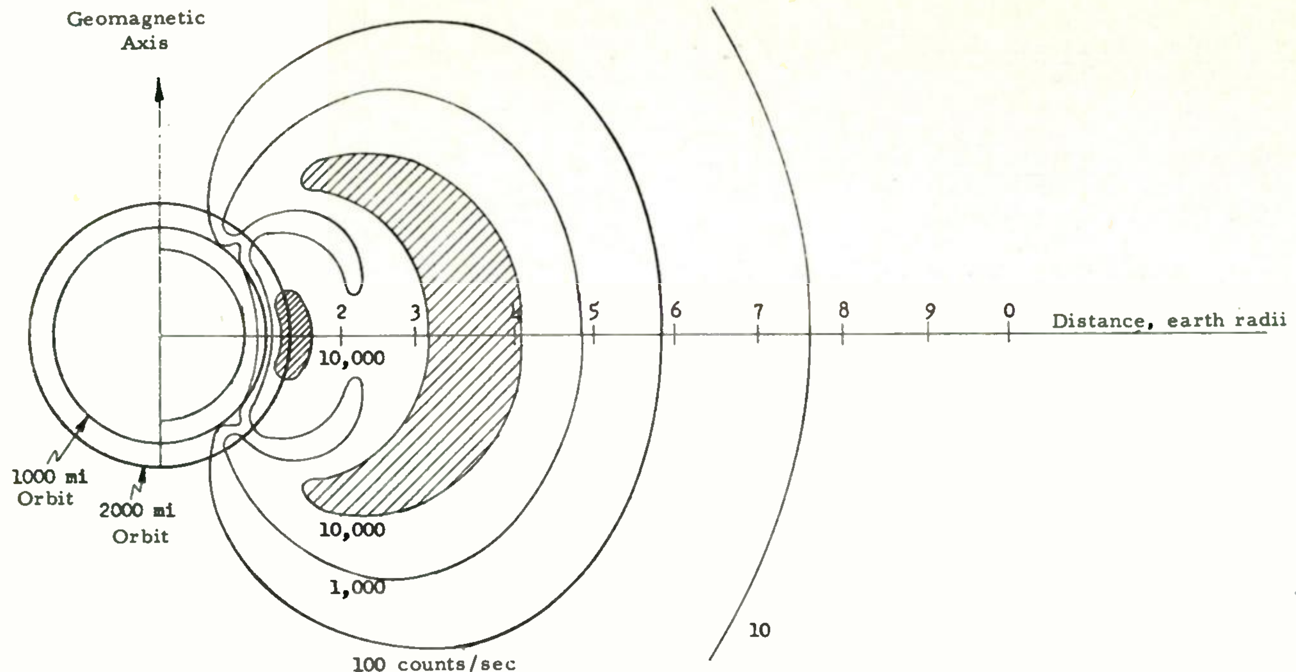


Figure 1 Suggested Contour Lines for Van Allen Radiation Belts Based on Explorer and Pioneer III Data with Trajectory of Pioneer III
J. A. Van Allen and L. A. Frank

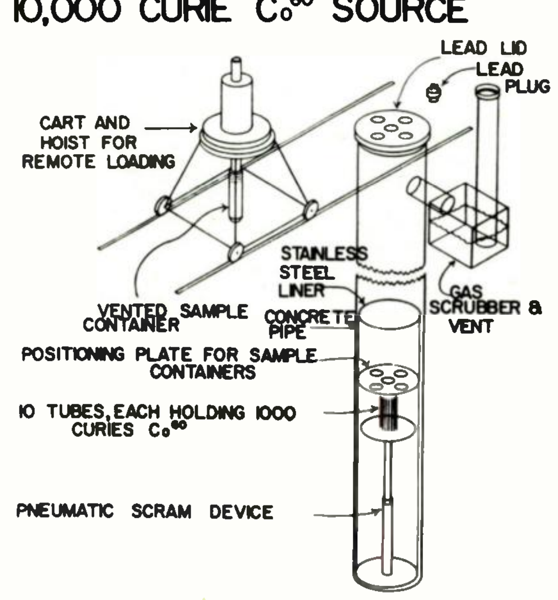
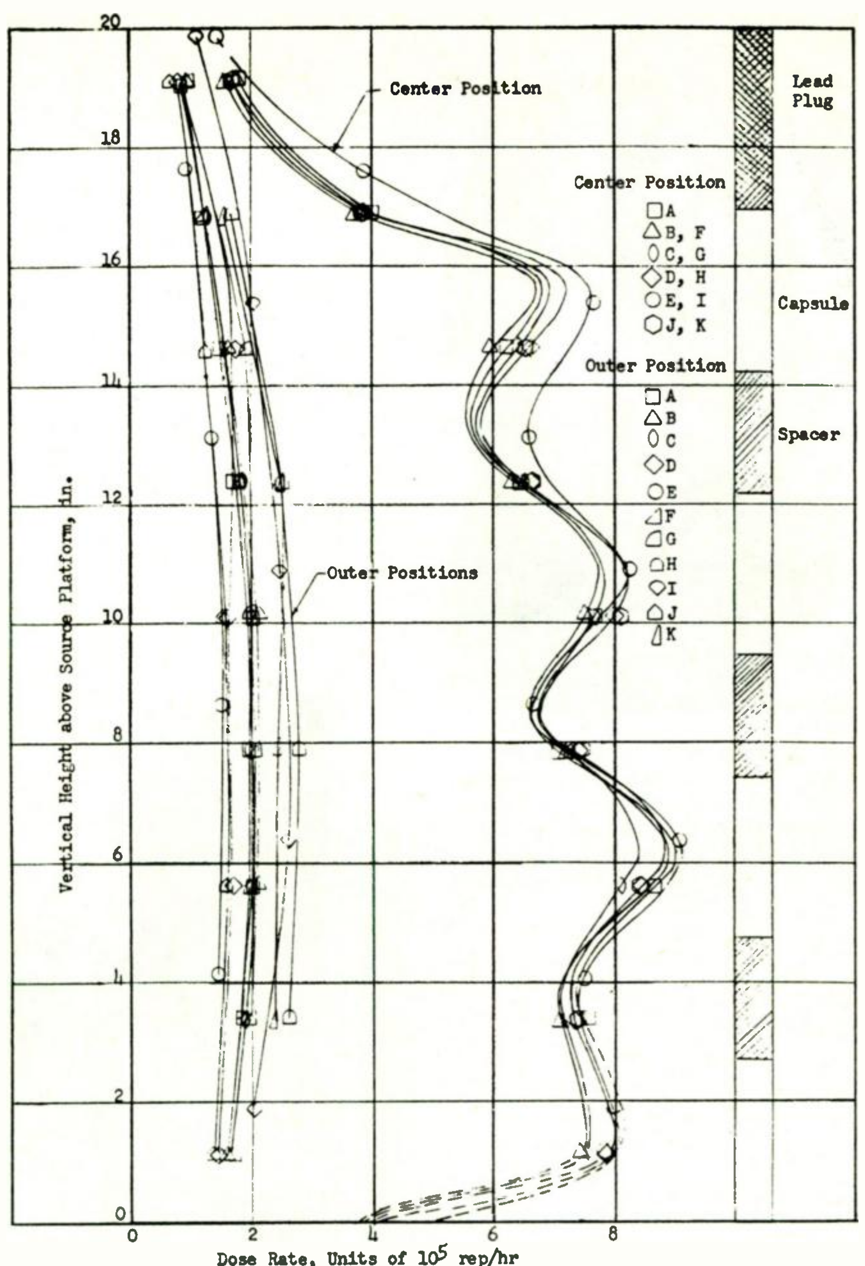


Figure 2



VARIATION OF DOSE RATE WITH LOCATION IN 10,000 CURIE COBALT-60 SOURCE

Figure 3

Fig. 4. Dr. Lazier, Vice President Engineering, Sprague Electric, examines container pinched and soldered vacuum seal prior to placing container in larger capsule. Messrs. Vilter, Brown and Doshay of Aerojet facing the container with Sheldon Howard of Sprague on Dr. Lazier's right.

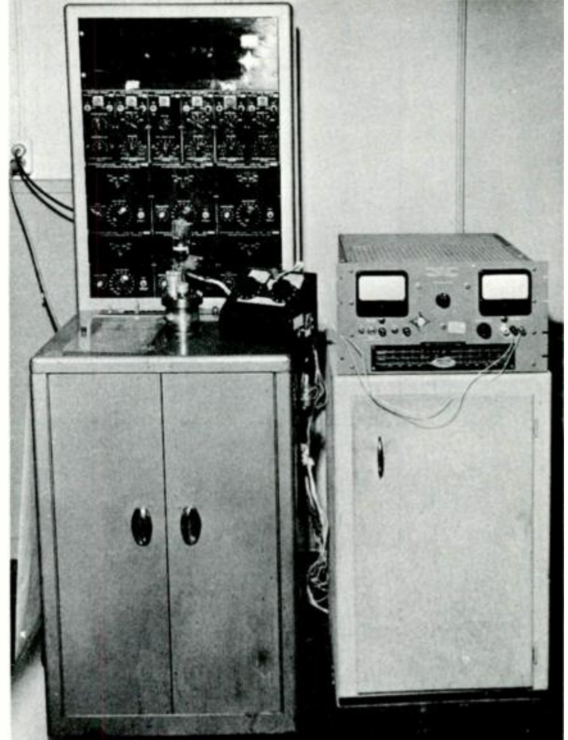


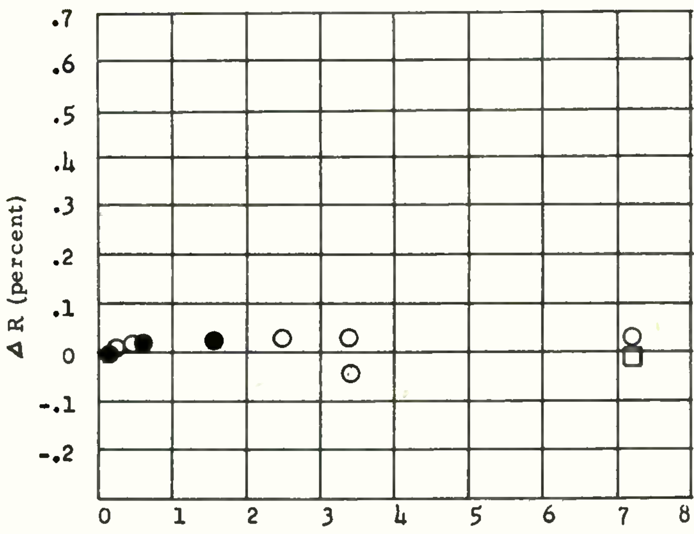
Fig. 5. Instrumented container rests on top of Offner recorder during instrumentation checkout prior to moving equipment to radiation facility for dynamic test.

Legend

● - Dosage rate = 1.5 X 10^5 Roentgen/hour

○ - Dosage rate = 6.0 X 10^5 Roentgen/hour

□ - Sample measured at 95° F. (all other samples measured at 72 ±1° F.)



Total Radiation Exposure (10^7 Roentgens)

Figure 6

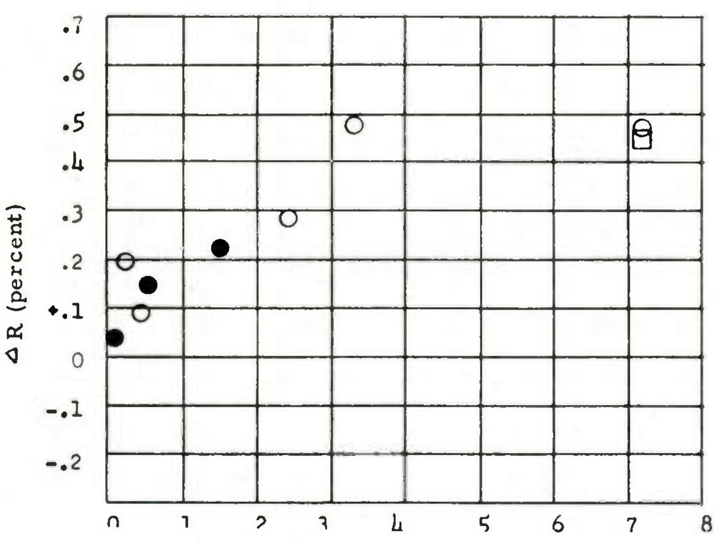
Resistance Measurement Deviation from Origianl Values of Resistor Type RN65B, 1KΩ ±1%, as a Function of Radiation.

Legend

● - Dosage rate = 1.5 X 10^5 Roentgen/hour

○ - Dosage rate = 6.0 X 10^5 Roentgen/hour

□ - Sample measured at 95° F. (all other samples measured at 72 ± 1° F.)



Total Radiation Exposure (10^7 Roentgens)

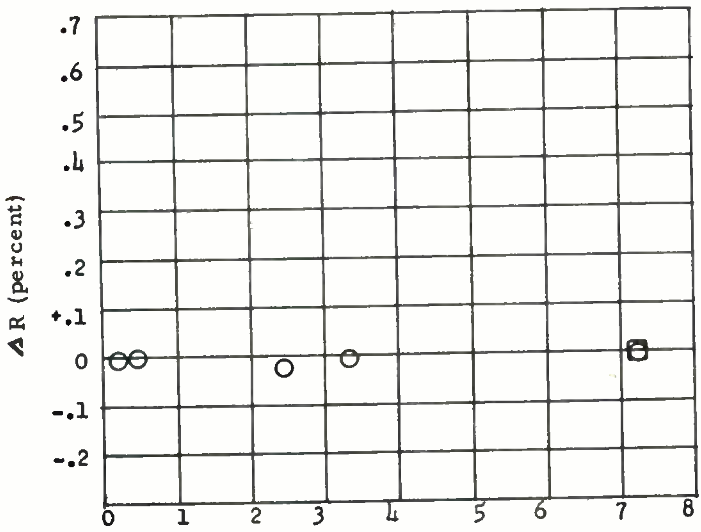
Figure 7

Resistance Measurement Deviation from Original Values of Resistor Type RN65B, 1 MΩ ±1%, as a Function of Radiation.

Legend

○ - Dosage rate =6. 0 X 10^5 Roentgen/hour

□ - Sample measured at 95°F. (all other samples measured at 72° ± 1°F.)



Total Radiation Exposure (10^7 Roentgens)

Figure 8

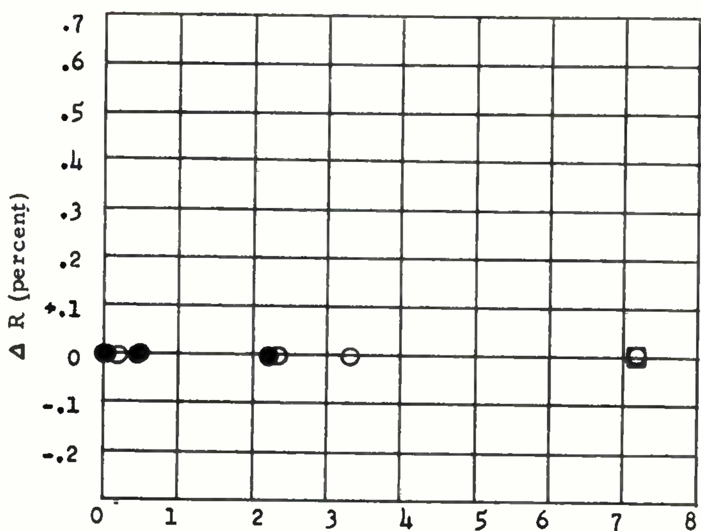
Resistance Measurement Deviation from Original Values of Resistor Type RN65C, 1KΩ±1%, as a Function of Radiation.

Legend

● - Dosage rate = 1.5 X 10^5 Roentgen/hour

○ - Dosage rate = 6.0 X 10^5 Roentgen/hour

□ - Sample measured at 95°F. (all other samples measured at 72 ±1°F.)



Total Radiation Exposure (10^7 Roentgens)

Figure 9

Resistance Measurement Deviation from Original Values of Resistor Type RN65C, 750KΩ ±1%, as a Function of Radiation.

Sprague Resistor Type RN65B
Deposited Carbon, 1M ±1%
Sample Size - 26
$\bar{x}$ = Mean Value of Resistance
s = Standard Deviation

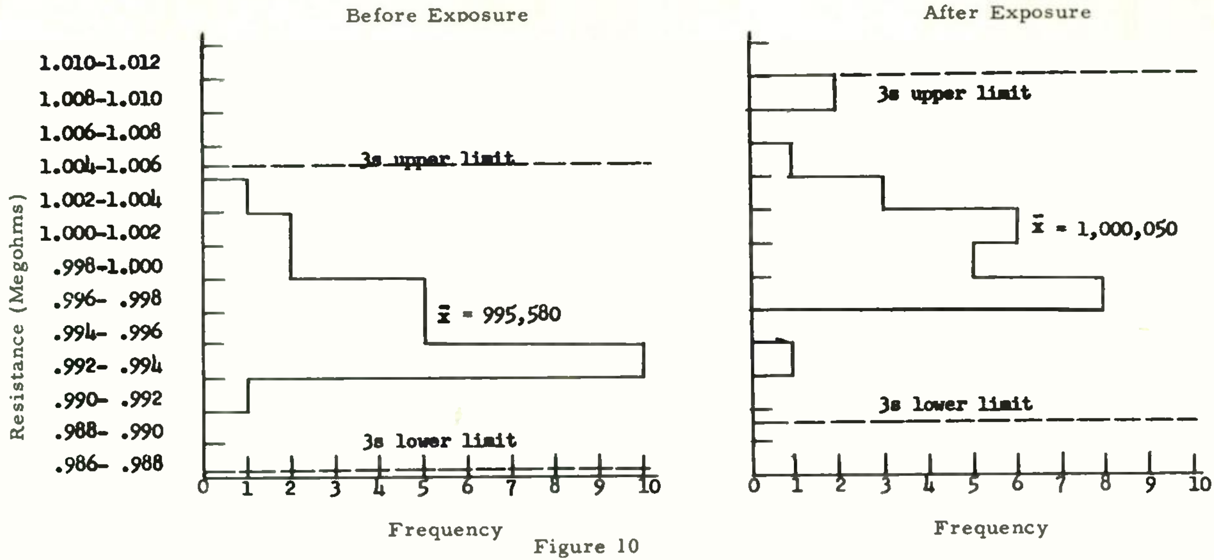


Figure 10

Histograms of Resistance Values Before and After Exposure to 7.2 X 10^7 Roentgens of γ Radiation

Sprague Resistor Type RN65B
Deposited Carbon, 1M ±1%
Sample Size - 26

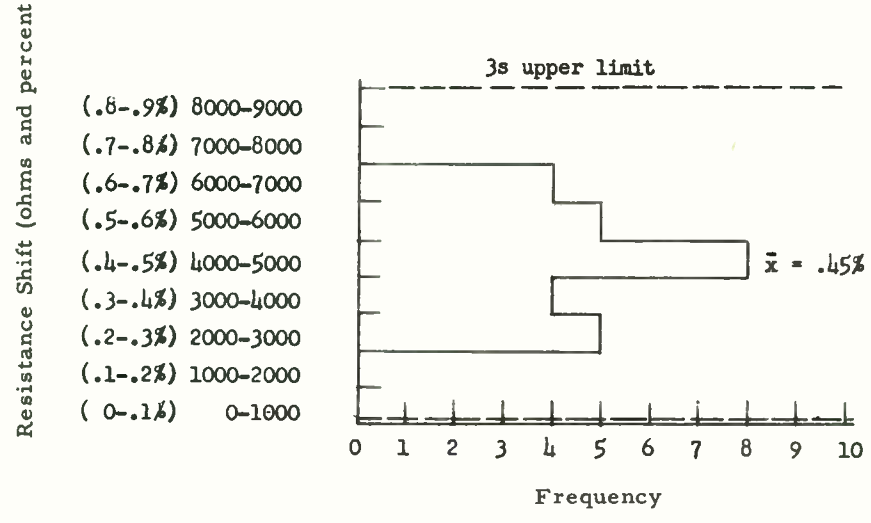


Figure 11

Histogram of Permanent Resistance Increases Caused by Exposure to 7.2 X 10^7 Roentgens of γ Radiation.

THE EFFECT OF RADIATION ENVIRONMENT ON FILM RESISTORS

Stanley O. Dorst and Leonard H. Wurzel
Sprague Electric Company
Nashua N. H.

Introduction

In the preceding paper Irving Doshay reported that a facility has been devised for providing a simulated space environment in which to evaluate electronic components. We have taken advantage of that facility to test the suitability of Sprague precision film resistors for use in satellite equipment. Mr. Doshay has related some results of this project in regard to the resistance stability and the weight loss resulting from exposure to this environment.

This paper proposes to explore the effects on other properties of the resistors. Some of these may be important to the proper functioning of the resistor in the satellite equipment. Further, we shall investigate any less significant degradation by stressing the samples physically and electrically to determine where changes have occurred in comparison with unexposed controls.

Selection Of Test Resistors

When the test plan was formulated, little was known concerning the effect on a resistor of the combined exposure to radiation, vacuum, and the thermal and electrical conditions of operation. In fact, it was conjectured that a resistor with molded enclosure might be entirely unsuitable for such conditions of operation.

The design of the experiment, therefore, required a survey of a wide range of dose rate, and time in the environment. This necessitated a sufficient number of resistors to make frequent withdrawals of test resistors for the final evaluation. Ten combinations of dosage rate and time were used, giving a range of total dosage from 1.05×10^6 to 7.2×10^7 Roentgens gamma radiation. Additionally, 12 resistors were operated at rated wattage for 135 hours for a total dosage of 1.35×10^4 Roentgens. The former range of dosage, according to Mr. Doshay's estimate, represents an accelerated testing equivalent to many years of Van Allen Belt radiation. The latter experiment, in which the resistors were operated at rating, is equivalent to the estimate of actual dosage conditions to be encountered (100 Roentgens/hour) in satellite use.

Since we wished to include two types of film resistors and cover the resistance range, the survey mentioned above and the equipment limitation of 324 resistors restricted the experiment to the following selection:

Test Resistors

MIL Designation	Resistance Value	Sprague Type	Resistance Element
RN65B	1KΩ	406E	Pyrolytic carbon
RN65B	1MΩ	406E	Pyrolytic carbon
RN65C	1KΩ	419E	Metal Film
RN65C	750KΩ	419E	Metal Film

These resistors were taken from manufactured stock. After suitably tagging, they were weighed and measured for resistance. Part of the group was forwarded to the space radiation facility, and the remaining were retained for controls.

Resistor Construction

Any consideration of the effects of environment on a component involves the materials of which it is constructed. Generically, the resistors came under the classification of accurate film resistors with molded enclosure. The resistance element is a very thin film of conducting material possessing a high degree of stability. They are manufactured to a precise resistance value, 0.1% to 1.0%, and their success in service depends on the reliability with which they maintain this value.

The MIL Type RN65B, as embodied in the Sprague 406E, has a pyrolytically deposited carbon film element. This is produced at high temperature, about 1800 F, by the cracking of methane gas on special electrical porcelain rods [1]. Termination is made to the resistance element using bands of highly conducting silver paint with organic binder. Silver-plated brass end-caps with attached lead wires are pressed onto the core over the silver bands. The desired resistance value is obtained accurately by helically cutting groove through the film[2]. Two coatings are applied for protection of the film. The first is a layer of mica-loaded silicone, and the second is a layer of filled epoxy resin. The final, over-all encapsulation on the resistor body is a molded diallyl-phthalate compound with fiberglass filler for strength. The diallyl-phthalate is of the ortho structure. In this construction a resistance range of 10 ohms to 2.49 megohms is obtained in a size of 0.600" length by 0.200" diameter.

The other type accurate film resistor under consideration is the metal film resistor. This comes under the MIL Type RN65C, with the lower temperature coefficient of resistance and higher stability requirements of characteristic C of MIL-R-10509D[3]. It is represented in our evaluation by the Sprague Type 419E.

In this metallic element film resistor the substrate core material is also the electrical porcelain composition. These cores are metallized with gold bands for electrical termination. The resistance film is a nickel-chromium alloy applied to the substrate by evaporation in a vacuum. Again, the accurate value is obtained by helixing the film. End caps with wire leads complete the termination to the noble metal bands on the element. Coatings and a molded enclosure make up the external protection. But in this case the coatings are silicones and silicone alkyds, and the molding material is diallyl meta-phthalate (also, fiberglass filled). The RN65C type affords higher temperature resistor performance.

A cross-section of a typical film resistor is depicted in Figure 1.

Consideration of Materials

The material constitution of these resistors (metals, ceramics, and plastics) is subjected in the simulated space environment to the combination of high vacuum (pressure lower than 10^{-5}mm), radiation, and the heat and electrical potential of resistor operation. Reviewing the list of materials in the above described construction we have the following:

Table 1

Metals and Alloys	Ceramics	Plastics
Brass	Electrical porcelain	Silicone
Gold	Fiberglass	Silicone alkyd
Silver	Mica	Diallyl phthalate
Nickel	Inorganic pigment	Epoxy
Chromium	Pyrolytic carbon	
Copper) lead wires		
Steel) lead wires		
Tin) lead wires		
Lead) lead wires		
Trace metals		

The changes in these materials due to radiation or vacuum environments and the possible damage to a resistor may be determined in a general way by a study of the literature. Table II is an outline summary of this information with opinions on how these changes could affect a film resistor.

Table II

Expected Effects of Vacuum or Intense Radiation

Material	Effects	Possible Damage to the Resistor
Metal	physical property changes	mechanical damage
	electrical conductivity change	resistance change
	Sublimination	resistance change change in temperature coefficient of resistance
Ceramic	induced conductivity	shunt resistance
	discoloration	none
Carbon	decreased conductivity	resistance change
	absorption of gas	resistance change
Plastics	breakdown of bonds & linkages	physical damage to the enclosure, gassing
	induced conductivity	shunt resistance
	vaporization of plasticizers	physical damage to the enclosure
Thin films	atomic displacement	resistance change

The metals and alloys would not be expected to deteriorate in this environment. The radiation dosage, even in our accelerated testing, is not sufficient to result in mechanical damage to the resistor. While at higher levels, mechanical strength and electrical resistivity are increased by radiation [4], in space environment the properties of metals do not appear to undergo any appreciable change[5]. Any sublimation of metals would not be expected to occur at the operation temperature of the resistor (125 to 175° C), and the vacuum of 10^{-5}mm, in our test. The zinc in the brass has the highest vapor pressure of the metals in the resistor. The protective coatings prevent the possibility of traces of this metal subliming and condensing on the resistance film, thus changing its electrical properties.

In ceramics, irradiation damage occurs principally from atomic displacement resulting in changes in density,

thermal conductivity, and electrical conductivity[5][6]. The conductivity of a good insulator, such as the high grade electrical porcelain resistor core, is sufficiently low to rule out any effect the change may have on the resistance value of the resistor in the range we are considering.

The radiation damage in organics results from the breakdown of bonds and linkages to form new compounds. The concentrations of these increase with increase in irradiation[5]. The vacuum environment may cause loss of plasticizers of relatively high vapor pressure, if any are present. The result of these changes would affect the ability of the organic coatings to perform their protective functions. Beyond their normal function of protecting the resistor element from humidity or mechanical damage, they appear to provide protection from radiation effects as well.

In the case of pyrolytic carbon, a study was made by Nakai and Sakakibara which illustrates this[7]. They found substantial changes in resistance in resistors irradiated with dosages corresponding to those in our accelerated environment. The paint coated resistors which they evaluated showed much less change than bare resistance elements, indicating that oxidized gases are absorbed in the unprotected film, causing changes in the resistance value.

Molded film resistors, protected by coatings and a molded enclosure, should be less subject to this effect, providing the coatings themselves are not sources of gas from their own degradation under radiation.

A recent investigation of the effects of intense nuclear radiation on resistors[8] draws the following general comparisons of film resistors. Molded resistors show less sensitivity to radiation than resistors with only a coating for protection. This reference states "the diallyl phthalate molded units seem to resist degradation better than glass-, epoxy- or the ceramic-enclosed resistors". Metal films show less sensitivity to radiation than deposited carbon.

The protective enclosures used in our tests should then be expected to be satisfactory in space environment radiation. In addition to the diallyl phthalate molded enclosure, the associated coatings of the resistor construction are reported to have good retention of their characteristics at levels of radiation above the dosage we are considering. Silicones retain their properties after dosages of 1000 mega-rad of gamma radiation[9]. The effects on epoxy systems are changes in the relative hardness and in dielectric properties[10].

Evaluation of the Results of the Environment on the Resistors

Mr. Doshay has reported the results of resistance stability and weight changes on the test resistors, noting that only one, the RN65B 1 megohm resistor, appeared to be significantly affected by gamma radiation exposure under accelerated conditions of space radiation. The resistors under operation on life-test under normal space simulated conditions showed no significant change.

We now turn to two other characteristics which may affect the resistor's usefulness in normal satellite equipment operation. These are electrical current noise and temperature coefficient of resistance.

Current Noise

Changes in the crystal structure of the film on the substrate, resulting in damage to the film, or an increase in contact resistance to the film, could show up in the current noise of the resistor. This property was of particular interest in our evaluation because of the Nakai and Sakakibara findings[7] that increased dosage produced a random and fairly large variation in noise level.

Our noise evaluation was made on a commercially available noise meter operated according to the method of Conrad[11]. Table III gives the results of the noise test in terms of a noise factor, micro volts/volts in a decade, versus the dosage.

Table III

	Total Dosage Roentgens	Noise Factor, μV/V Average	Min.	Max.
RN65B (Carbon Film)				
1K	0 x	.041	.035	.043
	2.4×10^6	.068	.053	.078
	7.2×10^7	.046	.041	.065
1Meg	0 x	.39	.28	.45
	4.8×10^6	.74	.68	.80
	3.36×10^7	.33	.24	.45
	7.2×10^7	.83	.30	1.60
RN65C (Metal Film)				
1K	0 x	.011	.011	.012
	2.4×10^6	.013	.010	.019
	7.2×10^7	.012	.010	.019
750K	0 x	.060	.012	.140
	2.4×10^6	.022	.012	.026
	7.2×10^7	.052	.010	.125

These test samples did not show the random degree of variation with rise in dosage noted in the literature cited above. In the 1 megohm carbon film group, there was indication of this at the maximum dosage. The metal film resistors have a lower noise level and the exposed samples were within the variation in noise of the unexposed controls. Other than the RN65B 1 megohm, there was no significant effect on the current noise in the film

resistors which could be attributed to the environment.

Temperature Coefficient of Resistance

Table IV gives the temperature coefficient range of the samples measured according to MIL-R-10509D. Listed first is the extreme value of temperature coefficient to the low temperature and second the extreme value from room temperature to the high temperature.

The results show no pattern of change in temperature coefficient of resistance with exposure to the simulated space environment.

Table IV

Type	Resistance Value	Radiation Hrs. at 6 x 10^5 r/hr.	Temperature Coefficient Range
RN65B	1MΩ	0	(-354to-268)ppm/°C
	"	120	(-332 to -267)
RN65B	1KΩ	0	(-246 to -219)
	"	4	(-239 to -213)
	"	40	(-241 to -212)
	"	120	(-249 to -214)
RN65C	750KΩ	0	(- 9 to 35)
	"	40	(- 15 to 8)
	"	56	(11 to 34)
	"	120	(- 18 to 27)
RN65C	1KΩ	0	(- 38 to - 16)
	"	120	(- 37 to - 17)

Physical Tests of the Resistor Materials

To evaluate the physical properties of materials adequately, it is necessary to use prepared samples of dimensions appropriate for the test involved. In the philosophy of testing we are using here, we test the materials in the place, form, and amount that they are used in the resistor. Thus, if one part changes in a physical property it may not effect the overall performance if that particular property is not an important function of the part.

This approach makes it difficult to get a significant measure of the change in these physical properties. A general indication can, nevertheless, be obtained.

Micro-Structure

Figures 2 and 3 show the micro-structure, at X 125 magnification, of an irradiated resistor (dosage 7.2 X 10^7 roentgens) compared with an unexposed control sample. These views are taken in the vicinity of the intermediate layer (the organic coatings) between the filmed ceramic rod and the outer molding. The micro-structure of each resistor is quite similar. A spheroidal phase with occasional concentration of rod shaped particles (fiberglass filler) is uniformly distributed in the molding material. Sporadic porosity is observed at the interface of the molded layer and the intermediate layers. The matrix comprising the intermediate layer contains particles of varying size and regularity. The diamond shaped impressions in this layer were produced during hardness tests.

The fine grained structure of the ceramic substrate is shown in figures 4 and 5. On a linear bases, the ceramic porosity distribution is similar for both specimens.

Hardness Measurements

Hardness measurements were taken in the three distinct regions — ceramic substrate, the intermediate layer, and the outer molded layer. The hardness data are tabulated in Table V.

Table V

Specimen	Diamond Pyramid Hardness		
	Ceramic Core	Intermediate Layer	Outer Layer
Control	819	24.2	51.2
Irradiated Sample (7.2 X 10^7 r.)	823	24.0	51.3

No significant hardness differences are apparent between corresponding regions of the specimen resistors.

Infra-Red Spectrophotometry

Infra-red absorption curves were taken of the molding material from both a control and an irradiated sample. The comparison of these curves showed no evidence of changes of a chemical nature. Any structural changes resulting from the radiation exposure would not be revealed by this technique.

X-ray Diffraction

An attempt was made to determine if there was any change in the molding material with respect to the presence of crystallization before or after irradiation. The patterns developed from the amorphous material were not definitive.

Comparison of Physical Appearance

A visual examination of the control and exposed samples indicated no difference in appearance except for the samples that were operated at their rating in the simulated space environment. The RN65B samples had a slightly darkened area around the longitudinal

center of the resistor. This corresponded to the hottest portion of the resistor surface. The RN65B size is molded of tan colored material and a slight darkening on load is normal for this resistor.

Stressing the Resistors to Locate Physical Changes

The tested resistors were stressed mechanically by temperature cycling and electrically by overload. The first test was to determine if the molding material had become embrittled, and the second to determine any incipient damage to the resistance element due to the accelerated simulation of space environment.

The differential expansion between the ceramic core and the molding material will cause stresses in the molded enclosure. If it is brittle or of reduced strength, cracks may occur as a result of temperature cycling. This has been demonstrated on resistors without sufficient curing of the molding material required to bring it to its full strength. Any embrittlement or loss of strength due to the radiation or vacuum exposure would be expected to have the same effect. This may not show up as visual cracks in the material, but as micro-cracks, or as deterioration in adherence and seal to the leads. A test in humidity can detect such situations.

Figures 6 and 7 illustrate the results of the thermal cycling and moisture resistance tests on the resistance value. These tests were conducted according to specification MIL-R-10509D. No cracks were observed in the test resistors. The performance in terms of change in resistance value shows one significant effect in the RN65B 1 megohm resistors. These appear to have been stabilized by radiation and the resistance changes appreciably decrease with increased dosage.

Electrically stressing by overload, followed by operation on load, was done on test samples to determine any possible damage to the film. If this damage is localized, it occurs without a necessarily large change in resistance value. The localized damage would become evident on overload and long time operation. These tests were run on exposed resistors and on controls following the procedures of MIL-R-10509D.

No noticeable change was found in the samples on overload, and the results of 1000 hour load-life are depicted in Figure 8. In two cases the exposed sample average curve departed substantially from that of the controls. Whether or not the effect of the radiation is involved in the difference is not conclusive.

Other standard MIL tests were run, each of which would have a tendency to bring out substantial differences, if such were present, in the physical properties of the resistor materials. These tests, high frequency vibration, insulation resistance in humidity, and terminal strength, offered no evidence of deterioration due to the exposure.

Conclusion

Sprague molded metal film and pyrolytic carbon resistors are practically unaffected by simulated space environment even where the gamma radiation is increased to the equivalent of many years in the Van Allen radiation belt.

A slight positive resistance change with increased dosage was noted in the 1 megohm pyrolytic carbon resistors in Mr. Doshay's paper. This accomplishes a stabilization which becomes evident in lower positive changes on other testing, such as temperature cycling.

The random variation in noise with increase in radiation dosage, noted in earlier literature, on pyrolytic carbon shows some evidence of being present in the irradiated high resistance carbon samples.

Acknowledgement

The authors acknowledge the assistance of Mr. M. L. Arel in organizing and executing the test program, with the assistance of other members of the Sprague Electric organization.

Bibliography

1. M. D. Rigterink, U.S. Patent 2,386,633 (1945).

2. R. O. Grisdale, A. C. Pfister, and W. Van Roosbroeck, "Pyrolytic Film Resistors", Bell System Technical Journal, Vol. 30, P. 271 (Apr. 1951).

3. Military Specification for Resistors, Fixed, Film (High Stability), MIL-R-10509D.

4. Richard E. Bowman, "How Radiation Affects Engineering Materials", Materials in Design Engineering Manual No. 173, P. 119, July 1960.

5. S. N. Lehr and V. J. Tronolone, "The Space Environment and Its Effect on Materials and Component Parts", Proc. of 1961 Conference on Electronic Components.

6. Arthur Bradley, "Insulation for a Radiation Environment: Part 1 - Inorganics", Insulation, P. 23, Oct. 1961.

7. T. Nakal and T. Sakakibara, "Effects of Gamma-Ray and Neutron Irradiation on the Pyrolytic Carbon Film Resistors", Journal Scientific Research Inst. Tokyo, Vol. 54, P. 88, No. 1 (1960).

8. Hamman, Chapin, Hanke, and Wyler, "The Effect of Nuclear Radiation on Electronic Components",

REIC Report No. 18, June 1961, Battelle Memorial Institute.

9. Javitz and Jacobs, "Electronic Materials and Components for Extreme Environmental Problems", Electrical Manufacturing, Nov. 1958.

10. Charles A. Harper, "Electronic Packaging with Resins," P. 225-229, McGraw-Hill Book Company, New York, N.Y.

11. Conrad, Newman, and Stansbury, "A Recommended Standard Resistor-Noise Test System", IRE Transactions of the Professional Group on Component Parts, Vol. CP-7, No. 3, Sept. 1960.

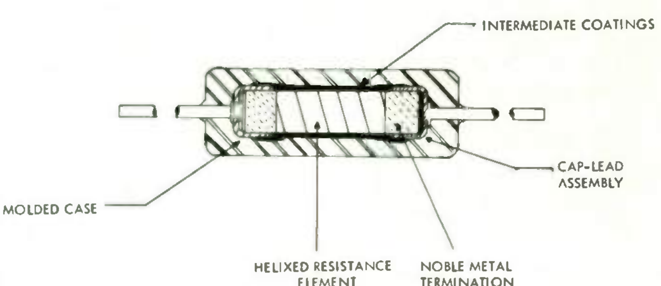


Fig. 1. Film resistor construction.

Fig. 3. Micro-structure resistor cross section. Irradiated resistor. Etched: X 125.

Fig. 2. Micro-structure resistor cross section. Control resistor. Etched: X 125.

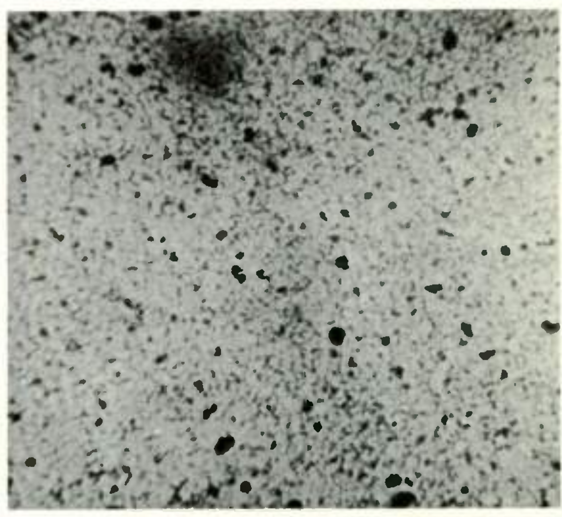


Fig. 4. Micro-structure ceramic core. Control resistor. Etched: X 125.

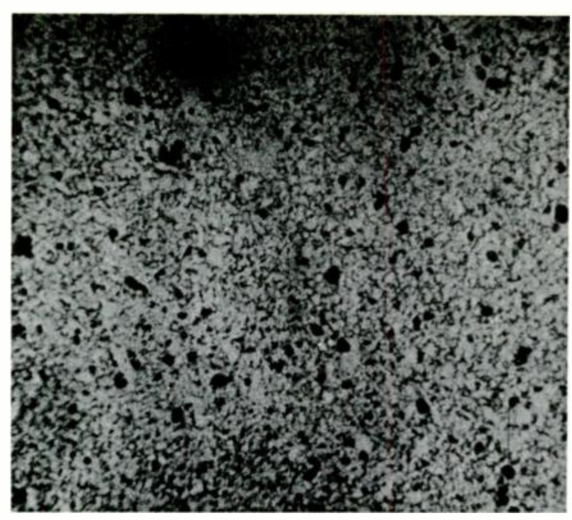


Fig. 5. Micro-structure ceramic core. Irradiated resistor. Etched: X 125.

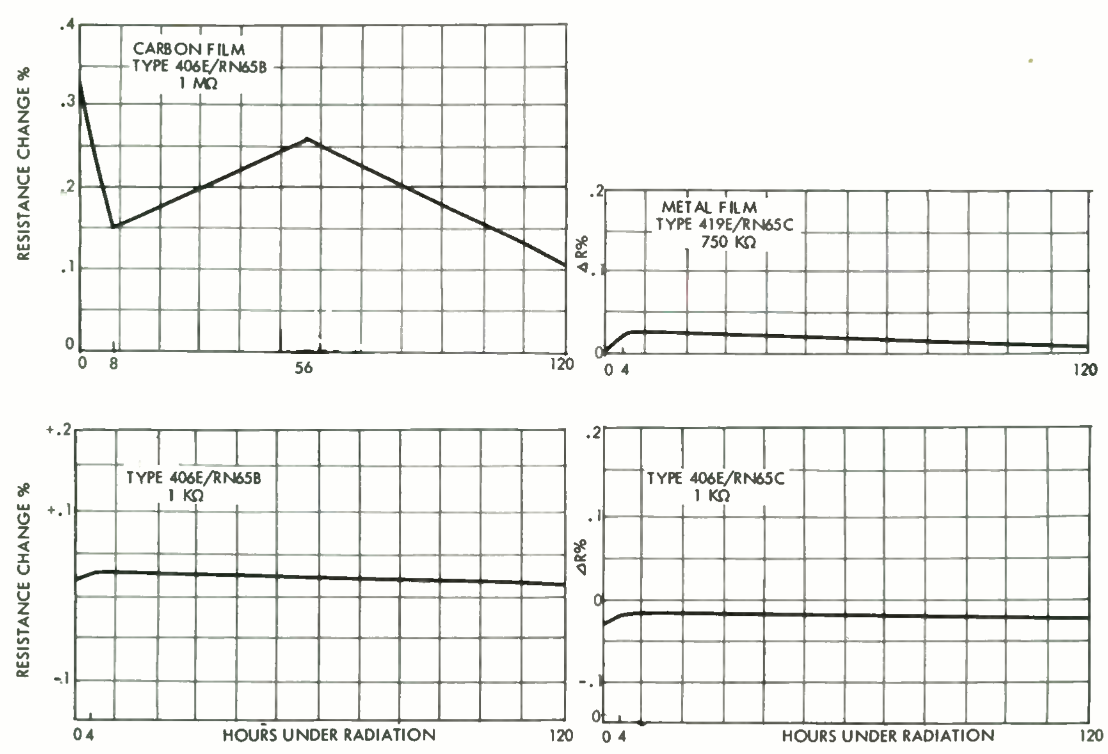


Fig. 6.

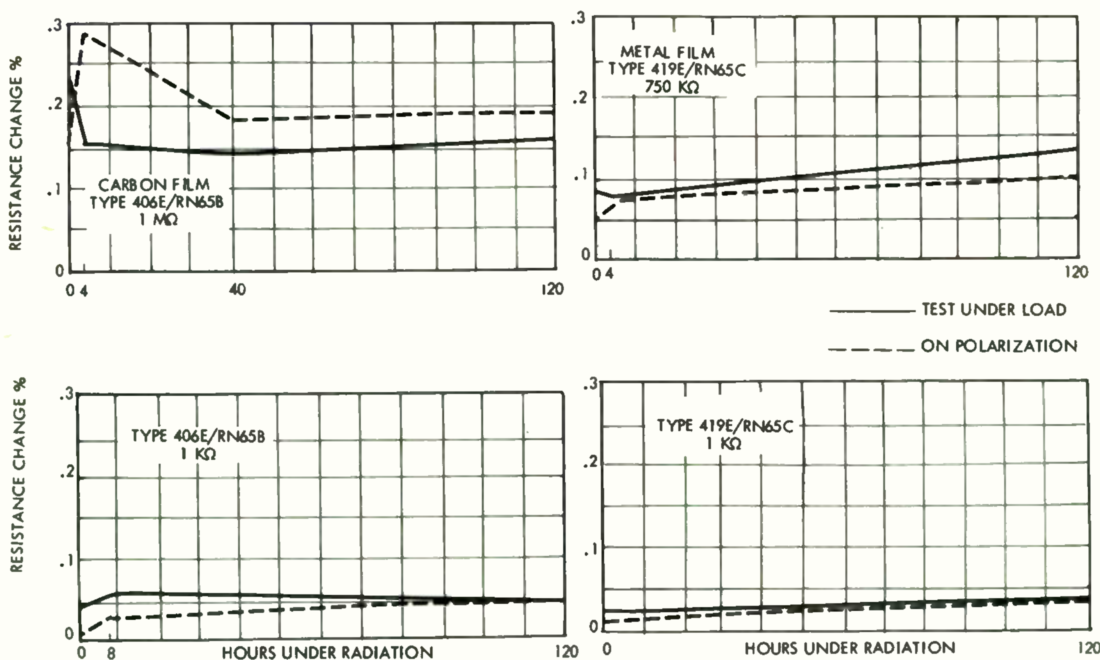


Fig. 7.

IRRADIATED RESISTORS
LOAD-LIFE PERFORMANCE

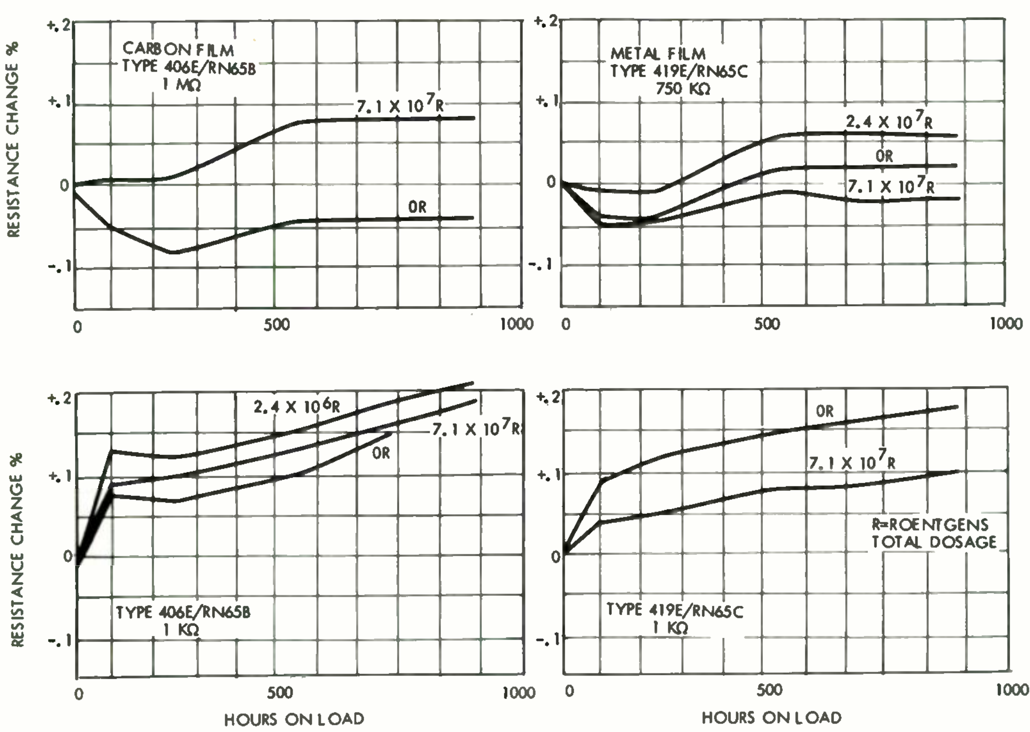


Fig. 8.

RELIABILITY STUDIES OF MESA TRANSISTORS

Frank Aubin

Airborne Instruments Laboratory
A Division of Cutler-Hammer, Inc.
Deer Park, Long Island, New York

Summary

The problem of system reliability has prompted the undertaking of an extensive parts testing program at AIL. Probably the parts most sensitive to environmental extremes in present systems are semiconductors. For this reason, considerable emphasis has been given to testing transistors. Among the types of transistors tested are grown junction, alloy junction, MADT, diffused mesa, and diffused planar. It is the purpose of this paper to present a general picture of the methods employed in the testing of mesa and planar types, and to present in detail the results of some of the tests.

The general philosophy of the test program has been to use sequential stressing or step-stressing to failure In the first case, a representative sample is subjected to the stresses of shock, vibration, thermal cycling, humidity, and life testing, usually in that order. Characteristics are measured after exposure to each of the environments, and the resulting data analyzed with a view to prediction of characteristics variations. In the second case, the sample is subjected to repetitive stresses, either mechanical or electrical, of increasing severity, until failure occurs. The step-stress method yields data which indicates the strength of a part type and the available safety margins. The results may also be used to determine what, if any, selection procedures are necessary when purchasing particular transistor types to eliminate weak units and potential failures.

The most common mode of failure observed in these studies was that of failure of thermocompression bonds. Failures of this type are generally caused by inadequate control of this portion of the manufacturing procedure. Nearly all semiconductor manufacturers have experienced this trouble at one time or another, and many still experience it periodically. This is in spite of the fact that most of them have learned what troubles were present in the assembly techniques and have taken steps to correct them. Transistors that fail in this mode or other modes are generally not of high density in the end products. Though a given manufacturer's product may pass the requirements of military specifications on a lot-by-lot basis, units which are potential failures may still be delivered. Because the few potential failures cannot be tolerated under a high reliability requirement, techniques for culling out the worst of these have been developed.

The functioning of a space system is dependent upon more than the elimination of parts that are potential catastrophic failures. It is dependent also upon the proper functioning of the many parts that make up the system after these parts have been subjected to severe mechanical environments and to more-or-less severe electrical environments. If the circuits in which the parts are used do not tolerate the variations in part characteristics resulting from environmental exposure, the system can fail to perform its intended function. For this reason, the part characteristics important to circuit design are observed throughout testing, and design tolerance limits are established, giving the designers information that will allow them to consider the worst possible combinations in their circuit design.

The utilization of the information generated in this testing program has resulted in the development of several highly reliable space systems, such as Topside Sounder.

I. Introduction

The reliability of any electronic system is, in the last analysis, dependent upon the reliability of the parts used in its construction, and upon the method of parts application. To ensure the design and manufacture of systems with the highest attainable reliability, several kinds of information are necessary. The design engineer must know, with a high degree of confidence, the variations that will occur in part characteristics as a result of the environmental conditions experienced when parts are installed in operational equipment. By knowing the extremes between which a part characteristic will vary, it is frequently possible to design circuits that will tolerate these variations or in cases where this is not possible, to incorporate compensating elements to correct variations, permitting circuits to function properly.

The packaging engineer must know which environments--ambient, mechanical and/or electrical--will cause catastrophic damage to the parts. With this knowledge, parts that are sensitive to the system environments can be isolated by shock or vibration mounting, heat sinking, pressurizing, or other means.

The production engineer must know the precautions to be taken in storing and handling parts for a system to eliminate possible operational

failures induced by improper handling during manufacture, as well as minimizing rework.

To predict system reliability, system effectiveness, and other important parameters, the reliability engineer must have information about part failure rates resulting from mechanical environments (shock, vibration, temperature, etc.), and hazard rates. With this information available to the specialist engineering groups, it is possible to design and fabricate a system possessing a high and predictable reliability. At Airborne Instruments Laboratory (AIL), one of the functions of the Reliability Test Laboratory is to provide the specialist engineering groups with this information where it is not otherwise available.

II. Test Program

With the rapid advancement of transistor technology state-of-the-art, there is little time for semiconductor manufacturers to generate adequate reliability data. Therefore, considerable emphasis has been placed upon transistor testing at AIL. To most efficiently generate the necessary information, three major types of test programs are used. The first type--exploratory testing--is used to investigate areas where no information is available, and to provide a preliminary look at parts to be subjected to further testing. By first performing an exploratory test, it is sometimes possible to anticipate failures and failure modes.

The second type--evaluation testing--is performed on most semiconductors of interest to AIL, and is concerned with determining the inability of a part to meet the reliability requirements of a complex system. The parts are subjected to a profile of the environments and operating life they will be expected to survive in equipment. For one project, the program at AIL consists of sequential exposure to shock, vibration, thermal cycling, moisture resistance, vacuum, and operating life. Environmental levels are summarized in Table I. Data are taken initially, after each of the environmental exposures, and at several points during the operating life test. These data can then be analyzed to show how the parts characteristics varied as a result of exposure to the test profile. Design centers can be established; and 3σ limits, based on these and on the spread of characteristic values observed, are used to estimate the maximum deviations of these characteristics. In the case of environmental exposures, it has been found that linear regression analysis can often be applied to indicate the effects of the individual environments. These data, together with information about other expected variations, provide the designer with the information needed to ensure circuit operation under all conditions. The data also provide the reliability engineer with information necessary for the prediction of system reliability, including failure rates under environmental stresses, as well as hazard rates.

The third type--step-stressing--is used to compare equivalent parts from two or more manufacturers to determine which, if any, manufacturer's product is the least susceptible to damage from mechanical, electrical, and ambient environments. One or more environments are selected as the most likely to cause failure, and the test samples are subjected to successively higher levels of stress in these environments. It is assumed that those parts having the greatest resistance to the higher stress levels will have the lowest failure rate at lower stress levels. This, of course, is true only for the mode of failure under study.

All three test programs provide information valuable for establishing inspection tests to ensure purchase of highly reliable parts. The acceptance testing procedures established by military specifications are designed to ensure delivery of a high-quality product; however, these tests are based upon statistical sampling techniques and though the general quality of the delivered items is usually high, the passed group may contain units that are potential failures. This is particularly true of units having failure modes classed as critical under specific environmental conditions. AIL's test program has helped to identify many of these critical failure modes, and causes of failure, that exist in mesa and planar transistors. Once these causes were identified, tests for culling out potentially bad units were designed and written into AIL purchase specifications.

III. Test Results

The most critical and widespread failure mode observed throughout the studies of mesa and planar transistors has been that of thermal compression bond failures. Every manufacturer's product, and almost every type of mesa and planar transistor tested, has exhibited this failure mode. Indeed, the first test samples of a prominent manufacturer's rather exotic mesa transistors gave singularly poor results--the failure rate in this mode after environmental exposures was almost 100 percent.

Two basic causes for failure in this mode have been observed. During the very early stages of development of mesa transistors, it was found that most failures were caused by a lead whisker having been cut nearly through by the bonding tool, or by a whisker lifting off the junction. The implication here is that in the early development stages, the manufacturers had not yet established sufficient quality controls during the manufacturing process. This failure still appears from time-to-time; however, today it is a much less serious problem than in the past. The second basic cause of thermal compression bond failure is the formation of the brittle silicon-gold eutectic--the so-called "purple plague." Other failure modes, common to all transistors, were observed with much less frequency in the studies of mesa and planar transistors. These include surface contamination, leaks in cans or in glass-to-metal seals initially present or induced by exposure to environments, junction degradation caused by uneven diffusion, and loose particles enclosed in the can. An additional failure mode, peculiar to

mesa and planar transistors, is detachment of the semiconductor chip from the header. Table II summarizes the failures observed, and indicates tests selected to stress units for the particular failure modes and to detect failures when they occur. By subjecting planar and mesa transistors to 100 percent processing through these tests, the majority of infant mortality units are eliminated before they get into equipment; thus assuring more reliable equipment.

The 2N1132 is one mesa transistor type subjected to evaluation testing at AIL. This transistor exhibited no failures throughout the test program. If it is assumed that failures in environments occur binomially, it can be said with 90 percent confidence that the lower limit for the probability of a unit successfully surviving the environmental exposures is 0.91.

Figure 1 is a graph of the linear regression equation for the post-shock value of BV_{CBO} on the initial value of BV_{CBO}. The dashed portion of the curve is an extrapolation showing initial values less than the manufacturer's specified minimum of 50 volts. The points plotted on the same graph give a scatter diagram of the values observed, and verify the regression equation. Similar curves can be plotted for each of the environmental exposures.

Figure 2 is a plot of the normalized BV_{CBO} versus test points with the associated 3 σ limits, and shows how design tolerances are established. It can be seen that the median value did not change appreciably throughout the test. However, the 3 σ limits indicate that the range of values of BV_{CBO} underwent some change as the test progressed. In the case of BV_{CBO}, an increase in value is not significant from the designers' point of view, since it will not affect the operation of the circuit. On the other hand, a decrease in BV_{CBO} could cause failure of the circuit and must be taken into consideration. For the purpose of establishing a design tolerance for BV_{CBO}, the low limit is desired. Taking the manufacturer's specified minimum of 50 v, and correcting it to reflect the lowest value of the 3 σ limits throughout the test, the design tolerance becomes

$$0.46 \times 50 \text{ v} = 23 \text{ v}$$

A correction factor for temperature can also be added to this value; however, in the case of BV_{CBO}, this factor is small and can probably be safely ignored. It is interesting to note, that for this transistor, the derated value of 23 v arrived at after testing is close to the frequent blanket derating to 50 percent of the specified minimum.

Figure 3 is a graph of regression of the post-shock value of h_{FE} on the initial value for the 2N1132 transistor. This graph indicates that there is little effect on the pulsed current gain of the transistor due to shock. Figure 4 is a plot of the normalized h_{FE} versus test point, with the associated 3 σ limits and shows the behavior of h_{FE} throughout testing. For this characteristic, high and low limits are of equal interest to the designer, and are therefore calculated. The manufacturer's purchase tolerance for this transistor is 30 minimum and 90 maximum. The published specification shows that between -55°C and 100°C the value of h_{FE} varies from 0.51 to 1.55 times the value at 25°C. The lowest value of 3 σ calculated from the data is 0.53 times the initial value, and the highest 3 σ value calculated is 1.31 times the initial value. For the low temperature limit, the design tolerance is then

$$0.51 \times 0.53 \times 30 = 8.1$$

and for the high temperature limit, the design tolerance is

$$1.55 \times 1.31 \times 90 = 183$$

Increases in leakage currents caused by temperature increases and by aging are extremely important and must be taken into consideration for proper circuit design. Figure 5 shows the normalized median of I_{CBO} versus test point for the type 2N1132 transistor with the associated positive 3 σ limits. In this case, only the increases in I_{CBO} are of interest, and therefore only the upper design tolerance was calculated. The manufacturer's purchase tolerance on this characteristic is 1 μa maximum. A factor of 80 times the initial value is applied for temperature, and a factor of 93 for variation, giving a design tolerance of

$$80 \times 93 \times 1\ \mu a = 7.4 \text{ ma}$$

The curve of Figure 5 shows that, though the median value of I_{CBO}/I_{CBO} initial did not change appreciably throughout the test, there was a considerable amount of change in the values of some individual units. This is indicated by the wide variations in the 3 σ limits. There is no definite evidence as to the cause of these variations, but it is probable that a small amount of surface contamination is present in the worst of the units. No culling tests were performed on these units; however, the data indicate that a short burn-in period would identify the worst of the drifters, and these units could have been eliminated before testing.

A second transistor type subjected to evaluation testing at AIL was the 2N699. In this case, units from two manufacturers were tested. Manufacturer A's units survived the environmental portion of the test with no failures. After the units had been operating on life test for 48 hours, one unit was found to have failed because of excessive I_{CBO}. This was the only failure observed throughout the testing, though some marked variation in I_{CBO} was observed for other units. Based on the assumption that failures during operating life occur exponentially with time, it was calculated that the 90 percent confidence interval for the hazard rate, λ, is $0.0061 \leq \lambda \leq 0.133$ failure per thousand hours.

Design tolerance limits for operation between -55°C and 100°C were determined for this manufac-

turer's transistors. Figure 6 is a plot of the normalized BV_{CBO} for these units throughout test, with the associated 3 σ limits. Using the same techniques described in the discussion of the 2N1132 transistor, it was found that the design tolerance on BV_{CBO} for the 2N699 transistors from this manufacturer is

$$0.57 \times 120 = 68 \text{ v}$$

Again, the value calculated from the data agrees with the 50 percent rule for derating of BV_{CBO}.

Figure 7 shows the plot of normalized h_{FE} for this manufacturer's 2N699 transistors and the calculated design tolerance limits for operation between -55°C and 100°C. In this case, the temperature effects have been considered as well as the variations in h_{FE} throughout the profile test, and it can be seen that for operation at -55°C the design tolerance should be set at

$$0.61 \times 0.56 \times 40 = 13.7$$

For operation at 100°C, the design tolerance is

$$1.27 \times 1.66 \times 120 = 253$$

The transistor that failed after 48 hours of operating life because of an excessive I_{CBO} was excluded from the calculations for the plot of Figure 8, as it was from the calculations for Figures 6 and 7. The plot of Figure 8 shows that the median value of I_{CBO}/I_{CBO} initial did not vary appreciably throughout the test, but that there was considerable variation of I_{CBO} in some of the units. Taking temperature into consideration, it was calculated that for operation at 100°C, the design tolerance is

$$80 \times 11.5 \times 2 \ \mu a = 1.84 \text{ ma}$$

Again, there is no evidence from which to determine the cause of the changes in I_{CBO} among some units, but the prime suspect is surface contamination.

The sample of 2N699 transistors from manufacturer B had one failure after shock, giving a 90 percent confidence upper limit, for the probability of a unit successfully withstanding this environment of 0.996. One other failure was observed in this sample after 48 hours of operating life. It was calculated that the 90 percent confidence interval for the hazard rate, λ, is $0.021 < \lambda < 0.595$ failure per thousand hours. It should be noted that, though these figures are considerably higher than the figures for manufacturer A's sample, they are based on only 168 hours of life testing, and therefore cannot be directly compared.

Figure 9 shows the design tolerance limit on BV_{CBO} for manufacturer B's transistors to be 92 v. This is considerably higher than the value of 68 v established for manufacturer A's units, but again is based on 168 hours of testing, and cannot be directly compared. Figure 10 gives the design tolerance on h_{FE} as 9.3 to 232, and Figure 11 gives the design tolerance limit on I_{CBO} as 4.5 ma.

All of the test samples discussed here showed some variation in characteristics throughout the test program, particularly of the I_{CBO} characteristic. The failures observed were all excessive I_{CBO}, and occurred early in the testing or after a short period of operating life. The data and the plots of normalized I_{CBO} indicate that most of the units exhibiting large increases in I_{CBO} could have been eliminated from the test sample initially if the samples had been subjected to a short burn-in period.

IV. Conclusions

In summary, a technique has been presented for selecting the best available mesa and planar transistors of a given type by the use of culling tests. Also described is a test plan designed to subject the units being tested to a profile of environments that they would be expected to survive when installed in actual equipment. The results of testing two transistor types in such a profile of environments have been presented, and the techniques by which design tolerance limits are established have been explained. By applying these techniques to the selection of parts and to the design of equipments, AIL is developing complex electronic systems of high and predictable reliability.

TABLE I

ENVIRONMENT	LEVEL
SHOCK	60 G'S, 6 MSEC, 3 AXES
VIBRATION	TO 20 G'S, 5-3000 CPS, 3 AXES
THERMAL	−55°C TO +85°C
MOISTURE RESISTANCE	METHOD 106, MIL-STD-102 OR SIMILAR

TABLE II

FAILURE CAUSE	RELATED TEST
1. JUNCTION DEGRADATION (SURFACE CONTAMINATION, "HOT" SPOTS)	HIGH TEMPERATURE STORAGE, BURN-IN, DYNAMIC BREAKDOWN CHARACTERISTICS, LEAKAGE CURRENTS, LOW CURRENT BREAKDOWN VOLTAGE
2. WEAK OR BRITTLE COMPRESSION BONDS	THERMAL SHOCK, BURN-IN, ACCELERATION, LEAKAGE CURRENT, DYNAMIC BREAKDOWN CHARACTERISTICS
3. RESIDUAL STRESSES, POOR SEALS	THERMAL SHOCK, BURN-IN
4. LEAKS IN SEAL OR CAN (PRESENT INITIALLY OR INDUCED BY 3)	DETERGENT BOMB, RADIFLOW OR HELIUM LEAK, LEAKAGE CURRENT, DYNAMIC BREAKDOWN CHARACTERISTIC, LOW CURRENT BREAKDOWN
5. INSECURE MOUNTING OF CHIP	THERMAL SHOCK, BURN-IN, THERMAL RESISTANCE
6. LOOSE PARTICLES	ELECTROMECHANICAL STABILITY OR X-RAY

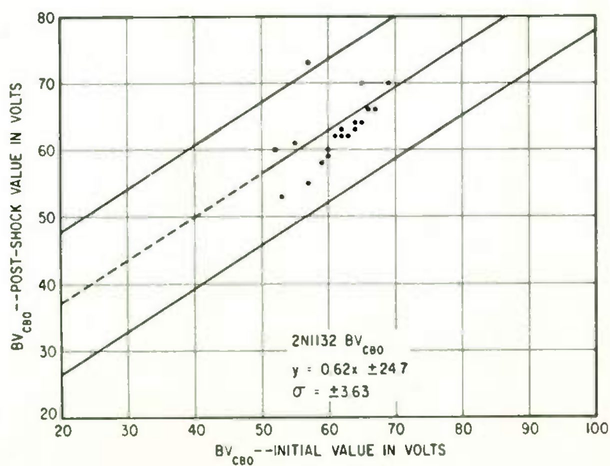


Fig. 1. BV_{CBO} Regression curve for 271132.

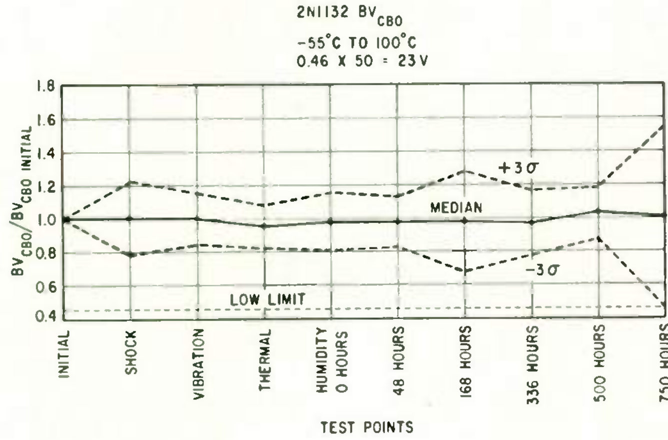


Fig. 2. 2N1132 BV_{CBO} Design tolerance.

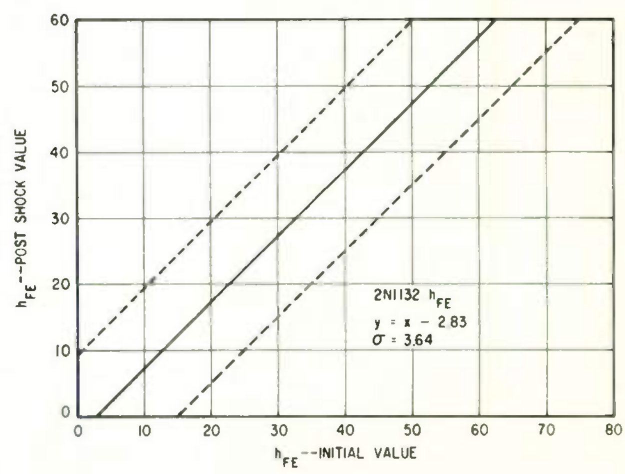


Fig. 3. h_{FE} Regression curve for 2N1132.

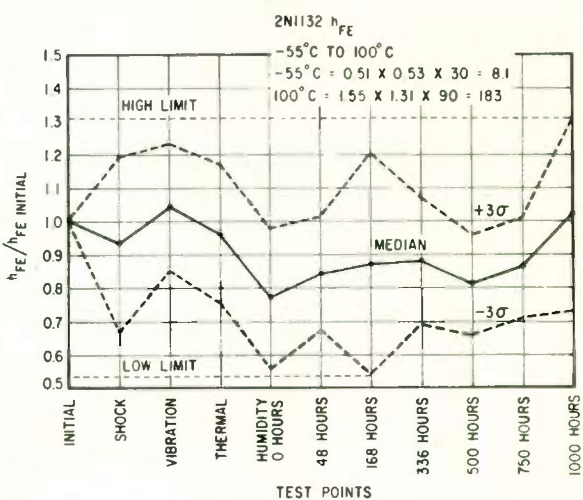
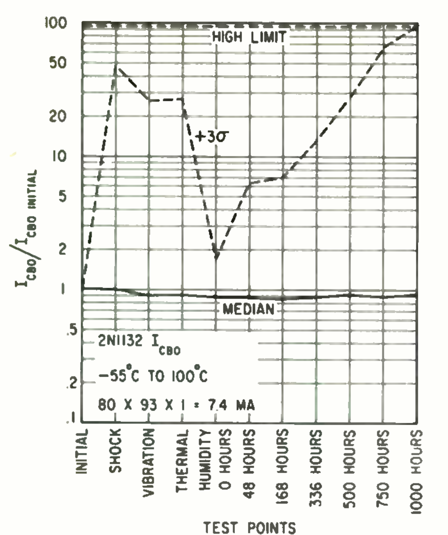


Fig. 4. 2N1132 h_{FE} Design tolerance.



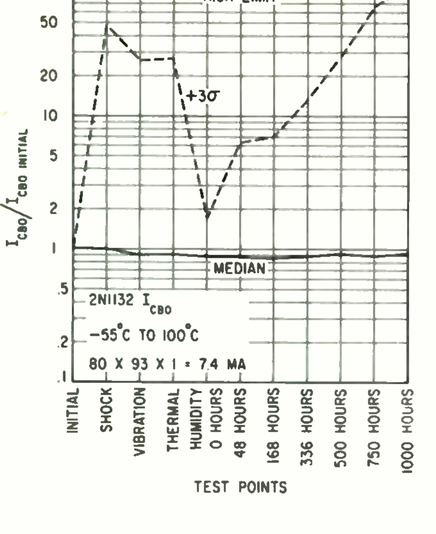


Fig. 5. 2N1132 I_{CBO} Design tolerance.

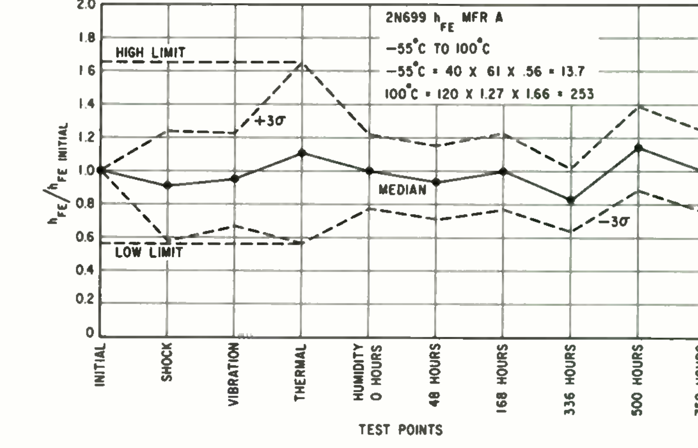


Fig. 7. 2N699 h_{FE} Design tolerance.

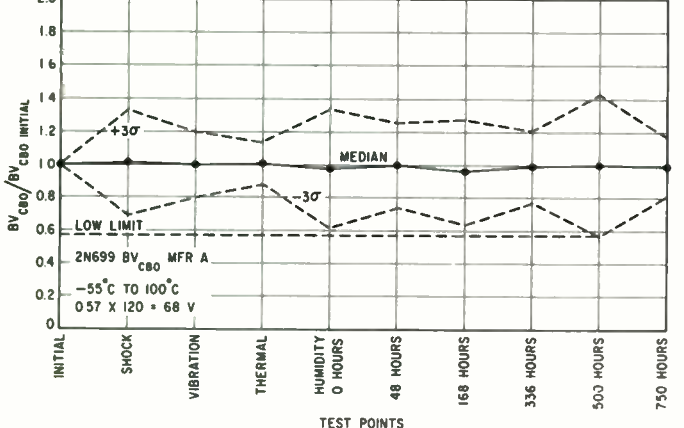


Fig. 6. 2N699 BV_{CBO} Design tolerance.

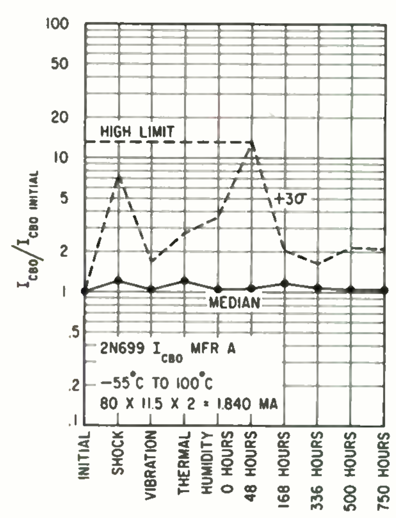


Fig. 8. 2N699 I_{CBO} Design tolerance.

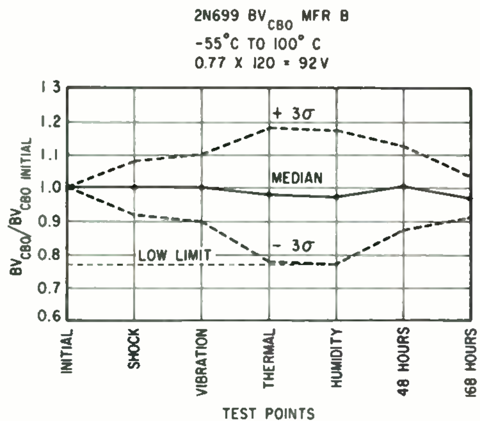


Fig. 9. 2N699 BV_{CBO} Design tolerance.

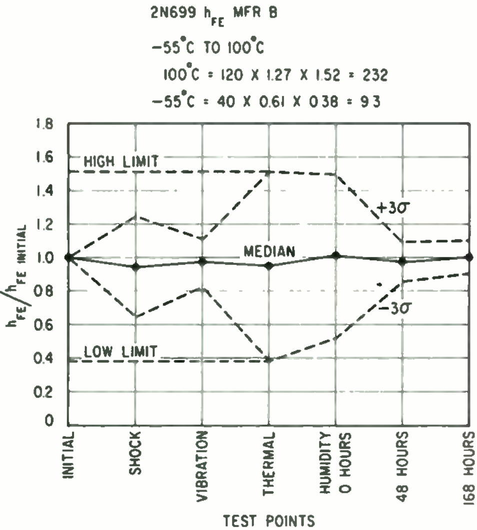


Fig. 10. 2N699 h_{FE} Design tolerance.

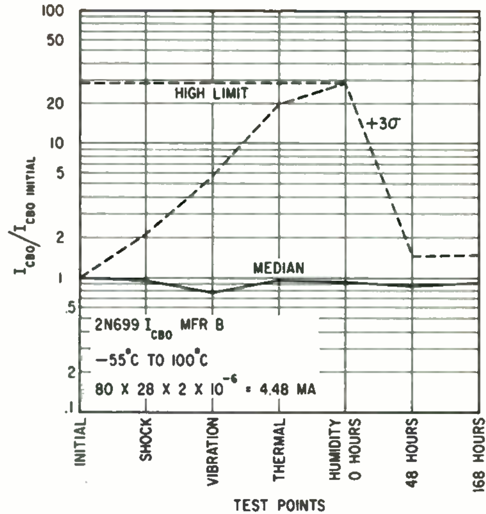


Fig. 11. 2N699 I_{CBO} Design tolerance.

ESTABLISHING RELIABILITY REQUIREMENTS

L. R. Diamond
General Telephone and Electronics, Inc.
Sylvania Electric Products Inc.
Sylvania Electronic Systems
Amherst, New York

INTRODUCTION

The bid requests which are issued for research and development projects act as triggers to set in motion a number of activities within the organization of each potential contractor. At Sylvania, one of these deals with reliability. The first action taken by the Reliability Engineering Group is to develop specific reliability requirements for the particular type of equipment specified in the bid request. Usually the goal is defined by either a general reliability statement or a specific design objective relating to reliability. To achieve the goal, a series of steps are taken, beginning with an investigation of the customer's real need. Does the bid request refer to mission reliability, equipment reliability, or both? The resulting hypothesis dictates the response, i. e. mission reliability generally is expressed in probability language whereas equipment reliability can be expressed at MTBF probability, or the allowable number of failures in a given period of operation.

To complete the evaluation, functional criteria are analyzed and reliability estimates are made to compare the feasibility of the proposed concept with the requirements. Some of the criteria considered are: physical environment, which includes the external and internal effects on the equipment as well as the use; maintainability; accessibility; skill level of the user; allowable downtime providing for necessary preventive maintenance; replacement of failed items; and availability. If the resulting evaluation fits the hypothesis, an optimum reliability statement can be written.

The possible design solutions are then postulated. After defining the basic system configurations and characteristics, functional block diagrams are developed. Factors influencing unreliability are screened carefully. Always bearing in mind the effect of failure on the system performance, an analysis is made of the effect of failure on each block of the diagram. Failure intolerant situations are corrected by circumventing specific modes of failure by the application of various levels of redundancy. Total capability vs reduced capability is analyzed for a feasible and realistic solution.

At this point in the design, a rough reliability estimate is made utilizing current prediction techniques. Component part failure rates suitable for the design application and physical environments are evolved. In addition, this prediction includes an estimate of degradation factors reflecting design application and fabrication methods. If comparison of this estimate with the postulated requirements shows areas where reliability improvement may be necessary, the design concept is re-evaluated in order to determine where the increase in reliability can be made.

Finally, the design is reviewed for its relative capabilities versus economic factors, technological factors, and customers need. Such an approach demonstrates the practicality of the solution.

DISCUSSION

In establishing reliability capability, proven prediction techniques are used on standard circuits as there is an abundance of adequate data available. Mathematical models are set up according to reliability requirements, the approaches used include the techniques of nonredundancy, redundancy, appropriate statistical distribution functions, and parts counting.

To clarify certain procedures which are used in the examples that follow, it should be mentioned that Sylvania recognizes and utilizes four levels of redundancy. These vary, in order of increasing complexity, and are termed alpha, beta, gamma, and delta. While only beta redundancy is used here, all four levels are defined in Appendix B.

The resulting calculations assess the capability of the system to function within the specified reliability requirements. Regardless of which approach is taken, care must be exercised to assure, with a high degree of confidence, that the one item will meet or exceed the reliability specified – and it must be economically feasible as well as in agreement with the specified physical characteristics regarding size, weight, shape, etc.

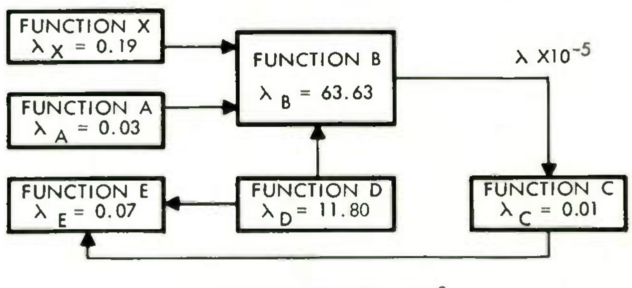
In answering bid requests, Sylvania endeavors to make the reliability statement both saleable and practical as well as responsive. The application of this philosophy is demonstrated in three examples, and each example is explained in sufficient detail to illustrate its use in communication systems.

EXAMPLE I

A communications system is desired which will perform in a supersonic aircraft, having a ceiling of 70,000 feet. Cooling inlet air will range from 0° to 80°F. It is required that the probability of failure during a mission of 20 hours be less than 1% on the average.

In order to fulfill this reliability requirement, a minimum MTBF of 2000 hours is needed.

After determining the method of system implementation (a vacuum tube configuration) which considers both known configuration of circuits previously developed for use in similar equipment and preliminary designs for new circuits, a failure rate prediction is made and shown in a reliability block diagram. A simplified portion of such a functional diagram is shown in Figure 1.



λ TOTAL = 75.73 X 10^{-5}
MTBF ≅ 1320 HOURS

FIGURE 1

The resulting prediction shows that the design will not meet the reliability requirement. The next step is to evaluate the design for:

(1) The modes of failure.

(2) Areas of greatest complexity.

(3) Effect of changes in the conceptual configuration.

(4) Major contributors of failure.

A quick glance of the functional diagram points out Functions B and D contribute approximately 99% of the estimated failures. Further evaluation of Function D indicates that little improvement could be gained.

Initially, the failure rate of Function B was calculated on the basis of continuous operation where, in fact upon further investigation, the duty cycle was found to be less. The next approach is to recalculate the failure rate based on the lesser duty cycle. However, stresses introduced by repeated turn-ons and shut-downs appear to overcome such differences. The transients introduced are sufficient to increase the failure rate rather than reduce it.

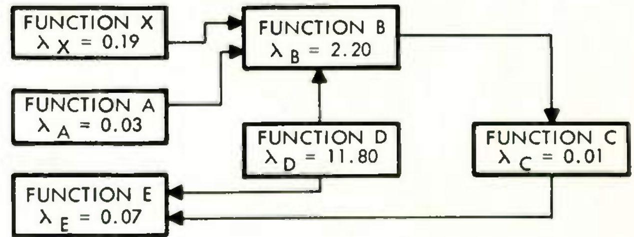
The next consideration is an alternate design whereby tube filaments will be kept warm and plate voltages reduced during standby so that a minimum plate current will flow. This approach introduces greater complexity, meaning that only a slight improvement in reliability can be achieved. If, however, the design incorporates redundancy, it might prove to be the optimum choice, with respect to cost effectiveness.

Another approach could be the establishing of a preventive maintenance procedure whereby certain tubes which degrade in performance at a predictable point in time would be replaced at specified intervals. This approach, while sound, would be difficult to sell to those customers who need to minimize maintenance.

As a result, the long-range solution appears to be in the area of further development utilizing solid-state devices rather than electron tubes.

The short-range solution is to utilize redundancy of the tube circuits in Function B thereby gaining a reliability increase in the order of five times the original estimate. The resulting functional diagram is shown in Figure 2.

It must be noted that redundancy is not a panacea and does not always increase reliability. In this case, space and weight allowances are considered to be flexible.



λ TOTAL = 14.30 X 10^{-5}
MTBF ≅ 7000 HOURS

FIGURE 2

This second preliminary estimate not only demonstrates that the MTBF requirement is attainable, but that it can be, in fact, exceeded. Because these estimates represent the upper bound on the achievable MTBF, extremely good reliability control and monitoring techniques will be required in the basic design and production phases of the program.

EXAMPLE II

In this example, it is required that a communications system operate continuously for one year in a synchronous orbiting satellite with the probability of no failure equivalent to 96%.

The steps taken to determine the possible design solution with respect to reliability are basically as follows:

(1) Evaluation of requirements.

(a) Physical environment - Van Allen belt radiation, ambient temperature at 75°C, launch acceleration of known value.

(b) Continuous operation.

(c) No maintenance in use.

(d) Maximum of 200 hours of intermittent operation prior to launch.

(e) Controlled temperature, pressure and humidity.

(f) Minimum power output.

(g) Frequency specified.

(2) The concept evolves to a functional block diagram which is a simple series string.

(3) Evaluation made of block diagram with respect to design requirements.

(4) The series string will meet the design requirements with the exception of power output and reliability.

(5) By utilizing various levels of redundancy, it becomes possible to approach the requirement specified.

The basic configuration contains the following elements (see Figure 3). Assumptions used in making the reliability calculations are:

(1) Estimate probability of success (P_s) utilizing the method outlined in MIL-STD-441.

(2) Base part failure rates are determined using RADC Reliability Notebook Supplement 1 PB-161894-1 and Sylvania developed failure rates.

(3) Use stress values calculated for each application.

(4) Part ambient temperature.

The resulting failure rates for each block are listed in Table I.

Various elements (blocks) of the basic configuration are combined in order to accomplish the design requirements, and the resulting reliabilities are calculated as shown in Table II. Interference between elements was considered with respect to worst case analysis of input to each stage. As a result, safety margins were incorporated in each stage.

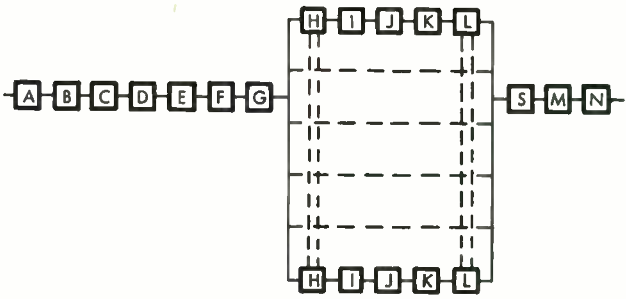


FIGURE 3

TABLE I

Blocks	Base Failure Rate % per 1000 hours
A	0.1518
B	0.1416
C	0.1120
D	0.0453
E	0.0554
F	0.2616
G	0.4585
H	0.266
I	0.4585
J	0.0649
K	0.07375
L	0.0673
M	0.00705
N	0.1550
S_1	0.5833
S_2	0.1480

TABLE II

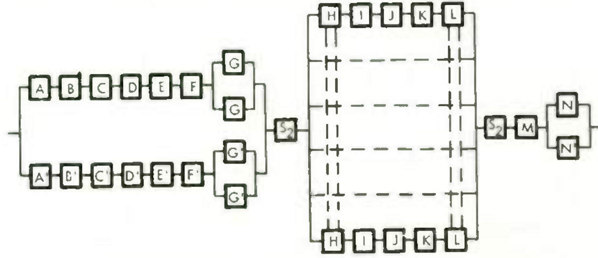
Conf. No.	Initial Power Req'd to Develop Final Power	Elements Used In Configuration	P_s (1 year)
1	X	ABCDEFG (HIJKL) 5 MN	0.6373
2	Y*	ABCDEFG 5 of 6 (HIJKL) MN	0.8750
3	Y	ABCDEFG 5 of 6 (HIJKL) (S_1 x 6)MN	0.6482
4	Y*	ABCDEFG 5 of 6 (HIJKL) (S_2 x 6)MN	0.8103

* Power output has degraded 15% due to lack of switching.

When further combinations of the elements are undertaken, reliabilities can be calculated for each of the configurations (see Table III).

At the same time that these configurations are being developed, estimates for each block are analyzed with respect to their failure contributions as well as mode of failure.

Because Block G has the highest probability of failure, a means must be found to reduce the predicted failure rate. Various redundant approaches can be taken, one of which is shown in the configuration of Figure 4.



BASE FAILURE RATES ARE THE SAME AS SHOWN IN TABLE I AND, IN ADDITION, A' = A, B' = B, ... G' = G, N' PERTAINS TO POWER CONNECTOR ONLY (REFER TO APPENDIX A "CALCULATIONS")

FIGURE 4.

It appears that configuration No. 9 is the optimum design, giving due consideration to the present state of the art and schedule commitments. The method for performing these calculations is combined with the calculations for configuration No. 9 which is described in Appendix A.

Using configuration No. 7 as the point at which design and development will start, breadboarding of the circuits begins. Then, by combining the various elements, reliability achieved can be measured during the entire program. As the state of the art progresses, the growth will approach the curve in Figure 5.

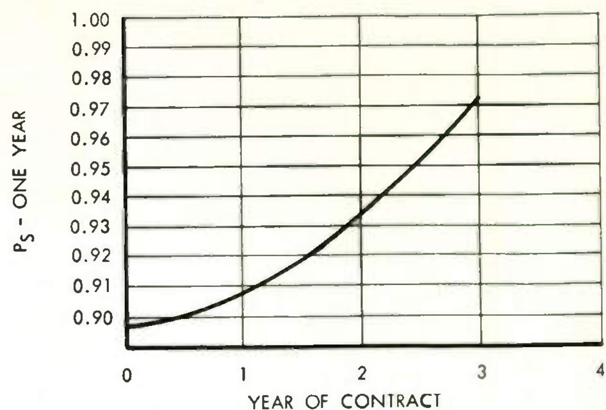


FIGURE 5

EXAMPLE III

An airborne warning system shall have an MTBF of at least 500 hours while being used in its operational environment.

An estimate of the inherent MTBF of the proposed design was calculated utilizing the parts count method. In this prediction, the following assumptions were made:

(1) 50% derating except for tantalum capacitors (70%) and carbon composition resistors (10%)

(2) Ambient temperature at 70°C

(3) Constant hazard rate (MIL-STD-441)

The resultant estimate of the MTBF was 864 hours or a failure rate of 115.724×10^{-5} per hour.

TABLE III

Conf. No.	Initial Power Req'd to Develop Final Power	Elements Used in Configuration	P_S (1 Year)
5	Y	ABCDEFG 1 of 2 (G) 5 of 6 (HIJKL) (S_2 x 6) MN	0.8423
6	Y*	Same as No. 5 except leave out switch	0.8969
7	Y*	Same as No. 6 except add N'	0.9078
8	Y	ABCDEF 1 of 2 (G) A'B'C'D'E'F' 1 of 2 (G') (S_2 x 2)	0.8862
9	Y*	Same as No. 8 except leave out switches	0.97270

*Power output has degraded 15% due to lack of switching.

Realizing that the purpose of this equipment was to warn the crew when their aircraft was illuminated by enemy pulse radar and to identify the type and quadrant location of this threat, Sylvania believed the mean-time-between-failure technique to be inadequate for describing the reliability capability of the equipment. Therefore, the following analysis was the optimum solution to the customer's needs. It is based on the probable operation of 6 threat detectors in each of four quadrants.

Since operation of all 24 threat detectors represents 100% capability, loss of any 1 detector means that 1/24 or approximately 4.24% of the total capability is lost. Furthermore, loss of 2 detection paths means that approximately 10% of the original capability is lost. Sylvania feels the capability should not be degraded by more than this amount to assure success of the mission.

To determine the probability that either failure event will not occur during a mission, the following probability formula is used.

Probability (A or B not occurring during a mission) =

Probability (A) + Probability (B) - Probability (A) Probability (B)

Where:

Probability (A) is the event, power supply doesn't fail (i.e. succeeds), and Probability (B) is the event, no more than 2 threat detectors fail.

It is desired that the resulting probability be no less than 90%. Assume mission duration is 15 hours.

Assume the system has failed if (1) power supply fails or (2) more than 2 threat detector capabilities have failed. There are 24 detector capabilities made up of 6 independent paths in each quadrant.

It is necessary to evaluate events A and B individually before combining them into the foregoing probability expression.

Event A - Probability of power supply failure during a mission.

(1)	Basic power supply failure rate less blowers, connectors, and solder joints	2.914×10^{-5}
(2)	Blower failure rate	1.78×10^{-5}
(3)	Connector failure rate	0.20×10^{-5}
(4)	Solder joint failure rate	0.10×10^{-5}
	Total failure rate (λ)	4.994×10^{-5} per hour

Compute the probability of successfully surviving a 15-hour mission for Event "A":

$$\text{Probability (A)} = P_s = e^{-\lambda t}$$

$$P_s = e^{-(4.994 \times 10^{-5} \times 15)} = 0.999251$$

Event B - Compute the probability of no more than 2 failures during a mission for Event B.

Refer to Figure 6A thru D for component allocation. Note that display modes are redundant since display is accomplished by either/both audio and visual.

Step One - Determine failure probabilities (P_f) of each of the 6-threat detection capabilities.

(1) Threat 1

$$\lambda = 4.420 \times 10^{-5} \text{ per hr}$$

$$P_{f_1} = 1 - e^{-\lambda t} = 0.000663$$

(2) Threat 2

$$\lambda = 4.420 \times 10^{-5} \text{ per hr}$$

$$P_{f_2} = 1 - e^{-\lambda t} = 0.000663$$

(3) Threat 3

$$\lambda = 4.420 \times 10^{-5} \text{ per hr}$$

$$P_{f_3} = 1 - e^{-\lambda t} = 0.000663$$

(4) Threat 4

$$\lambda = 4.824 \times 10^{-5} \text{ per hr}$$

$$P_{f_4} = 0.000724$$

(5) Threat 5

$$\lambda = 1.855 \times 10^{-5} \text{ per hr}$$

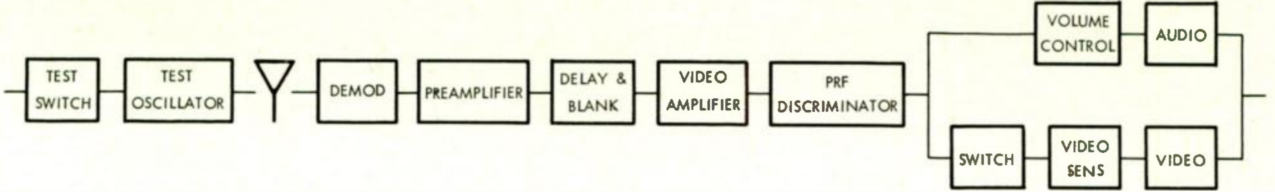
$$P_{f_5} = 0.000278$$

(6) Threat 6

$$\lambda = 7.149 \times 10^{-5} \text{ per hr}$$

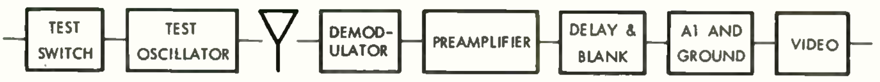
$$P_{f_6} = 0.001072$$

Step Two - The following logic will be used in determining the probability associated with event B.



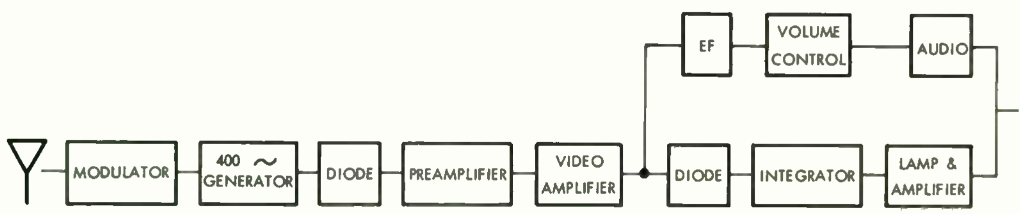
$\lambda = 4.420 \times 10^{-5}$ PER HOUR

FIGURE 6A



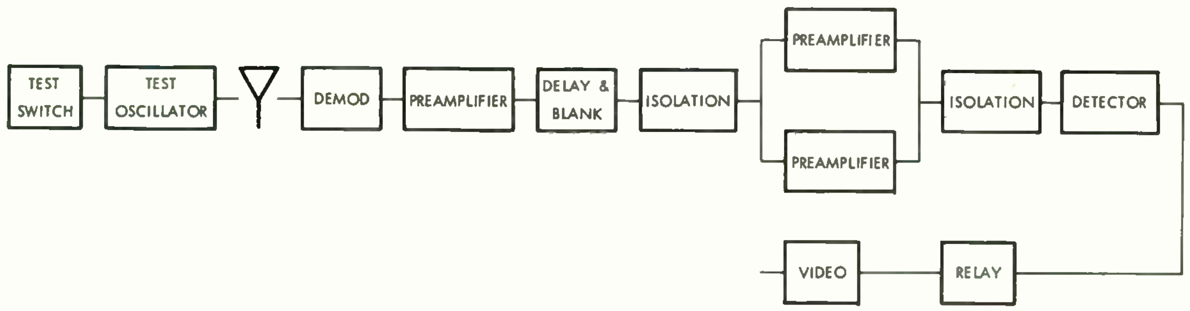
$\lambda = 4.824 \times 10^{-5}$ PER HOUR

FIGURE 6B



$\lambda = 1.855 \times 10^{-5}$ PER HOUR

FIGURE 6C



$\lambda = 7.149 \times 10^{-5}$ PER HOUR

FIGURE 6D

(1) Evaluate the ways 2 failures can occur. There are 123 different ways two detector paths can fail during a mission.

Determine the probabilities associated with each of these and add.

(2) Determine the probability of 1 failure occurring during a mission. Add the 24 ways this can happen.

(3) Add (1) and (2). This is the cumulative probability of 1 or 2 failures occurring.

(4) Solution of the above steps is the probability that no more than 2 failures occur. This is calculated to be 0.016308.

Having determined the separate probabilities of A and B occurring, the following formula is used in order to estimate the combined probabilities.

Probability (A or B occurring) = Probability (A) +

Probability (B) - Probability (A) Probability (B).

Probability (A or B occurring) = 0.999251 + 0.016308 - (0.99251 x 0.016380) = 0.999263.

Therefore, there is approximately one chance in a thousand that either the power supply or more than two detector capabilities will fail, on the average, during any given mission.

CONCLUSION

While this paper does indicate that redundancy is an excellent technique to be used in meeting established reliability requirements, it is not, for obvious reasons, the ultimate approach. It has its uses and, when properly applied, yields large dividends.

As demonstrated in the examples, reliability can be more readily optimized during the conceptual and design phases. Trade-offs are more easily effected and implemented. Design changes in hardware which are made after the design freeze are the most costly and can be minimized by utilizing reliability estimates as a basis for comparison throughout the program.

It is in essential in the establishment of reliability requirements that adequate controls be employed to determine whether deviations or variations will occur. No attempt was made in the paper to discuss these controls as they, in themselves, are a sizeable subject. The mathematic approach described, however, is a partner to these controls in that it pin points the control conditions while permitting a rapid assessment of the hardware.

Acknowledgement and appreciation is made of Mr. R. J. Costello of Hughes Aircraft and Messrs. H. McIntyre and B. K. Wolcott of Sylvania for their efforts which made this paper possible.

REFERENCES

Bazovsky, Igor, 1961, Reliability Theory and Practice, Prentice-Hall, Inc.

Herd, G. R., 1959, "Some Statistical Concepts and Techniques for Reliability Analysis and Prediction", Fifth National Symposium on Reliability and Quality Control Proceedings.

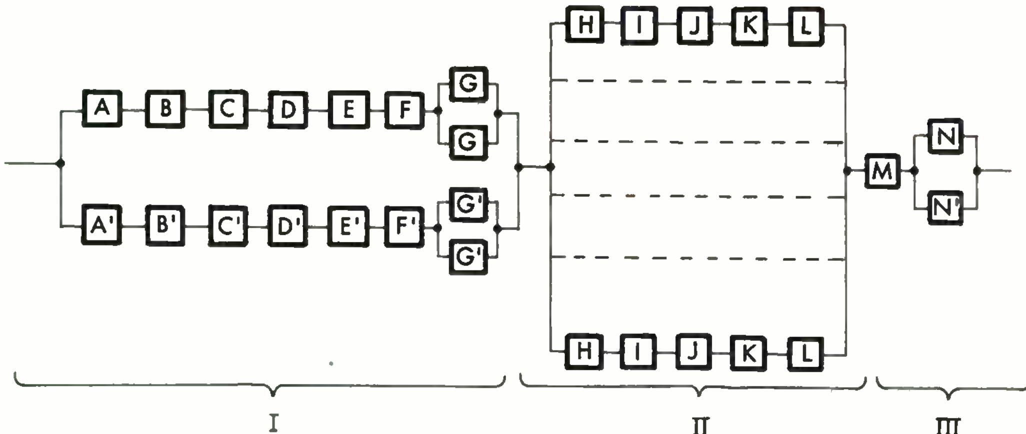
Rohn, W. B., 1959, "Reliability Prediction for Complex Systems", Fifth National Symposium on Reliability and Quality Control Proceedings.

Balabau, H. S., 1960, "Some Effects of Redundancy on System Reliability", Sixth National Symposium on Reliability and Quality Control Proceedings.

Sante, D. P., 1961, "Ultrareliability and Microminiaturization Study for UHF Receiver Design", ASD TR-61-669.

Sante, D. P., 1960, "Ultrareliability Techniques for Communication Equipment", WADD TR-60-702.

APPENDIX A



Calculations used to evaluate Conf. #9.

1. Calculate the probability of no failure for Function I (R_I).

(a) For block G - one out of two must operate. The combined probability of failure Q_{G_I} is equal to the square of the probability of failure of one element.

$$Q_{G_I} = Q_G^2 = \left(1 - e^{\lambda_G t}\right)^2$$

$$R_G = 1 - \left(1 - e^{\lambda_G t}\right)^2$$

(b) For string ABCDEFG

$$R_{A....G} = R_A R_B R_C R_D R_E R_F R_G$$

$$= e^{-t\,(\lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E + \lambda_F)} \left[1 - \left(1 - e^{-\lambda_G t}\right)^2\right]$$

(c) For accomplishment of Function I - one string out of two must operate. Again, the combined probability of failure Q_I is equal to the square of the probability of failure of one string.

$$Q_I = Q^2_{A....G} = \left(1 - R_{A....G}\right)^2$$

$$\therefore R_I = 1 - \left(1 - R_{A....G}\right)^2$$

$$= 1 - \left\{1 - \left[2e^{-t\,(\Sigma\lambda_{A....G})} - e^{-t\,(\Sigma\lambda_{A....F} + 2\lambda_G)}\right]\right\}^2$$

$$R_I = 4e^{-t\,(\Sigma\lambda_{A....G})} - 2e^{-t\,(\Sigma\lambda_{A....F} + 2\lambda_G)}$$

$$+ 4e^{-t\,(\Sigma\lambda_{A....G} + \Sigma\lambda_{A....F} + 2\lambda_G)}$$

$$- 4e^{-2t\,(\Sigma\lambda_{A....G})} - e^{-2t\,(\Sigma\lambda_{A....F} + 2\lambda_G)}$$

Substituting the values for the λ's listed in Table I it can be seen that

$R_I = .99562$

2. Calculate the probability of no failure for Function II (R_{II}).

(a) For string HIJKL

$$R_{H....L} = e^{-t(\lambda_H+\lambda_I+\lambda_J+\lambda_K+\lambda_L)}$$

(b) For 5 out of 6 strings must operate (R_{II}).

$$R_{II} = R^6_{HIJKL} + 6R^5_{HIJKL}\,Q_{HIJKL}$$

where $Q_{HIJKL} = 1 - R_{HIJKL}$

$$R_{II} = e^{-6t(\lambda_H+\lambda_I+\lambda_J+\lambda_K+\lambda_L)} + 6e^{-5t(\lambda_H+\lambda_I+\lambda_J+\lambda_K+\lambda_L)} \times \left[1 - e^{-t(\lambda_H+\lambda_I+\lambda_J+\lambda_K+\lambda_L)}\right]$$

Substituting the values for the λ's listed in Table I it can be seen that

$R_{II} = .97774$

3. Calculate the probability of no failure for Function III (R_{III})

(a) For block M

$$R_M = e^{-\lambda_M t}$$

(b) For block N - one out of two must operate. The combined probability of failure Q_N is equal to the square of the probability of failure of one block.

$$\therefore\ Q_N = \left(1 - e^{-\lambda_N t}\right)^2$$

$$R_N = 1 - \left(1 - e^{-\lambda_N t}\right)^2$$

$$R_{III} = R_M R_N$$

$$= e^{-\lambda_M t}\left[1 - \left(1 - e^{-\lambda_N t}\right)^2\right]$$

Substituting the values for the λ's listed in Table I it can be seen that

$R_{III} = .99922$

4. Since Functions I, II, and III are in series

$$P_S = R_I R_{II} R_{III}$$

Substituting the values calculated above:

$$P_S = (.99562)(.97774)(.99922)$$
$$= .97270$$

APPENDIX B

Definitions

A - Redundancy

Redundancy is employed as a means of averting loss of performance due to random failure. The paralleling (and in some cases, series) techniques, used by Sylvania, stress the utilization of the primary mode of failure of the part, circuit, or stage.

Four levels of redundancy are being utilized by Sylvania

(1) Alpha - Encompasses uncomplex circuits and consists of utilizing series and parallel where failure of either part is tolerable.

*(2) Beta - Encompasses medium complexity of circuits and is adaptable to the switching of parts or entire circuits in or out as the result of a failure.

(3) Gamma - In the event of a failure in a stage, the entire stage is bypassed to maintain continuity of operation of the equipment. An example of this would be a manual override.

(4) Delta - Utilizes two parallel equipment which may have alpha, beta, or gamma redundancy built into them.

B - Derating

Derating is an expression of the relationship between actual and/or expected stresses and rated stress levels for an item.

* Used in Example II.

DISTRIBUTIONS OF FUNCTIONS OF RANDOM VARIABLES

D.L. Heck and L.H. Tung

International Business Machines Corp.

General Products Division

Development Laboratory

Endicott, N.Y.

Summary — A system of 7090 computer programs is described for a method which obtains the probability distribution of a function involving the sum, difference, product, or quotient of independently distributed random variables. Continuous probability distributions are replaced by frequency distributions and the final distribution is obtained by an interval-by-interval combination of these frequency distributions, successively considering two distributions at a time. The system accepts a function expressed in an algebraic statement similar to Fortran, translates the function in accordance with an appropriate hierarchical ordering, and sets up the conditions necessary for the operator subroutines. Execution of the operators alternates with translation until the statement is completely evaluated. Close agreement has been obtained in a number of comparisons of the results from this numerical method with exact analytic results.

INTRODUCTION

This paper describes a system of IBM 7090 computer programs, called ELDIS, which has been developed by the authors to obtain the probability distributions of functions of independently distributed, continuous, random variables. In many physical systems, particularly those encountered in reliability studies, the variable of interest occurs as a function of component or subsystem variables, such as resistance, voltage, power, and time. For example, Figure 1 shows a simple series - parallel resistance circuit where the resistance values are assumed to follow a rectangular or uniform probability distribution. If the total resistance of this circuit is to be obtained, then, for a given set of values R_1, R_2, ..., R_6, the following equation can be used:

$$R_T = \frac{1}{\frac{1}{R_6} + \frac{1}{R_2 + R_3 + R_4}} + R_1 + R_5 . \qquad (1)$$

If a large number of such circuits as those in Figure 1 are manufactured and R_T is measured, then these measurements of R_T will tend to be different because of the variation of the resistances R_1, R_2, ...R_6 over their respective ranges. Hence, this leads to a distribution of values of R_T. Generally, in order to determine the adequacy of such a circuit for a given purpose, it is the frequency distribution of R_T (Figure 2) which is required. The circuit designer can then use the cumulative probabilities from this distribution to compare with failure criteria, and thus obtain a proportion or percentage of circuits that he might expect to be beyond specifications.

There are several methods for obtaining the distribution of a function of random variables such as R_T. One obviously impractical method is that just described, i.e., building a large number of systems, observing them in operation, and recording the value of the output variable so as to obtain a frequency distribution. Another method is to use Monte Carlo techniques and to simulate system behavior on a computer. But, in many instances, the system requirements or specifications are so stringent that this method becomes unfeasible due to the large number of trials that are necessary to obtain accurate information concerning the tails of the distribution. A third method is to obtain the analytic form of the resultant distribution using theoretical means, such as integral transforms; however, not only must the analytic form of the component distributions be known, but, in general this procedure becomes extremely difficult for problems involving a function of more than two variables. Other approaches include methods of piecewise approximation and the propagation of error technique, the latter requiring the assumption of normal distributions.[1]

The system of programs presented here, and the numerical method which it uses, possess several advantages over the approaches mentioned above. These are:

a) flexibility in the form in which the component distributions are specified

b) ease of specification of the equation or functional relationship to be evaluated

c) speed of computation

d) accuracy in the tails of the distribution

METHOD

The numerical method which is employed by the program is described below. First, represent the continuous probability distributions of component variables by means of discrete probabilities over a finite range of the variable. This is done by dividing the range of each variable into k intervals and then obtaining the probability associated with each interval. For example, let:

x_1 be distributed as $f_1(x_1)$ for $\alpha_1 < x_1 \leq \beta_1$, and

x_2 be distributed as $f_2(x_2)$ for $\alpha_2 < x_2 \leq \beta_2$

Dividing the range of x_1 into k equal intervals, with interval endpoints $L_{1,0}, L_{1,1}, \ldots, L_{1,k}$, where $L_{1,\ell} = \left[\alpha_1 + \left(\frac{\beta_1 - \alpha_1}{k}\right)\ell\right]$ for $\ell = 0, 1, 2, \ldots, k$, the distribution $f_1(x_1)$ is described by giving:

$P_{1,1}, P_{1,2}, \ldots, P_{1,k}$, where $P_{1,j} = \text{prob}\left\{L_{1,j-1} < x_1 \leq L_{1,j}\right\}$ for $j = 1, 2, \ldots, k$.

Similarly, define $L_{2,0}, L_{2,1}, \ldots, L_{2,k}$, where $L_{2,\ell} = \left[\alpha_2 + \left(\frac{\beta_2 - \alpha_2}{k}\right)\ell\right]$ for $\ell = 0, 1, 2, \ldots, k$

and $P_{2,1}, P_{2,2}, \ldots, P_{2,k}$, where $P_{2,j} = \text{prob}\left\{L_{2,j-1} < x_2 \leq L_{2,j}\right\}$ for $j = 1, 2, \ldots, k$

for the distribution $f_2(x_2)$.

Two distributions are then combined according to one of the operations, "sum", "difference", "product", or "quotient", and, by means of an interval-by-interval combination procedure, a third or resultant distribution is obtained which is also represented by discrete probabilities. Let Ω denote the arithmetic operations and consider $x = x_1 \, \Omega \, x_2$. The resultant distribution of x, say f (x), is then described by giving the limits $[\alpha < x \leq \beta]$, the interval endpoints $L_0, L_1, \ldots, L_k$ where $L_\ell = \left[\alpha + \left(\frac{\beta - \alpha}{k}\right)\ell\right]$ for $\ell = 0, 1, 2, \ldots, k$, and the k probabilities $P_1, P_2, \ldots, P_k$ where $P_j = \text{prob}\{L_{j-1} < x \leq L_j\}$ for $j = 1, 2, \ldots, k$. For example, for Ω = sum, $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$; for Ω = "quotient", $\alpha = \alpha_1/\beta_2$ and $\beta = \alpha_2/\beta_1$. Restrictions are placed on $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that α and β are finite.

To examine the underlying mechanics of the interval combining method more closely, let us consider the ith interval from $f_1(x_1)$ and the jth interval from $f_2(x_2)$, which are represented respectively by:

$P_{1,i} = \text{prob}\left\{L_{1,i-1} < x_1 \leq L_{1,i}\right\}$ and

$P_{2,j} = \text{prob}\left\{L_{2,j-1} < x_2 \leq L_{2,j}\right\}$.

The problem is essentially that of drawing an infinitely large number of pairs of observations from $f_1(x_1)$ and $f_2(x_2)$ and then combining each pair according to the operation Ω. Therefore, the probability of obtaining an observation from the ith interval of $f_1(x_1)$ and one from the jth interval of $f_2(x_2)$ is given by the product $P_{1,i} \, P_{2,j}$. When the operation Ω is applied to a pair of observations from these two intervals, the resulting value will fall (i.e., will be mapped) into some interval of x. This interval will be determined by applying operation Ω to the endpoints of the ith and j^{th} intervals in a manner analogous to that used to obtain the minimum (α) and maximum (β) values of the x distribution, as described in the previous paragraph. Denoting this interval by:

$$\left[L_{(i,j)\,\text{lower}}, L_{(i,j)\,\text{upper}}\right] = \left[L_{1,i-1}, L_{1,i}\right] \Omega \left[L_{2,j-1}, L_{2,j}\right] \quad (2)$$

it is found that, in general, (2) does not coincide exactly with the intervals $[L_{i-1}, L_i]$ of f(x), but may contain one or more of the latter intervals within its range. This problem is treated by the program by proportionally allocating the amount of probability represented by $P_{1,i} \cdot P_{2,j}$ to those intervals or portions of intervals of f(x) which are covered by (2). All k^2 possible combinations of pairs of intervals of $f_1(x_1)$ and $f_2(x_2)$ are processed, and the amounts of probabilities contributed to each of the k intervals of f(x) are accumulated to give the final values $P_1, P_2, \ldots, P_k$.

DESCRIPTION OF THE PROGRAM SYSTEM

The purpose of the program is to produce the probability distribution of a function of random variables, given the following input information:

1) A specification of the probability distributions of the random variables occurring in the function.
2) The equation defining the functional relationship between the random variables. This equation is written in FORTRAN-like notation, that is, it is a string of elements consisting of:[6]
   a) variable names one to five characters in length
   b) arithmetic operation signs (+ for addition, - for substraction, * for multiplication, / for division)
   c) left and right parentheses.

As an example of the manner in which the equation is expressed, the equation given by (1) could be punched on a card as follows:

RT = 1/ (1/R6 + 1/ (R2 + R3 + R4)) + R1 + R5

In this case "1", "R1", ..., "R6" are considered variable names.

The system of programs, which has been coded for a 32K IBM 7090 computer, consists of two major programs and their associated subroutines. The main programs are:

a) The input distribution generator, ELDIS1

b) The translator-operator routine, ELDIS2.

ELDIS1 is a FORTRAN coded program which generates a binary tape for input to ELDIS2; the leading records of this tape contain control and problem identification information as well as the equation which is to be evaluated. Succeeding blocks of records on the tape contain, for each variable in the equation, the name by which the variable is referred, the maximum and minimum values of the variable, the number of intervals in the tails of the distribution which contain probabilities smaller than a specified epsilon, and the k probabilities for the k class intervals.

A subroutine is provided for obtaining the probabilities for each of several different types of component distributions, and the input cards mentioned in 1) above determine which of these subroutines are to be used in a particular problem. Six subroutines are presently available, three of which are for the normal distribution; the rectangular or uniform distribution; and the Weibull distribution, which includes the negative exponential as a special case. For these distributions, the respective subroutine proceeds by picking up the minimum, maximum, and necessary location, scale, and shape parameters from the input card; divides the range of the variable into k class intervals; and performs an integration of the density function over each interval to obtain the k probabilities. It is also possible to make use of empirically obtained data which describes the input distributions, for two subroutines are included for handling histograms directly, as well as for polynomial smoothing of histograms. Still another subroutine takes care of the occurrence of constants in the equation by effectively considering the constant as a point distribution. The subroutines are included in the main program in such a manner that further subroutines, to evaluate additional types of distributions, could be easily incorporated. An optional feature which is available in this portion of the system is the output of the probability distributions which are generated.

ELDIS2 is coded in FAP and is an executive routine which performs two major tasks with respect to the information on the binary tape generated by ELDIS1.[2] The first task consists of the interpretation and translation of the algebraic equation with a simultaneous checking for violation of rules in writing the equation. As the equation is scanned, a reduced and coded expression describing the equation is set up internally for later use, and, in the event of an error in expressing the equation, a diagnostic message is printed on-line and the machine stops. If the first section is successfully completed, the program begins its second task. It reads in the control information and probabilities for the distributions whose names occur in the equation, sets up the necessary preliminaries, scans through the reduced expression, and calls up the operator subroutines in a sequence determined by a hierarchy of operations identical to that found in algebra. There are four such operator subroutines, one for each of the four arithmetic operations denoted by Ω in the previous section.

The executive program evaluates the equation by considering pairs of distributions, and, prior to entry to one of the operator subroutines, the program transfers the minimum and maximum values and the k probabilities associated with each of the two distributions $f_1(x_1)$ and $f_2(x_2)$ into a working storage area. The operator subroutine then proceeds to compute the minimum, maximum, and interval width of the resultant distribution f(x), as described in above, and to carry out the combination of the k^2 pairs of intervals in $f_1(x_1)$ and $f_2(x_2)$, allocating and accumulating the probabilities in the k intervals of f(x). Upon exit from the subroutine, the probabilities for f(x) are transferred by the executive program into the block of storage originally occupied by the probabilities of $f_1(x_1)$. Final reduction of the equation is effected by successive reduction of variable names occurring in the equation.

The final output is written on a BCD tape and results in printed information, which consists of the following:

a) identification information describing the function

b) the equation of the function being evaluated

c) the minimum, maximum, mean, and variance of the resultant distribution f(x)

d) the resultant distribution described by means of the endpoints of either 32 or 256 class intervals, the probability corresponding to each of these intervals, and the cumulative probability and complement of the cumulative probability corresponding to the upper endpoint of each interval.

The number of intervals, k, is presently fixed at 256, which was chosen to take advantage of the binary arithmetic of the machine. Computations within the operator subroutines for the limits and interval endpoints are performed by means of modified floating point arithmetic, while probability computations employ machine floating point arithmetic. A saving of computer time over straightforward methods results from a table lookup feature which is incorporated to link factors which occur in these two modes of arithmetic.

CHECKOUT AND ACCURACY OF THE METHOD

The procedure used to check out and assess the accuracy of the numerical method of the operator subroutines consisted of a comparison of the computed cumulative distribution function (c.d.f.) with the analytic c.d.f. for the same abscissa value. Functions of normal and rectangular variates were employed for this, and the results displayed in Table I were obtained by means of the Pearson-Hartley tables of the normal c.d.f.,[5] the formula for the convolution of rectangular variates given by Cramér,[3] the distribution of the ratio of two normal variates given by Kendall and Stuart,[4] and easily obtained formulas for the distribution of the product and quotient of rectangular variates. Table I shows the per cent relative error which occurs for several regions in the tails of the distributions of the stated functions, where

$$\text{Rel. error (\%)} = 100 \times \left| \frac{(\text{val. of computed c.d.f.}) - (\text{val. of anal.c.d.f.})}{(\text{val. of anal.c.d.f.})} \right|$$

Results pertaining to the "difference" operation are omitted since they are the same as those for "sum". The "sum" subroutine itself is used for the "difference" after appropriate modification of the limits is made by the program.

An indication of the manner in which the error builds up for a function of more than two variables can be observed from the results for the sum of two, eight, and sixteen rectangular variates.

RESTRICTIONS ON THE PROGRAM

The requirement that the variables which appear in the equation be independent imposes a restriction on the manner in which the equation is written. In general, a variable should not appear more than once in the equation unless it is intended that, each time the variable appears, it may take on values over its full range. For example, if we consider V_{out} in the circuit illustrated in Figure 3:

$$V_{out} = R_2 / (R_2 + R_1), \tag{3}$$

where it is understood that R_2 refers to the same resistance value. However, the program would mistakenly interpret this function as:

$$V'_{out} = R_3 / (R_2 + R_1), \tag{4}$$

where R_3 is considered as having the same distribution as R_2; consequently, the limits and hence the distribution of V_{out} would differ from V'_{out}. By algebraic simplification of (3):

$$V_{out} = 1 / (1 + R_1 / R_2) \tag{5}$$

which will be interpreted by the program to yield the proper limits and distribution. This problem arises in many circuit equations, but can be overcome if algebraic manipulation produces an equation in which each variable appears uniquely.

Further restrictions require that no more than fifty variables appear in the equation, with a variable being counted each time it appears. Also, it is necessary that the range of the variables be non-negative. This can generally be accomplished by a positive shift of the location parameter of the distribution. These are general restrictions governing use of the program. Detailed rules and instructions for preparing the input data cards are contained in a writeup of the program describing its operation.

RUNNING TIME CONSIDERATIONS

A number of runs were made in which estimates were obtained for the running time required for various sections of the program. Table IIa shows the times required by the ELDIS1 portion of the system to generate and write on binary tape the probabilities for the various types of distributions.

The program is presently set up to be run as a two link chain job within the 7090 Fortran Monitor System, and the approximate time required to make the transition between links is six seconds.

Table IIb shows the time required by the four operator subroutines to carry out the respective operations on a pair of distributions.

SAMPLE PROBLEM

As an illustration of the manner in which a problem might be set up and processed by the program, consider the situation depicted by Figure 1 and Equation (1). Figure 4 is a listing of the input cards which are required, together with notes describing the quantities appearing in various fields of the cards. (A full description is found in the program writeup previously mentioned). Figure 5 is an example of the output which may be obtained from the ELDIS1 section, while Figure 6 shows the final output. In reading Figures 4 and 5, a number which appears as floating point (e.g., .45000000E 02) is interpreted as $.45 \times 10^2$. In Figure 6, the floating point numbers 3 249677 and -6 821859 are interpreted respectively as $.249677 \times 10^3$ and $.821859 \times 10^{-6}$.

REFERENCES

1. Ashar, K.G., Evaluation of Circuit Failure Probability by Piecewise Approximation, TN 00.440, DSD Div., IBM Corporation, Poughkeepsie, N.Y. 1960.

2. "Bulletin, FORTRAN Assembly Program (FAP) for the IBM 709/7090", IBM Publication J28-6098-1, IBM Corp., White Plains, N.Y. 1961.

3. Cramér, Harold, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1951.

4. Kendall, M.G. and Stuart, A., The Advanced Theory of Statistics, Vol. I, Hafner Publishing Co., New York, 1958.

5. Pearson, E.S. and Hartley, H.O., Biometrika Tables for Statisticians, Vol. I., Cambridge University Press, Cambridge, 1956.

6. "IBM 709/7090 FORTRAN Programming System Reference Manual", IBM Publication C28-6054-2, IBM Corp., White Plains, N.Y. 1961.

TABLE I

Comparison of ELDIS Results With Analytic Cumulative Distribution Function

Operation	Function	Left Tail Probabilities (Value of c.d.f.)	% Relative Error	Right Tail Probabilities (Value of 1-c.d.f.)	% Relative Error
SUM	2 Normal $(50, \sqrt{2})$	$.1919 \times 10^{-5}$	.2	*	
		$.1591 \times 10^{-3}$	.2		
		$.4145 \times 10^{-2}$	.1		
	2 Rect (0, 1)	$.3052 \times 10^{-4}$	.0	*	
		$.1221 \times 10^{-3}$	.0		
		$.4395 \times 10^{-2}$	.0		
	8 Rect (0, 1)	$.2483 \times 10^{-5}$	2.1	*	
		$.1478 \times 10^{-3}$	.7		
		$.4234 \times 10^{-2}$	.3		
	16 Rect (0, 1)	$.2008 \times 10^{-5}$	4.5	*	
		$.1727 \times 10^{-3}$	2.2		
		$.4254 \times 10^{-2}$	1.1		
QUOTIENT	2 Rect (1, 2)	$.1357 \times 10^{-3}$	7.6	$.8607 \times 10^{-5}$	1.8
		$.5367 \times 10^{-3}$	2.0	$.2178 \times 10^{-3}$	.6
		$.4619 \times 10^{-2}$	.2	$.4440 \times 10^{-2}$	.0

TABLE I (Cont)

Operation	Function	Left Tail Probabilities (Value of c.d.f.)	% Relative Error	Right Tail Probabilities (Value of 1-c.d.f.)	% Relative Error
	2 Normal $(50, \sqrt{2})$	$.2488 \times 10^{-5}$	.6	$.2215 \times 10^{-5}$	.4
		$.1792 \times 10^{-3}$	.3	$.1812 \times 10^{-3}$	.3
		$.4599 \times 10^{-2}$	.2	$.4687 \times 10^{-2}$	.1
PRODUCT	2 Rect (0, 1)			$.7639 \times 10^{-5}$	.1
				$.1920 \times 10^{-3}$	.0
				$.4554 \times 10^{-2}$	.3

*No appreciable difference in per cent Relative Error was found between the two tails of these symmetrical distributions.

TABLE IIa —
ELDIS1 Timing

Type of Distribution	Avg. Running Time in seconds*
Normal	1.8
Rectangular	1.6
Polynomial Fit	3.0
Weibull	2.4
Histogram	1.6
Constant	1.4

*Includes output of the frequency and cumulative distributions on BCD tape for subsequent printout.

TABLE IIb —
ELDIS2 Timing

Operation	Avg. Running Time in seconds
Addition	9.1
Subtraction	9.1
Multiplication	21.8
Division	28.2

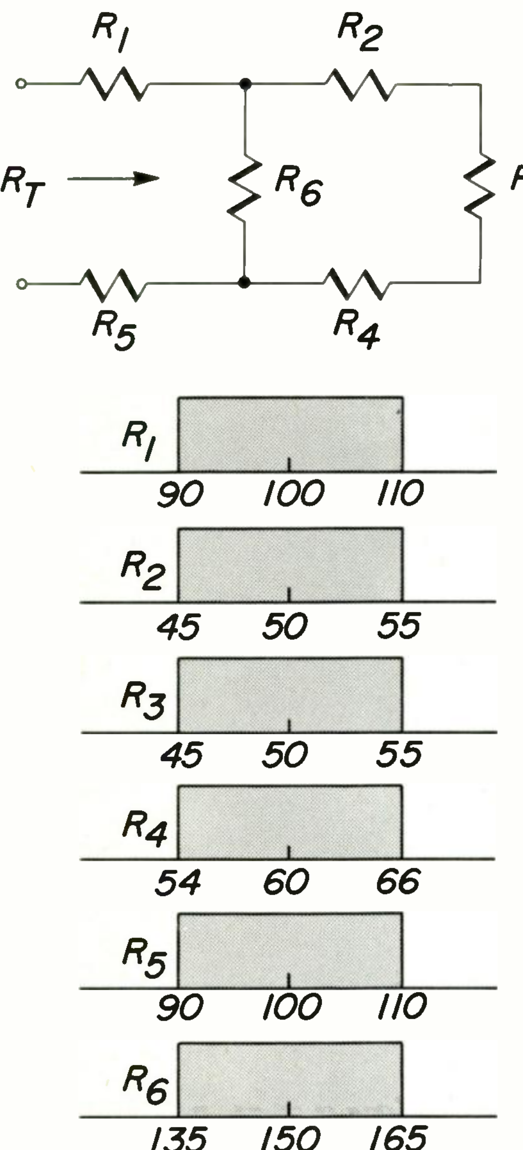


Fig. 1 - Series - Parallel Resistance Circuit

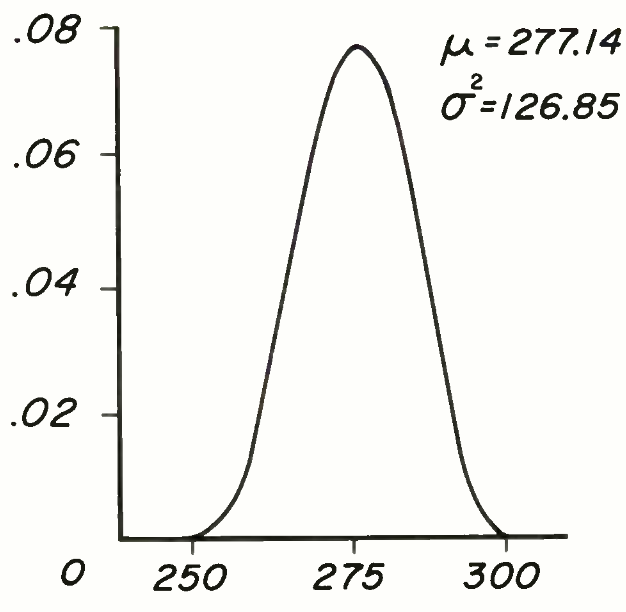


Fig. 2 - Frequency Distribution of R_T

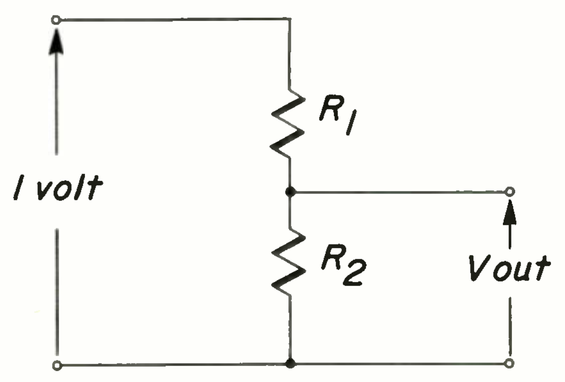


Fig. 3 - Example Circuit

```
      1 } Parameters required by
  1     } program

DISTRIBUTION OF TOTAL RESISTANCE FOR CIRCUIT IN FIGURE 1.  — Problem Identification
    1    7  — No. of variable names occurring in equation
RT =    1 / ( 1/R6 + 1/(R2+R3+R4) )   + R1 + R5  — Equation
 1    6.10000000E 01
R1    2.90000000E 02.11000000E 03
R2    2.45000000E 02.55000000E 02
R3    2.45000000E 02.55000000E 02
R4    2.54000000E 02.66000000E 02
R5    2.90000000E 02.11000000E 03
R6    2.13500000E 03.16500000E 03
```

Annotations:

- No. of cards required to express equation — 1
- Variable name
- Value of constant
- Minimum of Rect. Distn.
- Maximum of Rect. Distn.
- Refers to type of subroutine used in ELDIS1: 2 - Rectangular Distn.; 6 - Constant

Fig. 4 - Input to ELDIS Program

```
DISTRIBUTION OF TOTAL RESISTANCE FOR CIRCUIT IN FIGURE 1.
DISTRIBUTION NO.    6
NAME IS   R5   .  TYPE IS  RECTANGULAR.

LOCATION PARAMETER IS  .
SCALE PARAMETER IS     .
SHAPE PARAMETER IS     .

MINIMUM  IS  .90000000E 02 ,   MAXIMUM IS  .11000000E 03, DELTA IS  .78125000E-01
```

INTERVAL ENDPOINT	PROB	CUM PROB	INTERVAL ENDPOINT	PROB	CUM PROB
90.312500	15625000E-01	15625000E-01	100.312500	15625000E-01	51562500E 00
90.625000	15625000E-01	31250000E-01	100.625000	15625000E-01	53125000E 00
90.937500	15625000E-01	46875000E-01	100.937500	15625000E-01	54687500E 00
91.250000	15625000E-01	62500000E-01	101.250000	15625000E-01	56250000E 00
91.562500	15625000E-01	78125000E-01	101.562500	15625000E-01	57812500E 00
91.875000	15625000E-01	93750000E-01	101.875000	15625000E-01	59375000E 00
92.187500	15625000E-01	10937500E-00	102.187500	15625000E-01	60937500E 00
92.500000	15625000E-01	12500000E-00	102.500000	15625000E-01	62500000E 00
92.812500	15625000E-01	14062500E-00	102.812500	15625000E-01	64062500E 00
93.125000	15625000E-01	15625000E-00	103.125000	15625000E-01	65625000E 00
93.437500	15625000E-01	17187500E-00	103.437500	15625000E-01	67187500E 00
93.750000	15625000E-01	18750000E-00	103.750000	15625000E-01	68750000E 00
94.062500	15625000E-01	20312500E-00	104.062500	15625000E-01	70312500E 00
94.375000	15625000E-01	21875000E-00	104.375000	15625000E-01	71875000E 00
94.687500	15625000E-01	23437500E-00	104.687500	15625000E-01	73437500E 00
95.000000	15625000E-01	25000000E-00	105.000000	15625000E-01	75000000E 00
95.312500	15625000E-01	26562500E-00	105.312500	15625000E-01	76562500E 00
95.625000	15625000E-01	28125000E-00	105.625000	15625000E-01	78125000E 00
95.937500	15625000E-01	29687500E-00	105.937500	15625000E-01	79687500E 00
96.250000	15625000E-01	31250000E-00	106.250000	15625000E-01	81250000E 00
96.562500	15625000E-01	32812500E-00	106.562500	15625000E-01	82812500E 00
96.875000	15625000E-01	34375000E-00	106.875000	15625000E-01	84375000E 00
97.187500	15625000E-01	35937500E-00	107.187500	15625000E-01	35937500E 00
97.500000	15625000E-01	37500000E-00	107.500000	15625000E-01	87500000E 00
97.812500	15625000E-01	39062500E-00	107.812500	15625000E-01	89062500E 00
98.125000	15625000E-01	40625000E-00	108.125000	15625000E-01	90625000E 00
98.437500	15625000E-01	42187500E-00	109.437500	15625000E-01	92187500E 00
98.750000	15625000E-01	43750000E-00	108.750000	15625000E-01	93750000E 00
99.062500	15625000E-01	45312500E-00	109.062500	15625000E-01	95312500E 00
99.375000	15625000E-01	46875000E-00	109.375000	15625000E-01	96875000E 00
99.687500	15625000E-01	48437500E-00	109.687500	15625000E-01	98437500E 00
100.000000	15625000E-01	50000000E 00	110.000000	15625000E-01	09999999E 01

Fig. 5 - Output Results from ELDIS1

DISTRIBUTION OF TOTAL RESISTANCE FOR CIRCUIT IN FIGURE 1.

RT = 1 / (1/R6 + 1/(R2+R3+R4)) + R1 + R5

MIN 3 24967742 MAX 3 30516129 MEAN 3 27714062 VARIANCE 3 12685449

FROM	TO	PROB	CUM PROB	1 - CUM PROB
3 249677	3 251411	-6 821859	-6 821859	999306
3 251411	3 253145	-4 515775	-4 523994	999254
3 253145	3 254879	-3 510416	-3 562815	998744
3 254879	3 256613	-2 206889	-2 263170	996675
3 256613	3 258347	-2 515458	-2 778628	991521
3 258347	3 260081	-2 985628	-1 176426	981664
3 260081	3 261814	-1 160524	-1 336950	965612
3 261814	3 263548	-1 232028	-1 568978	942409
3 263548	3 265282	-1 306794	-1 875771	911730
3 265282	3 267016	-1 381866	125764	873543
3 267016	3 268750	-1 456999	171464	827843
3 268750	3 270484	-1 532079	224672	774635
3 270484	3 272218	-1 607049	285377	713931
3 272218	3 273952	-1 678820	353259	646049
3 273952	3 275685	-1 736191	426878	572429
3 275685	3 277419	-1 766419	503520	495788
3 277419	3 279153	-1 764810	580001	419307
3 279153	3 280887	-1 731221	653123	346184
3 280887	3 282621	-1 671770	720300	279008
3 282621	3 284355	-1 599598	780259	219048
3 284355	3 286089	-1 524616	832721	166586
3 286089	3 287823	-1 449536	877675	121633
3 287823	3 289556	-1 374402	915115	-1 841923
3 289556	3 291290	-1 299330	945048	-1 542593
3 291290	3 293024	-1 224607	967509	-1 317986
3 293024	3 294758	-1 153616	982870	-1 164371
3 294758	3 296492	-2 931153	992182	-2 712553
3 296492	3 298226	-2 477582	996957	-2 234971
3 298226	3 299960	-2 185527	998813	-3 494443
3 299960	3 301693	-3 447483	999260	-4 469597
3 301693	3 303427	-4 461812	999306	-6 778556
3 303427	3 305161	-6 778556	999307	000000

Fig. 6 – Final Output Results

A QUANTITATIVE APPROACH TO THE EVALUATION OF SYSTEM RELIABILITY

V. Lasewicz and S. Newman,
Federal Aviation Agency,
Atlantic City, N.J.
and

H. Thomas, Federal Electric Corp.,
Paramus, N. J.

SUMMARY

System reliability prediction, of necessity, must extend further than summation calculations performed to determine an overall index based on the design excellence of its constituent components. In an electronic complex the operational factors concerning personnel, maintenance, complexity, technical support and logistic efficiency greatly overshadow the effects of good design reflected in the inherent reliability and maintainability.

Thus, an evaluation index encompassing a complete system is correctly called an Operational Reliability Index (ORI) since it reflects the combined influences of all the above factors. The approach under study by FAA for use in its nation wide Air Traffic Control Systems calls for a new concept of reliability, one which not only encompasses the conventional hardware design factors used in equipment reliability prediction methods (usually expressed in MTBF's) but one which also completely describes and evaluates operational performance.

The ORI, when finally calculated becomes a system outage time in hours per thousand hours of operation. The expression converted to other terms is a measure of the system's percentage availability.

When compositely derived, however, it includes design parameters, population effects, maintenance and logistic efficiencies, plus a number of important personnel factors. It also accounts for factors and failure rates which are intimately allied to the complexity of computer-type sections of equipment.

The algebraic expression used in calculating the index is shown semi-diagrammatically in Figure I. Here we see the major (and final) breakdown of the various factors when they are separated into a composite system (Kc) sensitive factor and a composite hardward (Kh) factor, so that the final stage in the calculation for a particular ORI is a function of their product.

The system factors are design parameters derived from engineering analysis of the redundancy, equipment back-up, switching disposition and importance weighting influences. Each of these, derived on a system basis, is less than unity so that when multiplied the final overall expression represents performance improvement by way of a smaller ORI.

The composite hardware factor derived from summating individual component "black box" factors is predominant, being, in turn, derived (in each component) from a dozen or more subfactors which are obtained from recorded outage times, failure replacement rates, complexity failure rates, preventive maintenance influences, logistic performance figures and personnel-related time expenditures.

To successfully implement the present system fairly large quantities of maintenance and operational data are generated by means of standarized failure report forms from recordings made by a trained group of engineers, technicians and repair men. This data, after editing, is processed by accounting machines, analyzed mathematically and statistically by an IBM 7090 computer and subjected to final piece part, circuit and equipment analysis.

Final evaluation from calculated ORI's although serving to operationally monitor the overall system's progress with time or under modifying conditions does not fully illustrate many other benefits of the mechanism. Several of its most potent advantages are: (1) to indicate deficiencies or gains in major component performances with charges in operational modes (2) to show the worth of sectional circuitry interrelationships and (3) to guide and evaluate the course and type of various maintenance devices and procedures.

INTRODUCTION

Objective. Air Traffic Control Systems require speedy and efficient mechanisms to handle a great volume of constantly changing information.

In processing this mass of flight plan data the routines are logically handled by automated and computer-oriented equipment; they must also be accurate, continually self recording and updated and adaptable to rapid, decision making situations when conflicts or other emergencies arise.

The approach used in the subject material is aimed towards developing a flexible and universal

performance monitoring mechanism for system evaluation. It was tailored to the complexities of modern solid state electronic equipment and is capable of producing a meaningful index of system performance. It includes assessment of inherent system and hardware design factors, maintenance features and all other operationally sensitive qualities - particularly those concerning personnel, logistics, etc.

In common with other electronic complexes the reliability of an Air Traffic Control System suffers from the ills of its own complexity. Any evaluation must account for such design factors and in addition must constantly account for input and output operations. System Reliability,in other words, becomes operationally oriented in concept where all contributing influences must be accounted for. Any system analysis therefore, must include every one of these factors inasmuch as they reflect the excellence of the basic design as well as operation, maintenance, and support.

BACKGROUND

Reliability analysis by itself has progressed thru evolutionary stages paralleling those of the communication arts. New approaches and techniques are appearing while analysis of component malfunctions become concerned with system failures involving increasingly complex configurations. Analytical emphasis has become logically oriented toward operational factors.

System-wise we must now interpret the "reliability of equipment operation" quite differently from the old style phase relating to "equipment reliability". The latter, as has been used in conventional application, expresses pure hardware design reliability in a percentage figure aimed to approximate the probability of an equipment's survival when each and every piece is operated under its individual environmental and operational conditions. This percentage is used purely as a predictive figure aimed to give a "ball park" estimate of longevity (or more commonly the overall mean-time-between-failures) of the composite equipment when calculated or synthesized by summating the failure rates of all constituent piece parts. It gives no indication of maintenance or support influences and furthermore does not account for the importance of individual piece parts, modules or components in their various circuit functions. Since it ignores individual electronic locations and circuit arrangements, neither does it account for redundancy and other factors which contribute to operational longevity. Another fallacy in using piece-part failure data for prediction is the processing of failure prediction calculations as if all elements were completely interdependent (or effectively connected in series) when many times they should be considered individually and, at least in assemblies, should bear weighted importance values derived in logical manner to indicate their varying contribution to true reliability.

The attitudes toward reliability have naturally changed with technological advances. Modern reliability concepts justifiably focus attention upon the increased importance of operation and support functions where the major expenditures in money and manpower are made on any project. This trend is apparent thruout military electronic hardware and becomes particularly evident where system operation and efficiency has become of prime importance in weapons, control, navigational and other complexes. The conventional piece part and dependent component reliability thus becomes only a small part of overall availability which we have chosen to call Operational Reliability.

The trend toward the system approach to reliability and the emphasis on operational factors becomes doubly significant because of three technological advances:

(1) Piece part reliability has shown starting improvements in the past few years.

(2) Solid state piece parts (notably transistors) have changed the whole physical and electrical makeup of communication and control hardware in both military and commercial fields.

(3) Computer equipment and techniques have become important parts to most electronic systems. Complexity has mushroomed to the extent that full control of operational performance and analytical monitoring must also be done by computers.

To express the trend more vividly we might observe in instrumentation centers that single "black box" equipments existing by themselves are decreasing in number. In the majority of cases they are associated with other boxes in a system where the harmonization of maintenance, installation and performance between all boxes must be complete and precise and where the support, above all, must be just as complete and precise.

In addition, the operational complexity of most systems is compounded by the fact that measurements in support equipment require a number of complex boxes which are in turn used to evaluate a single operational black-box. Spacewise this compounding becomes a deterrent to operational efficiency and in addition requires many more and higher skilled operators to master the intricacies of the precise measurements usually required. The changing character of maintenance and support personnel organizations reflects this seriously worsening problem since it is more than directly sensitive to the increasing complexity of functionally operating electronic systems. It results in the increasing shortage of high technically skilled service and maintenance technicians and in the inordinate amount of human effort required to service and maintain modern complexes.

More intensive training of support crews

offers some relief in this area although maintenance operations and costs are still high. The computer oriented approach offers a breakthru for here we utilize, in effect, the mechanization of automatic self test, high degrees of redundancy, and highly sophisticated maintenance techniques. These factors are particularly significant when applied to Data Processing and Display Systems. Such operations must be heavily automated, efficiently serviced by highly skilled technicians, and statistically monitored with sufficient accuracy and speed by a computer controlled data processing system.

THE CONCEPT

Conceptually the evaluation of system equipment operation has come to mean something far more fundamental than scrutinizing the pure design reliability referred to above. Performance-wise any system-wide evaluation now has implications which really express total overall availability by combining all variances in design and all factors concerning operation. In simple terms a system-wide evaluation must arrive at the basic quantities of "how often" a system malfunctions and "how long" malfunctions last per offending incident.

Such an assessment of complete reliable equipment operation thus reflects all factors either direct or remotely affecting the delivery of functional performance.

Such an approach has been planned and being executed in the reliability system analysis of various Air Traffic Control Systems now under evaluation by the FAA at their NAFEC installation in Atlantic City. It is embodied in an evaluation index known as the Operational Reliability Index (ORI).

EQUIPMENT and MAINTENANCE CONSIDERATIONS

Types of Equipment

In order to synthesize a universal workable evaluation index, the basic types of gear applicable to the data processing central were categorized inasmuch as maintenance and longevity characteristics vary between the various types of equipment.

Physically a Data Processing and Display System can be divided into four basic groups of equipment:

A - Those concerning computer type of switching circuitry.

B - Those chiefly concerning radars and display circuitry.

C - Those which are predominantly mechanical such as typewriters, punch and printers, etc.

D - Those concerning all other functional circuitry types such as class "A" amplification, power supplies, etc.

(A) - Computer Type Equipment. The failure pattern and the maintenance characteristics of computer hardware are dependent upon two interrelated characteristics:

(1) Computer operation is basically pulse operated inasmuch as the gates and flip-flop circuit are always either "on" or "off".

(2) A considerable percentage of failures or interruptions consist of single lost bits or small parts of messages which are not generally considered to be normal catastrophic or slippage failures as in other electronic gear.

Everything else being equal, the active elements in pulse operated circuits in general exhibit better longevity characteristics than those in amplifiers, detectors, converters, or other electronic circuits. This is inherent in the duty cycle factor which basically infers that pulse circuits average, for maximum utilization, about a 50% duty cycle. On this basis pure wear-out will logically occur later in a pulse circuitry component's life span.

In addition, pulse operation in diodes and transistors are less susceptible to transitional or marginal operating conditions which inherently exist within an analog type circuit, where every effort is being made, to attain maximum utilization of gain or detection. Furthermore the detection of failure in pulse operated circuits is usually a more definite and clean cut indication than the process of checking for the maximum condition in an analog characteristic.

By the very nature of a computer's internal operation, message interruptions are to be expected but at the same time are not generally considered physical equipment failures. This rather special phenomenon is called complexity failure in view of the great number, nature and rapid recurrence of the message bits concurrently being processed by a rather complex control operation. As inferred above, complexity failure many times is due to program deficiencies, errors or "bugs" which consume a minor amount of repair time and strictly speaking are not piece part failures.

Such complexity failures must, however, be scored in the overall evaluation and accordingly are appropriately handled in the index calculations to follow. It is apparent that they should not be treated the same as marginal failures, say in a high gain amplifier, where deterioration is apparent as a continuous, noise sensitive measurement.

(B) - Radar and Display Equipment. Failures in this area constitute another type of failure dependent upon the quality of the performance or excellence of the display being measured and evaluated. Thus, we have the factor of deterioration which must somehow be tied in with system evaluation which has been made on the basis of

pure failure rate. In the subject system, deterioration is recorded, entered and measured on a weighted basis where the summated amount of partial or sub-standard performance constitutes failure and may be included in the overall performance calculations.

More specifically the premise is made that when there is enough sub-normal performance such as continued low light level display of radar "blips" or intermittently high noise level, the performance should be demerited in terms of failure rate.

(C)- Mechanical Equipment. Mechanical message transducing or message transmitting devices make up a great portion of the peripheral equipment associated with a communication and/or display complex. These, in general, feed input and accept output from the actively engaged electronic data processing sections and consist chiefly of teletypewriters, Flidens and punch and printer units.

From a message transmittal standpoint the mechanical devices used by any Data Processing and Display System are of a complexity comparable to the typical computer section. Even though they are mechanical they are pulse operated devices since they handle message bursts involving millions of bits or pulses of energy.

By this token such mechanical devices are subject to a modified complexity failure type of evaluation. The modification consists of using the bit error rate recorded on a given machine and converting it to the common denominator of the equivalent system failure rate. Since message counters are standard equipment on all teletypewriters and allied devices, and since the message input and output records are available, the error rate determination is readily available for reporting and entering into the evaluation system.

(D) - Other Equipment Types. Other electronic system equipment types are common to most instrumentation complexes. They will be subject to conventional reporting, data-processing and analysis. This area includes communication equipment, switching consoles, power supplies, analog type amplifiers, recorders, etc.

FORMULATION

Requirements

The synthesis of a workable system performance index or Operational Reliability Index (ORI) must include all factors, variants, and conditions discussed above. To briefly summarize, such an index must account for:

(1) System sensitive factors concerning redundancy, automated test, self test, etc.

(2) Inherent Reliability, Maintainability and hardware design factors.

(3) Operational, logistic and personnel factors accompanying all support activities.

To explain operation and synthesis of the index, Figure I expresses in mathematical and diagramatic form the System Index equation. Here all factors and their interrelationships are shown so that the basic mechanism for entering failure data can be followed. Each type of factor will be discussed as to its application, justification for use, and expected variances as they appear in the evaluation process.

The Basic Equation

In order to segregate factors pertaining to overall performance from those pertaining only to hardware, two major factors, Kh (for hardware) and Kc (for Systems) are used for calculating the final index.

The equation below this level is synthesized so as to enter system-wide performance and failure data of every type and magnitude concerning design operational, and personnel related data.

The reported failure data at the lowest level is entered into sub-factors concerning testing (CMFt), catastrophic failures, self test, etc. In ascending importance these are summed into the second level of factors of Outage Time (OT), Total Failure Rate (TFR) Preventive Maintenance (PM) Sensitivity (Ks) and Redundancy (Kr). The first three of these is calculated for each component black box in a sub-system and then multiplied by one another to give Kh. Summation of individual Kh's yields the major hardware factor $\sum$ Kh. Likewise, the product of the redundancy and importance sub-factors at the various lower equipment levels can be summed to give the major system factor, Kc. As indicated the product of Kc x Kh = ORI which in final numbers gives a proportionate index of the outage time for the system per thousand hours of operation on a weighted population, importance and system configurational basis.

SYSTEMS FACTORS

The major factor Kc is compositely derived from Ks and Kr. It is a quantity, less than unity, derived from their product (both also less than unity) which indicates how much improvement results (in lower ORI) by the combined effects of redundancy and the weighting of black box sensitivity.

Kr, the Redundancy Factor in the first stages of failure analysis is a calculated quantity derived from the system network configuration where the sum total of all active and passive redundancy has been included.

"Active" redundancy means that switching action is instituted by choosing between circuits performing the same function and thus providing essentially complete performance thru substitution. Passive redundancy, using no switching,

produces partial or incomplete effects by utilizing permanently installed parallel components or alternate paths. It generally produces less reliability than active redundancy altho trade-off studies relative to poorer performance versus lower initial cost may lead to choosing it in some cases.

The calculation for Kr includes a summing of the redundancy effects within component black boxes. This study in turn can extend into evaluation of partial redundancy within individual circuits.

The constant Kr, as established by overall network analysis is a modifying factor applied to components within the system to indicate true effects of circuit arrangements on the overall ORI. In this sense it is predictive in nature and in final analysis must be firmly established from comparison experiments using failure data with and without the redundant provisions.

Ks, the Sensitivity Factor, is a composite factor derived from logical analysis of each component black box on the basis of its failure importance. This factor is derived from system analysis aiming to apply weighting factors to each of the components with respect to the effect that their failures produce in operation of the overall system.

Ks, like Kr, is established by detailed circuit and functional study applied to each and all components with respect to each others' contribution to system efficiency. The composite Ks is predictive in early developmental stages and will be used to guide failure analysis of individual as well as overall hardware factors. Initially it serves to indicate allocations or proportionment of efficiencies within the various system areas. Experiments to verify the validity of the importance analysis area logical development once sufficient and stable failure data is available from system operations.

A number of other experiments involving both Kr and Ks can be conducted for checking various sub-system interrelationships and configurations. These determine such information as to where preventive maintenance is most advantageously applied, the best sequence of signal processing operations, the efficiency contribution of redundancy in the various importance zones, etc.

HARDWARE FACTORS

Basic Relationships between Major Factors

Overall Kh. The sum total of all hardware reliability performance, calculated at the system level, appears in the major factor Σ Kh. The individual Kh's describing single component black box performance are really where first significant comparisons are made from the directly reported failure data. It is this analysis area where all detailed hardware design, maintenance procedures, support inadequacies and personnel deterrents crystallize into performance figures.

Individual Kh figures, therefore must meet, in particular, the requirements set forth in the fundamental desiderata described in the introduction. Excluding, for the time, the preventive maintenance factor, Kpm, we see that TFR, the basis Total Failure Rate and OT, the Outage Time or Serviceability Factor for Corrective Maintenance, must combine to completely account for the following:

(1) Population differences

(2) Maintenance design factors

(3) Availability design factors

(4) True catastrophic failure rates

(5) Wearout factors

(6) Maintenance cycle factors concerning testing, repair, logistics, etc.

(7) Different types of failures and personnel influences reflected in the maintenance cycle.

TFR and OT must therefore both be partially sensitive to operating cycles and to population, but at the same time not include the rate of malfunctioning twice in the same expression since the two factors are multiplied to obtain final Kh's. The following breakdown and study of relationships within each factor's make-up discusses these influences in synthesizing the overall ORI.

Overall Outage Time

Inasmuch as OT is a major factor combining system performance with maintenance oriented conditions it is assigned as a measuring unit, the caption of hours per 1000 piece complexity per thousand hours of operation.

OT thus accounts, in part, for conventional reliability by combining a standardized population and a fixed number of operating hours with the maintenance cycle hours necessary to preserve operation. The OT is thus the "how long" factor applicable to an evaluated equipment. Note that the number and consequently the rate of failures are not included. The procedure thus reduces the measurements yardstick of outage hours to a common complexity denominator so that it applies equitably to different equipments having different populations. Past records show this quantity on ground equipment to vary from one hour to thirty hours, depending chiefly upon the depth of the support plus the nature and dispersement of personnel activity. In older electronic gear the factor does not vary greatly between radar, communication or navigational types of equipment. In modern transistorized computer-type equipment,

the figure is very low due to: (1) the high reliability of the solid state piece parts (2) to the high degree of modularization (which is, in essence, improved maintainability design) and (3) to the inherently high sophistication of the built in self-checking features.

The various sub-factors contributing to overall outage time are discussed in detail below.

Overall Total Failure Rate

The use of a rate for the TFR factor tells "how often" failures occur per 1000 hours and also arranges quantities mathematically so that when multiplied by "how long" in each thousand hours of operating time the result gives a total of hours down for a piece of equipment. TFR, as indicated is made up of two parts, normal Failure Replacement Rates, (FRR) and Complexity Failure Rates (CFR) described below:

The question now arises, how can the added proviso of "per thousand piece complexity be justified and not be included twice in the TFR. This is accomplished by minimizing population effects and as discussed above under equipment characteristics, by invoking a relatively new principle applicable to computer-type equipment. To repeat we include a type of failure known as complexity failures which are differentiated from the conventional failures encountered catastrophically, by wear out, association, etc., namely those producing the FRR.

Complexity Failures and their attendant rate (CFR) are peculiar to high population, computer oriented equipment. They may be defined as stoppages, interruptions, or incidents which are dependent upon an equipment's complex nature and which are not particularly related to or caused by piece part failure and above all which are readily remedied without the conventional maintenance cycle of diagnosis, repair and checkout.

Such failures usually occur in the form of missed message bits, jam incidents on mechanisms, such as teletypewriters, incorrect printouts, extra message data on tapes, erratic or sporadic performance in limited areas, such as on displays, control consoles, readers, etc.

These failures may be due to incorrect programming, sporadic noise, obsecure aberrations in logic and from many conditions which the maintenance or debugging personnel have no control over and which they may never have opportunity to solve since the incidents seldom reoccur consistently in consecutive program or maintenance re-runs. Thus we see that complexity failure rates are not closely related to population, but rather to simplicity and excellence of design. In another respect they are intimately related to preventive maintenance operations, although here the relationship is not rock-bound. In general, however, complexity failures should decrease as the amount of PM effort is increased.

Logically then, complexity failures being equipment and design deficiencies must be scored against performance. Since, however, they do not relate directly to outage time their place is rightfully under TFR and in combination with conventional failure rates (FRR's). In other words for computer-like devices they reflect excellence of inherent hardware and soft-ware design thru the care and quality applied to routine programming and to diagnostic preventive maintenance programming which is directly allied to it.

In non-computer-like devices the influence of complexity failures on the IFR will be small, the major portion being contributed by normally reported replacement failures (FRR's) which will in this case be determined by the conventional replacements in their various categories as discussed below.

Major Factor and Sub Factor Breakdown

The Serviceability Factor (OT). Continuing the overall discussion we see that OT is a partial link between reliability and maintainability since, as an overall index of the time consumed in the maintenance cycle, it places equipment down time on a normalized basis of hours per thousand piece-part population per 1000 hours of operation.

In the further breakdown of the maintenance cycle, recorded data is readily obtained allowing evaluation of the various sub-factors within the cycle.

OT (in hours) thus includes all corrective maintenance (CM) activities being made up of the proportionate part contributed by the following sub-factors:

CMFt - Testing

CMFr - Repair

CMFwp - Waiting and Personnel

We can thus directly obtain from failure reporting data on individual pieces of hardware an index of the excellence of maintainability design, the quality of the maintenance technician, the completeness of their training, facts concerning test procedures, facets of test equipment evaluation and finally the efficiency of the logistic and support organization.

Note that since logistic efficiency is chiefly reflected in waiting time the sub-factor CMFwp includes all other personnel factors which like-wise reflect themselves in waiting time.

In the actual processed data, each of these factors appears as a number of definitely reported hours altho for more penetrating analysis each factor is further broken down from recorded data into secondary sub-factors or percentage parts as discussed further on.

WRH

By calculation the total OT is calculated from their sum.

Thus:

OT (hrs) = CMFt + CMFr + CMFwp

The final breakdown of sub-factors serves two purposes when operated by means of recorded ratios. (1) It enables the formation of direct equipment performance analysis and (2) it gives a basis for stabilizing computer processed data thru recursive calculations when updating records or comparing to similar equipment.

CMFt - Test Factor. The test factor portion of the Outage Time Factor on any black box is a directly obtainable number of hours when extracted thru the data processing mechanism. Altho testing, in effect, consumes a number of steps, the major portion of testing time is spent in two sections of the test operation. These two parts are: (1) the isolation portion of diagnostic testing and (2) the checkout and alignment part. For simplicity these secondary sub-factors are reported as percentages of the total test time and are converted to actual hours by simple multiplication. In either form, these quantities aid in analysis of equipment performance. Too much diagnostic test time may mean, among other things, inadequate test equipment or again poorly trained technicians. Comparative analyses of this nature are used to show the benefits of modularization, or the effective operator confidence levels due to automatic or self test.

CMFr - Repair Factor. This factor is the total recorded time in hours spent in completing the repair on any hardware component. As part of the maintenance cycle it consists chiefly of time spent in (1) assembly and sub-assembly and (2) time spent in actual piece-part replacement.

When extracted for analysis this factor is available in either actual hours or on a percentage basis. As a ratio these figures are useful in evaluating maintainability design, in setting standards and comparisons of personnel effectiveness or in the analysis of manpower utilization.

CMFwp.- Waiting and Personnel Factor. Recorded outage time in this category is aimed to give an estimate of all personnel-related support time plus abnormal procurement intervals. We thus aim to extract logistic inefficiencies, moving or preparing equipment, reporting, consultation, and other unforseen delays.

TFR - Total Failure Rate Factor. As noted in preliminary discussion this major factor is an index of "how often" the system stops per thousand hours of operation from the combined effects of complexity and all other types of failures. It is thus a composite rate made up of the sum of Failure Replacement Rates plus Complexity Failure Rates.

The Failure-Replacement Rate (FRR). From the classical reliability standpoint this part of TFR is not an indication of "pure" catastrophic part failure for it includes replacements indicated by the secondary sub-factors caused by performance deterioration, wear-out, associated failures and other causes such as injury, misadjustment, handling, etc.

FRR is accordingly equal to a summation of the individual replacement rates as determined by recorded failure data on complete incidents involving the expenditure of repair, test and waiting time.

These factors, in rough approximation, a derived from typical ground electronic equipment make up an overall failure replacement rate according to the following percentages:

Catastrophic - 32%
(True Random)

Slippage - 36%
(Performance Deterioration)

Wear Out - 12%

Associated (Due to adjacent or interconnected parts) 10%

Injury, Physical Usage - 10%
Miscellaneous, Unknown

Each of these rates can, in one way or another be extracted or partially indicated from recorded coding on failure report forms. Special or controlled coding and analysis can be instituted and used in specific areas demanding investigation. A brief discussion of all individual factors is given below.

FRc - Catastrophic Failure Rates. These rates are based on true piece part random failure quantities utilizing the various equipment populations and accounting for all conditions relating to environment, derating, etc. Such rates become useful in design and piece-part evaluation when converted or compared to the Mean Time Between Failure indices developed by prediction calculations.

FRw - Wear Out Failure Rates. This type of failure is allied to the catastrophic type and must be extracted from it via coded report data. In this area the extraction process must be particularly selective since wearout in some parts is allied to pure catastrophic failure while at the same time being produced by symptoms that resemble slippage or deterioration.

FRs - Performance Slippage or Deterioration Failure Rates. This category reflects a non-homogeneous combination of wear-out complexity and equipment performance factors. Cross correlation of these factors is best attained by selective programming and other types of analysis aimed at penetrating the composite failure data. This rate

is an important one and it has a number variants depending upon the type of equipment, performance levels, accuracy of test procedures, etc.

FRa - Associated Failure Rates. These rates concern secondary failures caused or related to true primary failures. Their contribution to the overall FRR is minor although in some cases the failure analysis derives vital clues from this region.

FRm - Miscellaneous Failure Rates. This category includes all miscellaneous failure rates not included above. Most of these can be extracted by selective coding or in the editing. They are made up of failures due to handling, injury, shipping, storage, accident, environmental and physical usage effects.

Complexity Failure Rate (CFR). This failure rate is a composite of two predominant kinds of purely interruptive failures, namely, jamming failures (FRj) and parity errors (FRp). On the established thousand hour basis, the sum of these two rates equals the total CFR.

Jamming Failure Rate (FRj). This type of failure is particularly applicable to mechanical devices such as teletypewriters, punch and printers, etc. Such errors, although inherently mechanical can be in the case of the punch and printers the result of incomplete or erroneous data electronically generated by the computer or other associated equipment. Recorded failures of this type appear on the time-log sheet and are ultimately entered into the data processed information.

Parity Error Failure Rate (FRp). This 1000 hour rate of electronic type interruptions in the computer operation is a generalized indication of a number of design, personnel, maintenance and installation factors. Such failures can arise from electrical interference within digital information transmissions, operator errors, incorrect or obscure programming routines, or incomplete preventive maintenance diagnostic programming.
TFR Summary - The ultimate TFR of a given hardware assembly is thus a composite rate (FRR + CFR) both of which in turn are composites of their own contributing or secondary sub-rates.

TFR = FRR + CFR

= (FRc + FRw + FRs + FRa +FRm) +

(FRj + FRp)

The Preventive Maintenance Factor (Kpm)

This factor is an improvement quantity, which is numerically less than unity. When multiplied by the product of OT and TFR it shows how effective the self test and preventive maintenance procedures are in reducing the ultimate outage.

The determination of Kpm may be by circuit and system studies of test procedures and schedules when applied to self test and automatic evaluation devices, or by performance measurement of the effects of PM action or corrective maintenance results when taken over selected operational intervals.

The degree to which diagnostic programming affects overall efficiency yields another evaluation device aiming to establish the ultimate Kpm.

Sample Calculation

The following example demonstrates with typical values for the various constants the mechanism of calculating an ORI.

Given a system with the following constants:

Kr = .7 Redundancy and backup equipment studies show a net system improvement to the extent that when fully operating these devices reduce the equipment down time by 30%.

Ks = .8 Studies arriving at proper functional weighting and disposition of all sections of the system indicate that the inherent reliability failure rate can be battered by 20%.

Kc thus = .7 x .8 = .56

Now the total failure rate of this equipment (TFR) is once every 100 hours which includes all stoppages whether due to replacement failures or complexity reasons. Early in the equipments history this might run on a 70-30 % split between replacement and complexity failures.

Converting this rate (.01 per hour) to 10 failures per 1000 hours on a 70-30 % split:

TFR = 10/1000 hrs.

FRR =7
- Catastrophic (FRc) = 3
- Wearout (FRw) = 1
- Slippage (FRs) = 1
- Assoc. (FRa) = 1
- Miscellaneous(FRm) = 1

CFR=3
- Jam s(FRj)=1
- Parity(FRp)=2

Next the outage time (OT) for this equipment has been recorded at 30 hrs. per 1000 parts per 1000 hours operation. OT = 30. The breakdown for this time is:

Testing - (CMFt) = 4 hrs.

Repair - (CMFr) = 1 hrs.

Waiting and Personnel - (CMFwp) = 25 hrs.

Now experiment has proved that a full

Preventive maintenance program saturates at a point where outage time is reduced 30%. Kpm thus = .7

So for hardware the system factors are:

$$\sum Kh = Kpm \times TFR \times OT$$

$$= .7 \times 10 \times 30$$

$$= 210.\ \text{hrs.}$$

$$ORI = Kh \times Kc = .56 \times 210 = 117.6\ \text{hrs.}$$

The circuit availability is:

$$A = \frac{1000 - 117.6}{1000} = \frac{882.4}{1000} = .882$$

IMPLEMENTATION

The mechanism for evaluating a data processing and Display System is based on entering field failure data concerning time, incident and other information into a data processing system so that it can be mathematically combined and manipulated to deliver significant printed, operational, reliability, and maintainability quantities including the overall Operational Reliability Index described above.

Three groups or divisions of activity properly operate within the system for meaningful results. They are:

(1) A data collecting group

(2) A data processing group utilizing an IBM 7090 computer

(3) An engineering analysis group

The flow of data within the data collection and processing areas is shown in Figure IV. Engineering analysis proceeds in several directions in interpreting the computer printouts.

Data Collection

An incident failure reporting form and a combined incident and timelog form are the basis of information collection (See Fig. II and Fig. III). Upon these are entered all failure information in direct or coded form enabling entries to be made under all of the factors within the ORI. Suitable coding augments the direct entries so as to give complete information on every facet of a malfunction with a minimum of technician effort and confusion in recording the data.

Data Processing

Failure data from report forms is edited and then punched on three separate IBM card decks upon which are based the various stored analytical programs.

The information from these decks is sorted and processed under three categories concerning general information, system information and piece-part data. These groupings determine the final printouts basically concerned with hardware failure rates, outage time, and the various facets of system performance.

Some typical printouts plus a few special ones appear on the following list:

(1) Failure rates at all hardware levels (sub-system, unit, assembly, sub-assembly, card, etc.).

(2) Piece part failure rates, (resistors, capacitors, tubes, transistors, etc.)

(3) Outage times at all hardware levels

(4) PM and CM activity comparison at different hardware levels

(5) Outage time analysis based on repair, test, waiting time, components, etc.

(6) Comparative modularization evaluations

(7) Logistic efficiency analysis (spare part provisioning study)

(8) PM efficiency

Engineering Analysis

This activity concerns interpretation and statements covering the true meaning of the data processed quantities. Specifically the printouts must be converted or otherwise manipulated so that measured performance can be compared to specifications, so that failure rates can be evaluated against predicted values or so that actual maintenance costs can be compared to budget figures.

The engineering analysis explores all phases of system operation and is an activity encompassing the highest system levels down to such details as reliable circuit design or the specific derating on resistors.

Organization of the engineering analysis work follows the general functional layout of the equipment itself.

Conclusion:

The evaluation of complex electronic systems is practical and feasible when the failure reporting and analysis is carefully planned, correctly organized, and efficiently operated. Performance monitoring by such a reporting mechanism must penetrate every phase of system operation and hardware design to sufficient depths for developing a quantitative system-wide index of Operational Reliability. Such an index offers major indications of the electronic and physical health of a complex under a wide range of conditions. It also delivers details concerning

electrical performances and support conditions applying to its constituent parts at every level of operation and maintenance.

NOTE:

The methods and techniques described in the subject material was an approach suggested by three individuals in the course of a study program at NAFEC. No specific indorsement of the approach is made nor is it one that will necessarily be used by FAA.

REFERENCES

1. "Reliability Stress Analysis for Electronic Equipment," TR1100, PB131678 - RCA-1956.

2. "Reliability Stress Analysis for Electronic Equipment," TR59-416-1 - RCA-1959.

3. "Maintainability Measurement and Prediction," RADC-TN-60-221 - September, 1960.

4. "Reliability Requirements for Ground Electronic Equipment," RADC-2629 - October, 1958.

FIGURE I

SYSTEM

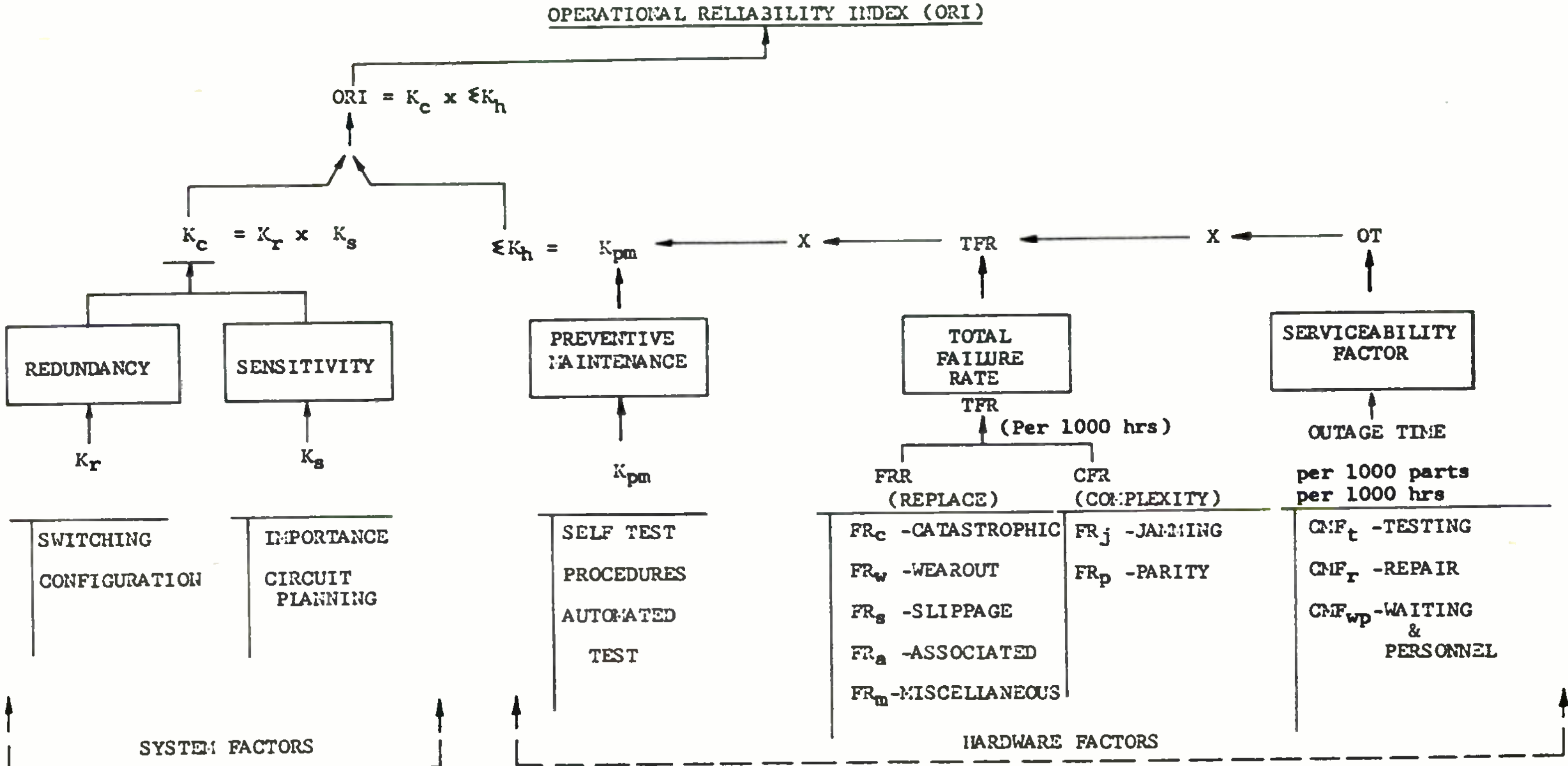


FIGURE II

FAA/ARDS ATC SYSTEMS BRANCH

INCIDENT REPORT

No. ____________

A	B	C	D	E	F	G	H	I	J	K
		LOG. INC. NO.	TIME FAILURE OCCURRED: MO. | DAY | YR. | HRS. | MIN.	TIME FAILURE CLEARED: MO. | DAY | YR. | HRS. | MIN.	TYPE OPER. (TASK)	GEN. TYP. FAIL	FAIL IND.	GROUP LOC.	GROUP SERIAL NO.	
1	2	3 4	5 6 7 8 9 10 11 12 13	14 15 16 17 18 19 20 21 22	23 24	25	26 27	28 29	30 31 32 33 34 35	36 37 38 39 40

L	M	N	O	P	Q	R	S
UNIT	UNIT SERIAL NO.	ASSY.	ASSY. SERIAL NO.	SUB. ASSY.	SUB ASSY. SERIAL NO.	TIME CODE: GEN. | DIAG. | WAIT. | REP.	
41 42	43 44 45 46 47 48	49 50	51 52 53 54 55 56	57 58 59	60 61 62 63 64 65	66 67 68 69	70 71 72 73 74 75 76 77 78 79 80

T	U	V	W	X	Y	Z	AA	BB	CC
	PART SYMBOL	PART TYPE NO.	FAIL. TYPE	FAIL. CAUSE	ASSOC. FAIL.	GROUP LOC.	DIAG. TIME HRS. | MIN.	WAITING TIME HRS.	REPAIR TIME HRS. | MIN.
1 2 3 4 5 6 7 8 9	10 11 12 13 14	15 16 17 18 19 20 21	22 23	24 25	26 27	28 29	30 31 32 33	34 35 36	37 38 39 40
			TOTAL TIME						

REMARKS

Date ____________

Signature ____________

FIGURE III

FAA/ARDS ATC SYSTEMS BRANCH - OPERATIONS LOG

SUB SYSTEM #________ GROUP________ GROUP SERIAL #________ DATE ________

TIME START	TIME STOP	TASK	OPERATING EFFICIENCY	PROGRAM I. D.	R. T. M.	OPER. CTR.	INCIDENT NUMBER	FAILURE INDICATION	EQUIP. AREA AFFECTED	COMMENTS	INITIALS OR OPERATOR NUMBER

GROUP LEADER SIGNATURE DATE

FIGURE IV

FLOW DIAGRAM FOR RELIABILITY IMPLEMENTATION

SOME ASPECTS OF TEST EQUIPMENT RELIABILITY

Frank A. Applegate
Light Military Electronics Department
General Electric Company
Utica, New York

Summary

Today's highly complex electronic systems have resulted in increased use of automatic test equipment for which electronic system customers require a simple means of determining reliability requirements. The reliability of such equipment is shown to depend upon both the desired availability and the desired reliability during check-out periods. The test equipment reliability requirement is actually a trade-off between the two.

A question often posed by electronic system customers is how large a system can be, or how many systems can be maintained by one test equipment. A solution based on a possible waiting line is presented.

Introduction

The growing use of highly complex electronic systems has necessitated the development of equally complex automatic test equipment. Electronic systems using automatic test equipment are represented diagramatically by Figure 1. A small number of EUT's are shown for problem simplification.

A system such as that represented in Figure 1 has a central test equipment, the Programmer/Evaluator, which is used to test or check out various electronic units through an adapter. This adapter, which permits the unit and the Programmer/Evaluator to communicate, is available for each type of unit to be tested.

Naturally, the customer for a new electronic system desires to know the test equipment reliability requirement. The electronic system reliability is usually dependent upon a weapon system success probability and is often found, quite simply, by ignoring down time. It is sometimes reasoned--and rightly so--that test equipment reliability is a logistics problem, for the availability of the weapon system can easily be increased by adding more weapons. This solution, however, is expensive and still leaves the problem of the test equipment reliability requirement. In the end this requirement is quite often determined arbitrarily and may end up as more severe or less severe than the reliability requirement for the electronic equipment. It seems in order to seek a uniform approach.

Another problem often confronting the user of electronic systems is the determination of the number of test equipments needed to maintain the electronic system or, alternately, the number of systems which can be maintained by one check-out equipment. Because these two problems are related, a joint solution is plausible.

Required Test Equipment Reliability

To obtain a better appreciation of the test equipment reliability problem, consider the employment of test equipment, which generally must be available for use at two separate periods of time. One is the maintenance check-out period prior to operation, and the other is at some time following a system failure.

Because the desired time of operation is variable, it is essential that the test equipment be ready to operate at all times. Availability is usually found from the expression:

$$A = \frac{MTBF}{MTBF + MTTR} \qquad (1)$$

where,

A is the availability
MTBF is the mean-time-between-failure
MTTR is the mean-time-to-repair

We are seeking the parameter MTBF, but, at the same time, we want A to have a reasonably high value. A is maximum when MTBF is maximum, but both are asymptotic and, therefore, cannot be maximized. MTTR is relatively constant for test equipment and can be estimated quite accurately.

The problem can now be approached in two ways. Some criteria can be derived to fix A, or a means can be found to fix MTBF. For this study it was decided to fix MTBF.

It is of utmost importance that test equipment operate properly during test or check-out periods. This reliability is even more important than availability for test.

Reliability is usually found from the expression:

$$R = e^{-t/MTBF} \qquad (2)$$

where

R is the reliability for time t

t is the time interval of interest (check-out period)

MTBF is the mean-time-between-failure

e is the base of natural or naperian logarithms

Reasonably good data exist on check-out periods for various types of equipment. Check-out periods, like MTTR, are relatively inflexible and, hence, estimates are quite accurate.

Using the estimates of check-out periods, arbitrary reliability can be entered into the expression to calculate the required MTBF. Figure 2 is designed to facilitate such calculations. Read the graph at the bottom for the desired check-out period (for example, the longest check-out period in the system). Read up the scale to the desired reliability for that particular check-out period. Then read to the left to find the MTBF required of the test equipment which will give the desired reliability during the check-out period used.

A required MTBF is now known, but before accepting this as final the availability should be determined. Note that the availability as defined is an availability for operation of the test equipment and does not consider that the test equipment may be in use.

Figure 3 displays the availability expression given by equation (1). This set of curves may be used either to determine the availability after the MTBF is found or to find the required MTBF by entering the desired availability, in the event that the user feels availability to be the deciding criterion.

By using Figures 2 and 3 together, it is a simple matter to find an MTBF which provides satisfactory reliability and availability for the desired check-out period. Admittedly, the solution is somewhat arbitrary, but this procedure does consider the system parameters.

The availability expression used in this study considers only two states--operating and failed. For items such as test equipment a third state--non-operating--must be considered. While the equipment is non-operating, the failure rate will be drastically reduced, resulting in a higher availability during non-operating periods. This study does not consider such a case, but work of this nature has been done.[3]

Test Equipment Employment

So far the reliability requirement for a proposed test equipment has been determined by estimating some system parameters. A further problem is how many of the proposed test equipments will be needed for field employment. To determine this answer, estimates of still more system parameters are needed.

As indicated earlier, the test equipment must operate at two specific periods of time. The first is at some time following a failure of any equipment which the test equipment checks out. This then relates demand on the test equipment to system failure rates. For example, if a piece of equipment fails once per hour, there will be a demand on the test equipment once per hour.

The second demand is placed on test equipment by the maintenance policy. For example, a piece of equipment may undergo a maintenance check out once each 24 hours to assure satisfactory performance. Studies have been performed to determine check-out intervals precisely.[1]

Each demand entails a finite holding time when the equipment under test and the test equipment itself are unavailable.

The problem now appears simply as a question of queuing theory. Several texts and many articles are available for the solution of such problems under various distribution assumptions of input and holding times.[2] The basic solution is to determine the probability of a waiting line and then to regulate the demand on the test equipment so that this probability is kept below a specified level.

Using queuing theory for arbitrary input and holding time distributions, the probability of a queue greater than zero can be calculated by:

$$P(n > 0) = m/M \qquad (3)$$

where

m = mean number of inputs per unit time

M = mean number of outputs per unit time from an occupied P/E (reciprocal of holding time)

n = queue length (includes item being tested)

Figure 4 can be used for some approximate solutions of equation (3). From equation (3), or Figure 4 the answer can be obtained.

Since the test equipment itself can fail, the probability of a queue forming differs according to the "up" and "down" periods of the test equipment. Therefore, Figure 4 must be checked to find the probability of a waiting line while the test equipment is operating. When the test equipment is down, the probability of a waiting line is merely the probability of a demand because any demand constitutes a waiting line. The two probabilities are then weighed by the availability and unavailability respectively. These

weighted probabilities indicate the probability of a waiting line's forming.

Figures 5 through 12 show the probability of a waiting line for various inputs, test or holding times, availability, unavailability (with its associated down time), and the level of availability. Instead of making separate curves for input from failure rate and maintenance, one input suitable for either one was used.

Mention should be made that the terms "queue" and "waiting line" are used interchangeably in this study. Although this is not technically correct, it is insignificant when considered with such things as failure rate estimates, which are imprecise themselves. In the present study the probability of a queue being greater than zero was found. Some finite length could be used although the computational effort would have increased. In short, the accuracy of the present method should be adequate for all practical applications.

Example

Taking the hypothetical system in Figure 1, assume that the estimates shown in Table I are available. Although the figures have been chosen arbitrarily, they are representative. For the sake of simplicity, the following assumptions will also be made.

1. Any failure is a total failure
2. There is only one fault per incident
3. No time is lost in switching between adapters
4. No testing except when failure is suspected
5. MTTR for P/E is estimated to be four hours

Problem - What MTBF is required in the P/E to operate either an adapter or an adapter plus EUT whenever they require a test?

To serve its function, the P/E must be available at some instant following each failure of either an adapter or EUT. As Table I shows this will occur on the average of .0157 failures per hour or about one failure every 64 hours.

Table I shows the P/E operating time to be .9 hours for EUT4. This is the maximum P/E "tie-up time" and will therefore provide a safety factor in our calculation of the required P/E mean-time-between-failure. A more accurate approach would be one using a distribtuion of P/E tie-up time.

Using Figure 2 with a test time of .9 hours, an arbitrary reliability of .99 shows a required mean-time-between-failure cf about 90 hours. We now have a situation where a P/E with a mean-time-between-failure of 90 hours will test successfully 99 times out of a hundred an EUT requiring a .9 hour test. This seems to be a reasonable reliability without a prohibitive mean-time-between-failure requirement.

Any desired reliability could be used instead of .99, but notice that a .999 reliability would require an MTBF of 870 hours. This would be an extremely difficult requirement for a test equipment as complex as that required for the system described. In the final analysis, the higher reliability most likely would cost more than another P/E. If such a requirement is needed, therefore, the best approach might be a standby-redundant test equipment.

Figure 3 shows the availability of the P/E to be roughly .953. Remember, however, that the availability used in this study is pessimistic for test equipment. The availability of .953 applies only to periods of test equipment operation, which occur approximately once every 64 hours and last for about one hour. A second availability is required for the remaining time when the P/E is non-operating. Very little data exists on non-operating failure rates, but assuming these to be one-tenth of the operating failures rates, the availability during the non-operating period increases to .995. Since the P/E is non-operating most of the time, actual availability is close to .995.

In the event that this availability is considered to be too low, the MTBF can be increased accordingly. The final decision must be based on user requirements.

With the MTBF requirement established at 90 hours, a decision can now be made as to the operational load for the test equipment. As was indicated previously, the approach is to find the probability of a queue greater than zero.

Using the availability estimate of roughly .95, examine Figures 7 and 8. From Table I the system failure rate is found to be .016. As the operational load for the test equipment is increased, each new system brings an increase in failure rate of .012. (Additional adapters will not be required). For simplification the worst case test time of .9 hour is used.

Consider first the case where the only check out is after a failure. Entering Figure 7 with a test time of .9 hour, the probability of a queue greater than zero is about .09, with a system failure rate of .1. Figure 8 shows a probability of .02 for the same failure rate and a P/E down time of four hours. The probability of a queue greater than zero, therefore, is .11 for a P/E having availability .95

and MTTR of four hours with a system having a failure rate of .1. It can be shown then, that some seven systems can be checked out by this P/E.

Consider now the case of maintenance check out. Assume we have a maintenance rate of .1. With seven systems similar to Figure 1, this is equivalent to a maintenance check-out rate of about once every 13 days per equipment. With this check-out rate and the check out following failure, the probability of a queue greater than zero is roughly .22.

This probability may seem excessive and at the same time the rate of maintenance check out may seem insufficient. If this is the case, the number of systems maintained by the P/E must be varied accordingly. But in any event the curves do present a simple and rapid solution to the questions posed in the example.

Conclusion

This paper has developed a rationale for determining test equipment reliability requirements based on such usage factors as availability and check-out time.

The discussion includes a study of test equipment in relation to demand rates, reliability, check-out time and waiting lines. A method is presented for arriving at an employment scheme for a proposed test equipment.

No attempt has been made to optimize any of these criteria although this would be the next logical step. Optimum reliability and use, for example, could be determined in reference to cost. The method itself could be refined to allow for distributions of demand and check-out time as they exist in an actual situation. Even without such refinements, however, the present approach provides effective, workable solutions.

Acknowledgment

The author is indebted to George Siegel, Light Military Electronics Department, for posing the problem and encouraging a solution.

References

1. Kamins, M., "Determining Checkout Intervals for Systems Subject to Random Failures", The Rand Corpration, Research Memorandum RM-2578, 15 June 1960.

2. Morse, P. M., "Queues, Inventories, and Maintenance", John Wiley & Sons, 1958.

3. Hosford, John E., "Measures of Dependability", Operations Research, Vol. 8 #1, p. 53, Jan-Feb 1960.

TABLE I

Item	No. of Components	Failure/ Hour	MTBF Hrs.	MTTR Hrs.	Fault Isolation Time (P/E Tie-up)
EUT_1	25,000	.0025	400	2.4	0.3
EUT_2	15,000	.0015	700	2.0	0.3
EUT_3	5,000	.0050	200	3.0	0.4
EUT_4	3,000	.0030	300	5.0	0.9
EUT	48,000	.0120	83		
A_1	20,000	.0020	500	1.0	Neg.
A_2	10,000	.0010	1000	0.8	Neg.
A_3	5,000	.0005	2000	0.6	Neg.
A_4	2,000	.0002	5000	0.6	Neg.
A	37,000	.0037	270		
Total	85,000	.0157	64		

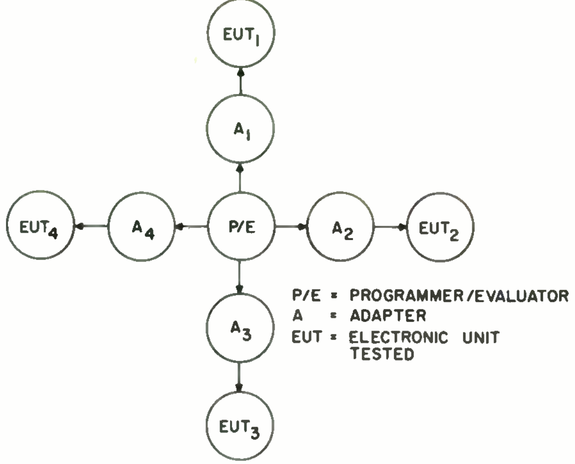


Fig. 1. Electronic equipment under test.

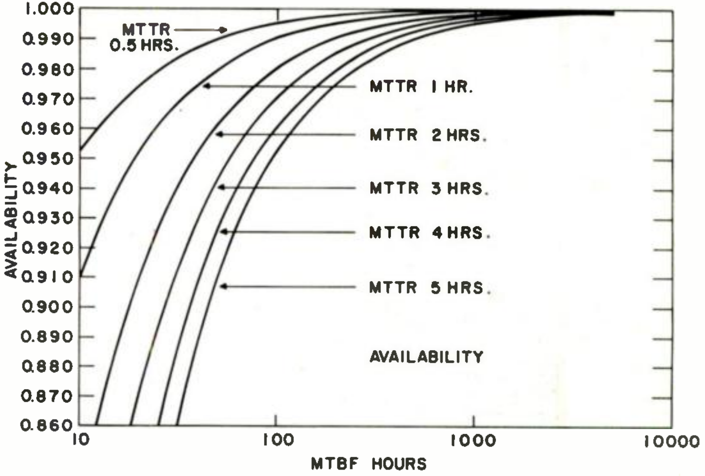


Fig. 3. The availability expression.

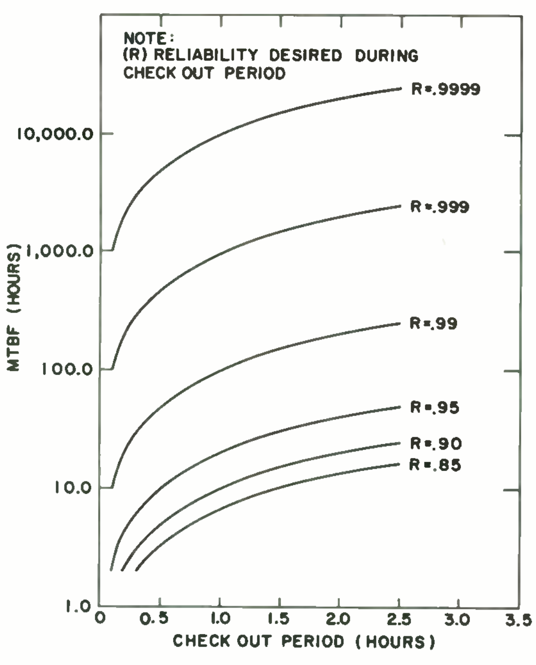


Fig. 2. MTBF as a function of check out period.

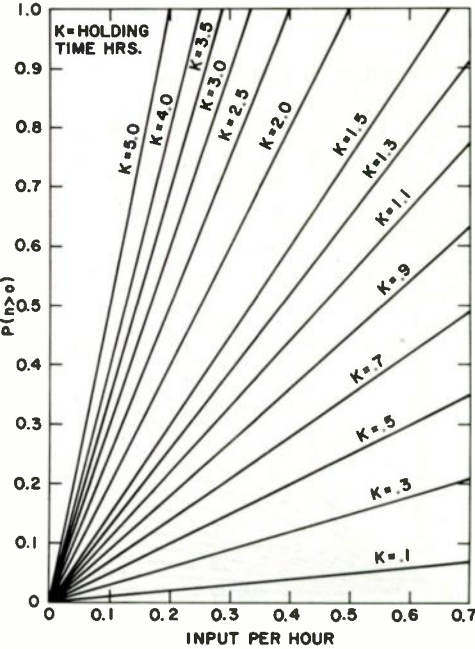


Fig. 4. Probability of a waiting line.

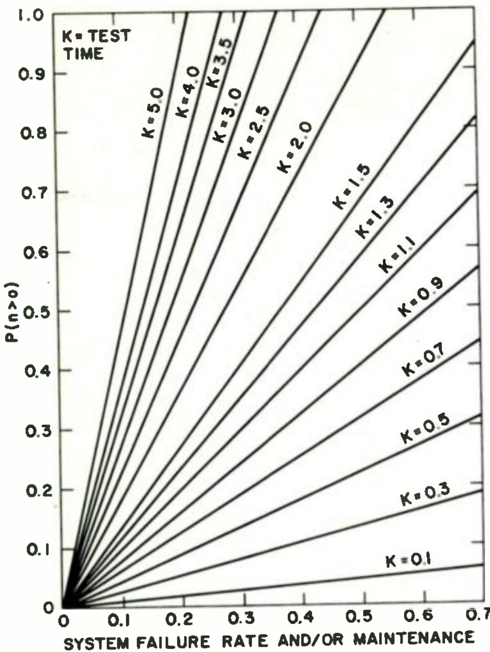


Fig. 5. Probability of a waiting line P/E operating (availability 0.90).

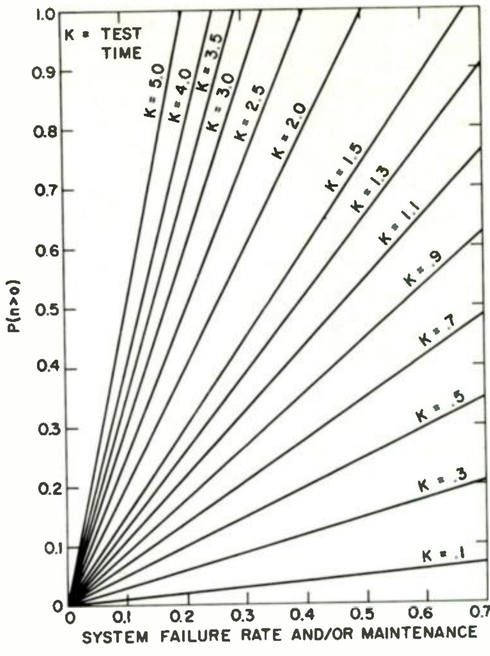


Fig. 7. Probability of a waiting line P/E operating (availability 0.95).

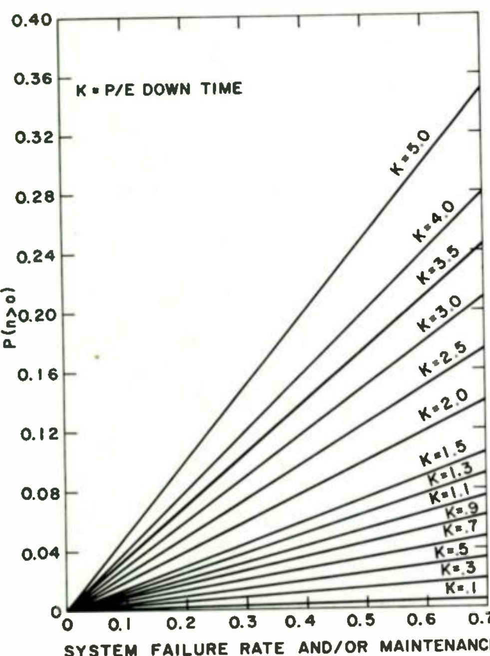


Fig. 6. Probability of a waiting line P/E down for failure (availability 0.90).

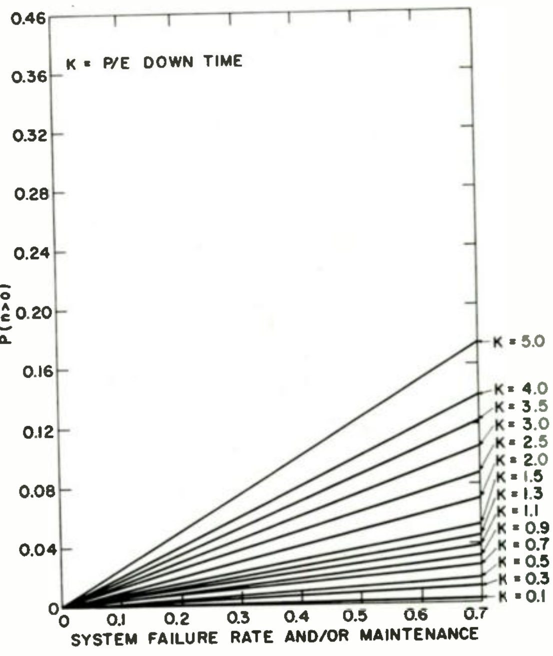


Fig. 8. Probability of a waiting line P/E down for failure (availability 0.95).

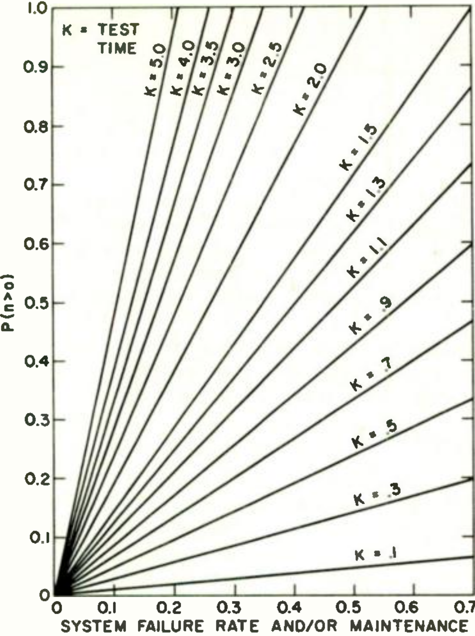


Fig. 9. Probability of a waiting line P/E operating (availability 0.99).

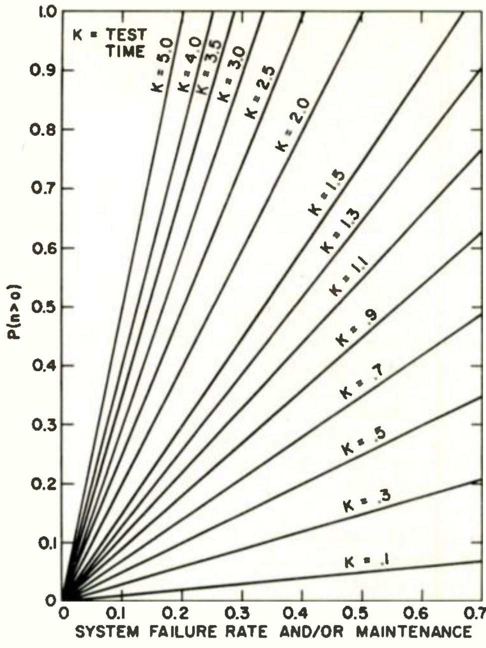


Fig. 11. Probability of a waiting line P/E operating (availability 0.999).

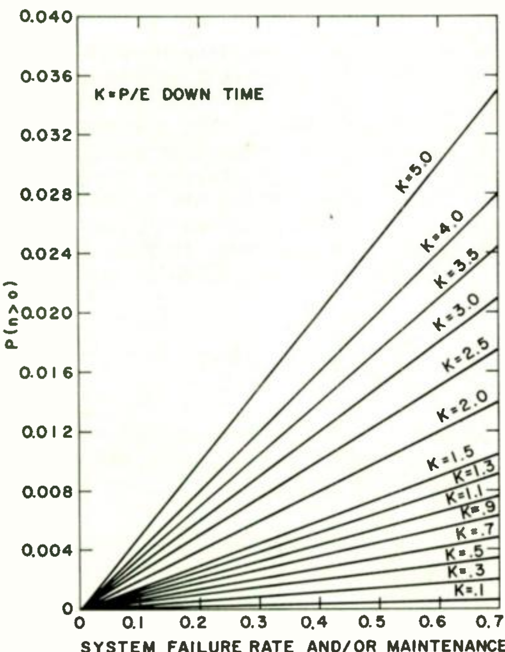


Fig. 10. Probability of a waiting line P/E down for failure (availability 0.99).

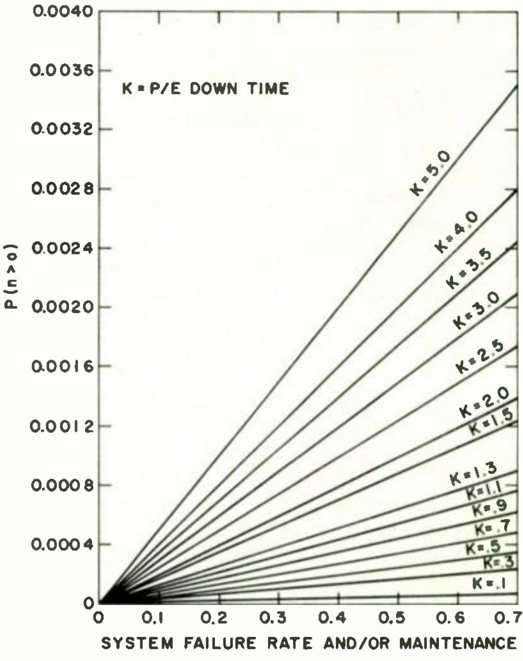


Fig. 12. Probability of a waiting line P/E down for failure (availability 0.999).

SYNTHESIS OF SOLID STATE DISTRIBUTED PARAMETER FUNCTIONS

W. W. Happ, P. S. Castro and W. D. Fuller
Microsystems Electronics Department
Lockheed Missiles & Space Company
Sunnyvale, California

Summary

Film type distributed parameter circuits present the circuit designer with a wide range of function shaping networks. To exploit fully the potential of distributed parameter film type circuits, alternative techniques of shaping distributed parameter networks are proposed; these include variations in morphology and electrical parameters of multilayer structures to obtain network functions not readily available from lumped parameter networks.

An effective procedure of obtaining a wide range of network functions consists of generating subnetworks for multiterminal structures. Guide lines for establishing figures-of-merit of distributed parameter networks are proposed and illustrated by representative examples.

Introduction

Recent work[1,2] in designing network response functions by shaping of distributed parameter circuits has established the feasibility of a new and useful technique for "synthesis" of network functions. However, the full potential of this powerful technique of network synthesis appears to have been overlooked by these investigators. The purpose of this communication is to draw attention to three broad areas of investigation, where thin film distributed parameter techniques lead to improvements in network flexibility and in operational characteristics, namely: alternative methods of network shaping, "synthesis" of functions by subnetwork generation from multiterminal structures, and the judicious use of figures-of-merit for distributed parameter networks.

Network Shaping

Geometrical Variations

Distributed parameter networks made with rectangular geometry and uniform resistive and dielectric films have been thoroughly analyzed. Some control over the network functions obtained may be exercised by varying the geometry of the distributed parameter structure. Linear tapering has been investigated experimentally in the case of null networks by Smith,[3] and he shows that null characteristics can be improved by tapering.

Figure 1 shows several geometrically tapered structures. In general, an analytic solution for even the simplest tapered structure is difficult. An exception is the case of the circular distributed parameter networks.[4] However, even in this case, the network functions are expressed in terms of modified Bessel functions with complex argument and defy an easy interpretation.

By assuming an equivalent network as shown in Figure 2 the equation determining the voltage in the structure can be derived as

$$\nabla^2 V = j w r c V$$

where ∇^2 is the two-dimensional Laplacian operator, V is the two-dimensional voltage function, r is the ohms per square of the resistive film, and c is the capacitance per unit area between the resistive film and the conductive film. Both r and c are assumed constant. The solution to this equation, subject to the appropriate boundary conditions, is in general extremely difficult and recourse must be made to experimental techniques. Details of some techniques which have met with success will be mentioned in a later section.

Parameter Variations

In addition to controlling network functions by using different geometrical shapes, some control can be obtained by variation of the resistive and capacitive parameters. The situation is shown in Figure 3. A rectangular geometry and a one-dimensional variation of resistance and capacitance, obtained by varying the thickness of the resistive and dielectric films, is assumed. Using the equivalent circuit shown in Fig. 3, the equation which must be satisfied by the voltage V is

$$\frac{d^2V}{dx^2} - \frac{1}{r}\frac{dr}{dx}\frac{dV}{dx} = j w r c V$$

where r and c are the resistance and capacitance per unit length and per unit width and are in general functions of x.

Subnetwork Generation

Matrix Approach

From a specified multi-terminal network it is possible to derive a variety of functions by imposing appropriate boundary conditions upon a network.

Functional flexibility in network realization results from the large number of possible alternatives in connecting the terminals of a basic film

type structure. The interaction between adjacent films is described mathematically by laws governing multi-terminal systems[5,6]. If the parent network is described by a matrix relating voltages and currents as shown in Table 1 two types of formulations are possible:

Y - matrix - based on the concept of terminals. The voltages are independent variables and resultant currents are the dependent variables.

Z - matrix - based on the concept of ports. The currents are the dependent variables and the resultant voltages are the dependent variables.

Imposing constraints on terminals or ports corresponds to a well-defined set of matrix operations[5,7], summarized in Table 2. Therefore, if the matrix elements are known, circuit functions of any subnetwork can be derived. On the other hand, the inverse operation is far more difficult, namely to synthesize a desired function from a multi-terminal network by imposing suitable constraints.

As a first step to realize a circuit function inherent in a multi-terminal structure, it is necessary to explore all non-redundant subnetworks. An example using a distributed parameter network will illustrate the procedure of listing all possible one-ports and two-ports which can be generated from a four terminal distributed parameter structure.

Non-redundant Subnetworks

Table 3 shows two representative distributed parameter four terminal networks, referred to as R-C-NR and C-R-NC. To derive the set of non-redundant subnetworks it will be assumed that the following are known:

- matrix elements describing the multi-terminal network uniquely
- matrix operations which correspond to the generation of subnetworks by imposing successively constraints on the parent network
- techniques of obtaining response characteristics from matrix elements

To assure that all subnetworks are enumerated, it is advisable to group networks preferably by categories which can serve as a useful criterion in selecting a subnetwork for a specific application. In Fig. 4 to Fig. 7 the following criteria were considered:

- two types of three layer structure, namely R-C-NR and C-R-NC
- one-ports and two-ports, no distinction was made as to three-terminal and four-terminal two-ports
- capacitive coupling or resistive coupling between input and output of a two-port or between the two terminals of a one-port
- network symmetry or asymmetry, for example, an interchange between NR and R may not produce a new response in a one-port. Similarly interchange of input and output may or may not yield a new two-port.

Recognition of criteria, such as symmetry may substantially aid in the selection of the appropriate subnetwork to realize a desired transfer characteristic. In a somewhat similar way, the type of coupling will determine the response at low frequencies. Similar criteria for the high frequency response as well as for the transient response result from a classification which has in mind the synthesis objectives in terms of desired network response.

Systematics

Response characteristics of subnetworks can be predicted from the response functions of the parent network. Since, frequently, a wide range of functions are realizable by subnetwork generation, the crucial problem becomes the recognition of the complete range of useful network functions which can be generated.

Present techniques of network analysis are, with a few exceptions, limited to three terminal two-port systems. Thus, greater flexibility of design of film type circuits appears predicated upon more effective utilization of subnetwork generation or similar techniques of realizing functions inherent in a multi-terminal system.

Three-layer distributed parameter networks were seen to lead to four-terminal networks, which generated one-ports and two kinds of two-ports, three-terminal and four-terminal. Four-layer distributed parameter networks are capable of generating a far greater variety of networks.

A systematic assessment of the number of subnetworks generated from an N-terminal parent network is presented in Fig. 8. To illustrate, a four-terminal network yields subnetworks denoted as follows:

2-P/4-T: two-ports/four-terminal - 3
3-P/4-T: three-ports/four-terminal - 12
2-P/3-T: two-ports/three-terminal - 30
1-P/2-T: one-ports/two-terminal - 33

The number of subnetworks thus generated are assumed to be free of constraints, so this number must be reduced to account for symmetry by the methods indicated above. Nevertheless, the trend toward multilayer structures implies multi-terminal networks with thousands or even millions of possible subnetworks. Consequently present concepts of network synthesis must be extended to include function synthesis by subnetwork generation.

Conventional network synthesis may be looked upon as function synthesis from a large universe of individual components by imposing suitable constraints on the terminal or ports of all components. Clearly this is a necessary - but not a sufficient - condition for a definition of conventional network synthesis, within the meaning of present terminology. The new concept of function synthesis here proposed would no longer be based solely upon a large universe of individual components, but include and even emphasize function synthesis from a single multi-terminal structure. Thus subnetwork generation may be regarded pari passu with conventional techniques of network synthesis.

Experimental Structures

Characteristics of RC distributed parameter networks may be investigated on a modelling basis, both to confirm the mathematical analysis and to explore network shapes not easily expressed analytically. Modelling, as shown in Fig. 9, is an extremely simple approach to experimental networks for evaluation. Networks are fabricated from sheets of Teledeltos+ or similar paper laminated with Mylar film and aluminum foil using polystyrene cement as the interlayer adhesive. Rather large assemblies may be constructed in this manner and then cut with scissors into the required network shapes. Teledeltos Type L paper has a resistance of approximately 2000 ohms/square, and Type H has a resistance of 20,000 ohms/square. Using 1/4 mil Mylar as dielectric results in a capacitance of approximately 300 pf/in^2.

The physical structures of distributed parameter RC components have two forms. The first as shown in Fig. 10 is a thin film assembly which uses a high dielectric constant substrate with resistive film patterns on both sides. Such R-C-NR structures may be fabricated by most of the thin film technologies and represent an important by-product of the thin film integrated circuit technologies.

The second form as shown in Fig. 11 is a solid state structure which may be based upon many different single crystal materials. The C-R-NC structure as an example is easily developed from a bar or wafer of silicon which has been thermally oxidized to produce a dielectric layer on the outside over which counterelectrodes may be formed. A wider range of RC values may be developed from the single crystal refractory oxides which through chemical reduction can vary in resistance from 10^{16} ohm-cm to 10^{-1} ohm-cm. The dielectric layer on the outside of the crystal is produced after reduction by anodization and in the case of titanium dioxide the dielectric constant approaches 100.

The solid state structures allow extra dimensions in both physical and parametric shaping that are not easily obtained in thin film structures.

+ Western Union

Figures-of-Merit

Figures-of-merit have proven of value to the circuit designer as an effective means of assessing the performance of a circuit element, of a device or a network. The illustrative examples below show the distinct advantage of thin film circuits and the possibility of a more sophisticated device design. This permits the designer to optimize figures-of-merit according to his needs more effectively than is generally possible with ready-made, pig-tailed components.

For example, for a phase-shift oscillator similar to those discussed by Wilson and Wilson[1] or by Edson,[2] figures-of-merit found most useful in this laboratory are shown in Fig. 12. The same graph also shows the corresponding N-stage lumped parameter networks to illustrate the asymptotic character of distributed parameter networks.

As a second example illustrating the judicious use of the concept of figures-of-merit, consider the fourteen null networks listed in Table 4. In this case the most appropriate figures-of-merit for null networks were found to be:

- the null frequency (w_o), normalized in terms of R and C of the structure
- the sharpness of the notch at the null frequency measured by the slope of the output voltage (V_o) as a function of frequency (w) and normalized with respect to input voltage (V_i) and null frequency
- the change in phase at the null frequency
- the range of the above three criteria for each type of network as a function in structural parameters
- an optimum "operating point" in this range denoted by (x)

Closer inspection of the range of the figures-of-merit will reveal that in every case the limits of the ranges consist of a set of six figures-of-merit. This set corresponds to the set of six networks shown in Fig.13, and from it figures-of-merit of null networks can be predicted. This leads to the important conclusion that figures-of-merit are a guide in the choice of networks required to perform a desired network function. Conversely, the synthesis of an appropriate null network can be performed on the basis of its figure-of-merit.

For synthesis purposes every distributed parameter structure thus may have a different figure-of-merit depending upon the specific application. By similar procedures it is possible to establish appropriate figures-of-merit for other film type circuits.

The use of a single figure-of-merit is, in general, an over-simplification and often misleading, hence the circuit designer is seeking a design compromise based on more than one figure-of-merit. The design flexibility obtained by shaping distributed parameter structures meets requirements of this type admirably and provides the stimulus for a continuing development of the thin film circuits.

Conclusion

To exploit the potential of distributed parameter film type circuits, other techniques must be used in conjunction with "shaping". Circuit parameters have been evaluated for several types of multi-layer structures as well as for circular and tapered patterns. Thus a "catalogue" of circuit functions in terms of the film morphology can serve as a basis of synthesis procedures. Figures-of-merit for a given circuit function provide guide lines in the synthesis of a film type network.

One significant new aspect in the synthesis of film type circuits is the important role played by multi-terminal networks. The synthesis of lumped parameter "pig-tailed" components was based on the appropriate interconnection of a large number of two-terminal or two-port networks. The synthesis of functions from a single multi-terminal (or multi-port) structure results from constraints imposed upon this structure.

Techniques of analysis and synthesis will therefore differ greatly between "multi-function" systems and "multi-component" systems.

On the other hand, a common feature emerges by considering a multi-component system simply as a large number of accessible terminals (or ports). If this assumption is made, network synthesis, as the term is now generally understood, can be regarded as the result of constraints imposed upon accessible terminals. In short, synthesis may be looked upon as the realization of circuit functions by two nearly equivalent operations, namely by subnetwork generation from a given film type structure, and by selection of components from a specified class or universe of elements.

Acknowledgement

This work is part of a continuing Independent Research Program in the synthesis of thin film integrated circuits performed in the Microsystems Electronics Department at Lockheed Missiles & Space Company.

References

1. L. B. H. Wilson and R. B. Wilson, "Shaping of Distributed RC Networks," Proc. IRE, Vol. 49, pp. 1330-1331, August 1961.

2. W. A. Edson, "Tapered Distributed RC Lines for Phase-Shift Oscillators," Proc. IRE, Vol. 49, June 1961.

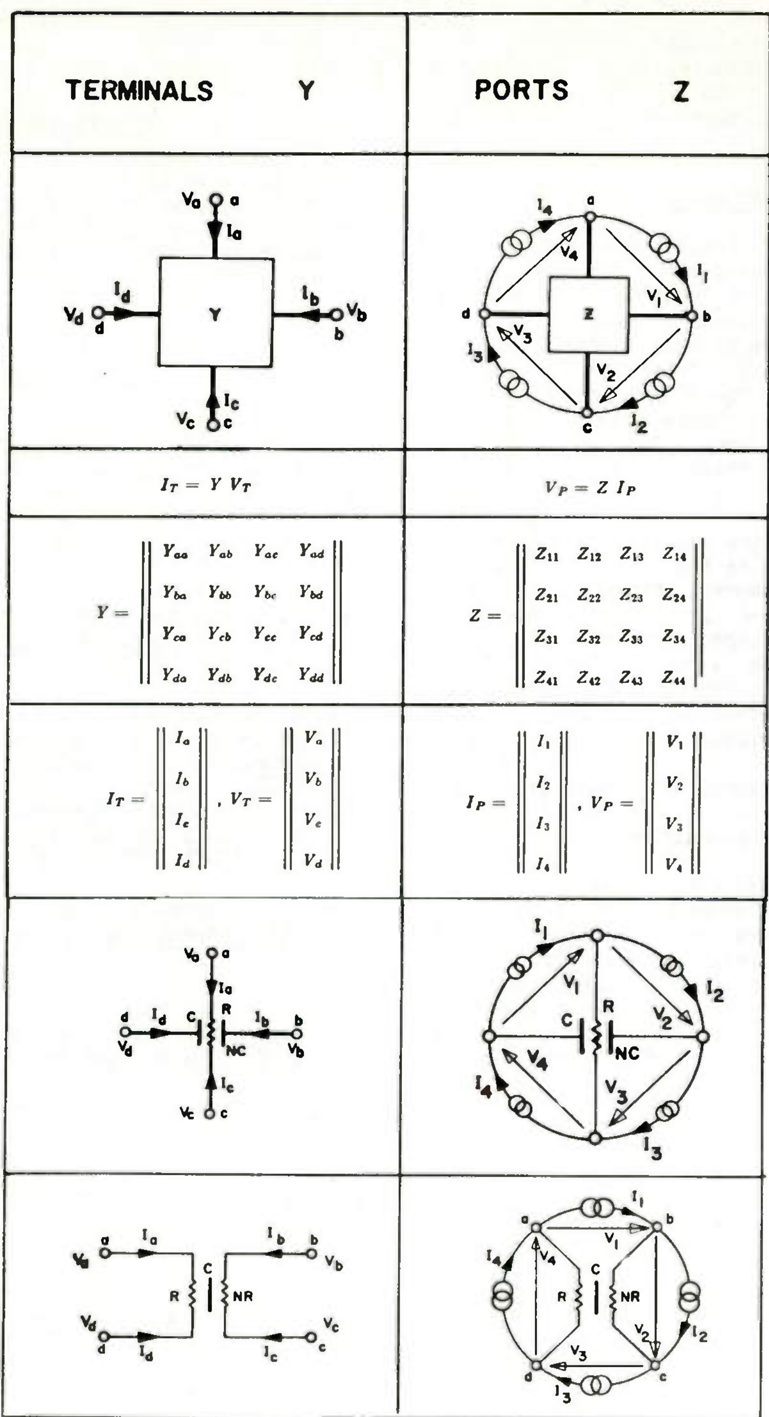
3. A. Smith, "Rejection Filters with Distributed R and C," Proc. 1960 Elec. Comp. Conference, May 1960.

4. W. W. Happ and P. S. Castro, "Distributed Parameter Circuit Design Techniques," NEC Proc., Vol. 17, pp. 44-70, 1961.

5. P. S. Castro, "Microsystem Circuit Analysis," Electrical Engineering, Vol. 80, No. 7, pp. 535-542, July 1961.

6. P. S. Castro and W. W. Happ, "Subnetworks," IRE International Conference Record, Vol. 10, 1962.

7. W. W. Happ and P. W. Castro, "Distributed Parameter Circuits and Microsystems Electronic," Proc. NEC, Vol. 16, pp. 448-460, 1960.



TERMINALS Y	PORTS Z
$I_T = Y\, V_T$	$V_P = Z\, I_P$
$Y = \begin{Vmatrix} Y_{aa} & Y_{ab} & Y_{ac} & Y_{ad} \\ Y_{ba} & Y_{bb} & Y_{bc} & Y_{bd} \\ Y_{ca} & Y_{cb} & Y_{cc} & Y_{cd} \\ Y_{da} & Y_{db} & Y_{dc} & Y_{dd} \end{Vmatrix}$	$Z = \begin{Vmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{14} \\ Z_{21} & Z_{22} & Z_{23} & Z_{24} \\ Z_{31} & Z_{32} & Z_{33} & Z_{34} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} \end{Vmatrix}$
$I_T = \begin{Vmatrix} I_a \\ I_b \\ I_c \\ I_d \end{Vmatrix}, \; V_T = \begin{Vmatrix} V_a \\ V_b \\ V_c \\ V_d \end{Vmatrix}$	$I_P = \begin{Vmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{Vmatrix}, \; V_P = \begin{Vmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{Vmatrix}$

TABLE 1 Matrix Formulation of Four-Terminal Network

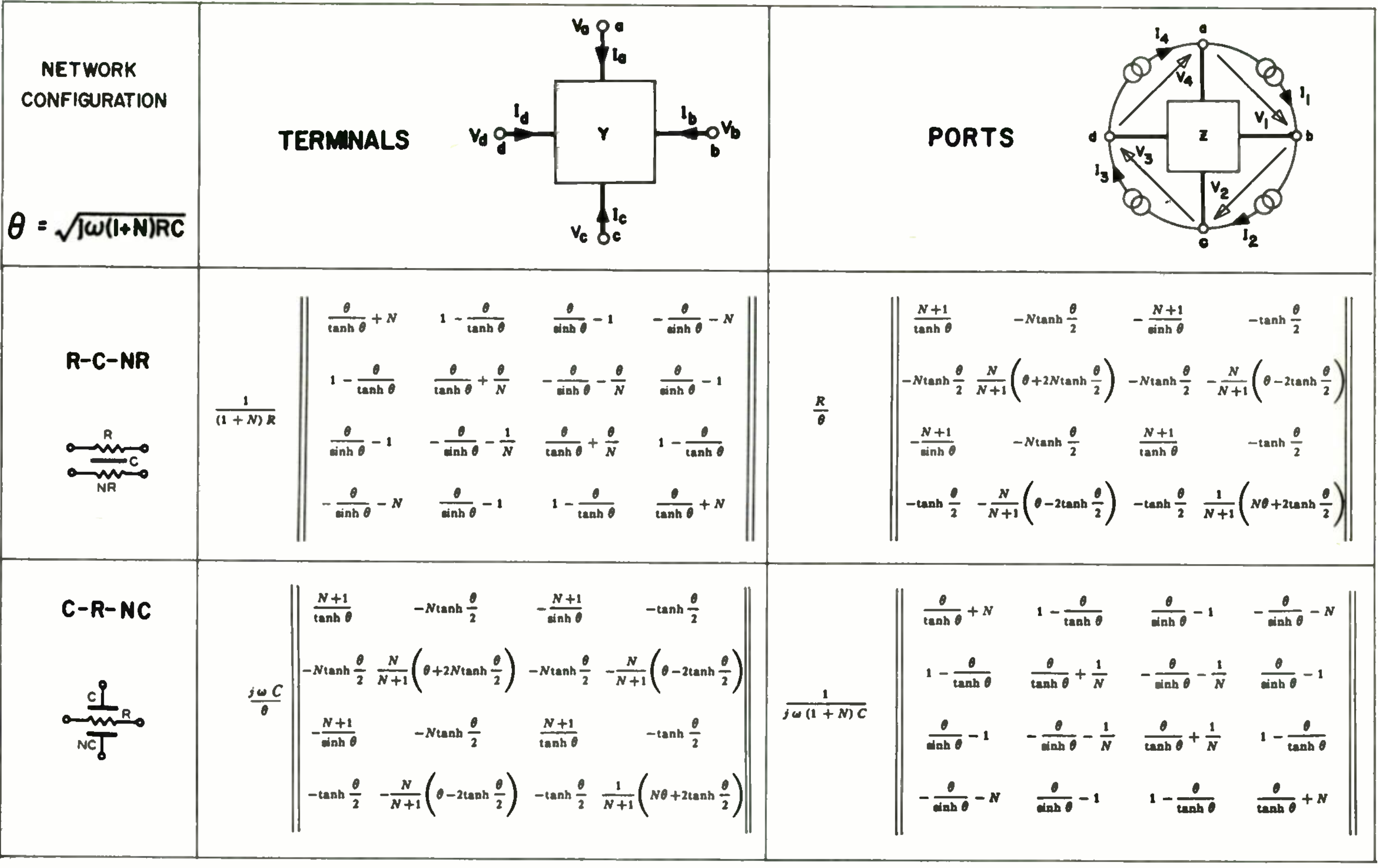
OPERATION ON NETWORK		OPERATION ON MATRIX	
Terminals	Ports	Admittance	Impedance
Float	Disconnect	Add Rows and columns	Invert Row
Connect	Short	Invert Row	Add Rows and Columns
Use as Reference		Cross Off Rows and Columns	

TABLE 2 - MATRIX OPERATIONS

Normalized Parameter		Rectangular	Linear Taper Approximate	Exponential Taper Approximate	Resistive Disk
Resistance	$\frac{a}{\rho} R$	$\frac{\ell}{w}$	$\frac{(\ell)}{w} \frac{n \log}{(n-1)}$	$\frac{n-1}{aw}$	$\frac{\log n}{2\pi}$
Capacitance	$\frac{t}{\epsilon} C$	ℓw	$(\ell w) \frac{n+1}{2n}$	$w \frac{(n-1)}{an}$	$\pi \frac{w^2}{n^2} (n^2 - 1)$
Time Constant	$\frac{at}{\rho\epsilon} RC$	ℓ^2	$(\ell^2) \frac{(n+1)\log n}{2(n-1)}$	$\frac{(n-1)^2}{na^2}$	$\frac{w^2}{2n^2} (n^2-1)\log n$
Impedance Level	$\frac{\epsilon}{\rho} \frac{R}{C}$	$\frac{1}{w^2}$	$(\frac{1}{w^2}) \frac{2n^2 \log n}{n^2 - 1}$	$\frac{n}{w^2}$	$\frac{1}{2\pi^2} \frac{n^2 \log n}{w^2(n^2-1)}$

EXAMPLES OF TWO-DIMENSIONAL FILM-TYPE STRUCTURES
WITH CORRESPONSING FORMULAS RELATING CIRCUIT PARAMETER
R AND C TO MATERIALS PARAMETERS

FIGURE 1 - GEOMETRICALLY TAPERED STRUCTURES



NETWORK CONFIGURATION $\theta = \sqrt{j\omega(1+N)RC}$	TERMINALS	PORTS
R-C-NR	$\frac{1}{(1+N)R}\left\| \begin{matrix} \frac{\theta}{\tanh\theta}+N & 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\sinh\theta}-1 & -\frac{\theta}{\sinh\theta}-N \\ 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\tanh\theta}+\frac{\theta}{N} & -\frac{\theta}{\sinh\theta}-\frac{\theta}{N} & \frac{\theta}{\sinh\theta}-1 \\ \frac{\theta}{\sinh\theta}-1 & -\frac{\theta}{\sinh\theta}-\frac{1}{N} & \frac{\theta}{\tanh\theta}+\frac{\theta}{N} & 1-\frac{\theta}{\tanh\theta} \\ -\frac{\theta}{\sinh\theta}-N & \frac{\theta}{\sinh\theta}-1 & 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\tanh\theta}+N \end{matrix} \right\|$	$\frac{R}{\theta}\left\| \begin{matrix} \frac{N+1}{\tanh\theta} & -N\tanh\frac{\theta}{2} & -\frac{N+1}{\sinh\theta} & -\tanh\frac{\theta}{2} \\ -N\tanh\frac{\theta}{2} & \frac{N}{N+1}\left(\theta+2N\tanh\frac{\theta}{2}\right) & -N\tanh\frac{\theta}{2} & -\frac{N}{N+1}\left(\theta-2\tanh\frac{\theta}{2}\right) \\ -\frac{N+1}{\sinh\theta} & -N\tanh\frac{\theta}{2} & \frac{N+1}{\tanh\theta} & -\tanh\frac{\theta}{2} \\ -\tanh\frac{\theta}{2} & -\frac{N}{N+1}\left(\theta-2\tanh\frac{\theta}{2}\right) & -\tanh\frac{\theta}{2} & \frac{1}{N+1}\left(N\theta+2\tanh\frac{\theta}{2}\right) \end{matrix} \right\|$
C-R-NC	$\frac{j\omega C}{\theta}\left\| \begin{matrix} \frac{N+1}{\tanh\theta} & -N\tanh\frac{\theta}{2} & -\frac{N+1}{\sinh\theta} & -\tanh\frac{\theta}{2} \\ -N\tanh\frac{\theta}{2} & \frac{N}{N+1}\left(\theta+2N\tanh\frac{\theta}{2}\right) & -N\tanh\frac{\theta}{2} & -\frac{N}{N+1}\left(\theta-2\tanh\frac{\theta}{2}\right) \\ -\frac{N+1}{\sinh\theta} & -N\tanh\frac{\theta}{2} & \frac{N+1}{\tanh\theta} & -\tanh\frac{\theta}{2} \\ -\tanh\frac{\theta}{2} & -\frac{N}{N+1}\left(\theta-2\tanh\frac{\theta}{2}\right) & -\tanh\frac{\theta}{2} & \frac{1}{N+1}\left(N\theta+2\tanh\frac{\theta}{2}\right) \end{matrix} \right\|$	$\frac{1}{j\omega(1+N)C}\left\| \begin{matrix} \frac{\theta}{\tanh\theta}+N & 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\sinh\theta}-1 & -\frac{\theta}{\sinh\theta}-N \\ 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\tanh\theta}+\frac{1}{N} & -\frac{\theta}{\sinh\theta}-\frac{1}{N} & \frac{\theta}{\sinh\theta}-1 \\ \frac{\theta}{\sinh\theta}-1 & -\frac{\theta}{\sinh\theta}-\frac{1}{N} & \frac{\theta}{\tanh\theta}+\frac{1}{N} & 1-\frac{\theta}{\tanh\theta} \\ -\frac{\theta}{\sinh\theta}-N & \frac{\theta}{\sinh\theta}-1 & 1-\frac{\theta}{\tanh\theta} & \frac{\theta}{\tanh\theta}+N \end{matrix} \right\|$

TABLE 3 - MATRIX FORMULATION OF DISTRIBUTED PARAMETER NETWORK

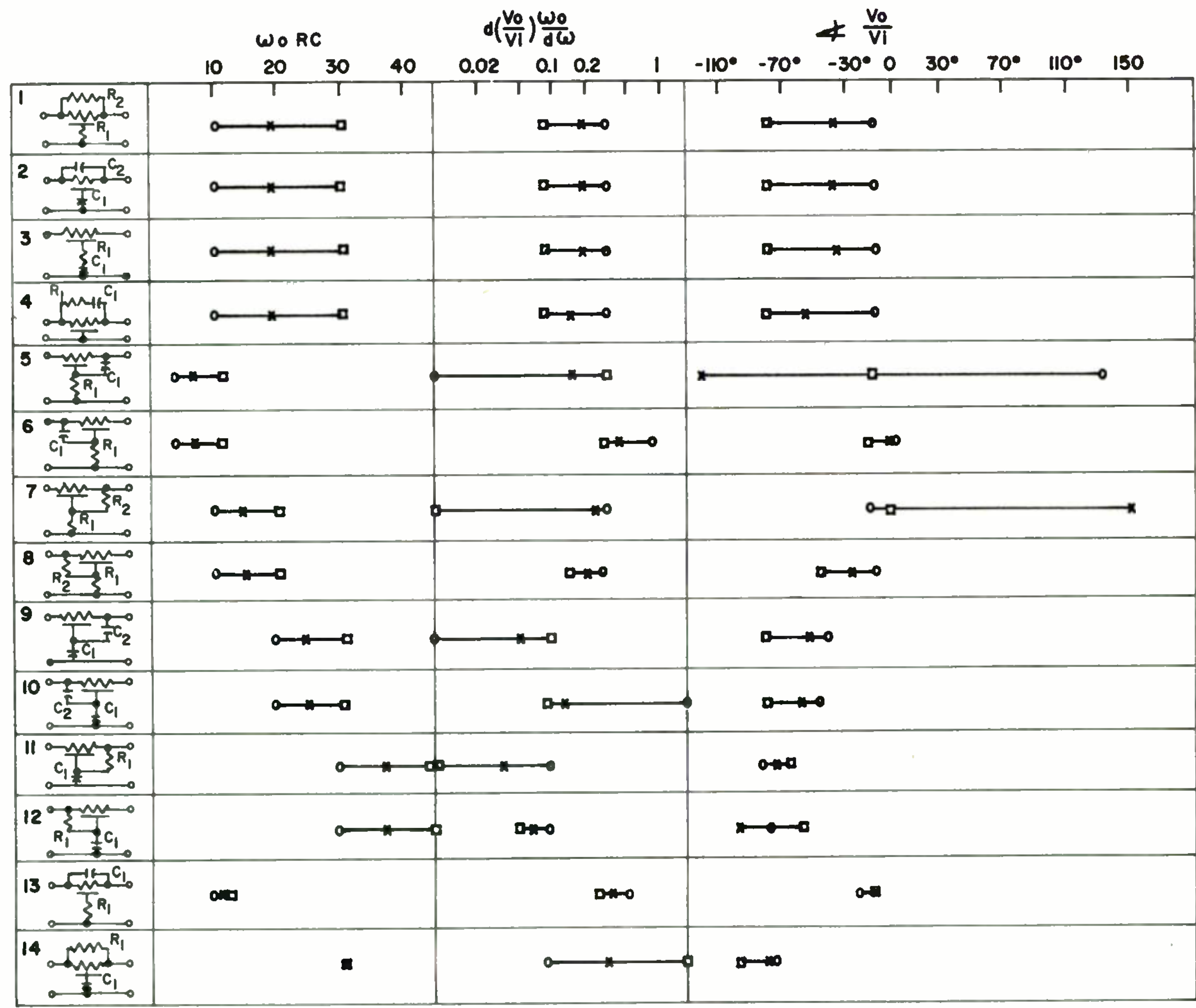


TABLE 4 - TABLE OF NULL NETWORKS

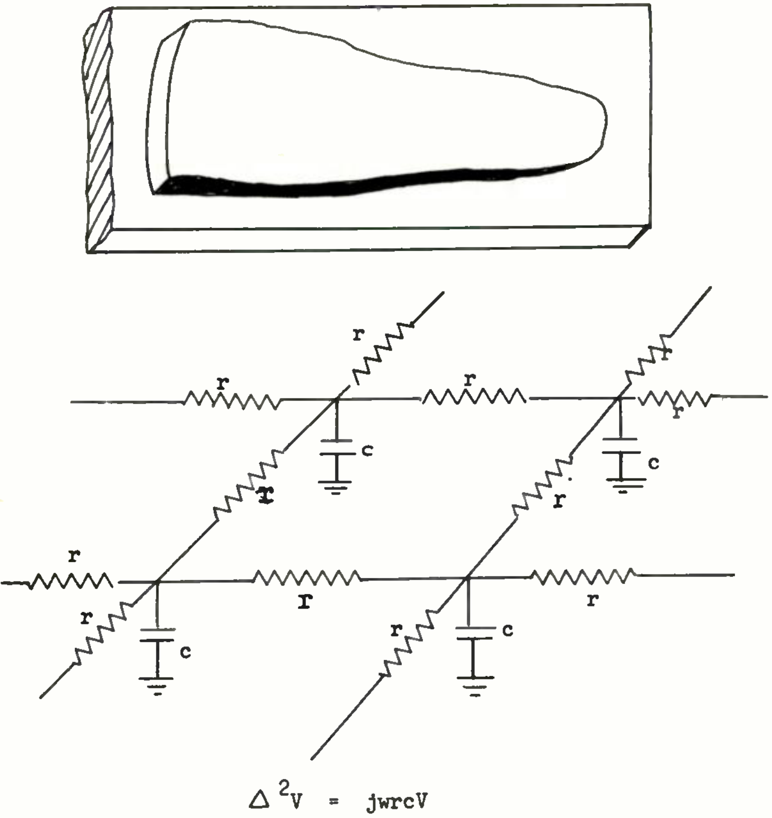


FIGURE 2

ANALYSIS OF GEOMETRICALLY TAPERED STRUCTURES

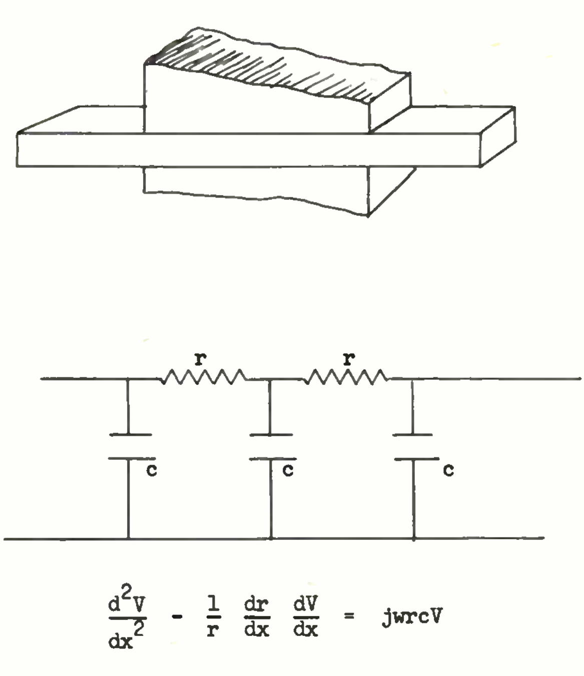


FIGURE 3

ANALYSIS OF STRUCTURES WITH PARAMETER TAPERING

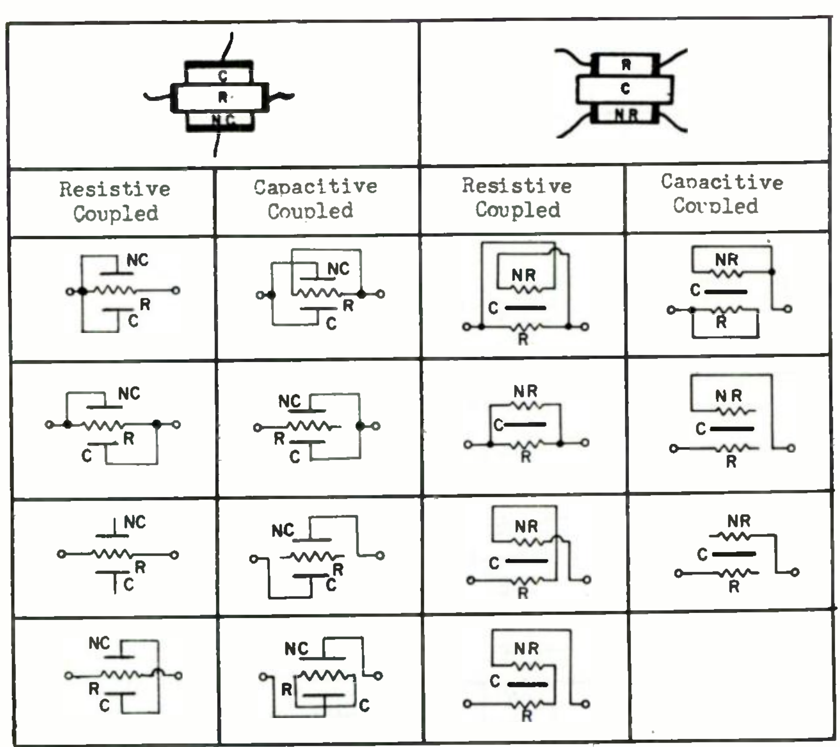


FIGURE 4 - Symmetric One-Ports

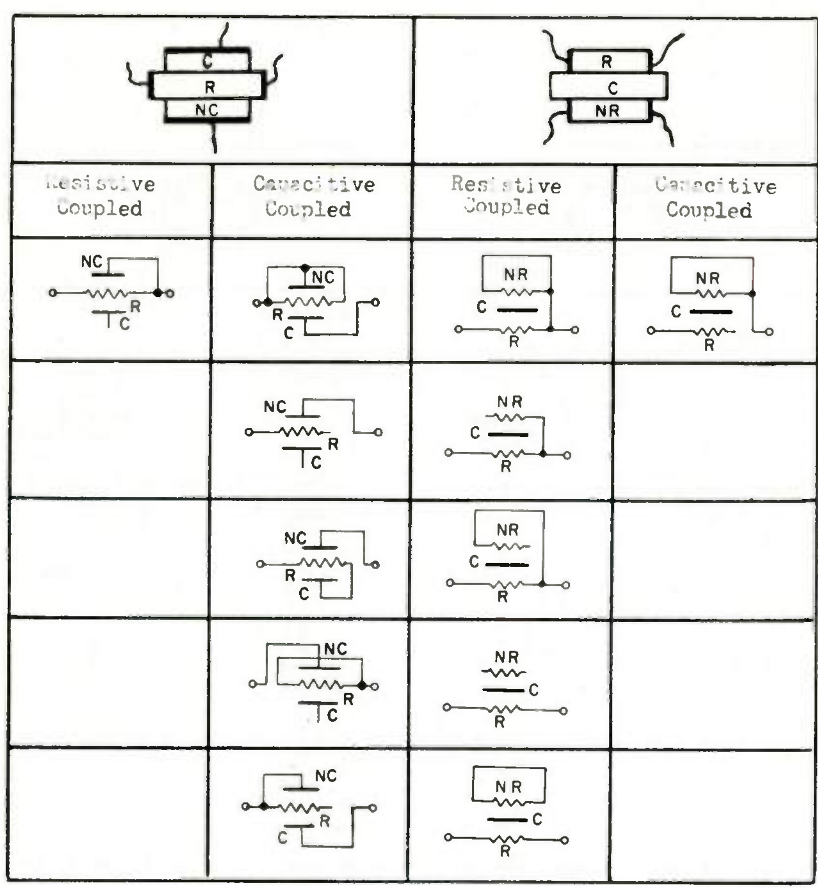


FIGURE 5 Asymmetric One-Ports

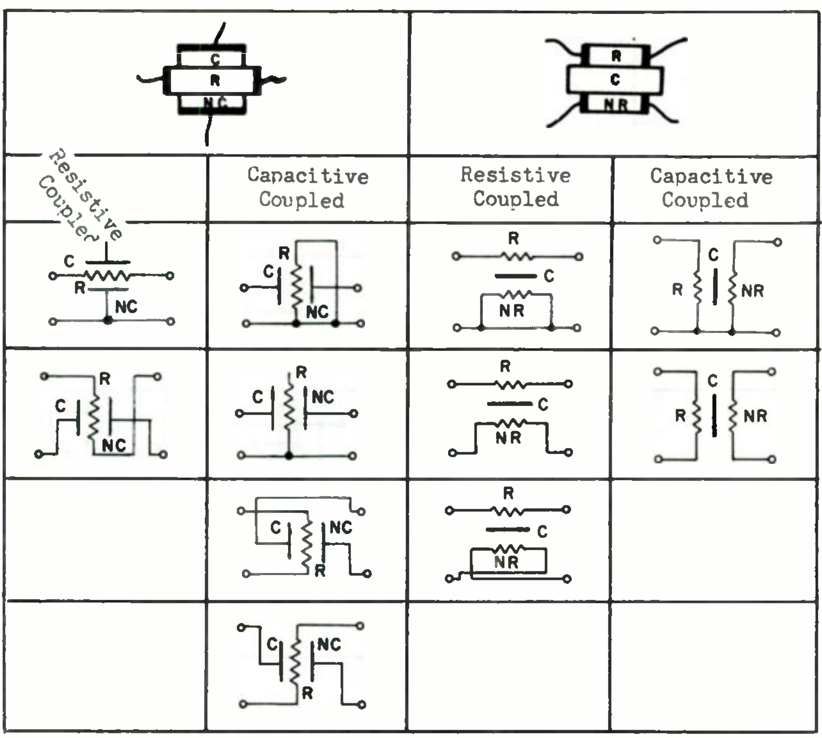


FIGURE 6 - Symmetric Two-Ports

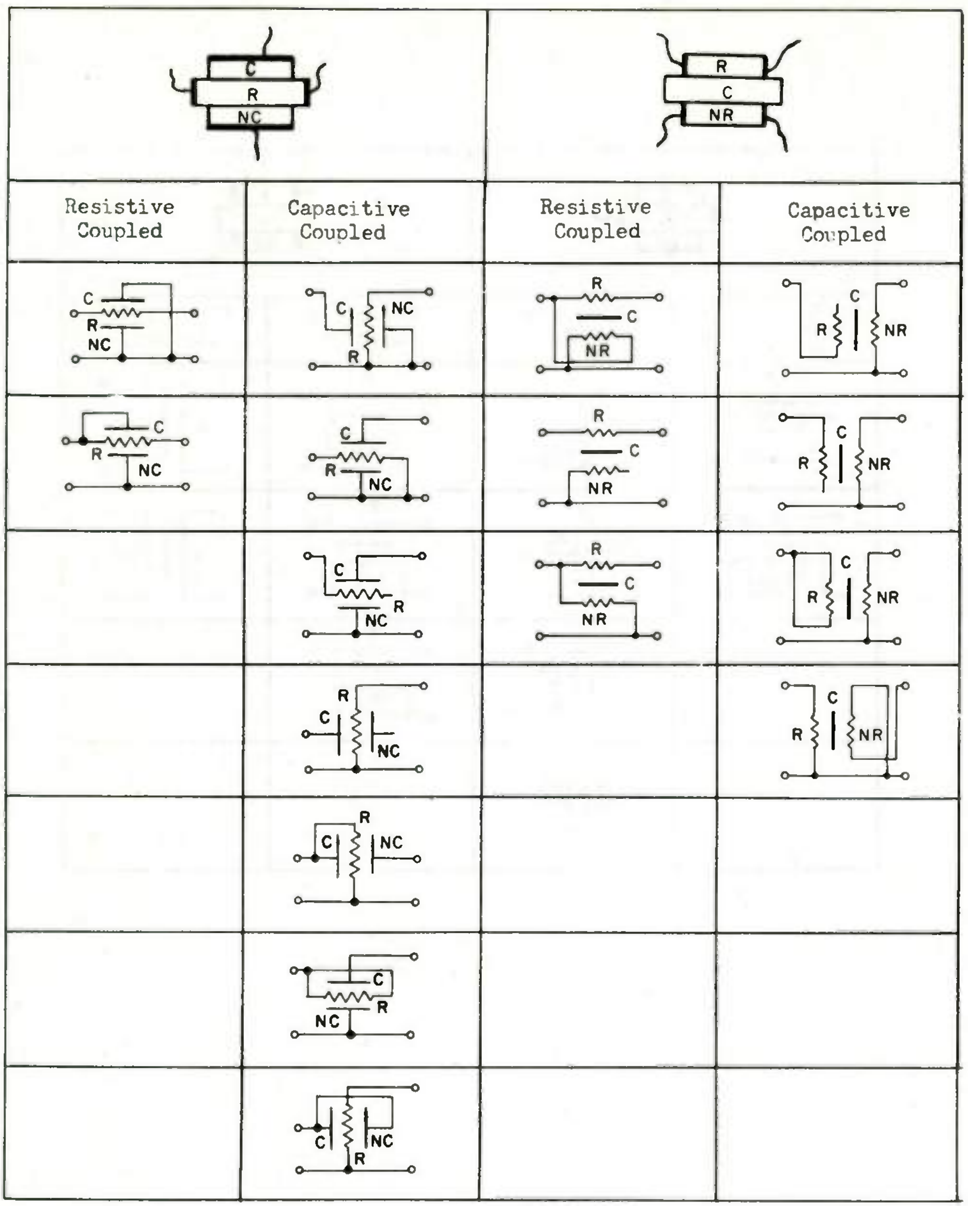


FIGURE 7 – Asymmetric Two-Ports

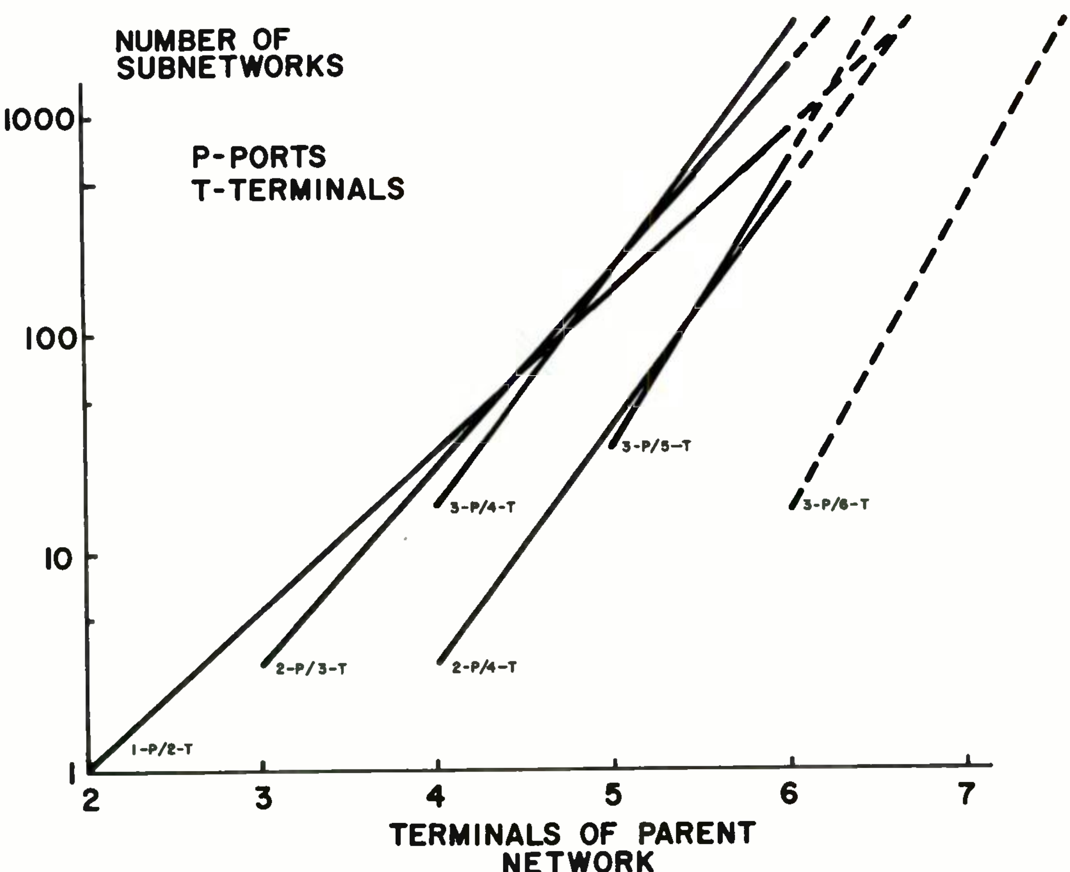
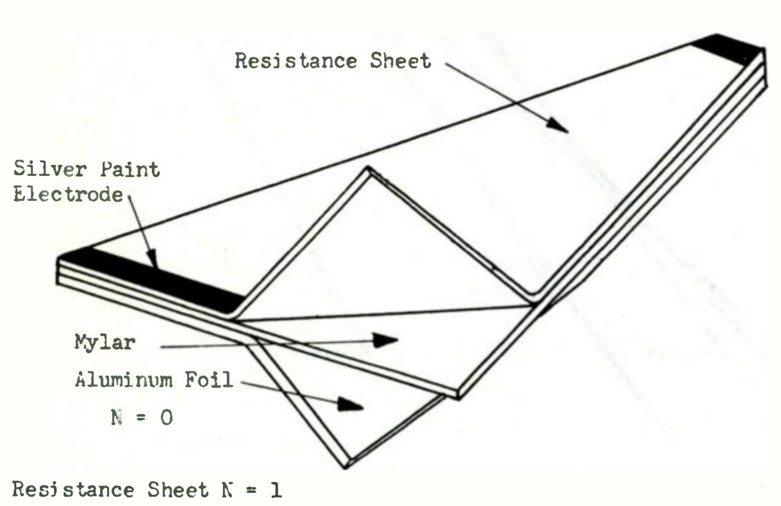


FIGURE 8 - NUMBER OF SUBNETWORKS FROM AN N-TERMINAL PARENT NETWORK



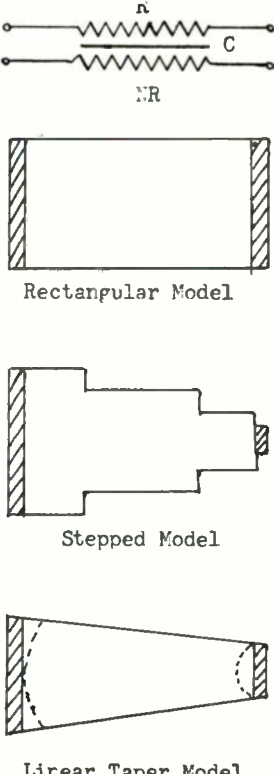


FIGURE 9

MODELLING WITH TELEDELTOS PAPER

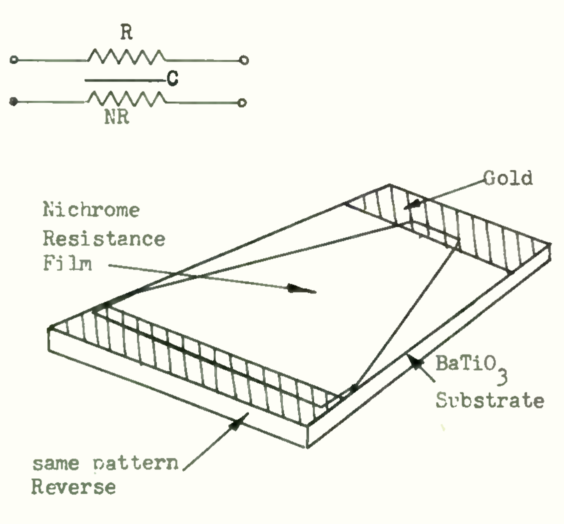


FIGURE 10

R-C-NR STRUCTURE ON HIGH DIELECTRIC CONSTANT SUBSTRATE

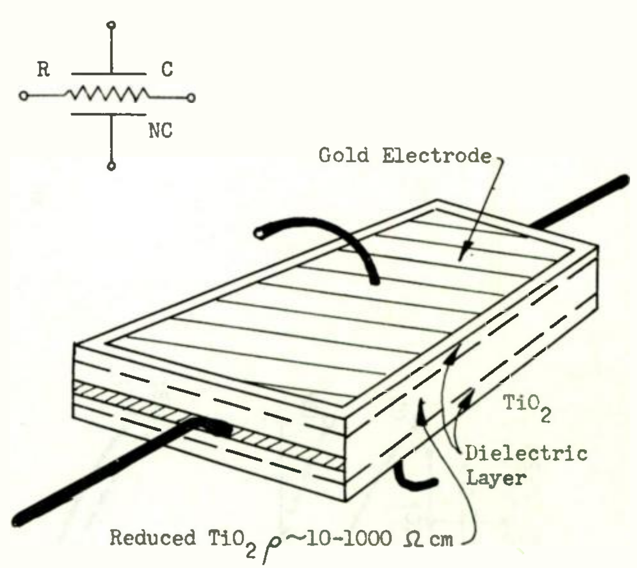


FIGURE 11

SINGLE CRYSTAL FORM OF D-P STRUCTURE

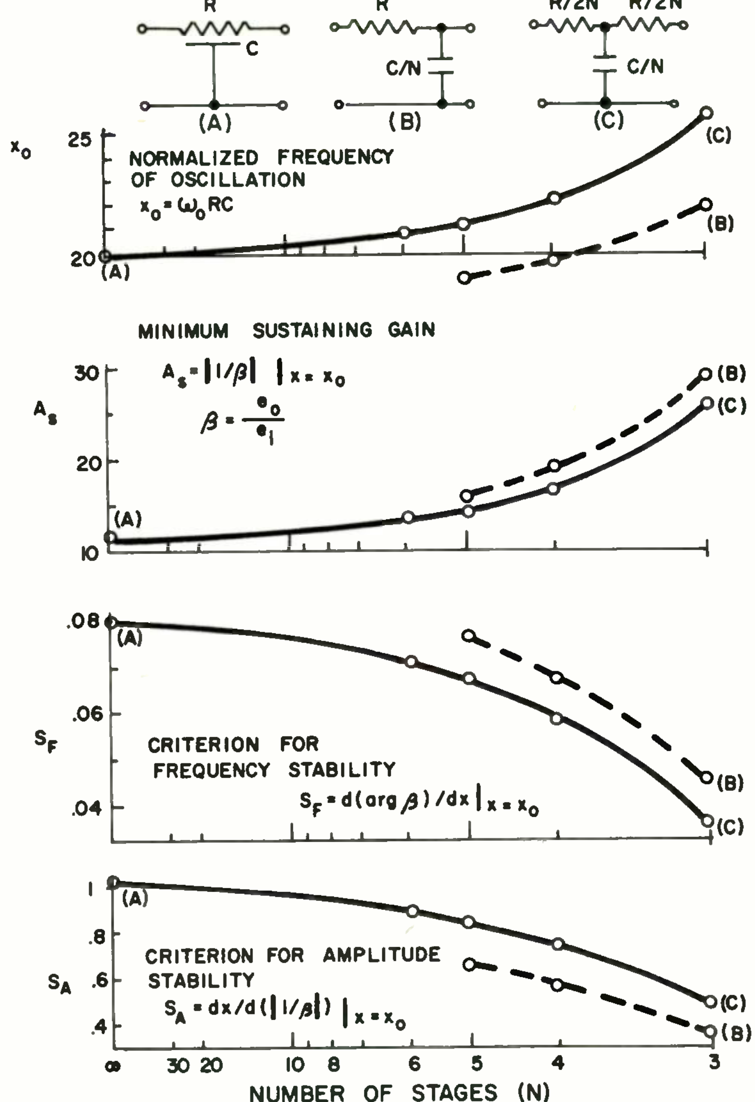


FIGURE 12 - FIGURES-OF-MERIT FOR PHASE-SHIFT OSCILLATORS

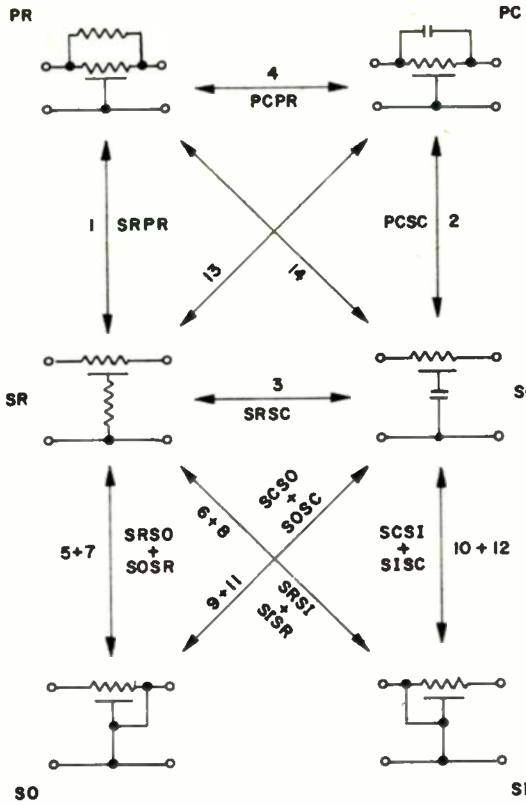


FIGURE 13 - LIMITING NULL NETWORKS

A THEORETICAL COMPARISON OF DOUBLY LOADED DISTRIBUTED BRIDGED T AND LUMPED TWIN T RC NOTCH FILTERS

Joel L. Ekstrom

Applied Research Laboratory
Sylvania Electronic Systems
Waltham 54, Massachusetts

Introduction

There exists quite a large body of literature[1-9] on the design of distributed and lumped parameter passive RC null networks, but most of these analyses, especially in the case of the distributed RC null network, are concerned with behavior under unloaded conditions. In today's low impedance transistor circuits, open circuits are often hard to find, so that the behavior of null networks under loaded conditions is of some interest. We will show that under these conditions the lumped Twin T has some advantages over the distributed bridged T null network.

Theory

Consider first the Twin T network in Figure 1, where the input and output loads are chosen to be identical for the reason that if they are not chosen this way, both the Twin T and the bridged T have to be asymmetrical networks to obtained optimum performance, which in turn complicates the analysis unnecessarily.

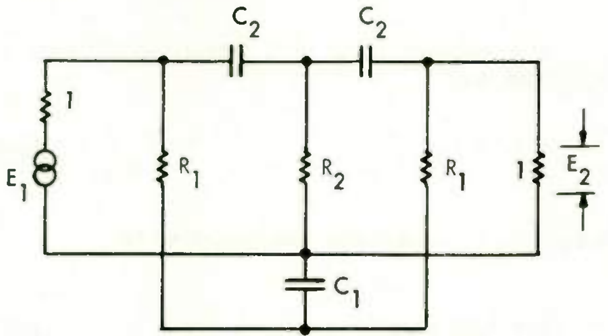


FIGURE 1 LOADED RC TWIN-T

Under the conditions that the notch frequency is 1 radian per second, that the responses at zero and infinite frequencies are equal, and that the insertion loss is minimized for a given bandwidth, the voltage transfer function of the above device is given by[3]

$$T_L(S) = \frac{E_2}{E_1} = \frac{T_1}{2\left[T_1 + \sqrt{T_1^2 + 1}\right]} \frac{S^2 + 1}{S^2 + 2\sqrt{T_1^2 + 1}\, S + 1} \quad (1)$$

where

$$T_1 = 2R_2C_2 = \frac{R_1C_1}{2} \quad (2)$$

$$T_2 = \frac{1}{T_1} = R_1C_2 \quad (3)$$

$$R_1 = \frac{\sqrt{T_1^2 + 1}}{T_1} \quad (4)$$

The bandwidth of the notch filter is defined by the frequencies at which the magnitude of the response is down 3 db from its value at zero and infinite frequency. For the Twin T the bandwidth is

$$\Delta\omega = 2\sqrt{T_1^2 + 1} \text{ radians} \quad (5)$$

From (1) and (5) it is seen that one may trade bandwidth and insertion loss by adjustment of the parameter T_1. The minimum bandwidth of the Twin T occurs when $T_1 = 0$, at which value $\Delta\omega = 2$ and the insertion loss is seen from (1) to be infinite. To find the behavior with real frequency, first define a normalized magnitude function

$$\left|T_L(j\omega)\right|_n = \frac{2\left[T_1 + \sqrt{T_1^2 + 1}\right]}{T_1} \left|T_L(j\omega)\right| \quad (6)$$

From (1) it is easy to show that

$$\left|T_L(j\omega)\right|_n = \left|\frac{1 - \omega^2}{\sqrt{\omega^4 + (T_1^2 + 2)\omega^2 + 1}}\right| \quad (7)$$

Next consider the Bridged T notch filter (Figure 2). In order that the responses at zero

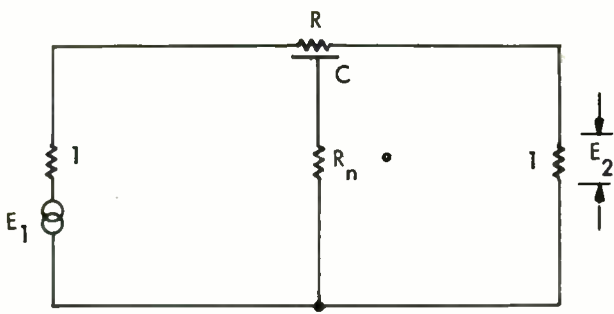


FIGURE 2 LOADED DISTRIBUTED BRIDGED - T

and infinite frequencies be equal, it is necessary that

$$R_n = \frac{1}{R} \tag{8}$$

To find the voltage transfer function, we see from network theory[10] that

$$\frac{E_2}{E_1} = \frac{z_{12}}{(z_{11}+1)^2 - z_{12}^2} = T_D(S) \tag{9}$$

This can also be written

$$T_D(S) = \frac{\frac{1}{2}}{z_{11} - z_{12} + 1} - \frac{\frac{1}{2}}{z_{11} + z_{12} + 1} \tag{10}$$

where[11,12]

$$z_{11} - z_{12} = \sqrt{\frac{R}{SC}} \tanh \frac{1}{2}\sqrt{RSC} \tag{11}$$

$$z_{11} + z_{12} = 2R_n + \sqrt{\frac{R}{SC}} \coth \frac{1}{2}\sqrt{RSC} \tag{12}$$

To secure rejection of a single frequency ω_n we must have

$$R_n = -\sqrt{\frac{R}{j\omega_n C}} \operatorname{csch} \sqrt{j\omega_n RC} \tag{13}$$

Using the well known expansions of hyperbolic functions of complex arguments,[11] together with (8), we find two conditions which must be satisfied

$$R^2 + \sqrt{\frac{\omega_n RC}{2}}\left[\sinh\sqrt{\frac{\omega_n RC}{2}}\cos\sqrt{\frac{\omega_n RC}{2}} - \cosh\sqrt{\frac{\omega_n RC}{2}}\sin\sqrt{\frac{\omega_n RC}{2}}\right] = 0 \tag{14}$$

$$\sinh\sqrt{\frac{\omega_n RC}{2}}\cos\sqrt{\frac{\omega_n RC}{2}} + \cosh\sqrt{\frac{\omega_n RC}{2}}\sin\sqrt{\frac{\omega_n RC}{2}} = 0 \tag{15}$$

There are many values of the product $\omega_n RC$ which will satisfy (15); we are interested only in the smallest product. Since we have chosen $\omega_n = 1$ as the notch frequency in both cases, we can put

$$\sqrt{\frac{\omega_n RC}{2}} = \sqrt{\frac{RC}{2}} = \lambda_{min} = \lambda_m \tag{16}$$

and write (14) and (15) in the form

$$\tanh \lambda_m + \tan \lambda_m = 0 \tag{17}$$

$$R^2 = 2\lambda_m \cosh \lambda_m \sin \lambda_m \tag{18}$$

The smallest value of λ satisfying (17) is approximately

$$\lambda_m = 2.365 \tag{19}$$

whence R, R_n, and C are determined to be

$$R = 4.22 \text{ ohms} \tag{20}$$

$$R_n = 0.237 \text{ ohms} \tag{21}$$

$$C = 2.66 \text{ farads} \tag{22}$$

Returning to the transfer function, we have from (10), (11), and (12)

$$T_D(S) = \frac{\frac{1}{2}}{1 + R\dfrac{\tanh\frac{1}{2}\sqrt{RCS}}{\sqrt{RCS}}} - \frac{\frac{1}{2}}{1 + \frac{2}{R} + R\dfrac{\coth\frac{1}{2}\sqrt{RCS}}{\sqrt{RCS}}} \tag{23}$$

WRH

whence it is seen that the equal zero and infinite frequency gain is not adjustable, as with the Twin-T, and is given by

$$G = \frac{1}{R + 2}$$

$$= 0.161 \tag{24}$$

Finding the magnitude of the transfer function of the Bridged T is more difficult than for the Twin T. First define

$$\left|T_D(j\omega)\right|_n = (R + 2)\left|T_D(j\omega)\right|$$

$$= \left|\alpha + j\beta\right|$$

$$= \sqrt{\alpha^2 + \beta^2} \tag{25}$$

α and β are given by

$$\alpha(\omega) = \frac{\frac{R+2}{2}\left[1 + \frac{R}{2X}\frac{\sinh X + \sin X}{\cosh X + \cos X}\right]}{\left[1 + \frac{R}{2X}\frac{\sinh X + \sin X}{\cosh X + \cos X}\right]^2 + \left[\frac{R}{2X}\frac{\sinh X - \sin X}{\cosh X + \cos X}\right]^2} - \frac{\frac{R+2}{2}\left[\frac{2}{R} + 1 - \frac{R}{2X}\frac{\sin X - \sinh X}{\cosh X - \cos X}\right]}{\left[\frac{2}{R} + 1 - \frac{R}{2X}\frac{\sin X - \sinh X}{\cosh X - \cos X}\right]^2 + \left[\frac{R}{2X}\frac{\sinh X + \sin X}{\cosh X - \cos X}\right]^2} \tag{26}$$

$$\beta(\omega) = \frac{\frac{R(R+2)}{4X}\left[\frac{\sinh X - \sin X}{\cosh X + \cos X}\right]}{\left[1 + \frac{R}{2X}\frac{\sinh X + \sin X}{\cosh X + \cos X}\right]^2 + \left[\frac{R}{2X}\frac{\sinh X - \sin X}{\cosh X + \cos X}\right]^2} - \frac{\frac{R(R+2)}{4X}\left[\frac{\sinh X + \sin X}{\cosh X - \cos X}\right]}{\left[\frac{2}{R} + 1 - \frac{R}{2X}\frac{\sin X - \sinh X}{\cosh X - \cos X}\right]^2 + \left[\frac{R}{2X}\frac{\sinh X + \sin X}{\cosh X - \cos X}\right]^2} \tag{27}$$

with

$$X = 2.365\sqrt{\omega} \tag{28}$$

and R given by (18).

Examples

To compare the performance of the two networks on an equal insertion loss basis, we must have for the Twin T

$$\frac{T_1}{2(T_1 + \sqrt{T_1^2 + 1})} = 0.161 \tag{29}$$

This may be solved for T_1 to yield

$$T_1 = 0.539 \tag{30}$$

Using (5), the Twin T bandwidth is

$$\Delta\omega = 2.272 \text{ radians} \tag{31}$$

Using (2), (3), and (4), we find

$$R_1 = 2.11 \text{ ohms} \tag{32}$$

$$C_1 = 0.51 \text{ farad} \tag{33}$$

$$C_2 = 0.88 \text{ farad} \tag{34}$$

$$R_2 = 0.306 \text{ ohm} \tag{35}$$

Equations (7), (25), (26), and (27) have been plotted in Figure 3. Two facts are immediately apparent. First, the bandwidth of the Bridged T is about 16.73 radians, as contrasted with 2.272 for the Twin T. This is factor of about 7.4 to one. Second, the Twin T response displays excellent geometric symmetry, while the Bridged T response is very asymmetric, rising much more slowly on the high side of resonance. This slow rise of the response above the notch frequency compared to the behavior below has been noticed before in unloaded distributed notch filters.[5,6] This behavior may, for example, cause trouble when one attempts to construct wideband stagger tuned bandpass amplifiers using cascaded Bridged T feedback amplifiers. It might be possible to secure more symmetrical response in the region of the null at the expense of the behavior at extreme frequencies by changing the RR_n product, but this topic, together with the application to feedback amplifiers, is beyond the scope of this paper. It is also to be noticed that, for given responses at zero and infinite frequencies, the bandwidth is fixed and cannot be traded with insertion loss, as in the Twin T.

Conclusion

It has been shown that the analysis of even a simple device employing distributed RC networks may be a formidable task with results that are not necessarily superior in all respects to those obtainable with conventional lumped components. It may be possible to improve the performance by the use of different circuit configurations; Castro and Fuller[8] have derived other three and four terminal distributed constant null networks, but only a little information about them is given, and

their over-all selectivity characteristics are unknown. Tapered distributed networks might also be considered to see if they afford any improvement, but it is suspected that the analytical gymnastics required to analyze any of these networks is sufficient to make it more profitable to use lumped microcircuit components wherever possible.

References

1. L. Stanton, "Theory and Application of Parallel-T RC Frequency Selective Networks," Proc. IRE, July 1946, pp. 447-456.
2. L.G. Cowles, "The Parallel-T RC Network," Proc. IRE, December 1952, pp. 1712-1717.
3. Y. Oono, "Design of Parallel-T RC Networks," Proc. IRE, May 1955, pp. 617-619.
4. D.H. Smith, "The Characteristics of Parallel-T RC Networks," Electronic Engineering, February 1957, pp. 71-77.
5. A.B. Smith, "Rejection Filters with Distributed R and C," Electronic Components Conference, Washington, D.C., May 1960.
6. W.M. Kaufman, "Theory of a Monolithic Null Device and Some Novel Circuits," Proc. IRE, September 1960, pp. 1540-1545.
7. P.S. Castro and W.W. Happ, "Distributed Parameter Circuits and Microsystem Electronics," Proc. 1960 NEC, pp. 448-460.
8. W.D. Fuller and P.S. Castro, "A Microsystems Bandpass Amplifier," Proc. 1960 NEC, pp. 139-151.
9. J.J. Jones, "Distributed Parameter RC Immittance," Sylvania ARL ARM-245, May 1961, unpublished.
10. E.S. Kuh and D.O. Pederson, "Principles of Circuit Synthesis," McGraw-Hill, 1959.
11. H.B. Dwight, "Tables of Integrals and Other Mathematical Data," Third Edition, MacMillan, 1957.
12. C.K. Hager, "Network Design of Microcircuits," Electronics, September 4, 1959.

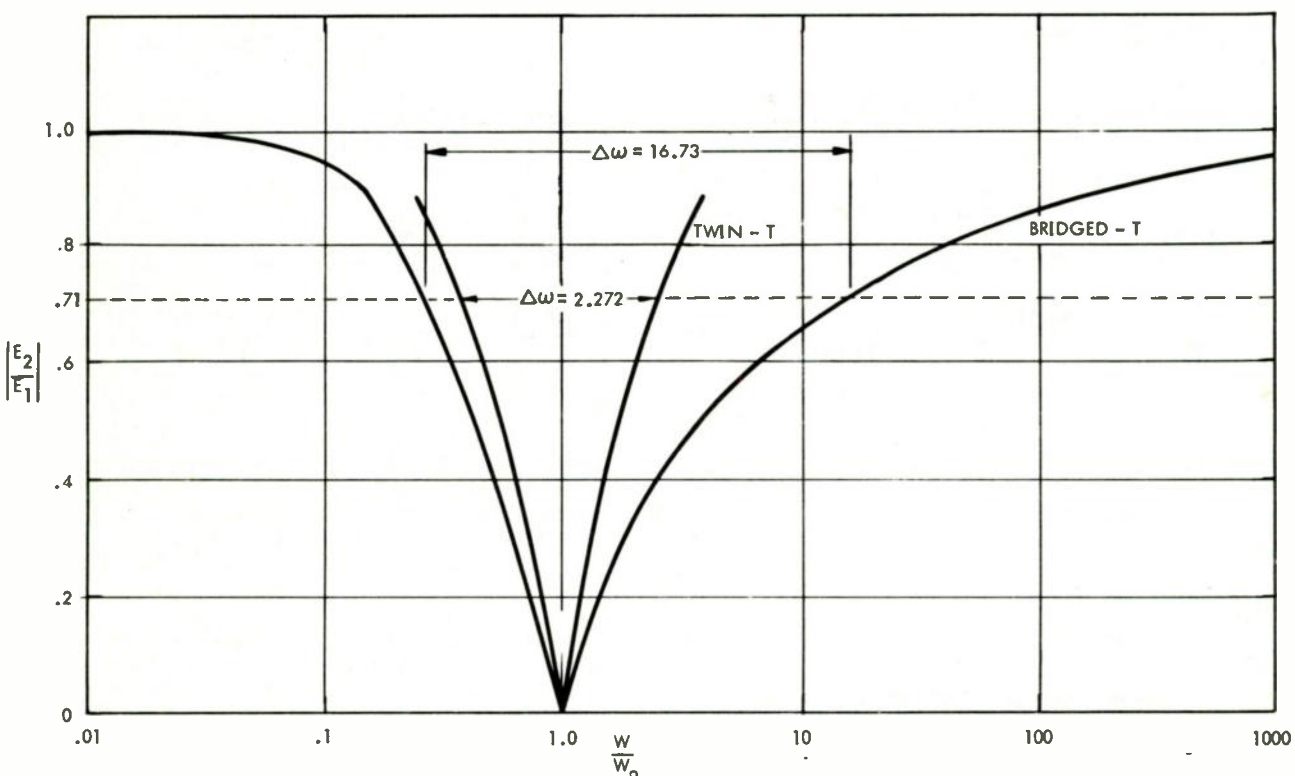


FIGURE 3 SELECTIVITY CURVES FOR TWIN - T AND BRIDGED - T

THE ELECTRICAL PROPERTIES OF PORCELAIN ENAMELS AND CERAMIC COATINGS, AND FACTORS AFFECTING THEIR PERFORMANCE

Gerald Geltman
Isomet Corporation
Palisades Park, New Jersey

Porcelain enamels and ceramic coatings are defined. Methods of manufacture and differences and similarities are discussed. Porcelain enamels and ceramic coatings are considered analogous, respectively, to glass and glass bonded dielectrics except that they are generally thinner and have one electrode attached prior to any further processing. Inherent coating characteristics such as thickness, oxide diffusion, and pinholes are discussed in their relation to high quality coatings. The effect of coating composition, methods of varying composition, thermal treatment, process variables, and end-use are correlated to electrical properties. Typical examples of coatings used for various metal substrates are offered along with measurements of electrical properties.

Introduction

The origin of vitreous or porcelain enameling dates back to antiquity when the Egyptians and Chinese first applied glass to noble metal. The process rapidly developed into a fine art used for making small decorative objects of enameled gold, silver, bronze, and copper. Today, porcelain enamels and ceramic coatings can be applied to a wide range of commercially-available metals.[1,2] They have been used for kitchenware, sanitary ware, architectural applications, chemical process equipment and protection of aircraft and missile components at high temperatures. The uses for these coatings are being rapidly expanded in the electronics industry in answer to the need for stable, inert materials capable of operating at 500°C. Great potential exists in the use of these coatings as insulators for:

1. wires
2. cables
3. motor and transformer housings
4. transformer laminates
5. heating elements
6. induction coils
7. resistors
8. printed circuits
9. electro-luminescent devices

and as conductors for:

1. printed circuits
2. electro-luminescent devices .

Capacitors made by stacking coated foils are already a reality and will soon be practical.[3]

Traditionally, porcelain enamel is defined as "glass fused to metal." It is obvious that this definition is broad, covering any glass composition which develops adherence to metal. The adherence being largely the result of a compatible coefficient of expansion and the development of a chemical bond.

Enamels, which mature at temperatures between 1000°F and 1700°F, are made by melting refractory materials, such as quartz and feldspar, cryolite, lead, lithium, and zinc oxides and with adherence producing or promoting oxides such as cobalt, nickel, and manganese oxides. The smelting is done at about 2200°F to 2800°F from four to six hours. The melt is usually quenched in cold water which chills the glass and causes it to break up to form a "frit." The frit is then finely ground in a mill and held in suspension by the use of suit-

able electrolytes for the purpose of applying to either a cast or sheet metal form. Certain inorganic oxides may be included in the mill to impart color, opacity, and a variety of other characteristics. After preparation of a slip the coating is applied to the previously cleaned base metal. The coatings are then fired at a temperature and for a period of time suited to the particular system of glass and metal. Coatings must be applied at temperatures below the deterioration point of the base metal. Coatings for aluminum, for example, must mature at temperatures less than 1050°F. Ceramic coatings differ from porcelain enamels at the "mill" stage. Mill additions of up to 15 percent may be added in the case of porcelain enamel. Ceramic coatings,on the other hand, may have mill additions of up to 50 to 60 percent that drastically alter the properties of the coatings. Porcelain enamel is essentially glass whereas the ceramic coating is glass with a large volume of dispersed, inorganic, particulate matter.

General Discussion

Glass-Enamel Analogy

In the porcelain enamel type of coating, it can be assumed that the electrical properties of the coating, unaffected by interaction with the base metal, are analogous to the properties of equivalent compositions of glass. Similarly, it is useful to extend the analogy by considering ceramic coatings as high-glass-bonded dielectrics.

Effect of Interface

The effect of the metal-coating interface is of course significant.[4] The oxidation of the base metal and the solution of that oxide in the coating may be more or less serious depending on the alloy. However, oxidation and solution of the oxide in the glass is important in the development of adherence.

In a simple glass (.18 Na_2O, and .82 SiO_2), aluminum oxide decreased while iron oxide increased the specific resistance.[5] It is true that a homogeneous glass presents a picture that is very different from a heterogeneous graded system which is typical of enamels on iron.

The effect of copper oxide generated at the copper-coating interface on the electrical properties of copper enamels was studied.[6] Platinum was coated with glasses of varying copper oxide additions and correlated with coated copper plates. With less than 5 percent copper oxide, marked changes occurred in the coating as a function of percent copper oxide. The changes were accentuated at elevated temperatures. At 5 to 20 percent copper oxide, for coatings fired at 1600°F, room temperature properties were unaffected. At elevated temperatures, as copper oxide content increased there was a decrease in the dielectric constant and dissipation factor. It was recommended that copper oxide be added to the mill to stabilize the electrical properties.

Although no comprehensive study of the electrical properties of glasses with silver or silver compound additions has been reported, it has been stated[7] that silver migrates into glass with an applied d.c. current. The implication is obvious: silver electrodes applied to glass or coating surfaces may be responsible for erratic measurements especially at elevated temperatures.

Effect of Composition

Morey[8] states that at all temperatures glasses are electrolytic conductors. Resistivities at temperatures of 25° to 1200°C may range from 10^{19} to 1 ohm. Surface resistance measurements are difficult due to absorbed moisture and are not always reliable. The surfact conductivity may be very high in alkali glasses, particularly so at high humidity. Electrical conductivity of glass depends on the composition, temperature, and to some extent on the surrounding atmospheric conditions.

Surface conductivity of glass may be ascribed to a film of water condensed on its surface.[9,10] The conductivity varied with humidity and the ratio of Na_2O/K_2O. Volatile constituents effect the electrical resistance, especially

water which is probably most prevalent. Heating in vacuo to 350°C increased the volume resistivity six-fold.[11]

In terms of electrical strength, such constituents as Na_2O and Li_2O should be eliminated. PbO, CaO, B_2O_3, on the other hand, are suitable for reducing the conductance.[12] In an 0.18 Na_2O-0.82 SiO_2 glass substitution for various oxides show that the alkalies greatly increase conductivity, Na_2O more so than K_2O. Al_2O_3 slightly increases the conductivity whereas Fe_2O_3 up to 35 percent decreases the conductivity. CaO increased the resistivity considerably while BaO, MgO, PbO, and ZnO did so to a lesser degree.[13]

Dielectric losses in alkali-silicate glasses were observed with additions of cations (Ba, Sr, Cd, Pb, Mg, Zn, Al, B, Sn, Zr, Ti).[14] In borate glasses, conductance increased with monovalent cation concentration (Ba, Sr, Ca). Oxides of divalent cations produced large decrease in loss.[15]

The relationship of chemical stability to electrical performance is an interesting one. It was shown that high acid resistance compositions had higher resistivity than low acid resistant materials at room and high temperatures[16]. In another study of glaze and overglaze ("on glaze"), high SiO_2 and PbO were found to improve the overglaze chemical resistance. High alkali or alkaline earth additions lowered chemical resistance. It was shown using radio isotopes that sodium ions migrate from the glaze to the overglaze.[17] It can be assumed that one coating applied for adherence containing sodium ions covered by a non-sodium containing coating would not be a satisfactory approach since, at high use temperatures, the sodium ions would migrate.

Hirayama and Berg[18] showed that low temperature resistivity is dependent on composition and previous thermal history. Activation energy for ionic motion is larger in the high temperature regions. It was demonstrated that the high mobility of alkali ions also gives high dielectric losses.

Moore et al[19] cited the superiority of an alkali free ceramic coating containing almost 30 percent mill added chromium oxide (Cr_2O_3) over a conventional porcelain enamel. (Compositions are shown in Table 1.)

Table I

Oxide Formulas of Various Coatings
(Parts by Weight)

Frit	(1)	(2)	(3)	(4)	(5)
SiO_2	49.48	55	28.6	44.1	28.40
B_2O_3	15.24	10	1.0	13.3	4.46
Al_2O_3	7.37	-	-	-	1.90
PbO	-	-	41.7	-	-
BaO	-	-	-	-	30.29
CaO	5.21	-	-	-	2.75
ZnO	-	-	-	-	3.44
Na_2O	13.75	15	9.4	9.9	-
K_2O	3.54	5	8.8	3.3	-
F_2	3.44	5	-	5.0	-
MnO	0.76	-	-	-	-
CoO	0.56	-	-	-	-
NiO	0.65	-	-	-	-
TiO_2	-	10	9.7	18.9	-
Li_2O	-	-	2.9	-	-
P_2O_5	-	-	-	2.2	-
Mill Addition					
Cr_2O_3	-	-	-	-	28.76
TiO_2	-	-	0-30	-	-
CuO	-	1-20	-	-	-

Resistivity measurements were in agreement with the Rasch-Hinrichsen Law:

$$\ln R = \frac{A}{T} + B$$

where:
R = electrical resistivity of the material
T = absolute temperature
A and B = constants

It is apparent that certain chemical or physical properties can be obtained only at the expense of an electrically undesirable chemical addition to the composition.

Thermal Treatment

In addition to the obvious problem of producing a glass with excellent adherence to metal, but which possesses good electrical properties, is the effect of

thermal treatment on performance. Generally at the point of maturity, the enameled ware is abruptly withdrawn from the furnace. Specimens are subjected to a thermal treatment, more often, required to produce a sound coating in all respects except electrical.

The heating and cooling of two materials with different coefficients of expansion will create stresses in the system. These stresses will have various effects.[20,21] For example, annealed specimens of glass showed a decrease in resistance with time while samples cooled in air from 450°C showed an increase in resistance with time. This, it is reported, is due to the chilled glass corresponding to an equilibrium condition at temperatures higher than that of the experiment and the annealed glass corresponding to a temperature lower than the experiment. Annealing decreases the conductivity of the glass due to removal of mechanical strain.[22,23]

Rinehart[24] showed that irreversible changes in dielectric losses occur as a result of heat treatments of glass below the transformation range.

Coefficient of Expansion

Harman[25] has pointed out the loss factor in steatites appears to go up with increasing difference between the coefficients of expansion of the ceramic and glass phases. It is obvious that if the coefficients of expansion of glass, metal, and filler differ widely, the dielectric strength and loss characteristics will be adversely affected.

Thickness of Coating

The dielectric strength given in volts per mil (rms) varies with the thickness of the test specimen. In general, the dielectric strength decreases with the increase in thickness, being proportional to the two-third to three-quarters power of thickness. Tests of different thicknesses are not comparable.[26] The coatings are applied through the use of suspension (usually in water) of finely-divided material, in some cases sprayed and in other cases dipped. Precise control of thickness is difficult and uniformity of coating on irregularly-shaped objects impossible.

It is well to keep in mind that the majority of ceramic coatings and especially coatings on complex shapes and wire are usually less than 5 mils in order to improve mechanical strength of coating.

Increasing the coating thickness or decreasing the bend radii of wire winding was reported to increase cracking of coatings on wires.[27]

Coatings over 5 mils are also usually the result of several applications and firings, a process which may not be practical.

Bubble Pinholes and Cracks

Many other factors besides composition have an effect upon the dielectric properties: mechanical stress as a result of thickness, gas bubbles, pinholes, and interface oxidation and interaction. The density of porcelain enamels for steel varies from about 2.3 to 3.0 gm/cc. With lead oxide and other heavy compounds, the density will be higher. Significant variations in a particular composition, however, are possible due to trapped gas bubbles. The degree of bubble entrapment will effect capacitance and breakdown. In an opaque coating, it may be impossible to detect a subsurface bubble which reduces the effective coating thickness 50 percent or more. It is, therefore, an additional stumbling block to obtaining consistent results. The amount of bubbles depends on the nature of electrolyte, application vehicle which must be evaporated, gasses liberated from the base metal during firing, and firing temperature.

Bubbles trapped in the coating and interaction are probably more important with thin coatings while stress causing cracks is probably more a problem with thick coatings. Bubbles are more likely, however, in thick coatings due to the tendency to underfire and the difficulty of trapped gas to reach the surface in less than a minute of lowered coating viscosity. The partially coalesced hole resulting from a burst gas bubble at the surface leaves a fine imperceptible pinhole.

Careful microscopic examination of numerous transparent coatings has shown that small cracks appear to originate at bubble sites. In addition to the effect

on capacitance and other electrical properties, bubbles increase the dielectric loss by initiating cracks and contributing to the stress.

Electrical Properties

Table I lists typical compositions for various base metals. Enamel No. 1 is a typical porcelain enamel used for steel. In this case, it had been applied to Inconel. Ceramic Coating No. 5, high in chrome oxide, was also applied to Inconel.

Moore et al[28] reported for alkali free Ceramic Coating No. 5 at 100°C resistivities of $10^{12.4}$ ohms cm. and at 400°C 1.6×10^{8} ohm cm. Enamel No. 1 at 100°C was $10^{10.5}$ ohm cm. and at 400°C was measured at 2.5×10^{5} ohm cm. No significant change in resistivity with thickness ranging from 5 to 20 mils was observed.

Enamel No. 2[29] was used for coating copper. Resistivity measurements were shown without copper oxide additions to go from 10^{14} ohm cm, at 100°F to $10^{3.5}$ at 500°F. With 1 percent CuO at 700°, resistivity was 10^{9}.

Compositions 1, 2, 4, and 5 were fired at 1500°F or higher. Composition No. 3 was used on aluminum and copper with and without mill additives and was fired at 1000°F. Room temperature resistivity values calculated from insulation resistance measurements were made of identical surfaces by first applying colloidal graphite and then air drying silver. The results are shown in Table II.

Table II
Properties of Enamel No. 3 on Copper Plates

Electrode	Colloidal Graphite		Air dried Silver	
Coating Thickness	5 mils	3	5	3
Insulation Resistance (KMΩ)	2350	1200	2550	1370
Breakdown (KV rms)	5.2	2.0	4.3	2.6
Resistivity ($\times 10^{14}$)	2.68	2.81	2.65	2.52
Dielectric Strength (volts/mil)	1040	680	850	880

Average resistivities of approximately 2.5 to 2.8×10^{14} were calculated. Dielectric strengths went from 500 to a high of 1370 volts per mil. The average was 860 volts per mil. Any differences in the use of colloidal graphite versus silver can be attributed to experimental error. Overfiring of enamels on copper plates, thereby increasing the copper oxide content, was found to improve the dielectric strength.

The coatings used were high in lead in distinction to the type of coatings reported by other investigators.

Table III gives capacitance, loss tangent, and dielectric constant for low-melting enamels on copper plates.

Table III
Dielectric Properties of Enamel No. 3 2.5 Mils Thick on Copper

Frequency (cps) KC	Capacitance (uuf)	Loss Tan.	Dielectric Constant (K)
1000	256	0.060	7.5
100	255	0.002	7.4
1.0	258	0.006	7.5
0.1	263	0.009	7.7

Although ceramic materials, coatings applied to metal, will resist the destructive effects of prolonged high temperatures or high temperature cycling, their electrical properties are varied at performance temperatures. It is plain from Figures 1 and 2 that dielectric strengths drop off rapidly at 300°F. It is interesting that at about 300°F the average breakdown strengths frequently improve. This would, of course, be due to a relaxation, or stress relieving, of the coating and metal system. It should be noted that low temperature compositions deteriorate sooner or more rapidly with increasing use temperature than compositions fired to maturity at higher temperatures. In Figure 1, Coating No. 3 with 15 percent TiO_2 was applied to aluminum 0.025 inches thick.

In Figure 2, Coating No. 3 applied to aluminized steel in thicknesses varying from 3 to 5.5 mils gives a somewhat different picture. It may be due to the thickness variation, but there is no precise correlation with thickness and breakdown, as previously stipulated.

Bergeron[30] reports for Coating No. 4 which contains a high percentage of smelt added titania a maximum dielectric strength of 485 volts per mil. Titanium dioxide is used as an opacifier in kitchenware enamels. For the best opaque blue-whites, the anatase form of TiO_2 is desired. Rutile tends to give a yellowish color and is usually due to high temperature firing which lowers coating viscosity and increases crystal growth. The opacifying effect of smelter added TiO_2 is due to a tendency of the dissolved TiO_2 to recrystallize during the reheating of the enamel in firing. Anatase predominates in low temperature enamels. Where TiO_2 is smelter added, high anatase enamels have highest reflectance.

An anomoly in the results of breakdown tests of titanium oxide mill additions to a low melting (No. 3) porcelain enamel was reported.[31] It has been subsequently established that the crystal form appears to have an important effect on breakdown. Figure 3 shows the effect of the incremental additions to the mill. Two plots are shown, one of the effect of mill added rutile up to 15 percent, and the other of mill added anatase.

Although only limited data was collected, mixtures of rutile and anatase gave results higher than any other values achieved.

Dielectric constants of coatings with titania smelter and mill additions tend to be high, ranging from 7 to 22. The dielectric constant of titanium dioxide is approximately 75. Increases in Coating No. 3 of the lithium and boron content to 2.5 and 3 percent respectively resulted in lower dielectric constants and higher breakdown strengths. Coatings with 15 percent mill added TiO_2 had resistivities of 7.5×10^{13} ohm cm.

In work performed to provide high dielectric low constant coatings for copper wire, Friedberg et al[32] added silica and boron nitride. The results have been selected in such a way as to illustrate the effect of selected mill additions on electrical properties. Table IV clearly illustrates the advantages and improvements possible in mill modification of porcelain enamels. Boron nitride was added and resulted in the lowest dielectric constants achieved.

Table IV

Electrical Properties of Porcelain Enamel* and Ceramic Coatings on Copper Wire

Coating Type	R.T.	Frequency in KC .4	10	100	Coating Thickness in in.
Porcelain Enamel	DF%	9.2	1.8	2.9	0.0002
	C(uuf)	462	4437	425	
	K	8.11	7.64	7.44	
Ceramic Coating-Porcelain Enamel Plus 40% mill added Silica	DF%	1.64	.52	.40	0.00055
	C(uuf)	151	146	145	
	K	5.7	5.6	5.5	
Ceramic Coating-10% mill Added Boron Nitride	DF%	0.48	1.0	1.0	0.0007
	C(uuf)	37.0	35.8	33.8	
	K	3.43	-	3.13	

* After Ref. 32

Practical Applications

Weaver and Terry[33] described the successful coating and sealing of transformers. Martin[34] described construction of an electroluminescent lamp in which a porcelain enameled steel plate is one electrode and a dielectric second coat of enamel contained the dispersed luminescent phosphor. The phosphor is included as a mill addition. The electroluminescent lamp is presently a commercial product. Plastic types are in use, but are not as stable or long lived as the ceramic type.

Coatings similar to Composition No. 5 were used as the dielectric in an electromagnetic clutch arrangement for precision tape machines. In order to guarantee uniform thickness, the coatings were made oversize, ground to size, and refired to seal the surface.

The decorative finishing of electric appliance housings with enamel or ceramic coating provides an additional safety feature.

Porcelain enameled or ceramic coated metal plates makes an inexpensive, structurally sound base for printed circuitry. Portions of the circuit can then be coated with a dielectric. The examples cited are only a few of the possibilities inherent in the properties of porcelain enamels and ceramic coatings.

Conclusions

It is expected that porcelain enamel and ceramic coatings will take their places along with other coating methods in the rapidly expanding development of new materials and application. These coatings can be varied over a wide compositional range to possess properties formulated specifically for use in electrical devices.

References

1. A.I. Andrews, Porcelain Enamels, Garrard Press, Champaign, Ill., 1961.
2. W.E. Pierce, "Properties of Porcelain Enamels," Materials in Design Engineering, 97-102, July 1960.
3. A.E. Javitz, "Research Horizons - Capacitor Research," Electro-Technology, 11, Sept. 1961.
4. W.H. Fischer, "Electrical Resistivity of Ceramic Coatings," Jour. Electrochem. Soc. 105, 201-3, April 1960.
5. G.W. Morey, The Properties of Glass, 483, Reinhold Publishing Corp., New York, Second Edition 1960.
6. Bergeron, Schwarzlose, Friedberg, "Electrical Insulating Properties of Porcelain Enamels on Copper," Jour. Amer. Ceram. Soc., 42 (10) 451-459, (1959).
7. Hirayama and Berg, "Electrical Resistivity of Glasses at Low Temperatures," Bull. Amer. Ceram. Soc. 40 (9) 551-554 (1961).
8. G.W. Morey, op. cit. 465.
9. G.W. Morey, op. cit. 474.
10. P. LeClerc, Silicates ind., 19, 237-242, (1954).
11. Bush and Connell, Jour. Franklin Inst., 194, 231-240, (1922).
12. G. Glaser, Ber. Dent. Keram. Ges., 28, 404-409, (1957), (Ceramic Abstracts 36,1952).
13. G. Morey, op. cit., 489.
14. Rinehart and Bonino, "Dielectric Losses of Some Simple Ternary Silicate Glasses," Jour. Amer. Ceram. Soc., 42 (3) 107-112, (1959).
15. Shartsis and Shermer, Jour. Amer. Ceram. Soc., 37 (1) 544-551, (1954).
16. Salmang and Holler, Sprechsaal, 66, 474, (1933).
17. Franklin, Tindall, Dinsdale, Trans. Brit. Ceram. Soc., 59 (10) 401-423, (1960), (Ceramic Abstracts 44 (9) (1961).
18. Hirayama and Berg, op. cit.
19. Strauss, Richards, Moore, "Effect of Temperature on the Electrical Resistivity of Several Ceramic and Silicone-Type Coatings," ASTM Special Tech. Pub. No. 153, Jan. 1954.
20. Hirayama and Berg, op. cit.
21. Morey, op. cit. 486.
22. Morey, op. cit. 483.
23. Zhdanov and Kyzmetsov Doklady, Akad. Nauk S.S.S.R., 85, 586-589, (1952), (Chem. Abstracts, 49, 2806 (1955).
24. D. Rinehart, "Dielectric Loss and State of Glass," Jour. Amer. Ceram. Soc., 41 (11) 470-475, (1958).
25. C.G. Harman, "Method for Controlling Loss Factors of Dielectrics," Ceramic Industry, August, 1960.
26. ASTM Standard D149-55T, 1958.
27. King, Kelly, Duckworth, "Ceramic Coated Copper Wire," Bull. Amer. Ceram. Soc., 38, 251-255, (1959).
28. Strauss, Richards, Moore, op. cit.

29. Bergeron, Schwarzlose, Friedberg, op. cit.

30. C.G. Bergeron, "What Are The Electrical Properties of Porcelain Enamels," Ceramic Industry, 82-86, April 1960.

31. Geltman and Huppert, "Trends in Ceramic Coatings for Light Metals," Ceramic Age, Sept. 1959.

32. Friedberg et al, "High Temperature Electrical Insulating Inorganic Coatings on Wire," WADC Tech. Report 58-12, Parts I, II, III, March 1958 to May 1960.

33. Weaver and Terry, "Ceramic Materials Protect Hi-Temperature Transformers," Ceramic Industry, 73, August 1958.

34. A.V. Martin, "How to Light Up Porcelain Enamel and Glass," Ceramic Industry, 73, August 1958.

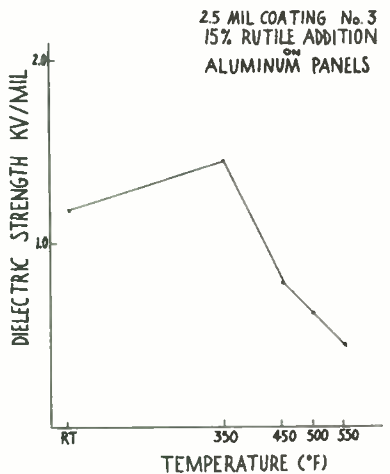


Fig. 1. Effect of temperature on dielectric strength of porcelain enamel with 15 percent rutile mill addition.

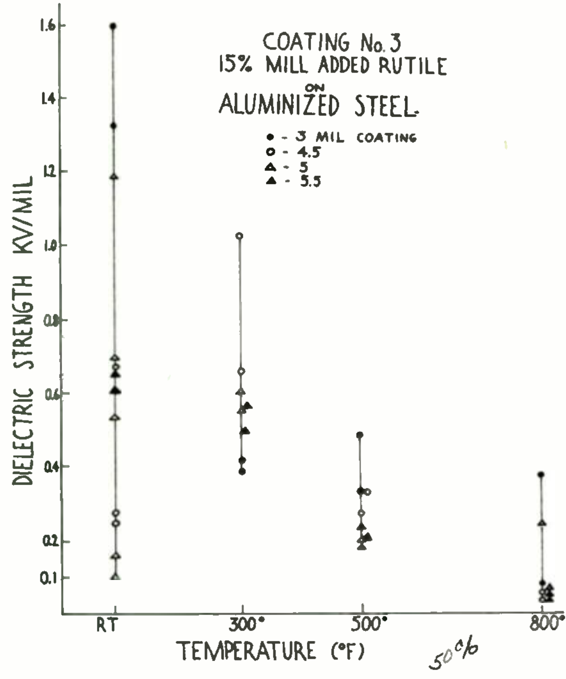


Fig. 2. Effect of temperature on dielectric strength of porcelain enamel on aluminized steel.

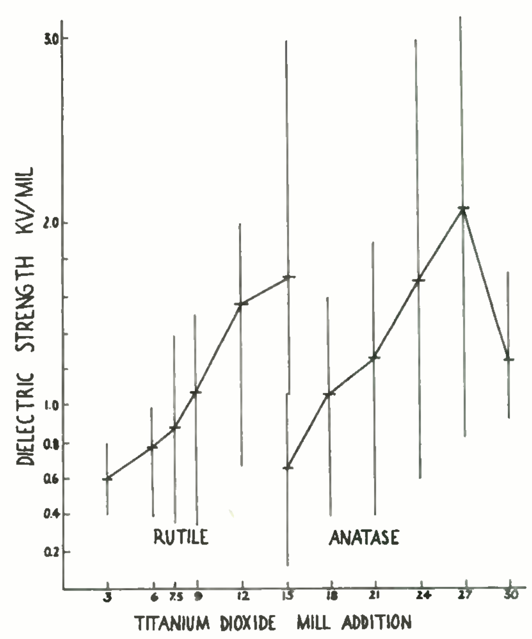


Fig. 3. Effect of rutile and anatase mill additions on dielectric strength of low melting coating on aluminum.

PROPERTIES OF THIN FILM AND SILICON SOLID STATE COMPONENTS - THEIR EFFECTS ON MICROCIRCUIT PERFORMANCE

J. A. Ekiss, J. Roschen and P. G. Thomas

Philco Corporation, Lansdale Division
Lansdale, Pennsylvania

Introduction

A large number of different microminiaturization approaches exist for synthesis of analogue and digital circuits. These approaches vary in degree of complexity from use of discrete microcomponents to fabrication of entire circuit function in a single block of semiconductor material.

Two approaches for fabrication of passive components have been extensively investigated. These are:

1. the sputtering, photoengraving, anodization and evaporation of thin metallic films to form resistors, conductors and capacitors;

2. photoengraving, diffusion, evaporation and alloying of selected regions of a silicon crystal is used to form resistors, conductors and capacitors.

We shall refer to these as the thin film and silicon solid state circuit approaches, respectively.

From the standpoint of the circuit designer, each approach results in a set of passive components with certain characteristics and capabilities. The choice of which technique should be utilized in a given application depends not only on the particular application but also on these characteristics and capabilities.

In this paper the use of thin film and silicon solid state passive components are compared with regard to applicability to digital and linear (analogue) circuits. We ask the question: "Given these two types of passive components, what are the advantages and disadvantages of each method from the standpoint of realizing a desired level of circuit performance?". The answer is provided by examining particular digital and analogue circuits.

Component Characterization

The intelligent design of microcircuits using either thin film passive or silicon solid state components requires detailed knowledge of the characteristics of these components. In general, the circuit designer requires information concerning several, if not all, of the following characteristics.

1. range of component values available
2. tolerance of components
3. statistical distribution of component values
4. temperature coefficient of component value
5. component model (for example, equivalent circuit or transfer relationship)
6. effects of various environmental stresses on component performance
7. component cost and how related to component value, reliability, tolerance and fabrication
8. component interaction (electrical and thermal)

Table I provides the basic information concerning range of component values available, tolerance and temperature coefficients being realized at present. Detailed information on component cost and reliability is beyond the scope of this paper.

The precise nature of the component model does in many cases require detailed knowledge of how the component has been fabricated. This is because of the effects of stray interactions between components within and on the same substrate. If it is assumed that parasitic elements are negligible we may define the circuit models for the "ideal" thin film and silicon solid state lumped components. These ideal circuit models are shown in Figure 1.

The thin film resistor is very nearly an ideal element except for the presence of a stray capacitance, the magnitude of which depends on the resistor geometry. There is, of course, inductance in series with R but it is negligibly small.

The thin film capacitor, as fabricated from Ta_2O_5 or Si O, departs from an ideal capacitor because of the presence of series and parallel resistance. The form for the thin film capacitor model has been deduced from measurements of capacitor quality factor (Q) as a function of frequency. Typical values of the parasitic elements are $10 \leq R_S \leq 150\ \Omega$ and $R_p > 1$ meg Ω.

The model required to represent the silicon solid state resistor is highly dependent on how it is fabricated in the silicon substrate and how it is isolated from the other components within and on the substrate. If the resistor is isolated by means of a pn junction, the equivalent circuit for the solid state resistor takes the form shown in Figure 1. The distributed capacity Δ C represents the capacitance of a reverse biased pn junction. The distributed resistance Δ r represents the resistance of the bulk semiconductor material.

The silicon solid state capacitor is formed from a reverse biased pn junction. The circuit model for the capacitor shown in Figure 1 is conventional. R_S is the series bulk resistance or spreading resistance of the diode, R_p is the leakage resistance. The temperature sensitive current generator shunting the capacitor is the diode saturation current which doubles for every 10°C increase in junction temperature. The range of the parasitic components in this model are $1 \leq R_S \leq 100$, $10^7 \leq R_p \leq 10^{10}$, and $10^{-10} \leq I_r \leq 10^{-6}$ (at 25°C).

Relation Between Properties of Passive Components and Circuit Performance

Digital Circuitry

In this section we are interested in determining how the properties of thin film and silicon solid state circuits effect the performance of digital data processing circuits. The general problems faced by the designer of such circuits are as follows:

1. Power dissipation per stage
2. Temperature range
3. Choice of active components
4. DC design criteria
5. Reliability
6. Circuit speed
7. Cost

In order to discuss more specifically how the above factors are effected by the use of thin film and silicon solid state circuit techniques, three forms of basic digital building blocks (gates) will be discussed. These circuits were chosen because they typify pulse circuits in general use.

The most important DC figure of merit for the logic gate is the maximum number of inputs and outputs to the gate. The number of possible inputs and outputs is dependent upon the specification of the active device (especially current gain), the tolerance, temperature coefficient, changes in parameter values with life and the value of power dissipation per stage.

Figure 2 shows a curve of the maximum number of outputs (with 3 inputs) versus variation in resistor values (change in resistance due to tolerance, aging and temperature) for a (RTL) resistor transistor logic NOR gate[1]. Thus this curve shows the DC performance capability of the gate and illustrates the necessity of holding resistance variations to a minimum. Since in this type of logic an M input gate requires M + 2 resistors per transistor and since the resistance variations due to tolerance, aging and temperature are lower for thin film resistors, this technique seems to hold a strong advantage for RTL. Notice that for resistor variations greater than 20% the gate shown in Figure 2 is rendered useless. The specification of the transistor is as follows:

h_{FE} - 40
I_{CO} - 10 μa
V_{BE} - 0.95 V
V_{SAT} - 0.2 V

Figure 3 shows how the required current gain (h_{FE}) varies with the % change in resistance values for a (DTL) diode transistor logic NAND gate[2]. The plot is given for the normally encountered range of logic that keeping the resistor variations low is doubly important; first because the required h_{FE} is lower and secondly because of the range of R_K to achieve minimum h_{FE} is larger.

This curve dramatically shows the trade off between the specification of active and passive components in a microelectronic block. Since the number of resistors is three regardless of the fan-in and fan-out this type of logic would lend itself to either the thin film or solid silicon approach so long as the

resistor tolerance can be kept below 10%. The following transistor and diode specifications were used in the above design.

$V_{D(forward)} = 0.85$ max

$V_{D(forward)} = 0.30$ min

$I_{DO(reverse)} = 10$ µa

The transistor specification is the same as given above.

The next form of logic considered, (LLL) low level logic[3], contains only one resistor per stage; thus one would expect that the change in circuit performance as a function of resistance variation would be less than that observed for RTL and DTL. Figure 4 shows this to be the case. Here curves are given for fan-out (DC drive capacity) required h_{FE} (transistor specification) and the required dissipation per stage versus % resistance variation.

Notice that the fan-out curve shows a smaller change than for RTL, the required h_{FE} curves shows a smaller change than for DTL and that the resistance variation must reach almost 30% before the required power dissipation per stage is doubled. In this form of logic the ratio of diodes and transistors to resistors is high which would probably dictate a silicon solid state approach.

Other forms of logic in which the transistor is used as the gating element as well as for amplification, such as (DCTL) direct coupled transistor logic, will also yield a high ratio of active to passive components, and thus dictate a silicon solid state approach.

The departure of thin film and silicon solid state capacitors from that of the ideal component will effect the AC design of digital circuits. Since the leakage currents of silicon solid state capacitors may be as high as 1 - 10 µa at +125°C, the DC design may also be effected.

Both stray and distributed capacitance are very important in determining the transient response of the inverter stage.

As an example of the effect of parasitic elements, the curve in Figure 5 shows how the turn off time of a transistor varies with the series resistance of the speed up capacitor. The series resistance sets the rate at which clean up charge may enter the transistor. An approximate expression for the turn off time is given as follows:

$$t_{off} \approx \frac{R_S + r_b'}{V} \times Q_T \qquad (1)$$

where R_S = capacitor series resistance
r_b' = base spreading resistance of the transistor
Q_T = total charge stored in transistor when saturated
V = voltage change across capacitor.

From the curve of Figure 5 it is seen that one should expect a 10 to 50% increase in transistor turn off time over that obtained by using discrete capacitors.

Linear Circuitry

Linear circuits form the second broad classification of circuits types for which some form of microelectronics techniques will prove useful. Specific circuits within this category are oscillators and linear amplifiers (tuned and untuned). These circuits process analogue (as opposed to digital) information.

The general problems faced by the circuit designer, arising from active and passive component characteristics as well as system requirements, may be summarized as

1. bias point stability
2. choice of quiescent operating point
3. power dissipation
4. circuit performance should be independent of temperature and aging
5. a wide range of component values and characteristics are required
6. component characterization (needed for equivalent circuit models)
7. reliability
8. cost

A circuit which illustrates how thin film and silicon solid state capacitors influence design and performance is the RC phase shift oscillator[4] of the form shown in Figure 6. The resistor R_C, R_{B1}, R_{B2} and R_E set the DC operating point. C_E must by-pass R_E at the operating frequency.

The stability of the bias point may be influenced by the change in I_{CO}, collector base leakage current, with temperature. With the capacitors, realized with reverse biased pn junction diodes, there is the additional factor of the capacitor

leakage to consider. The stability of the operating point with respect to changes in temperature may be found from

$$-V_{CE(T)} = V_{CC} - \left[I_{C(T)} + I_{1(T)}\right]\left[R_E + R_C\right]. \quad (2)$$

The collector current, I_C, is related to the thermal stability factor, S, by

$$S = \frac{\partial I_C}{\partial T} = \frac{R_{B1}\, R_{B2}}{R_E\,(R_{B1} + R_{B2})} \quad (3)$$

$$I_{C(T)} = I_{C(T_o)} - SI_{CO(T_o)} + SI_{CO(T)} \quad (4)$$

where T_o is the reference temperature (25°C).

$I_{CO(T)}$ approximately doubles for every 10°C increase in junction temperature. The capacitor leakage, $I_1(T)$, may be expected to do the same. Operation of the phase shift oscillator at relatively low frequencies (500 cps - 5 kc) requires large values of capacity. These are obtained by use of a large area pn junction, thus resulting in large values of leakage current and subsequent instability of operating point. With increasing temperature, V_{CE} decreases which may result in a decrease in transistor current gain to such an extent that the oscillator stops. The required common emitter current gain (neglecting phase shift through the transistor) is

$$h_{fe} \geq 23 + 29\frac{R}{R_C} + \frac{4\,R_C}{R} \quad . \quad (5)$$

This must be satisfied over the entire temperature range.

An important consideration in any oscillator is frequency stability with respect to temperature changes and aging. The nominal frequency of oscillation of the phase shift oscillator is

$$\omega = \frac{1}{\sqrt{6}\;RC\left[1 + 2R_C/3R\right]^{1/2}} \quad . \quad (6)$$

In general the change in frequency with temperature will be due to changes in R_C, R and C. For sake of illustration we will evaluate the effect of changes in R and C only. If we define the temperature coefficient of R_C and R as

$$\alpha = \frac{\partial R}{\partial T}\bigg|_{T = T_o} \quad (7)$$

We may explicitly determine the temperature dependence of the oscillation frequency as

$$\frac{\omega(T_o)}{\omega(T)} = \frac{(1 + x)\left[1 + \frac{2k}{3}\left(\frac{1 + x/k}{1 + x}\right)\right]^{1/2}}{\left[1 + 2\,k/3\right]^{1/2}}, \quad (8)$$

where $k = \dfrac{R_C(T_o)}{R(T_o)}$,

and $x = \dfrac{\alpha\,(T - T_o)}{R(T_o)}$.

The normalized variation of frequency with temperature is shown in Figure 7 for $K = 0.5$ and $10^{-4} \leq \alpha/R(T_o) \leq 2.5 \times 10^{-3}$. This covers the range of temperature coefficient of resistance that may be encountered in thin film and silicon solid state circuit resistors. In order to provide a reasonable frequency stability, the resistor temperature coefficient must be less than 100 ppm/°C. This indicated the use of thin film resistors.

The phase shift oscillator made with thin film or silicon solid state circuits is not readily tunable (adjustable). Changes in frequency may be made by variation in the supply voltage if reverse biased pn junctions are used for the capacitors. The degree to which oscillator performance will approach design center depends on the tolerance of the passive components. The designability of the oscillator depends on the degree to which the components may be represented by ideal elements. While it is true that this equivalent circuit of the capacitor may be used to analyze the oscillator operation (frequency of oscillation and starting conditions), the analysis and subsequent design are quite complicated.

It is well known that the inductance function is difficult to synthesize in microminiature form. Various schemes that have been employed include use of an appropriately constructed semiconductor diode, thin magnetic films, extremely small toroidal cores and use of non-planar RC networks.

The key IF amplifier specifications directly influencing the design and fabrication of the microcircuit inductance function will be

1. amplifier center frequency
2. amplifier bandwidth
3. allowable change in amplifier center frequency and bandwidth due to changes in temperature and aging of components
4. allowable perturbation in amplifier bandwidth and center frequency due to adverse effects of component tolerance.

In order to determine the practicality of the null network method for fabrication of a selective amplifier and to gain some insight into the particular component characteristics required in a practical system, the Wien bridge null network was selected as a representative circuit[5].

The ideal Wien bridge circuit is shown in Figure 8 together with a more realistic equivalent circuit of the network that may be realized in practice. The series resistance R_S that has been added in the lumped model representation arises from the capacitor series resistance.

The ideal Wien bridge has a null frequency given by

$$\omega_o = \frac{1}{RC} \qquad (9)$$

The transfer function of the ideal network is

$$\frac{E_o}{E_i} = \frac{1/3}{1 - j\,\dfrac{3\omega/\omega_o}{\omega^2/\omega_o^2 - 1}} \quad . \qquad (10)$$

In order to determine the effect of the series resistance on the null network, the bridge circuit as modified by capacitor series resistance was analyzed. As might be expected, the parasitic series resistance can seriously alter the center frequency from its design value. The quantitative results for the center frequency and transfer function of the practical null network are

$$\omega_o^* = \frac{1}{RC\left[\dfrac{2R_S}{R} + \left(\dfrac{R_S + R}{R}\right)^2\right]^{1/2}} \qquad (11)$$

$$\left(\frac{E_o}{E_i}\right)^* = \frac{\dfrac{1}{3} + j\omega RC\left(\dfrac{1 + 2\gamma}{3} - 1\right)}{1 - j\omega RC\left[\dfrac{1 + 2\gamma}{1 - \left(\dfrac{\omega}{\omega_o^*}\right)^2}\right]} \quad , \qquad (12)$$

where $\gamma = \dfrac{R_S + R}{R}$.

The primary effect of (uncontrolled) parasitic series resistance on the center frequency is to lower it from design center. In addition another undesirable effect occurs in that variation of series resistance from network to network result in uncontrolled variations in center frequency. The undesirable effects of parasitic resistance may be eliminated if the ratio R_S/R can be made small. Unfortunately it is not possible to make R arbitrarily large. Since RC essentially fixes the nominal filter bandwidth, a large R dictates small C for a fixed center frequency. The capacitance C can not be chosen too small or parasitic (stray) capacitance will begin to effect the center frequency. For a typical tantalum thin film capacitor, R_S 50 Ω. The stray capacitance associated with R is C_S 1 pf. If R_S and C_S are not to effect the center frequency, choose

$$\frac{R_S}{R} \leq 10^{-2}$$

$$\frac{C_S}{C} \leq 10^{-2}$$

With these constraints on R and C (which determine center frequency), the maximum filter center frequency which can be realized is 160 kc.

The effect of series parasitic resistance on the transfer function is to increase the phase shift through the null network. There is also a slight, and probably non critical, effect on the amplitude. The increased phase shift through the network due to R_S should be considered when determining amplifier gain stability.

The change of center frequency with changes in temperature and tolerance of components is of extreme importance in the design of an IF amplifier, this aspect of null network design was investigated.

The effect of the finite temperature coefficient of resistance and capacitance can be evaluated by approximating the variation of these quantities with temperature as a linear function.

$$R = R_o \left[1 + \alpha_R \, (T - T_o) \right] \tag{13}$$

$$C = C_o \left[1 + \alpha_C \, (T - T_o) \right] \tag{14}$$

This is a good approximation for thin film resistors and capacitors and pn junction capacitors, but poor for silicon solid state resistors. The rather poor control of the capacitance of a pn junction and the large temperature coefficient of resistance of silicon solid state resistors, probably eliminates these components in this application.

The variation of center frequency with temperature is given explicitly by

$$\omega_o R_o C_o = \frac{1}{\left[1 + \alpha_C(T-T_o)\right]\left[1 + \alpha_C(T-T_o)\right]} \tag{15}$$

Since the temperature coefficients of thin film resistance and capacitance are of opposite sign, we may expect some "compensation" to take place. A variation of center frequency of less than 5% for a 100°C temperature variation would appear to be feasible.

The change of center frequency from network to network because of production tolerances may be evaluated in a similar manner. If the resistance and capacitance tolerances are taken as ${}_R$ and ${}_C$, respectively,

$$\omega_o = \frac{1}{R(o)\ C(o)\ (1 + \delta_R)(1 + \delta_C)}, \tag{16}$$

where R(o) and C(o) are the nominal values of resistance and capacitance, respectively. The effect of component tolerance on normalized center frequency is shown in Figure 10. There is a trade off between the allowable resistor and capacitor tolerances for a specified allowable variation of center frequency. At present, techniques are being developed to control resistance values to ±0.1%. This will allow a capacitance tolerance of approximately ±10% for a ±10% change in center frequency.

It would appear that the use of an RC null network is a practical way to build a selective amplifier. Because of the required close control of component values, the thin film approach provides the best fabrication technique.

Conclusions

The choice of thin film or silicon solid state components requires detailed knowledge of circuit requirements and characteristics of these components. In many cases the thin film approach will allow direct translation of existing (discrete component) designs to microcircuit form. As shown in this paper, however, some of the non ideal characteristics of these thin film components can result in degradation of circuit performance if care is not taken in circuit design. Thin film fabrication techniques are most applicable to those circuits which are characterized by a high ratio of passive to active components.

Silicon solid state passive components are less ideal in their behavior than thin film components. The distributed nature of these components complicates circuit design. For these reasons, silicon solid state components are most applicable to those circuits characterized by large ratio of active to passive components.

Each of these technologies has its place as a method for fabrication of passive components. Perhaps the most desirable technology for microcircuit fabrication will result from the combination of these two passive component fabrication methods.

Acknowledgements

The authors are indebted to P. M. Ansbro, J. A. Hiltebeitel and R. L. Kopski for calculations and J. Pierce and K. W. Roper for gathering technical information.

Bibliography

1. P. Thomas and J. Hiltebeitel, "A Worst Case Design Procedure for Resistor Transistor Logic NOR and NAND Gates" Philco Corp., Lansdale Division, Application Lab Report 719; October 20, 1961.

2. D. P. Masher, "The Design of Diode-Transistor NOR Circuits" PGEC, Vol. EC-9, No. 1; March, 1960.

3. P. M. Ansbro, "An Investigation and Analysis of Low Level Logic Circuitry", Philco Corp., Lansdale Division, Application Lab Report 761, February 27, 1962.

WRH

4. "Survey of the State of the Art in Microminiaturization", Electronic Design; January 18, 1962.

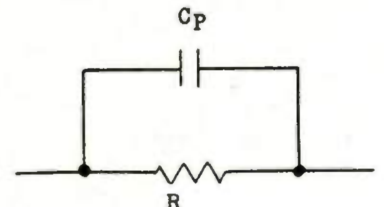
5. R. F. Shea, Transistor Circuit Engineering; John Wiley and Sons; November, 1957.

6. B. Chance, Waveforms, Radiation Laboratory Series, McGraw Hill Book Company; 1949.

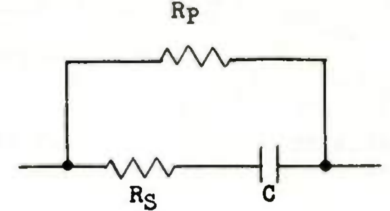
TABLE I

RANGE OF COMPONENT CHARACTERISTICS

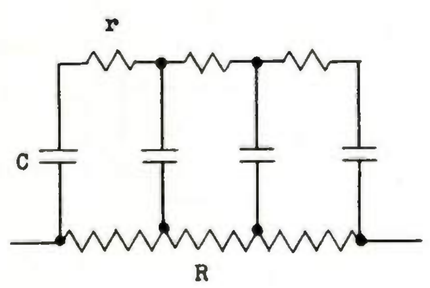
	Characteristic	Thin Film	Silicon Solid State
Resistors	range of values	50 Ω - 1 meg Ω	100 Ω - 50 K Ω
	tolerance	±(0.1 - 10)%	±(5 - 20)%
	temperature sensitivity	-(25 - 300)ppm/°C	50 - 5000 ppm/°C
Capacitors	range of values	0.01 - 5.0 μf/cm² (breakdown voltage 2 - 50 volts)	0.025 μf/cm² (breakdown voltage 1 - 100 volts
	tolerance	±(5 - 20)%	±(10 - 50)%
	temperature sensitivity	150 - 250 ppm/°C	200 ppm/°C



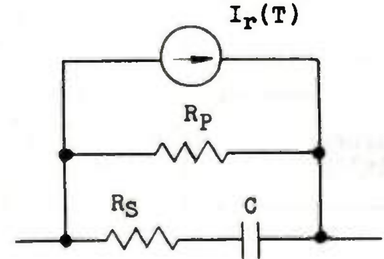
Thin Film Resistor



Thin Film Capacitor



Silicon Solid State Resistor



Silicon Solid State Capacitor

Fig. 1. Component circuit models.

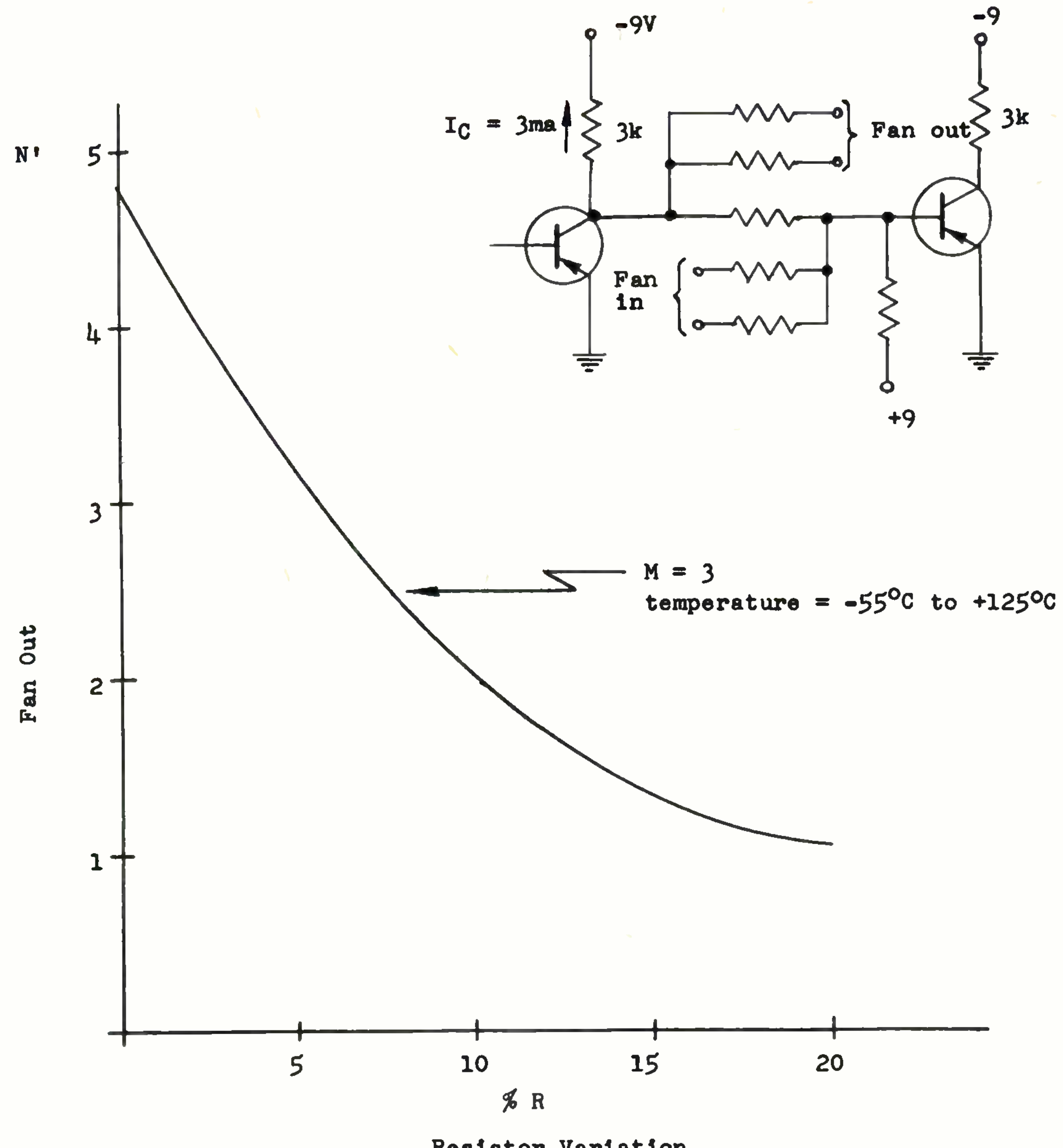


Fig. 2. Fan-out in RTL as determined by resistance variations.

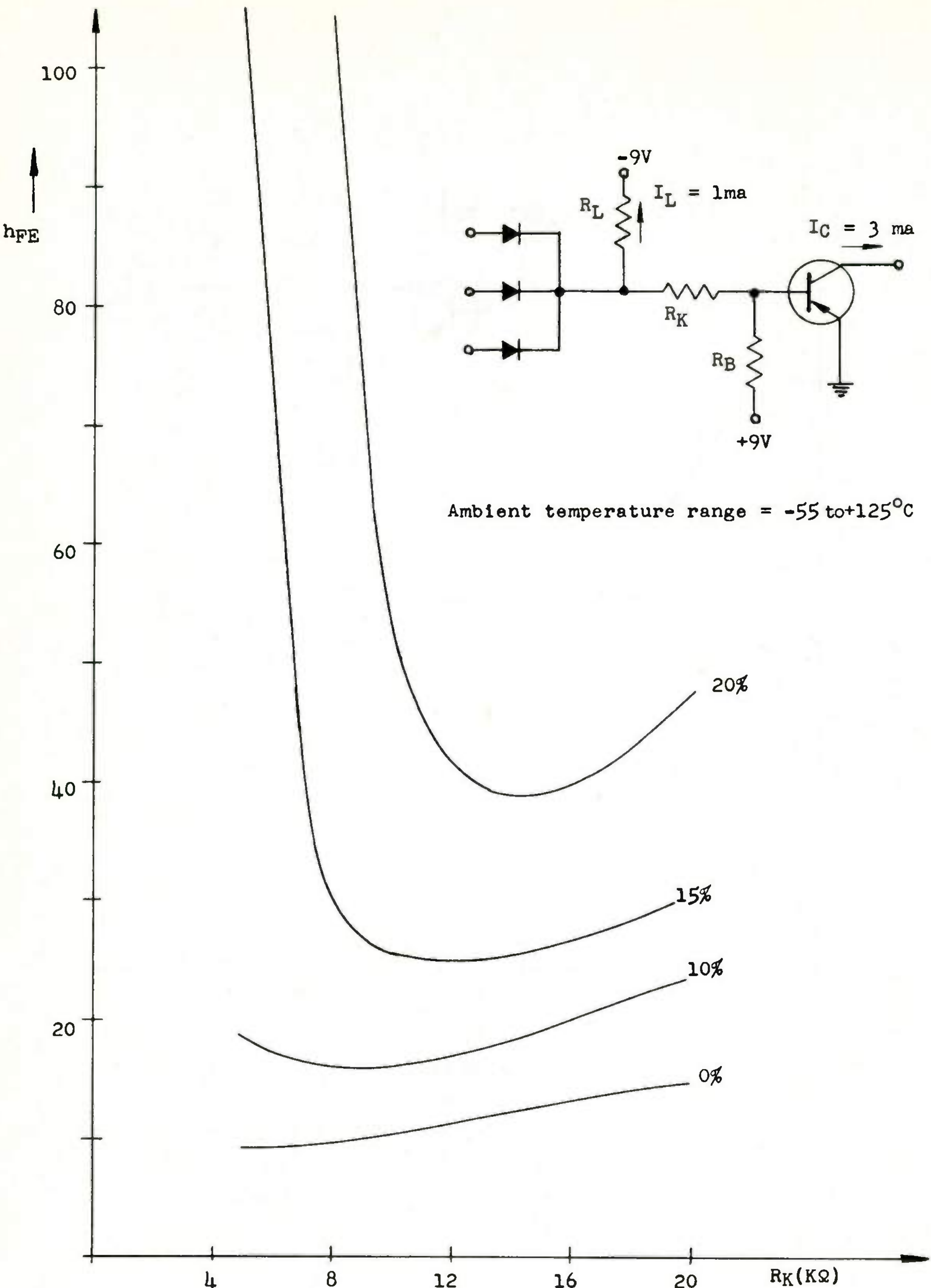


Fig. 3. Required minimum h_{FE} in DTL as determined by resistance variations.

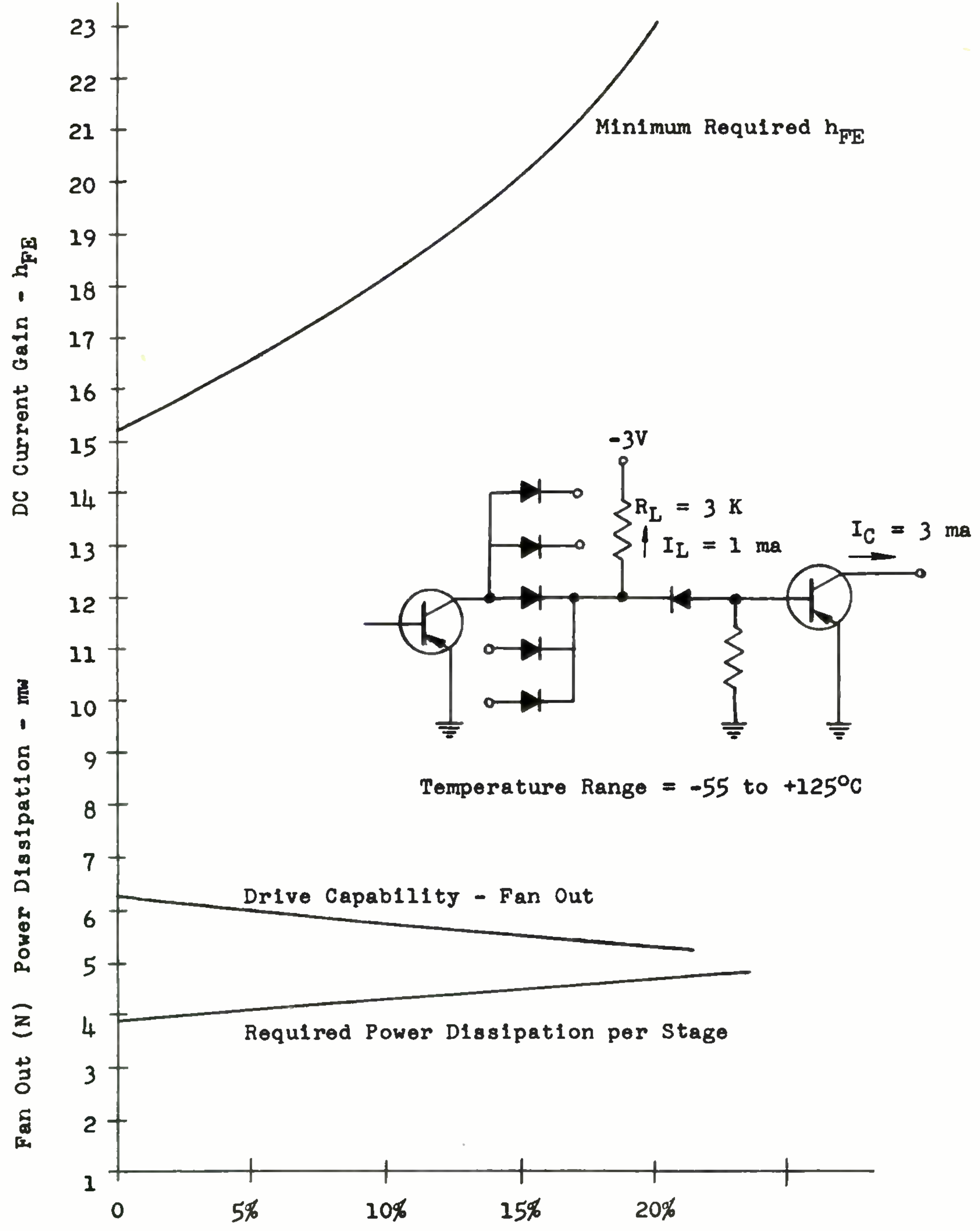


Fig. 4. Dependence of LLL performance on resistance variations.

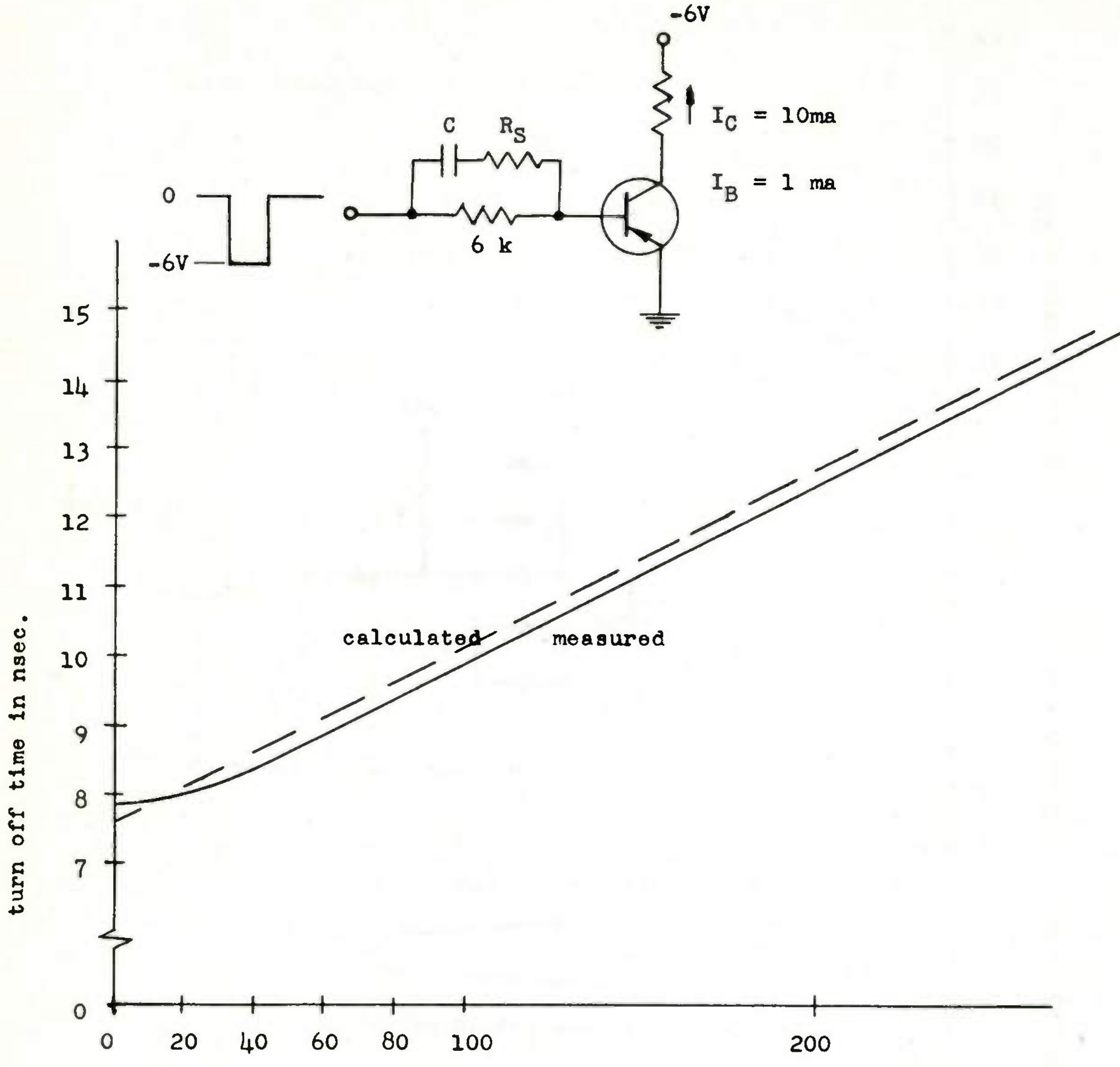


Fig. 5. Transistor turn off time as a function of capacitance series resistance.

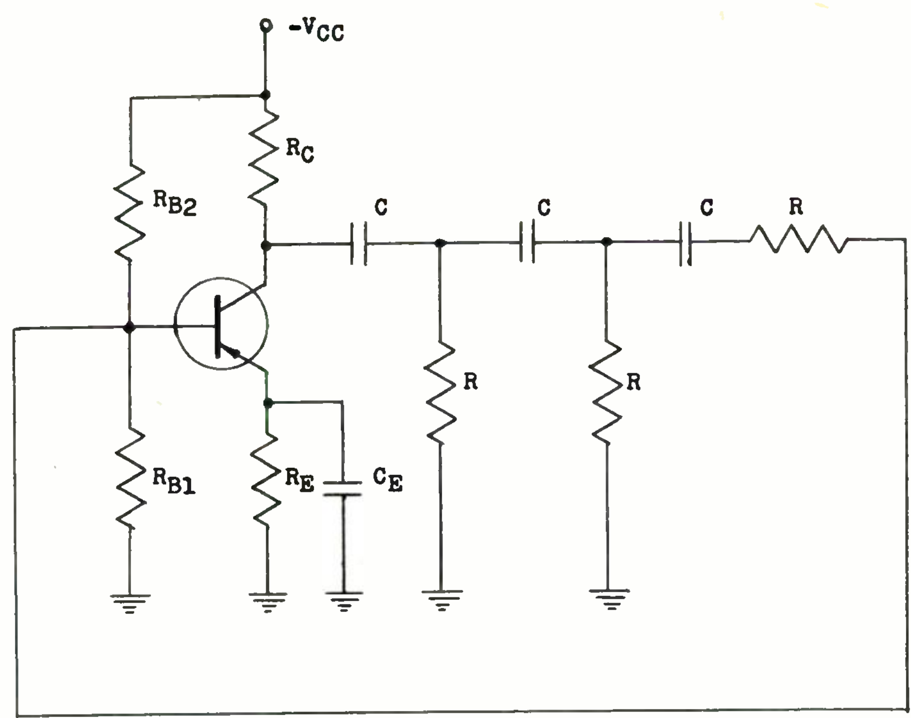


Fig. 6. RC phase shift oscillator.

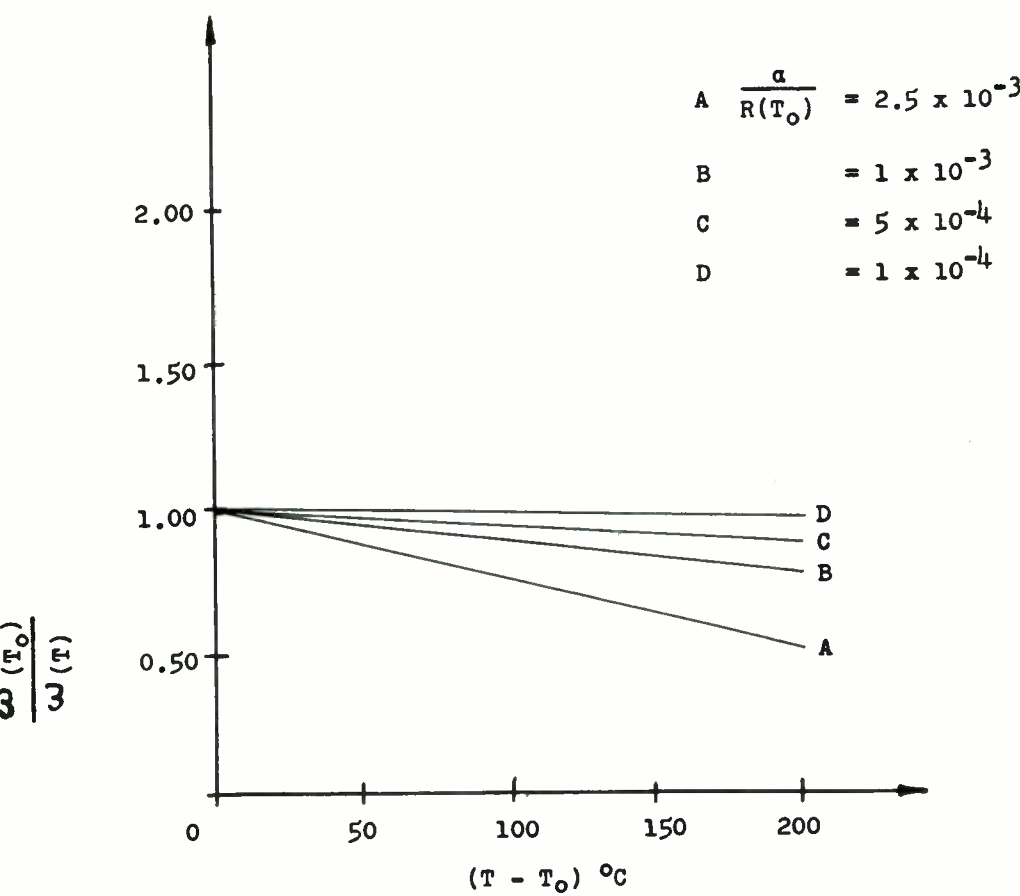


Fig. 7. Variation of Oscillator frequency with temperature.

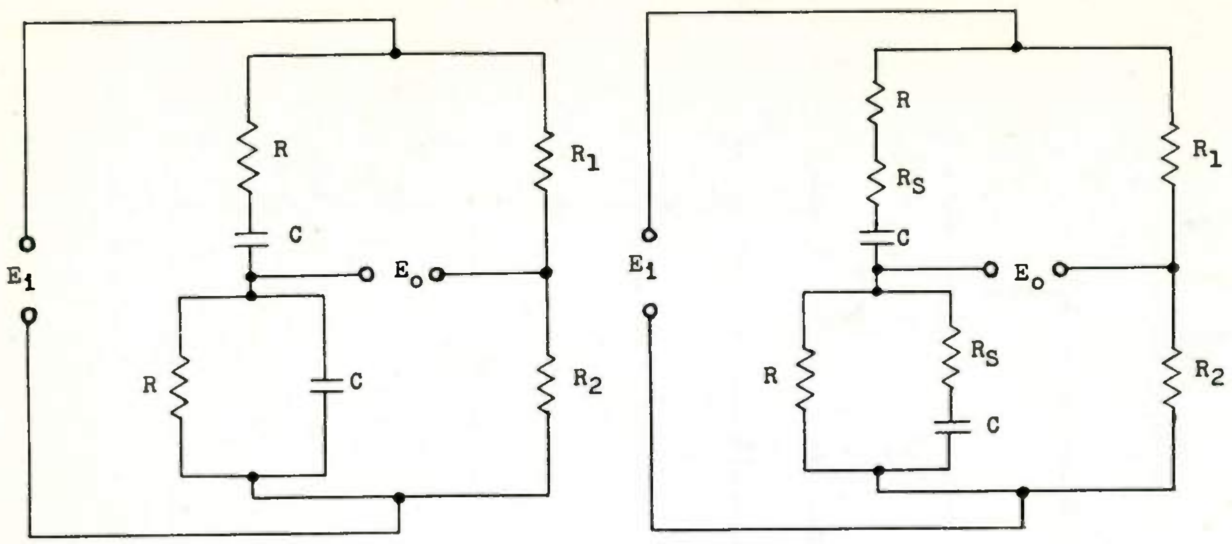


Fig. 8. Wien bridge null network.

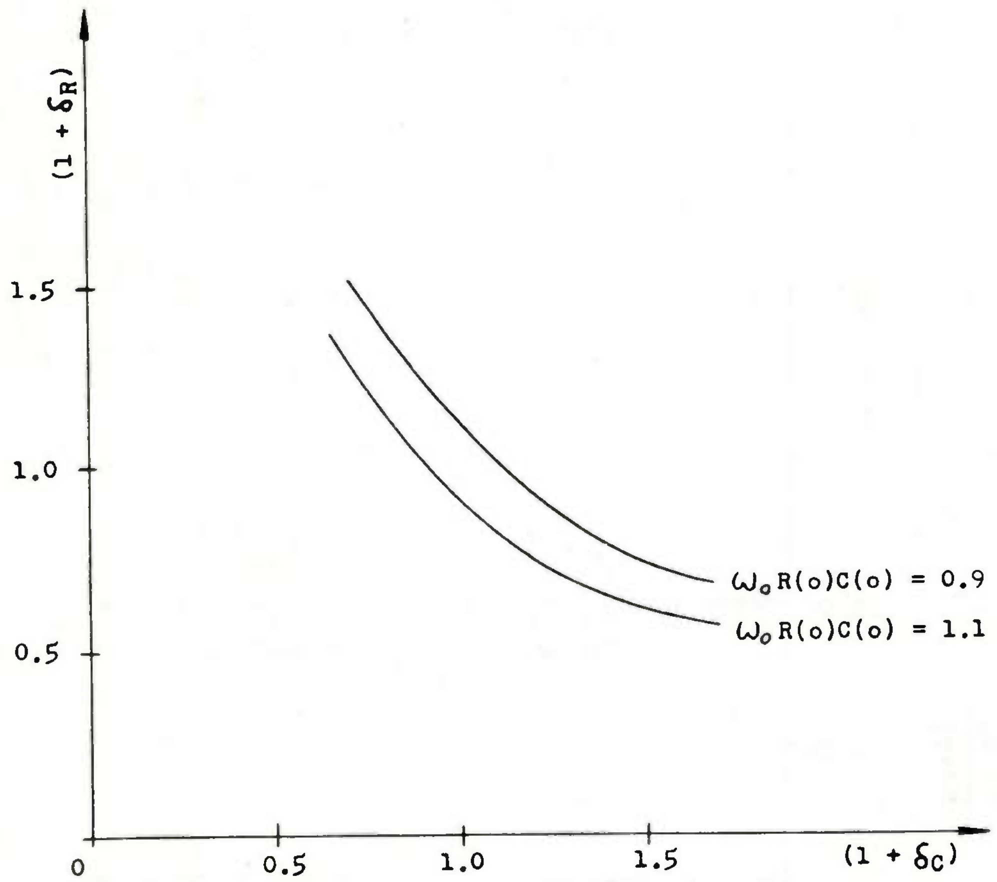


Fig. 9. Change of null network center frequency with tolerance.

TITANIUM INTEGRATED ELECTRONIC COMPONENTS

W. D. Fuller
Microsystems Electronics Department
Lockheed Missiles & Space Company
Sunnyvale, California

Summary

Titanium and titanium oxides have been used in the development of thin film integrated electronic components to meet the space age environmental requirements. Titanium is a refractory material with good conductivity and high radiation resistance. Titanium oxides have a wide range of electronic properties which include permittivity, semiconductivity, and metallic conduction. Chemical processes have been developed for the deposition of thin films of titanium on ceramic substrates, followed by anodic processes for the conversion of portions of the film into selected oxides with specified electronic properties. The combination of chemical processes and refractory materials to yield integrated electronic components will contribute greatly to the capability of system organizations to meet the requirements of reliability, producibility, product versatility, size reduction, economy, and control of system configuration.

Introduction

Technological evolution, particularly in the field of space vehicles, has created new demands on the electronics industry. New electronic systems must meet the ever-increasing requirements of reliability and extended operation under the extreme conditions of a space environment. In addition, demands for increased electronic task capability in limited vehicle volume have established a companion requirement for component size reductions. Product versatility and economy of production costs are still requirements in this space age. One solution to the electronic systems problem established by these requirements is a thin film technology.

The MICROSYSTEMS ELECTRONICS Program at Lockheed Missiles & Space Company is concerned with the development of stable electronic materials, reproducible processes and the application of the electronic phenomena in materials to meet the electronic requirements of the space age. In one phase of this program chemical processes have been developed to fabricate the passive structures of integrated components from the titanium-titanium-oxide system of materials as thin film structures. This program has resulted in an in-house capability to design and fabricate electronic systems to meet exactly the performance, environmental and configuration requirements of a specified task.

Materials and Processes

The environmental requirements for space electronics established the need for highly stable electronic materials, therefore refractory materials were investigated. Refractory materials are considered to be those with melting points in excess of 1500°C. In addition to metals such as tungsten, tantalum, molybdenum, zirconium, hafnium and titanium, this category includes many compounds such as the lanthanides, actinides and some alkaline earth metals. A wide range of electronic properties is found throughout this class of materials.

The Titanium-Titanium Oxide System

The oxides constitute, probably, one of the most interesting material systems. The refractory oxides have electronic properties covering the entire range of conductivity from metallic types to near perfect insulators. For a particular oxide the conductivity is largely a function of the oxidation state: Minimum oxidation states result in metallic conductors, maximum states are insulators, and intermediate states are in the semiconducting class.

Titanium oxides in the polycrystalline form have three well-known stable forms. Titanium dioxide (TiO_2) is an insulator with a resistivity in the order of 10^{14} ohm-cm and a dielectric constant of approximately 100. Titanium sesqui-oxide (Ti_2O_3) has a resistivity in the order of one ohm-cm with a band gap of approximately 0.1 e.v. Titanium oxide (TiO) has a resistivity in the order of 10^{-4} ohm-cm with a temperature coefficient approaching zero. Elemental titanium has a resistivity of 42 microhm-cm with a positive temperature coefficient.

Single crystal titanium dioxide has a vast range of non-stoichiometric forms with resistivity ranging from 10^{14} ohm-cm to 10^{-1} ohm-cm. The non-stoichiometric characteristics include an enhanced dielectric constant approaching 100,000 and an anisotropy of resistance greater than 1000 between the crystal axes. The conductivity characteristics of these two forms of titanium oxides are illustrated in Figure 1.

Substrates

The structural material of a thin film assembly must have physical characteristics compatible with the electronic materials and fabrication process. A comparison of some of the properties of several refractory metals and refractory oxides is shown in Table 1. Aluminum oxide (Al_2O_3) is the substrate material used in the titanium thin film technology. It, like titanium, is highly re-

sistant to degradation in a nuclear radiation environment.

Fabrication Processes

The fabrication, as outlined in Figure 2, of a thin film integrated component starts with a molten salt metallizing process that results in a complete coating of the ceramic substrate with a thick film (5000 Å) of very pure titanium metal. After cleaning to remove salt residues and other surface contaminants, an overcoating of copper is electroplated onto the metallized substrate. This coat is used to establish thick pads for termination areas, as well as a stable resist for subsequent pattern formation. The circuit pattern is then fabricated by typical photoetching processes. The copper remaining on the pad areas after the etching process is diffused into the titanium by thermal processes to yield an alloy to which leads may be brazed, welded, bonded or soldered. External connective leads have also been brazed to these pads during the diffusion process.

The next step in the fabrication process is the conversion of selected areas of the continuous titanium pattern into resistance and dielectric materials. An anodic process is used to produce both of these materials in which the electrolyte and electrical process parameters are the only variables. The resistance material is produced by converting a substantial portion of the thickness of the titanium film into a relatively nonconducting titanium dioxide. This results in a thin titanium film with a protective oxide coating

The dielectric material is produced by converting a small portion of the thickness of the titanium film into very pure titanium dioxide. This leaves a relatively thick titanium film for the bottom electrode of the capacitor. A conductive material is then painted over the oxide to form the top electrode of the capacitor.

The thin film integrated component is finally completed after area conversion by the assembly of active devices to the termination pads by soldering, welding or thermal compression bonding.

In the fabrication of both of these materials the electrolyte is restricted to the selected areas by plastic dams and very little volume of electrolyte is required. During the conversion process either the resistance or equivalent capacity of the processed area is measured on a continuous basis. Instruments have been developed to measure these parameters during the process and to stop the conversion when specified values of the parameters have been reached. This technique of process control has been termed "dynamic testing".

These processes with their control instrumentation have been incorporated into an experimental pilot plant wherein production process specifications are being developed and the effects of process variables on yield are being determined.

Component Characteristics

The characteristics of titanium thin film resistors prepared by these chemical processes are illustrated in Figure 3. The temperature coefficient varies with material thickness from + 120 ppm/°C at 50 ohms/square to -100 ppm/°C at 2000 ohms/square. A zero coefficient exists at approximately 1300 ohms/square. The voltage coefficient varies from 400 ppm at 1000 ohms/square to 1000 ppm at 2000 ohms/square. Combined life testing, both under load at approximately five watts per square inch and no-load at 110°C has shown approximately +0.5% change per 1000 hrs. over a 10,000 hour period. A test in gamma radiation resulted in less than ± 0.5% change in resistance at a dosage of 5.3×10^8 roentgens.

The characteristics of the titanium dioxide capacitor are shown in Figure 4. The capacity is 0.01 µf/cm², the dissipation factor is less than one percent and the leakage resistance exceeds 2×10^4 megohms at 50 vdc. The change in capacitance and dissipation factor with frequency is small. The temperature coefficient of capacity is + 800 ppm/°C and the change dissipation factor with temperature is +25 ppm/°C.

Both titanium and reduced (non-stoichiometric) titanium dioxide have been found to be useful electron emitters for the development of devices utilizing tunneling phenomena. The basic structure consists of an electron emitter, a space charge limited flow region (insulator) and an anode (metal).

In Figure 5 is shown the V-I characteristic of a polycrystalline titanium dioxide diode. The emitter is a layer of thermally produced oxide on the surface of a thin film of titanium on a ceramic substrate. The insulator region was produced by anodizing the surface of oxide film. The anode is silver paint.

In Figure 6 is shown the V-I characteristic of a two-terminal negative resistance titanium dioxide device. This device is fabricated similarly to the preceding diode, but the insulator region is much thinner. The value of the negative resistance is approximately -4000 ohms over a 6 volt range.

Integrated Components

An integrated component may be defined as the functional realization of many interrelated electrical concepts (ρ, ε, μ,) in a physically independent body which cannot be further divided without destroying its stated function. A flip-flop is an example of a functional realization of a prescribed interconnection of resistors, capacitors and active elements, and when it is fabricated as a thin film assembly on a single wafer, it cannot be separated into individual physical parts and still fulfill its stated physical and electrical functions. This same definition fits solid state circuits based upon semiconductor

technology.

Two integrated components fabricated by chemical processes from titanium and titanium oxides on alumina substrates are shown in Figures 7 and 8. The first component is a scale of 16 formed on a single one-inch square wafer. It contains 28 resistors, ranging from 4700 to 22,000 ohms, and 16 capacitors of 56 pf. Eight transistors and 8 diodes are attached by soldering to complete the assembly. Power requirements are 28 milliwatts at 6 volts. Typical operation over a -55°C to +100°C temperature range requires a 2.5 volt trigger input for operation at 1.0 mc/s.

The second integrated component is a NAND function. It contains five resistors and one transistor. The versatility of the thin film technology is illustrated by the two different physical sizes of the function. The smaller unit is formed on a 0.55 by 0.55 inch square substrate and has a dissipation of 100 milliwatts in its normal operation from a 6-volt supply. The larger unit is formed on a 1.5 inch diameter substrate. The same artwork was used for each function with only the photo reduction change. The performance of the integrated components are identical.

Conclusions

LMSC established the requirements for a thin film systems technology early in 1959 and funded through its Independent Research and Development Programs a Microsystems Electronics Program (MSE) with the basic responsibility of developing through research a thin film technology based upon missile and space vehicular electronic system requirements. This program has resulted in a thin film technology based upon refractory materials and chemical processes. In its present state of development it may be termed a "single metal" technology as it is based upon titanium and its oxides to furnish the electrical properties of conductance, resistance, permittivity and transistance. Recent research progress has shown that active, two- and three-terminal, devices operating by field emission can be created from titanium and its oxides. This program has furnished the research confirmation that a thin film system technology uniquely meeting system requirements can be developed.

Acknowledgement

The author would like to acknowledge the contributions of the entire staff of the MSE Department in the development of the technology that has been reported here.

References

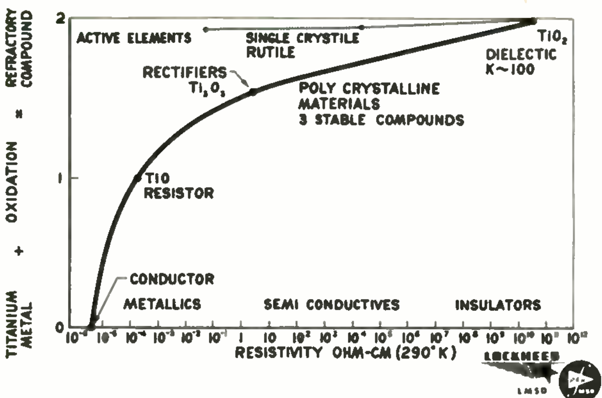
1. M. E. Sibert, Electrical Properties of Refractory Materials, LMSC Report No. 288214, April 1960.

2. O. Kubaschewski and E. LL. Evans, Metallurgical Thermochemistry, John Wiley & Sons, 1956.

3. C. A. Hampel, Rare Metals Handbook, 2nd Ed., Reinhold Publishing Corp., 1961.

TABLE 1 - MATERIAL CHARACTERISTICS[2,3]

CHARACTERISTIC	Ti	Ta	TiO_2	SiO_2	BeO	Al_2O_3
Density - grams/cc	4.54	16.6	4.24	2.32	3.03	3.96
Melting Point °C	1668	2996	1920	1728	2550	2040
Boiling Point °C	3260	6100	dec.	dec.	4120	dec.
Structure Type	Hex.	B.C.C.	Tet.	Tet.	Hex.	Hex.
Thermal Coefficient of Expansion cm/cm/°C x 10^{-6}	8.4	6.6	9.1	0.5	7.5	8.0
Thermal Conductivity cal/sec/cm²/°C/cm	0.041	0.13	0.01	0.003	0.20	0.01
Neutron Cross section barns/atom	5.8	21.3	5.8	0.16	0.01	0.23
Resistivity-ohm-cm	42×10^{-6}	12.5×10^{-6}	10^{14-25}	10^{14}	10^{18}	10^{11}
Dielectric Constant	-	-	100	4.5	11.8	10



ELECTRONIC CHARACTERISTICS OF THE OXIDES OF TITANIUM

Fig. 1.

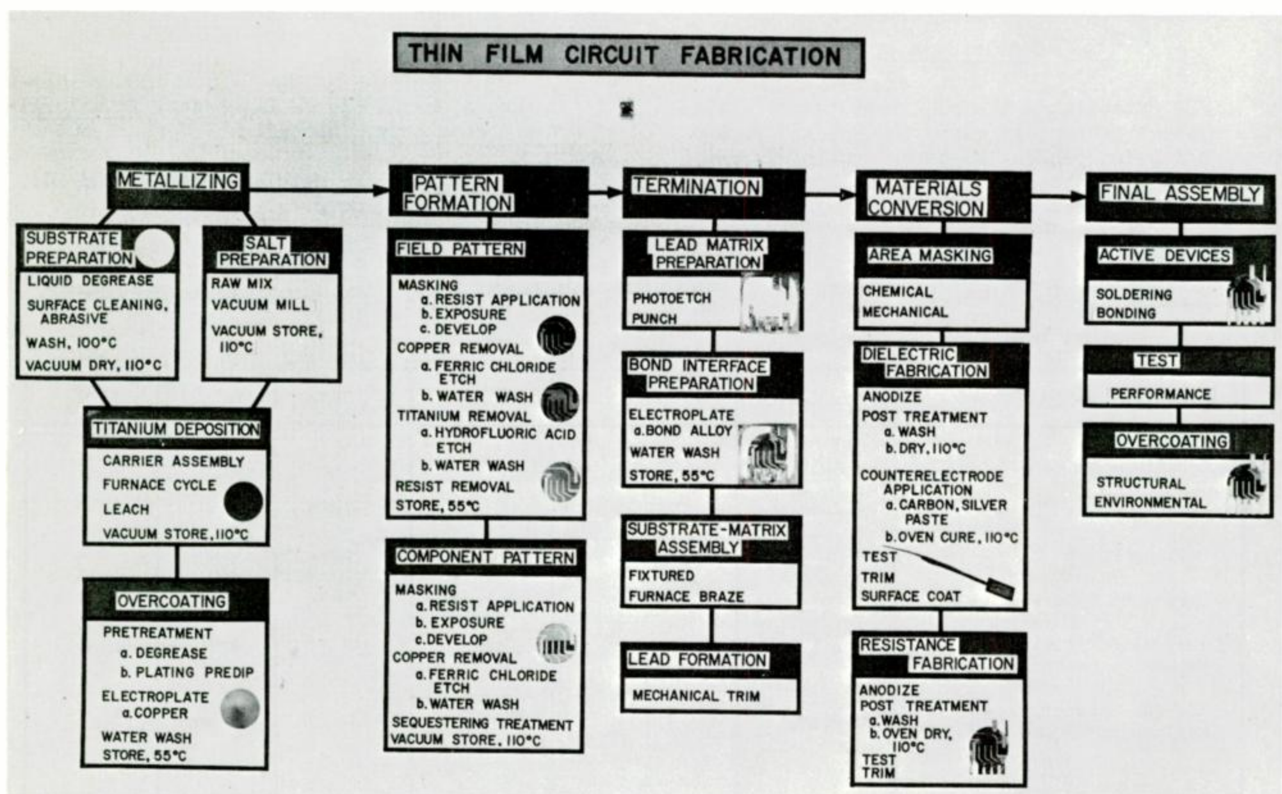


Fig. 2.

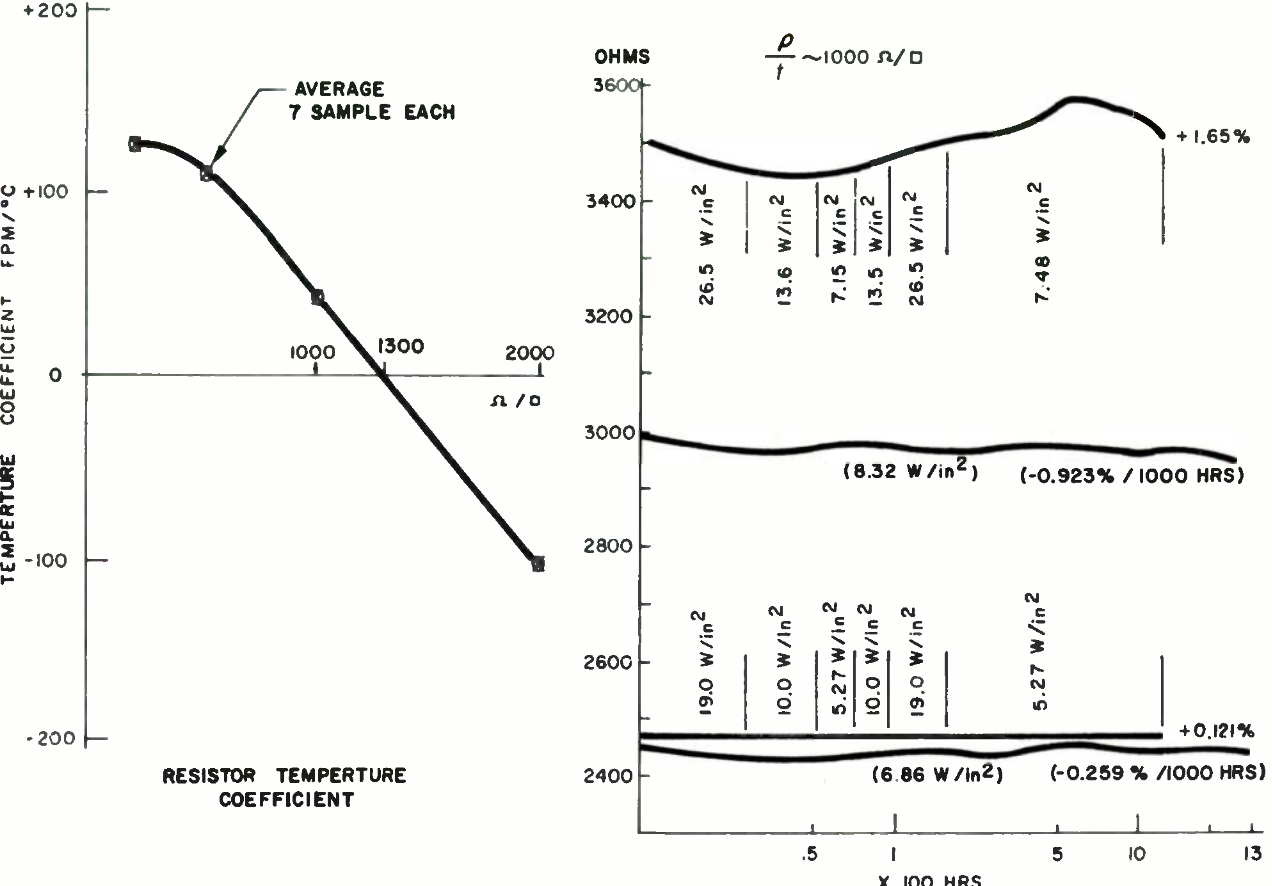


Fig. 3. Resistive characteristics.

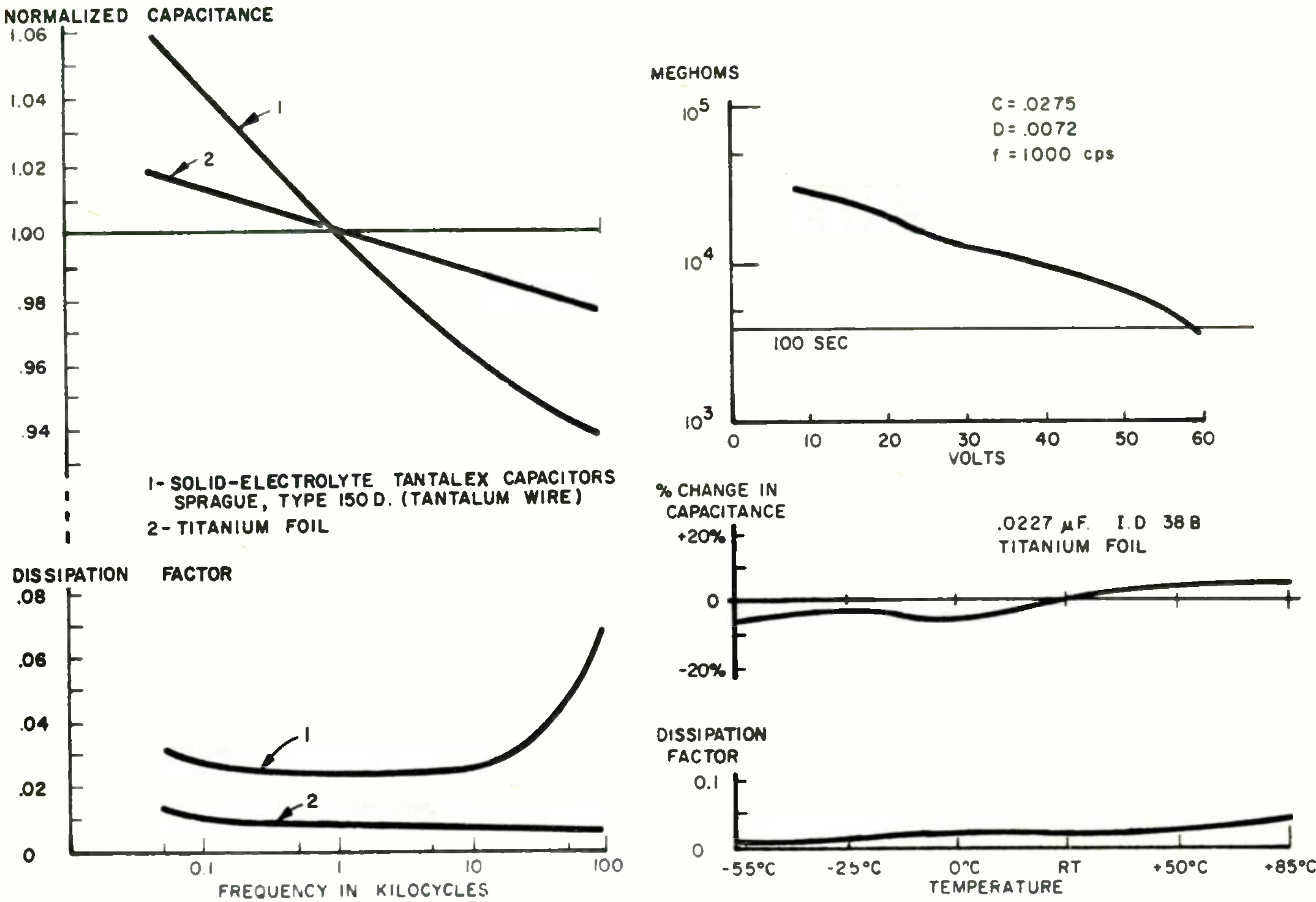


Fig. 4. Dielectric characteristics.

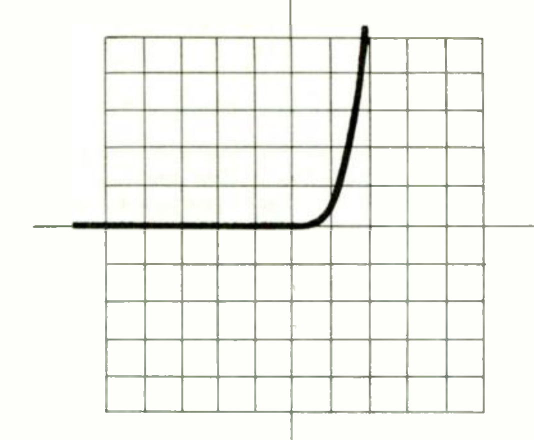


Fig. 5. TiO_2 diode.

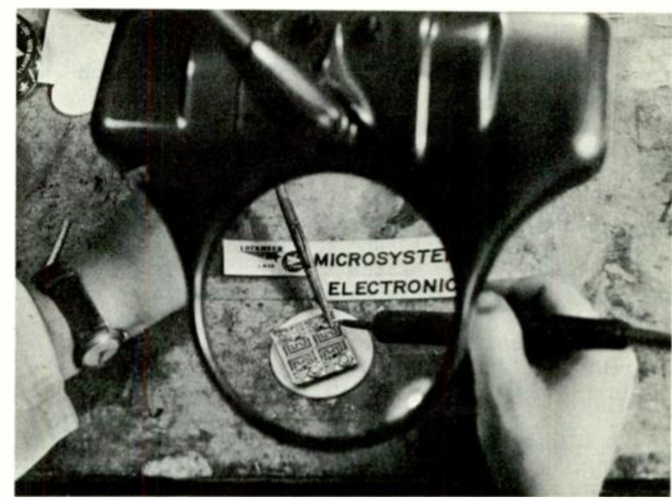


Fig. 7. Integrated component scaler.

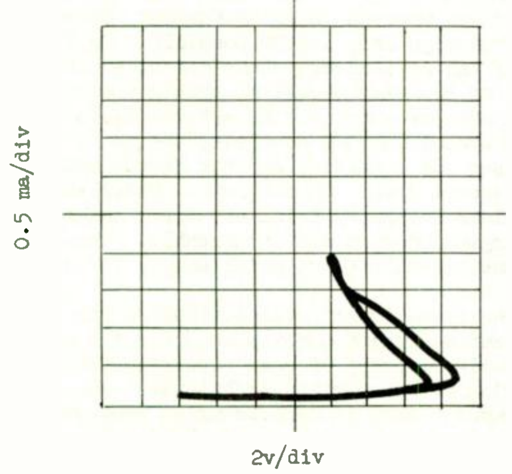


Fig. 6. TiO_2 negative resistance device.

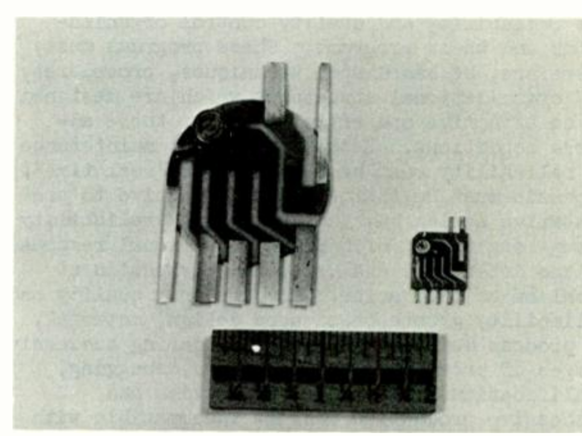


Fig. 8. Integrated component - NAND.

WRH

RELIABILITY IN REAL TIME

Thomas J. Scanlon
Quality Assurance Manager
Epsco, Incorporated
Cambridge, Massachusetts

INTRODUCTION

The design of a reliability program must take into consideration the environment in which the program is to operate. The challenge of continually increasing reliability and maintainability requirements must be met in the face of other trends in military electronic system procurement which adversely affect reliability, magnify the tasks of the reliability and quality control organizations, and create conditions which obsolete many of the time-honored techniques used to achieve satisfactory operational reliability. The rate of technological change and the necessity for time compression of research, development, and production have resulted in these trends by the Military towards:

1. Accelerated delivery schedules.
2. Single runs of fewer systems.
3. More sophisticated performance of more complex electronics instrumentation systems.
4. Reduced space and weight requiring advanced packaging and thermal design, and the development of new manufacturing techniques to attain the required component packaging density.

The conditions imposed by these requirements, in particular the requirement for extremely fast delivery, place severe stresses upon the reliability and quality control organizations and their programs. These programs must, therefore, be based upon techniques, procedures, and organizational structures which are designed to be effective and efficient under these adverse conditions. Establishment and maintenance of reliability must be performed in "real time". Emphasis must be shifted from corrective to preventative action and the quality and reliability groups organized for fast organizational response in the detection, analysis, and correction of problems as they arise. Conventional quality and reliability growth based upon design, material, or process deficiencies revealed during successive stages of prototype evaluation and debugging, qualification tests, pilot production and successive production runs is incompatible with delivery requirements and the limited quantities of systems involved. In the example of the digital computer described in this report the first system built was required to serve as a feasability model from which performance specifications could be evolved, an engineering prototype for the evaluation of circuit and logic performance, a preproduction unit for the evaluation of materials, parts and processes, as well as the first production system. The system was to perform these functions simultaneously, not sequentially, in a time compressed R. and D. program and was expected after arrival in the field, to comply with "state-of-the-art" reliability requirements. While the example described represents an extreme case of time compression it is not unusual today to find qualification tests postponed and performed in parallel with production rather than in series preceding quantity procurement and production. Similarly, where AGREE type proof of performance reliability tests are specified, we can anticipate that many systems will be in use in the field before conclusions can be drawn based upon these tests. Under such conditions the values of the qualification and reliability tests in protecting the interests of the manufacturer or the user are severely reduced. Alternatively consider the effects of major design defects revealed by qualification tests under conditions of concurrent, rather than successive, production of all systems on order. Under successive conditions drawings are not released to manufacturing and procurement is limited until the design has been proven by qualification tests; therefore major design changes, as a result of the tests, represent to the producer only an added expense and a delay whereas under concurrent conditions, major redesign many not be possible. In the case of the digital computers described in this paper, at the time the first system was completed and presented for Acceptance Tests the second system was 90% complete, the third system about 80% complete, and the fourth and final system about 50% complete. Under these conditions design by trial and error methods and marginal design are intolerable. The design must include substantial safety factors.

The trend toward miniaturization and high component packaging density requires the continuous development of more sophisticated manufacturing and inspection techniques. The continuous change in manufacturing processes, the development of high speed electronics assembly techniques, pilot type production, and crash scheduling impose severe stresses upon Quality Control organizations and methods. As the industry has progressed from terminal boards, to simple single sided printed circuit boards, to high density double sided printed circuit boards, to three dimensional potted modules and

now to three dimensional weld packs the "margin for error" has decreased. Each step along this road has reduced the freedom of the engineer to institute design changes and has reduced Quality Control's ability to detect and correct manufacturing errors while at the same time presenting more sophisticated process control problems. Your attention is directed to the photos of typical high density printed circuit card assemblies (Exhibits 1, 2, and 3). Compare these layouts in terms of reworkability with conventional vacuum tube circuit layouts with component parts mounted on terminal boards. During assembly the component leads are formed and cut by machine, printed circuits eliminate much of the inter-connecting wiring and speed the assembly process, and automatic flow-solder machines solder the completed boards - in quantity - in a few seconds each. The soldering operation represents a "moment of truth" in the life of the board. Once the board has passed this operation, the ability to rework the assembly without degrading quality or reliability is severely reduced. As large quantities of printed circuit boards may be subjected to this critical operation within a period of a few minutes, rigorous control of material and the assembly and test operations is necessary. The response time of the quality and reliability organization from the receipt of defect or failure data to the generation of corrective action must be short and the organization and procedures of Quality Control must be developed with speed of response as a cardinal requirement. The delays involved in failure information filtering slowly through a number of different levels of separate quality and reliability organizations awaiting analysis, recognition, and decision cannot be tolerated.

In summary, in time-compressed R and D programs reliability must be designed and built initially into the first system produced. Emphasis, must be placed on planning and preventative action rather than corrective action. This condition of time, quantity, and component density do not permit the evolution of a basically unsound design into a reliable status.

Conversely, as qualification and life testing programs are ruled out by the conditions described and system in-plant operating time limited, and considering the innumerable possibilities for reliability degradation inherent in design, procurement, manufacturing, test, installation, maintenance, and operational use, it is essential that the reliability planning contain provisions for a field follow-on program until the operational reliability and maintainability of the systems has been established. During this critical early period the high initial random catastrophic failure rate, combined with systematic failure mechanisms which may be present in the system coupled with interface and installation failures present a severe maintenance problem. The problem is aggravated by the lack of practical diagnostic experience at this point in time, of the customers maintenance personnel who, too often, must maintain the system without an adequate supply of spare parts, support equipment, or technical information due to delays in contracting for these essential support materials and services.

Reliability and Project engineering representatives of the manufacturer should be present during installation and early field use to perform rapid and complete failure analysis, assist in the resolution of interface problems, and supply technical direction and support to maintenance activities.

Epsco, Incorporated, faced with the task of designing and producing complex military digital control computers under the adverse conditions described, developed a "real time" reliability program -- a program designed to produce complex, high packaging density, solid state, technically sophisticated hardware capable of meeting severe operational reliability and maintainability requirements without benefit of models, prototypes or reliability tests. The following sections trace the history of the computers from initial planning stages to the field follow-on stage and describe typical reliability problems encountered at each stage.

INITIAL RELIABILITY PLANNING

Following preliminary logic design of the computers, a series of reliability planning conferences were held to evaluate the scope of the project, the conditions under which the systems were to be designed and produced, and the conditions under which the systems were to operate in the field. From these conferences emerged the policies and procedures which established the reliability program.

Derating Standards

The preliminary logic design was incomplete as detail input and output signal requirements could not be specified at the time, however, analysis of the logic indicated that the total number of components could approach 100,000. Space limitations would require high component density. The System MTBF goal was 100 hours, therefore, the inherent reliability of the system's component parts, if we omitted the degradation factor in the manufacturing and testing processes, would average out at 10,000,000 hours MTF. The degradation factor from manufacturing and testing processes could not be accurately predicted. We could conclude, however, that some degradation would definitely occur and that the normal loss of reliability in manufacturing and test would be magnified by the factor of high component density, accelerated

schedules and non-continuous production, therefore, it was estimated that an average component MTF of approximately 15 to 25 million hours was required to attain the goal of 100 Hours System MTBF. Analysis of component failure rate data at the time (1958) indicated that component parts operating within manufacturer's tolerances in conventionally designed circuits were not capable of achieving this level of reliability. The past experience of several of the project engineers in the design of similar solid state digital systems, however, convinced them that these levels could be attained providing the component parts were adequately derated. Severe derating standards were established for each component type which included derating for performance degradation with age and environment as well as electrical and thermal stress. The component count of the final design reached 90,000 parts and included approximately 14,000 transistors, 14,000 diodes, 38,000 resistors, 19,000 capacitors, and 2700 magnetic shift registers.

Design Review

Design rules were established requiring rigorous analytical circuit and logic design. The computations proving circuit performance under simultaneous worst--worst combinations of signals, voltages, loads and derated component tolerances were required by Design Review before drawings could be released for manufacturing.

Standardization

Standard components, circuits, modules, structures, layouts, and manufacturing processes were used repetitively throughout the system. Wherever practicable, circuits, structures, and techniques developed on previous projects were used or adapted for use. Standardization by bringing the benefits of past experience to the project and eliminating duplication of effort, reduced the cost of the project and substantially reduced the magnitude of the design, reliability, and quality control tasks.

Project Organization

The project organization contained a nucleus of senior engineers familiar with all design and construction aspects of the system. Specialized engineering skills were drawn for specific design or review tasks from the corporate engineering and divisional engineering groups and from technical consultants. These specialized skills flowed in and out of the job under the control of the project engineering group. The project contained a small pilot assembly group and a small module test group. As soon as the design drawings, processes, and inspection and test methods of a subassembly had been proven in the pilot assembly operation the job of constructing and testing the production quantities of the modules was farmed out to external vendors and internal corporate manufacturing groups. The project organization, by farming out the repetitive and routine and by drawing on specialized skills and facilities available throughout the rest of the corporation, was able to retain the flexibility and response time of a small organization.

Test Program

A master schedule based upon step testing of the system was developed and material procurement, subassembly production and test, and system wiring were keyed to the step testing schedules. As each step-lot of modules was delivered to the system the system tests would advance a step to include the system test of the new modules and the associated backboard wiring. As approximately 98% of the circuitry was in modular form, and as each module was a logical entity, it was possible to rigorously test each module prior to insertion in the system. Test procedures, specifications, and test equipment were supplied to each of the module manufacturing groups. Programmers, which adapted the individual module types to universal test fixtures, programmed marginal test under worst - worst combination of supply voltages, loads, and stimuli. A substantial margin of safety existed between the module test conditions and the range of dynamic conditions which the module could encounter in its operation in the system. Each module, following its production test and prior to its acceptance by the project, was subjected to accelerated aging and vibration conditioning to "burn in" the modules and advance them beyond the early failure period into the stabilized constant failure rate period of the modules life characteristic curve.

Serviceability

The system, a digital control computer, was to be an element in a number of feedback loops involving several peripheral systems and equipments. Fault isolation would be complicated by the fact that various sections of the computer would be controlling, and in turn, be controlled by external systems. The computer would be wholly maintained by enlisted military personnel who would also be responsible for the maintenance of a number of other complex electronics systems. The design of the system would determine the speed with which failures could be isolated and corrected and the degree of skill required to service the computer. Several of the designs for maintenance policies are described:

1. Modular Construction of Logical Entities

   Circuits were packaged in modular form (See Figure 1) and standardization was emphasized to keep the number of different module types to a minimum. The

modules were logical entities facilitating rapid isolation of system failures to a single module and simplifying the testing and troubleshooting of the modules.

2. Elimination of Field Adjustments

Field service adjustments and controls which compensate for broad tolerances and component drift were eliminated by the selection of stable components operating in circuits designed in accordance with the derating standards previously described. The component derating standards included tolerance derating for aging. Elimination of field service adjustments reduced maintenance complexity and eliminated performance and reliability degradation resulting from misadjustment of controls by maintenance personnel.

3. Self-Testing Features

Internal references supplied calibrating signals which could be switch substituted for the inputs from the external sources. Self-test problems based on these static references were developed to exercise the computer in each of its modes of operation. A bank of indicator lamps located on the front panel displayed the data contained by any register in the computer at any static program step. Individual registers and program steps were selectable for data display by front panel switches. All major input and output signals were brought to front panel test point jacks for easy access.

4. Accessibility

The modules were mounted in frames capable of sliding in and out of the system racks. In the maintenance position of the frames, the frames were slid forward out of the rack enclosure and fanned out on hinges, like the pages of a book, providing access to the modules and to the module connectors on the rear of the frames. In addition to the test point on the edges of the frames every connector pin was accessible from the front of the unit for signal tracing. Exhibit 6 shows the computer with the frames partially open.

5. Training

During the period the system was under construction a formal training program for military maintenance and Epsco field maintenance personnel was established. The training covered circuit, module, and system theories of operation, fault isolation techniques, and the use of specialized test equipment. The classroom study was supplemented with practical experience in troubleshooting failures introduced into one of the in-plant computers. The classes were taught by the project engineers who brought into the classroom their knowledge of theory of operation and circuit behavior coupled with the practical experience acquired in testing and developing diagnostic techniques for the in-plant computers. Epsco's reliability program stressed the importance of this in-plant training. It was essential that the responsibility for maintenance of these complex computers be turned over to technically qualified personnel. The location and application of the computers made them inaccessible for months at a time to Epsco Field Engineering personnel. During these periods the responsibility for maintenance would be vested solely in the enlisted shipboard personnel. The success of the computers, and of the weapons system, would be dependent upon their skill and knowledge.

IN PLANT HISTORY

During the early period of system growth, system construction proceeded in accordance with the step testing schedule; however, as the quantity of modules contained within the system increased, failures of these modules began to introduce delays in the step testing schedule. Analysis of the module failure reports revealed that failures were due to the following principal causes:

1. The majority of failures were attributed to printed circuit process defects which resulted in electrical intermittents. These intermittents were extremely difficult to isolate and, as troubleshooting was by module substitution, resulted in faulty diagnosis of many modules producing a fog factor in the failure reports.

2. Many failures were attributed to the high "infant mortality" failure rate of the component parts. This initially high failure rate was compounded by failures induced by module manufacturing and test processes which subjected the components to the temperature

shocks of soldering, the physical shocks of handling and cutting, and the stresses imposed by probing in test and inspection of the modules as well as secondary damage resulting from shorts and other component failures. Typical examples of this type of failure were resistors and diodes which physically cracked in the manufacturing process but failed after a period of time in the system.

3. A third major cause of failure were failures caused by system troubleshooting. In particular, intermittents caused by printed circuit process defects resulted in excessive probing and handling producing many test-induced failures in the system.

The corrective action taken to reduce the early failure rate of the modules in the system was the development of the environmental accelerated aging tests previously noted. The environmental tests consisted of temperature shock and temperature humidity cycling followed by vibration and were designed to "shake out" early failures attributable to component part "burn in" failures and latent manufacturing process defects. These tests were performed on each module following the production test of the module and prior to insertion of the module in the system. The modules already in the system were replaced with modules which had been subjected to these tests and routed back through the environmental test operation.

During this period the module manufacturing groups, as well as the project group, were plagued by process defects of the printed circuit cards. The majority of these defects could not be detected by visual inspection of the cards as received from the card vendors at our Receiving Inspection but were latent, and not detectable until after the card had been completely assembled and soldered. At this point in time the component parts could not be salvaged and the loss of the assemblies to the project resulted in delays in the step-testing schedule.

Several of the process defects that were encountered during this period are described below:

1. Contamination of foil solder plating resulting in non-solderable printed circuit foil conductors.

2. Contamination of base copper surface prior to nickel-rhodium plating of the finger contacts that resulted in fingers peeling after a limited number of insertions in a connector.

3. Undercutting of circuit foil which resulted in fine slivers of metal parting from the foil top edges and establishing themselves as intermittent shorts on the finished assembled cards.

4. Hygroscopic surface contaminants which, under conditions of high humidity, lowered surface leakage resistance between circuit paths and resulted in circuit failure.

5. Acidic surface residues that, combining with humidity, attacked the plating of transistor cases and jumper wires.

6. Printed circuit boards that warped when subjected to the temperature shock of dip or wave soldering.

7. Printed circuit foil that lost all adhesion strength when exposed to soldering temperatures.

The major problem that was encountered on the printed circuit boards, however, was due to the eyeletting process. The solder connection between the flange of the rolled-over brass eyelets and the printed circuit foil cracked as a result of board flexure in handling or as a result of stresses set up by the differences in expansion and contraction of the glass-epoxy laminate, copper foil, and brass eyelet when the card was subjected to temperature shock. The hairline fracture around the periphery of the eyelet flange was extremely difficult to detect by visual inspection and produced an electrical intermittent that was difficult to isolate in test. The brass rolled-over eyelets specified by MIL-STD-275 gave way to double funnel copper eyelets and eventually to plated through holes in our search for a reliable solution to this problem. The double funnel copper eyelets provided a sufficiently large interface between the eyelet and the funnel to produce a structurally strong connection and avoided the entrapped air pocket of the rolled-over brass eyelets. The funnel eyelets had the disadvantage, in our application, of increasing the heigth of the components and the soldering and could not be used on many of the printed circuit assemblies due to insufficient clearance between adjacent assemblies. In addition, the capillary action by which solder was reputed to flow from the bottom of the printed circuit board up the outside of the funnel and establish the connection between the funnel and the foil on the component side of the board was not reliable. It was necessary, therefore, to presolder the top foil connection prior to mounting of the component parts and subsequent dip or flow soldering in order to assure a reliable top foil connection.

The plated-through hole process eliminates the problem of making connection between the top and bottom foil of the double sided card but the integrity of the holes cannot be determined by visual inspection and more rigorous control of the fabrication and plating processes is required.

The rolled-over brass eyeletting process was intolerable, the double funnel copper eyeletting process could not be used on many of the printed circuit assemblies, however the recurrent latent process defects which we were encountering which were attributable to inadequate control of the fabrication and plating processes did not inspire confidence in the ability of our card vendors to control the more sophisticated plated-through process. Accordingly, the decision was made to establish an in-house printed circuit facility with a plated-through hole process capability which would enable us to exert the degree of process control which we believed was necessary. Once this facility was in operation our difficulties with printed circuit boards ceased. There are approximately 500,000 electrical connections on plated-through hole boards produced by this facility in the computers in the field. The computers now have an average operating time of 13,000 hours. We have never experienced a system failure due to the failure of a plated-through hole.

INSTALLATION PERIOD

The systems passed their acceptance tests and within a short period were installed and operating. During this period, and for a considerable time thereafter, secondary failures due to failures in peripheral systems and equipments, shorts in interconnecting cables, and failures due to damages caused during installation occurred in the system. As emphasis was placed on rapid system repair, troubleshooting during this period was often by massive, rather than selective, module substitution. As a result, failure analysis was clouded by numerous reported failures which were, in fact, diagnostic errors and numerous secondary or test induced failures. When the other components of the overall system became operational and effective failure analysis became possible it became apparent that the reliability and maintainability of the systems was not up to our expectations. The MTBF of the systems was approximately 24 hours and the average failure required 6 hours to detect, isolate and repair. Examination of the modules which had been returned for repair disclosed that approximately 40% of these modules had been incorrectly diagnosed as failures. Discussions with maintenance personnel revealed that the majority of system failures were intermittents which were difficult to isolate and that the speed of maintenance was further impaired by the external conditions under which maintenance was performed. Accordingly a field reliability and maintainability improvement investigation was initiated.

FIELD RELIABILITY IMPROVEMENT

Failure Analysis

The operating logs of the systems were not available for analysis, however, 87 cards which had been returned from the field for repair were available in-plant. The technique used to identify failure patterns within the group was by the following series of classifications:

1. Initial Classification
   a. Diagnostic Errors
   b. True Failures
2. Classification of True Failures
   a. Primary Failures
   b. Secondary or Test Induced Failures
3. Weighted Classification of Primary Failures
   a. By Card Type
   b. By Circuit
   c. By Component Type
   d. By Manufacturing Process

Analysis of the data obtained from the series of classifications revealed that 88% of the true primary failures were due to two causes. They were:

1. Intermittents due to eyelet failures on printed circuit boards. The failures were confined to specific production lots of four types of printed circuit assemblies which had been eyeletted with either the rolled-over brass or flat flange brass eyelets.
2. Intermittent malfunctions of magnetic shift register circuits.

Subsequent investigation and analysis revealed that eyelet failure was caused by fracture of the weak contour soldered joint existing between the top flange of the eyelet and the foil on the component side of the printed circuit board. The majority of these failures were flat flange as many cards with rolled over eyelets had been removed from the systems in-plant and had been replaced with plated-thru or double-funnel copper eyeletted boards. Significantly all of these modules had successfully passed the severe module environmental stress tests in-plant. Aging and the continual flexing produced by intermittent operation of the computers were required to produce the electrical intermittents.

Analysis of magnetic shift register failures was complicated by the fact that the static tests of the module testers would not

produce a malfunction and similarly, many of these registers would operate satisfactorily when installed and dynamically tested in the in-plant systems. The intermittent failures, when they could be detected, occurred as either the generation of a "one" (noise firing) or the loss of "ones" and was dependent upon the rate of change as well as the quantity of "ones" being processed by the register at any instant. The marginal tests of the module tester which varied the stimuli, loads, and voltages to the register had little effect on this particular type of failure and an independent variable was introduced to enable marginal evaluation. Failure of these registers was due to the increase in the effective series resistance with age of a wet foil tantalum capacitor used to bypass the stabistor which established the bias for the shift register blocking oscillators. This produced intermittent failures which were dependent upon the number of blocking oscillators which were firing at any instant. Aiding and abetting this fault mechanism was the voltage shift of the bias-setting stabistor as the base current of the blocking oscillators varied. Hysteresis of the stabistors due to the opposing voltage coefficients of IR drop and junction temperature with current changes produced intermittents which were dependent upon the rate of change of "ones" as well as the quantity of "ones" being processed.

Corrective Action

The corrective action taken to improve module reliability was rework of the small quantity of brass-eyeletted cards of the four types identified by failure analysis in order to obtain solder connections which were physically strong as well as electrically sound and replacement of the wet foil tantalum capacitors with solid tantalums. A quantity of "pump-priming" cards were obtained from in-plant systems and cards from the field systems were routed back to the plant for rework daily. During this period the field computers were kept in an operational status and card replacement was performed during scheduled "down" time.

Maintenance Environment

1. Operating Conditions

The temperature rise of the system exceeded the original predicted temperature rise due to the proximity of a bulkhead to the access ports of the computer and due to air leakage from the module frames. The temperature rise was reduced by sealing the frames which eliminated the air leakage and improved the efficiency of the convection cooling.

2. Maintenance Condition

In the maintenance mode, with the frames opened for servicing forced air cooling was necessary to limit the temperature withinin the frame. Forced air had not been provided and, as a result, maintenance personnel had to shut down every 30 minutes for a 30-minute period to let the frames cool down. This not only doubled "down" time but also subjected the system to excessive temperatures and temperature shocks. The corrective action taken was to provide snap-on blowers which mounted at the bottom of the module frames and supplied cooling air to the frames in the maintenance mode.

3. Computer Location

The computer was located in a passageway. When the frames were swung out for maintenance the passageway was partially blocked and traffic through the passageway interfered with maintenance personal.

4. Test Equipment Accessibility

Due to the cramped quarters it was difficult to monitor the computer output and simultaneously signal trace on the module frames. As the majority of failures which had been encountered were intermittents it was necessary to continuously monitor the output. A small scope was added to monitor the output and mounted so as to be visible from all servicing positions.

5. Service Manuals and Drawings

One of the factors affecting speed of maintenance was the lack of a work space upon which to place drawings, manuals, servicing data, etc. A pocket service manual was supplied containing test point information, logic information, and diagnostic technique descriptions.

6. Test Techniques

Improved testing techniques based upon the field maintenance history of the computers were established. Interconnections between computer sections were brought out to front panel test points to permit breaking into internal feedback loops and to permit bypassing of computer sections to aid in fault isolation.

Additional test problems were developed and equipment added to enable rapid measurement of computer accuracy independently of external systems.

7. Spares -- Type and Accessibility

Maintenance was affected by the quantities and types of spare modules available and by the lack of component parts when spares were exhausted. "Down" time was significantly affected by the fact that spares had to be obtained through paperwork from a central stockroom and were not available at the computer location. The corrective action taken was to advise the user of an optimum mix of spare modules and parts to be stocked and to provide a locked cabinet adjacent to the computer containing commonly used spare modules ready for installation.

OPERATIONAL RELIABILITY

Following completion of the reliability improvement program, the performance of Systems #1 was monitored for several thousand hours of operation. The following figures describe the reliability level of the computer during this period:

Initial Failure = 628 hours
System MTBF = 330 hours
Component Part MTF = 30 million hours
Average Down Time/Failure = 30 minutes

During this period the computer was maintained by military enlisted personnel. Analysis of the failure data at that time indicated that the system was still "wearing-in" from the rework program and, if properly operated and maintained, could approach a level of 50 million component hours.

SUMMARY AND CONCLUSIONS

Reliability goals can be consistently attained under the adverse conditions created by a time-compressed R and D program providing the producer utilizes a systematic plan of attack based upon planning and prevention while realistically recognizing that not all problems are predictable or preventable. The quality and reliability groups, to effectively cope with these problems as they arise must be organized for fast detection and response. The producer and user should jointly recognize the necessity for field follow-on reliability and project engineering support before systems arrive in the field. Note that the simple, rapid, and inexpensive field follow-on program described was capable of increasing the operational reliability of the computers by more than an order of magnitude.

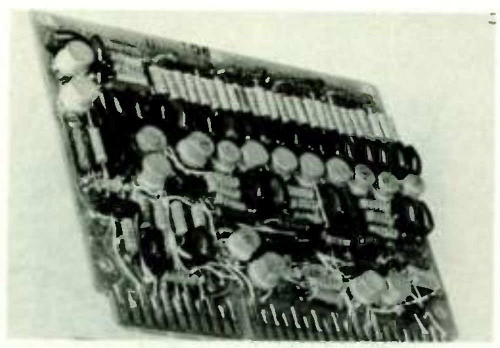


Fig. 1. High density printed circuit assemblies.

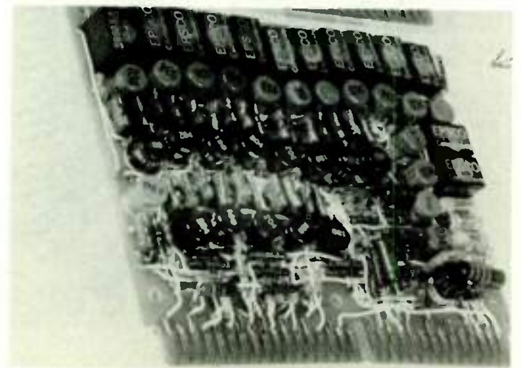


Fig. 2. High density printed circuit assemblies.

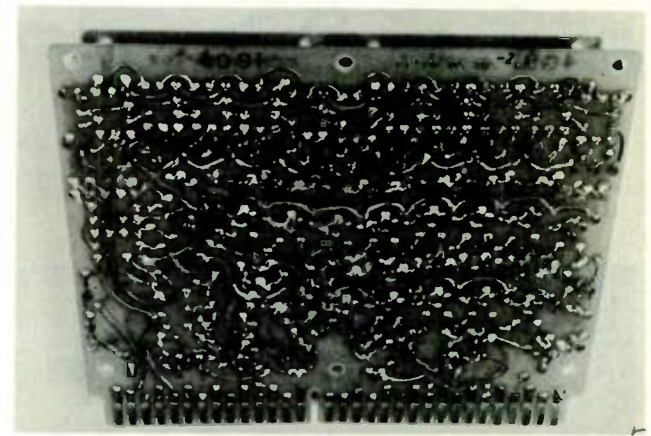
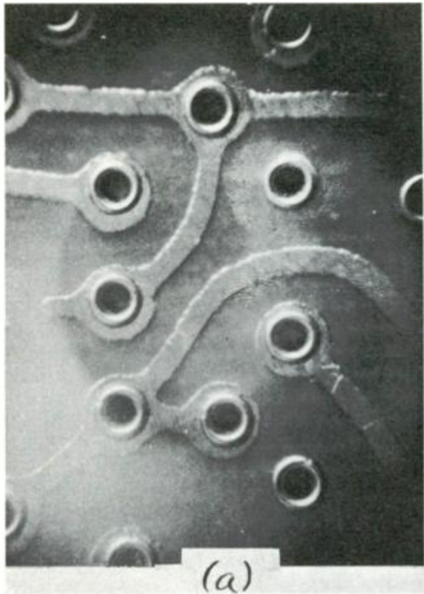


Fig. 3. High density printed circuit wiring.



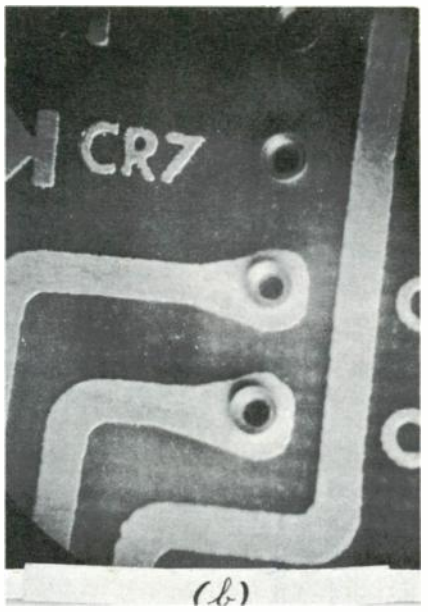


Fig. 4. Eyeletting processes.
(a) Brass rolled-over eyelets. (b) Double-funnel copper eyelets.
(c) Plated-through holes.

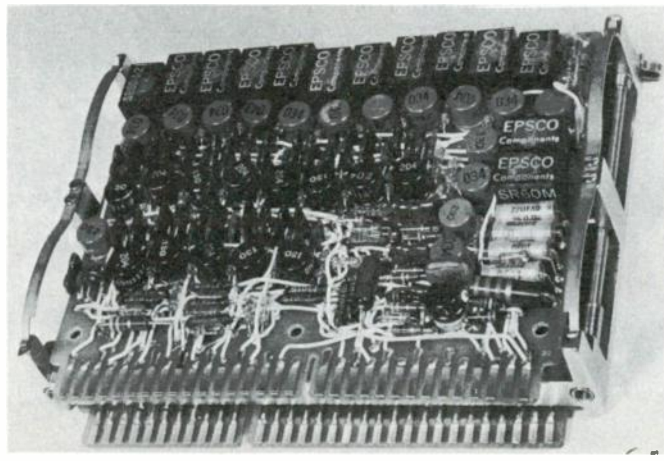


Fig. 5. Printed circuit module.

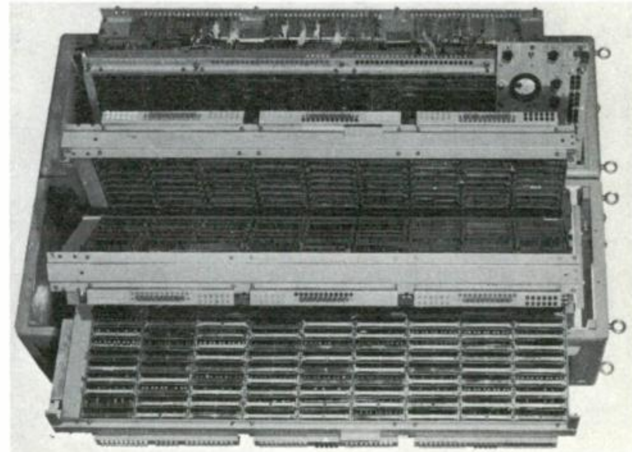


Fig. 6. Computer with sliding frames in maintenance position.

INDIVIDUALITY IN SMALL-PLANT RELIABILITY

Irvan J. Bearer
Martin Marietta Corporation
Baltimore 3, Maryland

Summary

Increases in performance requirements and environmental severity, coupled with customer demands for analysis, review, testing, demonstration, training and documentation have pushed reliability costs to a critical stage. To prevent short-cutting an adequate program, reliability manpower costs must be reduced. The contrasting situations between plants with independent reliability organizations and plants with the reliability tasks distributed among existing departments will be discussed. The work atmosphere that is characteristic of many small plants and its adaptation to the organizational technique of task integration will be described. Examples will be presented of reliability work assigned according to engineering planning and procedures to those departments with personnel capable of doing the work. Guidance, unity, strength and control of a task-integrated reliability program will be shown to be the responsibility of a management staff representative. Replies to a questionnaire on the distribution of reliability manpower will be analyzed with respect to the principle that, in small plants, reliability activities are performed by each of the major departments.

Introduction

Reliability problems grow increasingly difficult. In the Fifties, radio frequency interference and vibration hampered the development and qualification of many equipments, while in the Sixties space systems designers are concerned about the deterioration of electronic components caused by nuclear radiation. When such problems arise from the nature of the universe, scientists and engineers have found answers, but there are other laws that challenge reliability. That set of laws is called economics, and the results of its manipulations by man are frequently not predictable.

The reliability function must determine and provide for the economics of many activities--such as training programs for technicians, engineers and management; engineering effort for analysis and development of a mature design prior to production; and coordination of engineering and procurement in selection, application and purchase of components with sufficient derating and from suppliers with adequate quality ratings. In each case it must balance costs against gains in reliability and value. Reliability must answer other economic questions: What is the relative value of increased confidence versus the cost of continued testing? Will the design improvements resulting from expenditures incurred in collecting and analyzing factory and field failure data outweigh their costs? Does field maintenance cause more trouble to the customer than it is worth to him in design engineering effort to correct the problem?

Justification for additional reliability costs comes from the AGREE report. The comment of Arthur Malcarney, RCA executive vice-president, given at the Eighth National Symposium on Reliability and Quality Control, states that building and testing the AN/ARR-60 data link to conform to AGREE standards added approximately $2.5 million to the $100 million contract, but that the customer is expected to save $50 million.

In a Martin Marietta progress report, Don Gilmore cites another example of increasing costs to improve reliability. A missile-borne five-transistor a-c signal isolation amplifier was designed to satisfy MIL-E-5400 requirements. Mr. Gilmore recommended using parts from manufacturers who produce electronic components that fulfill the Minuteman reliability requirements. His evaluation of the proposed change for 23 of the transistors, resistors and capacitors used in the amplifier revealed that the mean time between failures (MTBF) increased from 637,000 hours to 3.2 million hours and that the cost of the parts also increased, from $32 to $173. Part procurement costs increased over 400%.

When the cost of an operation or of a part is high, value-cost relationships urge us to ask, "What is its function?" The need for high reliability is dictated by the performance, life and environment of our missiles, spacecraft, atomic submarines, commercial communication systems and many other products. The need is expressed in government specifications and contract requirements and is recognized by the many prime contractors who have authorized extensive reliability organizations. Staffing the reliability organizations has necessitated nationwide recruiting programs.

Strong men have been active in building the stature of reliability organizations--men whose presentations at proposal planning and contract meetings have directed the allocation of funds for design analysis and development testing; men who have had reliability training programs approved for high-paid engineers; men who have demanded and assumed authority and responsibility for functions that had been prerogatives of Engineering, Procurement, Testing and Quality; men who have settled jurisdictional conflicts and personal frictions inherent in any new organization. A separate reliability group should be efficient, in view of its specialized nature and

the repetition in its tasks, but protocol and protectivism have prevented many efforts to obtain design information, in collecting information on conditions and results of development tasks and in requiring action on recommendations.

The Small-Plant Picture

Small businesses offer a contrast in establishing a reliability program. There are no funds to recruit engineers for a separate reliability organization. Instead, the reliability work is divided among the existing departments to minimize clerical, supervisory and office costs and to provide a full work load for personnel whose capacities are not fully utilized.

Distribution Study

To give a factual basis for subjective evaluations of small business, a questionnaire, entitled "Reliability Personnel Distribution Study" (Appendix 1), was sent in January 1962 to 355 electronic part, component and equipment manufacturers throughout the United States. They supply over 100 types of electronic components. From the more than 100 replies (Appendix 2) 69 were suitable for tabulation.

Cell groupings shown in Table 1, "Reliability Manpower Allocations," indicate that the reliability assignment percentages of total employment follow a pattern in plants having 25 to 1500 employees. (Appendix 3 contains assorted questionnaire replies.) Figure 1, "Reliability Manpower Versus Employment," shows that the very small plants (1 to 25 employees) reported a much higher reliability assignment percentage of total employment and that the large plants (over 1500 employees) reported an appreciably lower reliability ratio. Figure 2, "Reliability Manpower Versus Employment (Less Test and Inspection)," emphasizes variations similar to those shown in Fig. 1.

The major fact reported is that reliability effort in small plants varies in the several groupings of plant sizes from 4% to 10% and from 2% to 7% when test and inspection assignments are discounted. (Throughout this paper, the term small plant refers to those electronic manufacturing plants having a total employment from 25 to 1500.) The average reliability assignment percentage in the 57 small-plant replies tabulated was 5.2%. A selection of 18 of these companies that make the higher complexity products showed practically the same ratio (5.3%). This fact indicates that complexity as a function of the number of parts does not require unusual reliability manpower. S. W. Lichtman[1], in a report on "Guided Missile Reliability Versus Complexity," gives examples to illustrate that reliability potential is not limited by the quantity of parts but rather by the quality of engineering reflected in the equipment or system.

Other small-plant observations made from the submitted data (Fig. 3) follow.

(1) Inspection and test personnel engaged in reliability activities comprise 2.2% of the total work force.

(2) Engineering effort in reliability (design, industrial, reliability, quality, and quality and reliability engineering) constitutes 1.2% of total employment.

(3) Design engineering personnel assigned to reliability work comprise 0.4% of the total plant employment.

Only one company mentioned training as a separate department. It provided 1/2 man-month per month for reliability education.

The questionnaires were completed by qualified representatives, mostly on the director and manager level. Fourteen of the representatives who returned the questionnaire had the word reliability in their job title.

Comments on the Study

If the reports are accepted at face value, a disproportionate number of Test and Inspection personnel are engaged in reliability work--40% of the total reliability effort. What does this mean? Many persons consider that during manufacturing, quality and reliability efforts are synonymous. Many expect reliability to be improved by failure analysis and design revision during pilot and production runs.

In contrast, consider the following reliability program principles listed in the specification, "Reliability Program Requirements for Aerospace Systems, Subsystems and Equipment" (MIL-R-27542).

(1) Reliability must be a major factor in planning, management and engineering.

(2) Inherent reliability is established by the basic design and can be improved only by design changes.

(3) Improvement of inherent reliability is achieved best in the early phases of development and the testing program.

(4) Control of reliability requires a planned production, quality control and testing program.

In principle, Quality Control, including inspection and test assurances to the customer that manufacturing is complying with applicable specifications, is part of reliability, but in practice quality activities are separated from reliability. Quality control has been established

TABLE 1

Reliability Manpower* Allocations
(Taken from questionnaire replies)

Supervising Department	Plant Size							Summation
	1-25	26-75	76-100	101-500	501-1000	1001-1500	1501-8000	
Design Engineering	3.8 (4.6)	7.6 (1.0)	7 (1.3)	48 (0.7)	9.5 (0.2)	23.3 (0.4)	57 (0.2)	156 (0.3)
Industrial Engineering	0.1 (0.1)	3.9 (0.5)	3.2 (0.6)	9.9 (0.1)	3 (0.1)	1.1	5	26 (0.1)
Reliability Engineering	0.6 (0.7)	2 (0.3)	1.6 (0.3)	11.3 (0.2)	23 (0.4)	8.5 (0.1)	11	58 (0.1)
Quality Engineering	0.1 (0.1)	2.4 (0.3)	0.5 (0.1)	19.3 (0.3)	14.5 (0.2)	11.5 (0.2)	15 (0.1)	63 (0.1)
Quality and Reliability	1.1 (1.4)	7.8 (1.0)	4.6 (0.8)	23 (0.3)	46 (0.8)	43 (0.7)	80 (0.3)	206 (0.4)
Quality Assurance	3.6 (4.4)	32.6 (4.2)	21.7 (3.9)	128.5 (1.8)	116.5 (1.9)	145 (2.5)	246 (0.9)	694 (1.5)
Manu-facturing	1.1 (1.4)	8.5 (1.1)	10.7 (2.0)	43.9 (0.6)	89 (1.5)	60 (1.1)	4	217 (0.5)
Production or Material Control	0.6 (0.7)	3.8 (0.5)	6.1 (1.1)	21.5 (0.3)	6 (0.1)	5 (0.1)	1.5	45 (0.1)
Sales or Contracts	0.1 (0.1)	2.7 (0.3)	3.2 (0.6)	19.5 (0.3)	8 (0.1)	5.1 (0.1)	0.5	31.9 (0.1)
Combined Employment	81	774	552	7,197	6,000	5,711	26,200	46,515 (3.2)
No. of Plants	5	14	6	26	7	4	7	69

*Reported in man-months per month; parenthetical figures refer to percent of total employment.

for decades, has developed its own techniques and has trained a nucleus of control-oriented personnel. Reliability's philosophy, techniques and personnel must be directed toward design and development if it hopes to withstand cost evaluations. Contracts can no longer permit the time or money for the sequence of design-produce-test-redesign. Reliability's job is to determine the quantitative inherent reliability possessed by a design; quality's job is to minimize the degradation of that quantity.

The Reliability Personnel Distribution Study verifies that small businesses recognize the need for reliability and have assigned manpower for reliability activities but in most cases have not organized a separate reliability group. In most companies, there is no one authority directing a comprehensive program. It is concluded that many small plants need a redirection of emphasis in their reliability efforts. A better understanding of reliability is needed. What is it? Where is it developed? How is it protected? Why do we worry about it? Is it self-supporting?

Answers to these questions come from several sources. Dr. P. H. Zorger[2] presented the extensive reliability training effort developing at Martin Marietta. Included are training manuals, films, digests, on-duty conferences and lectures, and after-hours classes at the technical and graduate levels in a program coordinated with the Drexel Institute of Technology. Other large companies can furnish their suppliers with reliability training information. Engineering consulting firms[3,4] will also give reliability service and guidance.

A perusal of the reliability specifications issued by the Government is an education in itself. MIL-R-27542 (USAF), "Reliability Program Requirement for Aerospace Systems, Subsystems

and Equipment," is especially recommended. The Contracts Technical Requirements Department at Martin Marietta has issued an excellent reference sheet listing outstanding government documents pertaining to reliability. It is included in "Contracting for Reliability" by R. H. Johnson[5]. Martin Marietta's expectations from subcontractors are given in its "Specification for Subcontractor Reliability Program."[6]

Reliability in Depth

Management's Contribution. Reliability activity must be considered at three levels. Top management defines the company reliability profile; one man directs the program and engineering and technical personnel implement the program.

The company position with respect to reliability policy should be stated in the form of a directive, manual, bulletin or series of procedures issued by the general manager or higher company official, because such documents establish general responsibilities and promote customer confidence in management's support of the program. To be convincing in the presentation of its reliability program, a company will have such documents not just on file but issued to responsible personnel, at least at the department head level, and immediately available.

"Mr. Reliability." The second level of reliability activity in small plants involves the man with full authority and responsibility for all aspects of the reliability program. His title is not important but his authority is. In small plants, a Mr. Reliability does not report to an engineering manager, to a quality manager or to a production manager, but instead to a general manager. His influence is active in every phase of reliability from proposal, contract or purchase order through engineering, production, test and delivery. In his staff capacity, Mr. R interprets principles and policies, concurs in contract requirements, determines and arranges necessary training, develops and issues techniques and procedures, and works with customer representatives at his plant to assure that the reliability requirements are being met. He provides unity, strength and continuity to reliability from one contract to another and prevents changes in principles, methods, techniques, practices, scope and manpower capabilities that would be detrimental to the program.

To control the entire program, Mr. R verifies the effectivity of all reliability efforts by reviewing formal reports from all departments, by informal and random surveillance, by monitoring critical activities in engineering, production and test, and by approving engineering documents and quality procedures. This audit and monitoring activity is in progress simultaneously with actual design, development and production.

Depending on the size of the company, Mr. R may or may not have other duties. He may also be chief engineer, quality engineering supervisor or an inspection and test supervisor. For the purpose of representing his company, his job title on reports, letters, contracts, bulletins and procedures should include his reliability responsibility.

When the general manager needs a representative for reliability, he should appoint a man with technical training and extensive experience in engineering and manufacturing. Mr. R should be personable, be able to communicate well, be acquainted with military specifications, and have a knowledge of statistics. Mr. R's influence on our country's electronic parts, components, equipments and systems makes his attributes worthy of a special study and report.

Task Integration. The third level of reliability activity is the "front line" area where the actual work is being performed. Either of two plans can satisfy this need--a reliability organization or task integration. A reliability organization assigns, controls, and supervises personnel assigned to perform reliability functions. In contrast task integration furnishes one man or a skeleton group to direct and monitor reliability activity without actually assigning or supervising those people engaged in reliability efforts. The effectivity and cost-saving aspects of this latter plan warrant a detailed analysis.

Task integration is an organizational technique that assigns units of work according to cost and capability considerations. The regular function of the worker or group, as opposed to job specialization, is of secondary importance. With task integration, production operators and assemblymen check and test their work to control quality, rather than to avoid disciplinary measures. Production operators may maintain inventory counts and production control records. Cost accounting and production control use copies of inspection and test source documents to tabulate scrap and rework quantities. Control and training are provided by the immediate supervisors.

With task integration, reliability calculations and analysis remain in the hands of the design engineering department. The determination of environmental and operating conditions, of tolerance and drift interactions and of tests required to evaluate or qualify a design also is kept as part of the design function, but prime responsibility for prevention of reliability degradation is given to the fabrication, handling and assembly workers rather than to a control department. As many employees as possible are trained to recognize and report discrepancies and failures. Such failure reporting provides a direct line of communications to engineering to eliminate a series of reviews and concurring signatures. Assembly technicians are used to record testing and usage data.

Task integration is not universally applicable in any one company. It does not eliminate the need for the regular departments, but it does keep specific manpower expenditures at lower levels. It requires a general interest in the quality, reliability and cost of the company's products. It must be carefully planned and controlled by industrial engineering process specifications and by interdepartment procedures.

By minimizing duplication of effort, by eliminating some unnecessary operations and material moves and by utilizing dormant knowledge and skills task integration increases efficiency in both production and research activities. It promotes greater interest in products and projects and provides added opportunities for professional and technical development along with a sound basis for good company morale.

Task integration is particularly applicable in reliability to keep organization costs from undermining the value that must be provided. Much of the reliability work belongs to functional groups that have been established for years. The tasks are not contradictory to the groups' basic training and experience, although the specific assignments and the required documentation may be. New techniques in engineering do not require a new breed of engineers; instead, some specialized training and reorientation is necessary to convince engineers to accept customer interest in the many details of their work--details which determine ultimate reliability.

A Conducive Atmosphere. The work atmosphere inherent in most small plants is especially suited to task integration. To meet the challenges of fixed price contracts and the competition in component technology and production, personnel at all levels have varied work activities. When problems arise, details of a job receive attention from several levels of management and supervision. A minimum of formality and short lines of communication provide for immediate recognition and correction of technical and scheduling difficulties. The close breath of bankruptcy dictates that costs be kept in line. "We will make the schedule" is a byword of small business.

Where management leadership and the necessities imposed by intense competition have developed a work atmosphere of flexibility and cooperative effort, reliability task integration can be instituted. Other preliminaries include the following.

(1) A study of the capabilities of the plant employees.

(2) A study of available laboratory, test and production equipment and of production flow.

(3) A determination of the reliability tasks that individuals or groups will perform effectively and economically.

(4) The issuing of instructions and procedures to define the purpose, responsibilities and details of the tasks.

Items (1), (2) and (4) vary considerably from plant to plant, so the following discussion will be confined to item (3). The reliability tasks are presented in Fig. 4.

A Job for Everybody

The first reliability task is to state the company's contractual obligations in brochures, engineering data sheets, proposals and contracts. The Sales and Contracts departments, in the continually advancing technology of electronics, need product and process engineering support for complete and accurate definition and description of company products and capabilities, of scientific and technical processes of mathematical and analytical techiniques and, of test equipment. When the plans for reliability demonstration are included in a proposal, the company's talent in statistics supplies the input, whether his regular assignment is in Quality, Cost Accounting or Engineering.

Design analysis prior to the start of production can save more money and develop more reliability than any other part of the program. Considerations and techniques for analysis include: establishing environmental stresses, both natural and equipment-produced; drift in part parameters; conflicts in tolerance buildups; derating requirements; complexity and redundancy relationships; modes of failure; tradeoff with maintenance, weight, cost, safety and human factors; reliability estimates; evaluation tests; and qualification tests. All of these are engineering functions and should remain there to avoid duplication of attention and work. Other departments can perform some of the tasks, but design engineering must review and act on the reports and the recommendations. The location of equipment, schedules and skills needed for operation of test equipment may suggest using an acceptance testing group to perform many of the evaluation tests. If the company has an experienced quality analyst, he may be far more effective in developing mode-of-failure charts than available design engineering personnel.

Design reviews need surveys and opinions by mature designers, who may come from top management in small companies. Additional study and comment should come from Quality, Inspection, Test, Production, Tool Engineering, Scheduling and Finance.

The selection of parts according to operational and environmental stresses is a Design Engineering function, but other departments may have more suitable wage scale levels and personnel more readily available for the related work in standard part usage, listing of vendors and performing the clerical function of obtaining customer approval or waiver on nonstandard parts. Procurement, Quality and Material

Control personnel should be considered for these tasks.

Recording of usage data should start in engineering evaluation testing, and continue through production. Test, inspection and assembly personnel can be trained and motivated to maintain the accuracy of these records.

The job of controlling the reliability of subcontracted components and equipment must take its authority from the procurement department, but a number of quality control functions are involved: tabulations of rejection rates, developing of statistical ratings of vendors, capability surveys, concurrence on sampling plans, and coordination of test and inspection tools and methods. Design engineering effort is required to establish test conditions and requirements at the vendor facility, to monitor the subcontracted equipment's reliability growth and demonstration, to review reports of reliability predictions, and to approve design changes recommended to correct reliability problems.

Control of reliability during production, shipping and installation must start with planning, i.e., methodizing and tooling issued by industrial engineering. Supplying detailed and current process and assembly instructions and the coordination of tooling and test equipment are vital contributions of industrial engineering, while calibration of tools and test equipment is a service that might come from Quality or Industrial or Product Engineering. Quality must issue procedures for inspection, test and failure analysis, but putting the control procedures into effect is the job of production operators, inspectors and testers. A reliability representative must audit or monitor these control activities.

Training in production, inspection and testing skills, primarily simple demonstration and instruction on the job, is handled most efficiently by the immediate supervisor, following patterns established by process instructions and procedures. Training for the purpose of upgrading or reclassification would require formal class sessions. Reliability indoctrination and orientation sessions and discussions should be provided at all levels by the staff reliability authority, Mr. R.

A distinguishing characteristic of reliability is that documentation is available for review and analysis by representatives of several departments and of the customer. In the past, a design was released without backup information; but now as a design is developed, it must be accompanied with reliability estimates, reports of problem areas, a listing of suppliers, the qualification status of parts and components and plans for improvement of design. All possible testing data on engineering evaluation of models, on reliability demonstration and production monitoring tests, and on acceptance testing must be kept on file for reference when design modifications or failure analysis is required. Inspection records provide degradation control. Operating time and cycles are recorded to provide maintenance and failure analysis information.

Economy dictates that documentation must be directed to the working, rather than control, groups. Each department must participate in putting this principle into effect.

Conclusions

Every technician, supervisor, engineer and manager in the small plants of the electronic industry has a duty to himself, to his company and to his country to learn more about reliability and to do more in reliability as part of his present job. This new learning should encompass more than motivation to excellence in his work. It should also provide the knowledge of principles, techniques and skills by which he can contribute either to the development or the prevention of degradation of the reliability of a design.

Management's thorough understanding of reliability will ensure a positive determination of scope, funding, and direction for the program; engineering training will provide for superiority in the achieving of reliable designs; industrial engineering's thorough knowledge of the opportunities in reliability task integration will give assurance to management and the customer that manufacturing plans and instructions will provide for efficiency in reliability control; quality engineering's continued study of the many factors in reliability will facilitate recognition and correction of weaknesses in materials, tools, equipment, personnel and working conditions that will affect the end product not only at the time of final acceptance but also during storage, transportation and operation.

Acknowledgments

I extend Martin Marietta's and my personal thanks to those representatives of the companies (Appendix 2) who contributed to the Reliability Personnel Distribution Study. Dr. Paul Zorger and the men of his Engineering Reliability group also have my sincere gratitude for the inspiration and assists they have given to this opportunity of expressing many of their opinions on the subject of reliability.

Bibliography

1. S. W. Lichtman, "Guided Missile Reliability Versus Complexity," 1957 IRE National Convention Record.

2. Dr. P. H. Zorger, "Effective Reliability--Through Education," Proceedings, Eighth National Symposium on Reliability and Quality Control, January 1962.

3. Directory of Consulting Members, American Management Association, 1515 Broadway, New York, N. Y.

4. Directory of Membership Services, Association of Consulting Management Engineers, Inc., 347 Madison Ave., New York, N. Y.

5. R. H. Johnson, "Contracting for Reliability," Proceedings, Eighth National Symposium on Reliability and Quality Control, January 1962.

6. Specification for Subcontractor Reliability Program, MB-1010, Martin Marietta Corporation, Baltimore, Md.

7. D. R. Earles, Reliability Analysis, RAD-TR-61-26, Air Force Ballistic System Division, Air Force Systems Command, U. S. Air Force, Inglewood, Calif.

8. J. D. Fernbach, "Integration of Quality and Reliability," Industrial Quality Control, November 1961.

9. G. R. Herd, Electronic Reliability--Estimation of Reliability Functions, ARINC Monograph No. 3, 1 May 1956.

10. Richard R. Landers, "Reliability Audits," Design, March 2, 1961.

11. S. N. Lehn, and J. Tronolone, "The Space Environment and Its Effects on Materials and Component Parts," Joint Reliability Seminar, Los Angeles, December 5, 1960, IRE Transactions, August, 1961.

12. R. G. Stokes, Handbook for the Prediction of Shipboard and Shore Electronic Component Reliability, Technical Report 133, NAV-SHIPS 93820, Vitro Laboratories, Silver Spring, Md., April 1961.

13. Richard G. Stranix, "Minuteman, Catalyst for Reliability," Electronic Industries, December 1960.

14. C. B. Tague and H. C. Bartels, "Manufacture and Control of High Frequency Transistors for Consumer Products," IRE Transactions on Reliability and Quality Control, November 1961.

15. AMC Pamphlet 74-1 A, February 1961, Reliability Evaluation Procedures for Pilot Production and Production, Wright-Patterson Air Force Base, Ohio.

16. Navy Reliability Design Handbook, PB-121 839, OTS, U. S. Dept. of Commerce, Revised 15 January 1961.

17. RADC Reliability Notebook, PB 161 894-1, OTS.

18. Reliability Engineering Handbook, RADC TR58-11, ASTIA AD148 868.

19. Reliability of Military Electronic Equipment, Report by Advisory Group of Electronic Equipment (AGREE), Office of the Assistant Secretary of Defense, June 4, 1957, U. S. Gov. Printing Office.

20. Reliability Stress Analysis of Electronic Equipment, RCA TR 59-416-1, 15 January 1959.

21. Prediction and Measurement of Air Force Ground Electronic Reliability, ASTIA AD 148 977, RADC TN-58-307.

22. Igor Bazovsky, Reliability: Theory and Practice, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1961.

23. David K. Lloyd and Myron Lipow, Reliability: Management, Methods and Mathematics, Prentice-Hall, Inc., 1961.

24. "Parts Specifications Management for Reliability--the PSMR-1 Report; Its Contents and Present Status" (The Darnell Report), Electronic Design, September 27, 1961.

Appendix 1

Reliability Personnel Distribution Study

Contributing Company

Name ______________________
Division or Plant ______________________
Address ______________________
____

Product Lines

(List such categories as resistors, capacitors, transistors, gyros, amplifiers, band-pass filters, thermionic equipment, power supplies, meters.) ______________________

Total Employment at Above Facility ______________________

Reliability Effort

Approximate man-month effort per month in one or more of the reliability activities (see attached sheet) under the direct supervision of presently established departments:

Design Engineering ______________________

Industrial Engineering ______________________

Reliability Engineering ______________________

Quality Engineering ______________________

Quality and Reliability ______________________

Quality Assurance (Inspection and Test) ____________

Manufacturing ____________

Production or Material Control ____________

Sales or Contracts ____________

Other ( ) ____________

The above information may be published in summary form and only as part of or in reference to the paper entitled Individuality in Small Plant Reliability, to be given at the IRE 1962 International Convention. The above company name may be used only in the listing of contributors to this study.

By ____________________________

Job Title ____________________________

Date ____________

Reliability Activities (reference)

1. Develop a reliability program.
2. Prepare reliability statements of work for proposals and contracts.
3. Analyze designs for reliability factors.
4. Make selections and applications with reliability considerations.
5. Make reliability predictions for equipments and subassemblies.
6. Provide for reliability growth, especially during research and development.
7. Plan, conduct and monitor reliability tests.
8. Conduct design reviews by groups independent of the responsible designer.
9. Record usage data.
10. Prepare and issue reports of reliability efforts and achievement.
11. Coordinate with the quality program.
12. Verify that quality assurance provisions are made.
13. Prepare reliability procedures, bulletins and directives.
14. Provide for control of reliability in subcontractor and supplier plants.
15. Maintain a reliability training program.
16. Provide for field maintenance and human factors.
17. File and distribute necessary documentation.
18. Provide for control of reliability during shipment and during installation.

Appendix 2

Contributors to the Reliability Personnel Distribution Study

Alfred Electronics
All-tronics, Inc.
Airflyte Electronics Co.
ACR Electronics Corp, Electronics Div.
AD-YU Electronics Lab., Inc.
Airpax Electronics, Inc., Seminole Div.
American Electronics, Inc.
American Electronic Labs., Inc.
Artisan Electronics Corp.
Ashland Electric Products, Inc.
Astron Corp.
Atohm Electronics

Bendix Corp., M. C. Jones Electronics Subsidiary
Bogue Electric Man. Co.
Boonton Electronics Corp.
Bradley Semiconductor Corp.
Bulova Watch Company, Electronics Div.
Burroughs Corp., Electronic Components Div.
Burton Manufacturing Company, Trans-Electronics

Computer Instruments Corp.
Coleman Electronics, Inc.
Cryogenics, Inc.
C and K Components, Inc.
Cannon Electric Company, Salem Div.
Centre Circuits, Inc.
CBS Electronic Products, Inc.
Circon Component Corp.
Clevite Corp., Clevite Transistor Div.
Collins Radio Company, Cedar Rapids Division
Columbus Electronics Corp., Silicon Div.
Compudyne Corp., Systems Div.
Continental Connector Corp.
Continental Industrial Electronics Corp.
Corning Glass Works, Electronic Components Dept.
Dale Electronics, Inc.
Daystrom Inc., Military Electronics Div.
Defense Electronics, Inc.
Delevan Electronics Corp.
Delta Coils, Inc.
Dickson Electronics Corp.
Dytronics Company, Inc., Main Factory

Ebert Electronics Corp.
Eldorado Electronics

Electronic Products Corp.
Epsco, Inc.
Erie Resistor Corp.
Federal Pacific Electric Co., Cornell Dublier Electronics Div.

General Atronics Corp., Electronic Tube and Instrument Division
General Dynamics/Electronics
Giannini Controls Corp., New Jersey Div.
Gray Instrument Company

Hercules Powder Company, Chemical Propulsion Div.
Hewlett-Packard Company
Hill Electronics
Hoffman Electronics Corp., Semiconductor Div.
Horman Electronic Laboratories, Inc.

Industrial Transformer Corp.

Johnson Electronics, Inc.
Jupiter Electronics, Inc.

Kings Electronics Co., Inc.

Lafayette Radio Electronics Corp., Industrial Electronics Division

Magnetics, Inc.
Merit Coil and Transformer Corporation
Metals and Controls, Inc., A Corporate Division of Texas Instruments, Inc.
Milo Electronics Corp.
Mid-Eastern Electronics, Inc.
Minneapolis Honeywell, Precision Meter Div.
Multi-Amp Electronic Corp., Defense Products Div.
National Research Corp., Equipment Div.
New England Instrument
North American Electronics, Inc.
Nuclear Corp. of America, Central Electronic Mfrs. Div.
Nuclear Measurements Corporation

Pacific Semiconductors, Inc.
Paradynamics, Inc.
Perkin Electronics Corp.
Perkin Elmer Corp., Vernistat Div.
Pomona Electronics Co., Inc.
P. R. Mallory and Co., Inc., Mallory Capacitor Co.

Resistance Products Co.
Robertson Instrument Co.
Royal Industry, Inc., Ideal-Aerosmith (Div.)

Sage Electronics Corp.
Sangamo Electric Co.
Schweber Electronics
Semi-Elements, Inc.
Servo Corporation of America
Solid State Products, Inc.
Speer Cardon Company
Spectrol Electronics Corp.
Sperry Rand Corporation, Sperry Electronic Tube Div.
Sunbank Electronics, Inc.

Telerad Div. of Lionel Corp.
Textron Electronics, Inc., Allegany Instrument Co.
The Siegler Corp., Hallamor Electronics Div.
Tracerlab, Div. of Laboratory for Electronics
Transformer Design, Inc. of Milwaukee
Transformer Engineering Corp.
Trygon Electronics, Inc.
United Aircraft Corp., Hamilton Standard Div.
Unitrode Transistor Products, Inc.
U. S. Semiconductor Products Div. of Nuclear Corp. of America

Vector Electronic Co., Inc.
Veeco-Vacuum-Electronic Corp.
Virginia Electronic Co., Inc.

White Avionics Corp.
Whittaker Controls and Guidance Division of Telecomputing Corp.
Winchester Electronics, Inc.

Appendix 3

Assorted Questionnaire Replies

The following replies were received in the responses labeled "Other," indicating the departments that supervised reliability activity.

	Man-Months	Plant Size
Repair and Calibration	5	1 to 25
Training	0.5	101 to 500
Packaging	1	101 to 500
Inspection	180	1501 to 8000
Quality Control Inspection	70	1501 to 8000
Preliminary Design	1	1501 to 8000

RELIABILITY MANPOWER VERSUS EMPLOYMENT

	Plant Size (personnel)						
	1 to 25	26 to 75	76 to 100	101 to 500	501 to 1000	1001 to 1500	1501 to 8000
No. of plants in grouping	5	14	6	26	7	4	7
Combined reliability effort (man-months)	16	71	59	283	316	303	672
Combined employment	81	774	552	7197	6000	5711	26, 200
Reliability effort	19. 9%	9. 2%	10. 6%	3. 9%	5. 3%	5. 3%	2. 6%

(Taken from questionnaire replies)

Figure 1

RELIABILITY MANPOWER VERSUS EMPLOYMENT
(LESS TEST AND INSPECTION)

	Plant Size (personnel)						
	1 to 25	26 to 75	76 to 100	101 to 500	501 to 1000	1001 to 1500	1501 to 8000
No. of plants in grouping	5	14	6	26	7	4	7
Combined reliability effort--less test and inspection (man-months)	12. 5	38. 7	36. 9	154. 5	199	157. 5	175
Combined employment	81	774	552	7197	6000	5711	26, 200
Reliability effort (less test and inspection)	15. 4%	5. 0%	6. 7%	2. 1%	3. 3%	2. 8%	0. 7%

(Taken from questionnaire replies)

Figure 2

Small Plant* Reliability Assignments

Plants tabulated	57
Total personnel	20, 234
Personnel Engaged in Reliability Work	1, 055 (5. 25%)
Inspection and Test Personnel Engaged in Reliability Work	443 (2. 2%)
Engineering Personnel Engaged in Reliability Work	240 (1. 2%)
Design Engineering Personnel Engaged in Reliability Work	85 (0. 4%)

*25 to 1500 employees
(Reported in questionnaire replies)

Figure 3

RELIABILITY TASK INTEGRATION

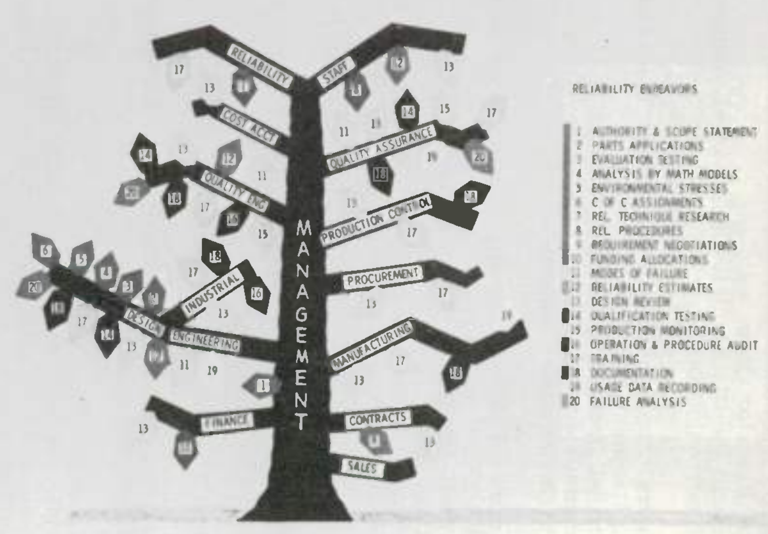


Figure 4

SHORT TERM PREDICTION OF HOOK-UP WIRE INSULATION COLD FLOW

Homer W. Hicks

International Business Machines Corporation

General Products Division

Development Laboratory

Endicott, N.Y.

Summary — This paper describes an investigation conducted to predict wire insulation cold flow which could result in short circuits between hook-up wires used in high-density wiring applications. The objective of the investigation was to qualitatively compare insulation resistance to cold flow. Approaches to solve this problem were both analytical and experimental. All data was obtained using a device (described in the paper) to stress the hook-up wire insulation. Although predictions were computed to 100,000 hours, they were checked experimentally to only 100 hours. No claims can be made for results to 100,000 hours except that, on a comparative basis, they follow the same trend as those checked to 100 hours. Experimental and predicted values agree within approximately 20 per cent.

INTRODUCTION

This paper describes a method for quickly predicting hook-up wire insulation cold flow. The prediction is made with a minimum of measured data combined with an empirical analytical approach.

The procedure employed can ultimately be refined for use as a quality control tool. In addition it is applicable to a wide range of commercially available insulations, and results can easily be compared on a qualitative basis.

This investigation was prompted to determine the criteria needed to prevent insulation cold flow of hook up wires used in high density back panel applications. Typical insulation damage due to wrapping is shown in Figure 1; by cold flow in Figure 2.

OBJECTIVES OF THE INVESTIGATION

One of the primary objectives of this investigation was how to quickly predict wire insulation cold flow. Other factors requiring analysis to completely describe a satisfactory wire were:

1. Wire wrapping machine compatibility.
2. Abrasion and slide resistance.
3. Dielectric characteristics.
4. Flammability requirements.
5. Costs.

Satisfactory results for these five factors were obtained by collecting data on available equipment or contacting wire vendors. Details of the wires we initially reviewed are:

a. All wires were tinned solid conductors, No. 24 gage.

b. Primary insulation varied from .006 to .012 inch, and jackets varied from .002 to .006 inch.

c. The insulation constructions were:

Primary Insulation	Jacket
PVC	Extruded Nylon Cotton Braid Rayon Braid Penton* Braid Nylon Braid Extruded Penton*
Nylon	None
Semi-Rigid PVC	Extruded Nylon Nylon Braid Extruded Penton* Dacron** Braid
Rigid PVC	None

* Registered Trademark, Hercules Powder Co.
** Registered Trademark, E.I. DuPont de Nemours & Co.

Teflon*	None
	Extruded Nylon
Rulan*	None
	Extruded Nylon
Low Density Polyethelene	None
	Extruded Nylon
High Density Polyethelene	None
	Extruded Nylon
Nylon	None

Typical cross sections of wire insulations are shown in Figures 3a, 3b and 3c. Figures 3a and 3b illustrate insulations with jackets. Figure 3c shows an insulation without a jacket.

EMPIRICAL DATA GATHERING METHODOLOGY

Insulation cold flow prediction data, particularly for our data processing machine back panel applications, was not readily available in the literature of the field. Prior to our own efforts, the methodology of published studies we examined consisted principally of

1. Hanging a weighted wire over a support and measuring the time to insulation cut through.

2. Recording data in a manner similar to the foregoing test, except using several weights. The various times to failure were measured and plotted. The graph was then extrapolated to the desired time to determine the load a particular insulation would withstand.

These empirical methods were admirably suited to the purposes for which they were designed. However we wished to examine at the outset approximately 35 wire insulations at two temperatures, and subjected to various loads. We wanted to compare these wire insulations under these simulated conditions in an environment as close to our system application as possible. The objective, therefore, of the study was to compare on a qualitative basis the resistance of various insulations to cold flow for our specific application. Results were required in a comparatively short time for planning production.

Although our predictions were computed for 100,000 hours they were checked experimentally for only 100 hours. The experimental and predicted values at 100 hours were within approximated ±20% agreement. No claims can be made for results to 100,000 hours, except that on a comparative basis they follow the same trend as those checked at 100 hours.

* Registered Trademark, E.I. DuPont de Nemours & Co.

MATHEMATICAL ANALYSIS OF CREEP

The literature was reviewed for equations previously used for predicting creep. The majority of equations used in previous cold flow studies in plastics required the evaluation of too many constants for our immediate needs on this project. Also the time required to collect data was too long. It also became apparent that while there were good theoretical reasons behind the derivation of many of the equations used, their application in general was strictly on an empirical basis. The typical equations were in the form of:

$$E = E_o' + \sin h \frac{\sigma}{\sigma_E} + m' \left(\frac{t}{t_o}\right)^n \sin h \frac{\sigma}{\sigma_m}$$

and

$$E = Ae^{-\frac{\phi}{RT}}$$

However, the literature regarding creep of metals contained expressions that required the evaluation of only two constants. One of these expressions seemed immediately applicable to our project and was therefore investigated. The equation and its dimensions are shown below:

$$\dot{E} = c\sigma^n \qquad (1)$$

where:

$\dot{E}$ = strain rate in./in./hour

σ = stress, lb./in.^{2}

c = experimental constant

n = experimental constant.

It should be emphasized that at the start of this investigation there was no valid reason for believing such a simple expression would describe plastic insulation cold flow behavior for a specific application.

Early objectives of this project were to obtain data experimentally, evaluate the constants in the equation, predict curves for long-term insulation failures, and complete the work within five days. Later numerous wires requiring evaluation were evaluated on a parallel basis.

Briefly the mechanics of the use of this equation are as follows:

$$\dot{E} = c\sigma^n$$

$$\sigma = \frac{P}{A} \qquad (2)$$

where:

P = load on insulation, lbs.

A = area load is applied, in.^{2}

$\dot{E}$ = strain rate determined from measured data.

The stresses and strain rates are found experimentally for several points, and the experimental constants are determined.

Numerous simplifying assumptions were:

a. The load-resisting stress is a straight compressive stress. Principal stresses were not analyzed.

b. The side effects of bulging were ignored.

c. The wire coating combinations of primary and secondary insulations were considered as a homogeneous material.

d. Load area is a linear function of insulation deformation.

e. The area used in calculating stresses is a projection of pin area on the wire surface.

DETERMINATION OF STRESS, STRAIN RATE AND TIME TO FAILURE

The exact manner in which the foregoing is used with experimental data is explained in detail.

1. Data is measured on the device shown in Figure 4. This is standard plastic laboratory equipment originally designed for determining heat distortion. The soft copper wire core is removed from a short piece of wire insulation and replaced with a hardened steel pin. The specimen is then mounted as shown in Figure 4.

The anvil of the heat distortion apparatus is fitted with a back panel pin so the corner of the pin rests on the insulation in the manner shown in Figure 4. This closely approximates the actual condition existing in a wired back panel shown in Figure 5.

Data measured directly on the heat distortion apparatus is the deflection of the back panel pin into the wire insulation for a known load in a period of time, also measured. This is illustrated in Figure 6.

The area of contact between pin and wire as a function of insulation deformation is:

$$A = (2Rd - d^2)^{1/2}\,(2F + \pi d) \qquad (3)$$

The deflection, d, for a given load at any time is expressed as the remaining insulation, ℓ, and plotted as shown in Figure 7.

Substituting in the area expression the deflection quantity, d, in terms of the original and remaining insulation thickness, we obtain:

$$A = \left[2R(\ell_o - \ell) - (\ell_o - \ell)^2\right]^{1/2}\left[2F + \pi(\ell_o - \ell)\right] \qquad (4)$$

This area is calculated for wires of various radii, R, and plotted for later use, as shown in Figure 8.

By rewriting the expression for stress $\sigma = \frac{P}{A}$ in terms of remaining insulation, ℓ, we have $\sigma(\ell) = \frac{P}{A(\ell)}$.

We are now in a position to determine the stress for a constant load, P, in terms of remaining insulation at any time, t.

Inasmuch as the deflection of the material was large compared to material thickness, the expression for true strain was used.

$$E = \ln \frac{\ell_o}{\ell} \qquad (5)$$

E = strain in/in.

An expression for strain rate is then derived:

$$\dot{E} = \frac{dE}{dt} = -\frac{1}{\ell}\frac{d\ell}{dt} \qquad (6)$$

where

$\dot{E}$ = strain rate $\frac{\text{in}}{\text{in hr.}}$.

$dt = \Delta t$ = time increment.

$d\ell = \Delta\ell$ = length increment.

Figure 9 is utilized to determine strain rate from measured data.

The strain rate is determined by substituting measured values from Figure 9 in the equation for strain rate (Equation 6).

The stress for the particular ℓ is determined by substituting values from Figure 8 in the equation for stress (Equation 2).

Stress and strain rates are now calculated for several values of ℓ for a number of different loads. One of two methods may be used to determine the experimental constants n and C. These are:

1. With the known values of

$\sigma_1, \dot{E}_1$ and

$\sigma_2, \dot{E}_2$ the

expression $\dot{E} = c\sigma^n$ may be solved for the two unknowns.

2. The preferred method is to plot various values of σ and $\dot{E}$, as shown in Figure 10. A straight line is fitted to the plotted points, since $\dot{E} = c\sigma^n$ will be a straight line, on a log log paper plot. Several values of σ and $\dot{E}$ can then be taken from the graph and the values n and c determined.

It is interesting to note that points falling in a straight line on log log paper show that measured data is following the law described by $\dot{E} = c\sigma^n$.

After experimental values for n and C are determined, the prediction of time to failure can be made. This is done in the following manner:

$$\dot{E} = c\sigma^n$$

$$\dot{E} = -\frac{1}{\ell}\frac{d\ell}{dt}$$

$$\sigma(\ell) = \frac{P}{A(\ell)}$$

$$-\frac{1}{\ell}\frac{d\ell}{dt} = C\left(\frac{P}{A(\ell)}\right)^n \qquad (7)$$

$$dt = \left(-\frac{1}{\ell}\right)\frac{1}{C}\cdot\frac{1}{P^n}\,[A(\ell)]^n\,d\ell \qquad (8)$$

$$t = \frac{1}{CP^n}\int_{\ell}^{\ell_o}\frac{[A(\ell)]^n d\ell}{\ell} \qquad (9)$$

In this expression t is time in hours as ℓ remaining insulation goes to zero. The expression for t was programmed and computed to very small values of ℓ for constant loads from five grams to 1500 grams, and plotted as shown in Figure 11.

Figure 12 lists the experimental constants determined for twelve wire insulations. Figure 12 also indicates the time to failure, with approximately .0002 inch insulation remaining, at two temperatures and two loads. No attempt has been made to correlate resistance to cold flow characteristics with the experimental constants.

A check of selected points using our data in:

$$^1E = E_o' \sinh\left(\frac{\sigma}{\sigma_e}\right) + m^1\left(\frac{t}{t_o}\right)^h \sinh\left(\frac{\sigma}{\sigma_m}\right)$$

E = creep rate

t, t_o = hours

E_o', σ_e, m^1, h and σ_m are constants for a given temperature and relative humidity showed very good agreement. This expression was used by Mr. W. N. Findlay[1] for a test run continuously for ten years and was found to be in good agreement with experimental data collected by Mr. Findlay over this period of time.

CONCLUSIONS

The analytical approach described in this paper for determining insulation cold flow in high density wiring applications has provided us with a promising method for determining, in advance, wire insulations that are satisfactory for our applications in machine back panels. As is often the case in a technology in evolution, previous approaches to determining creep and cold flow in a particular application provided only a limited, and therefore unsatisfactory, solution to our product needs.

We are continuing our studies of various other wire insulations which have recently become available. Preliminary results show a close correlation with the data presented in this paper.

REFERENCE

[1]Findlay, W. N., "Creep and Relaxation of Plastics" Machine Design, 32, 10, pp. 205 - 208, May 1960.

ACKNOWLEDGEMENTS

The derivation and use of Equation 9 was suggested by Dr. C. W. Mac Gregor, Manager, Advance Technology, the IBM Development Laboratory, Endicott, New York.

Contributions and assistance from the following colleagues on this project are gratefully acknowledged:

1. E.F. Stoddard
2. J.J. Rooney
3. R.G. Bayer
4. R.B. Turner
5. D.P. Griffiths

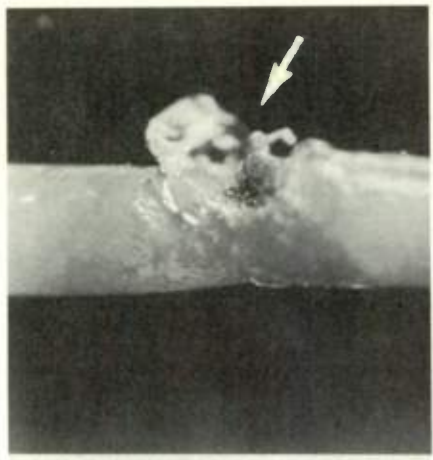


Fig. 1 - Typical wire insulation damage caused by wrapping.

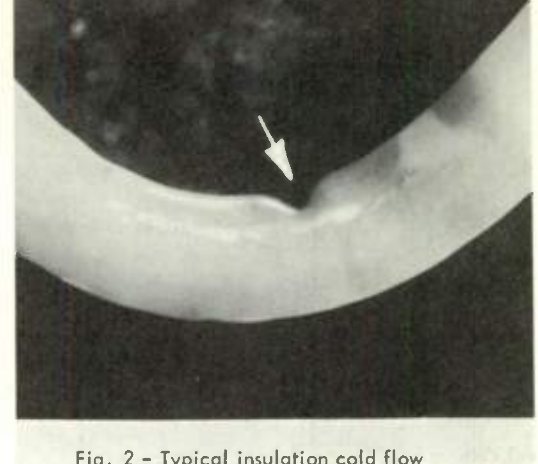


Fig. 2 - Typical insulation cold flow due to pin load after wrapping.

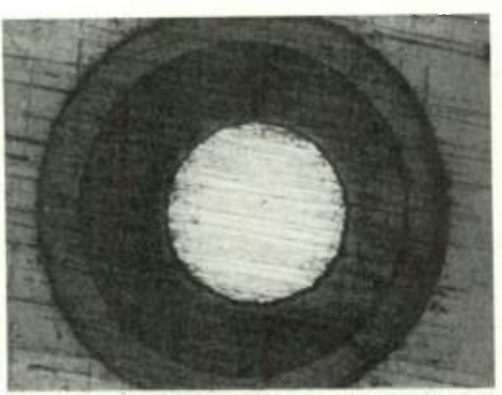


Fig. 3a - Typical cross section of wire and insulation with jacket.

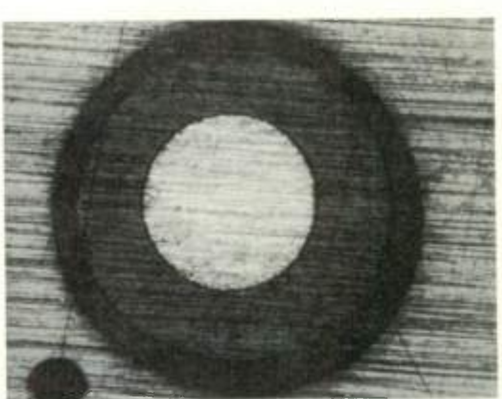


Fig. 3b - Typical cross section of wire and insulation with jacket.

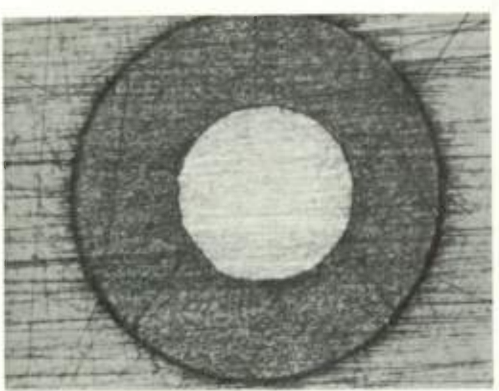


Fig. 3c - Typical cross section of wire and insulation without jacket.

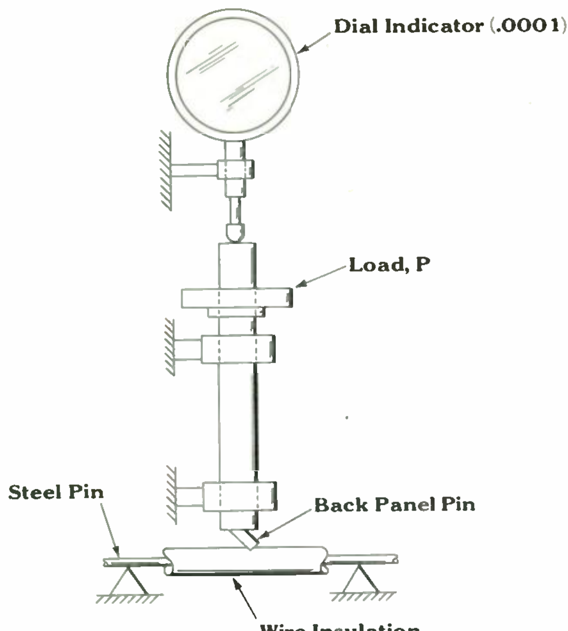


Fig. 4. Test device utilized to measure wire insulation deformation, with mounted wire specimen.

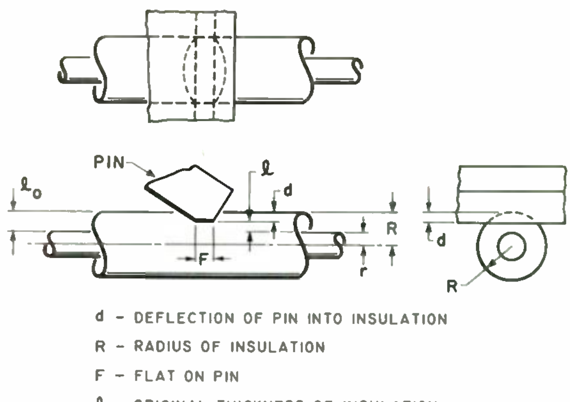


Fig. 6. Detail of measuring deflection of back panel pin for a known load in a period of time.

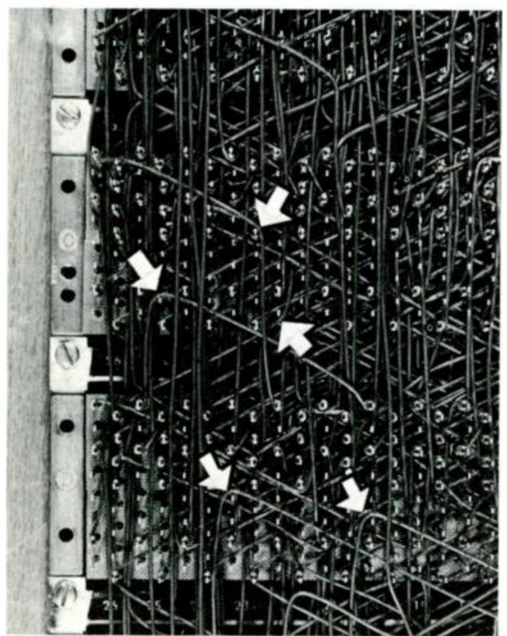


Fig. 5. Typical wired back panel showing wire insulation stress points.

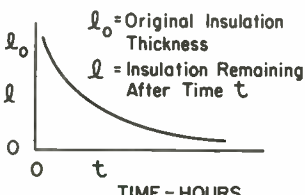


Fig. 7. Data obtained experimentally on device shown in Fig. 4.

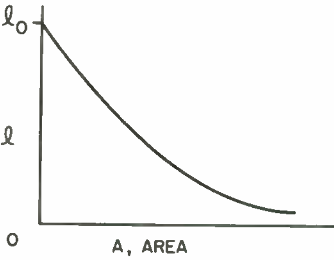


Fig. 8. Curve plot of remaining insulation for various wire radii.

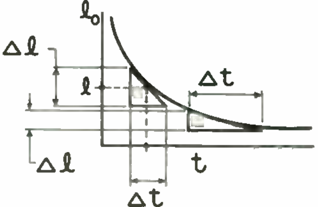


Fig. 9. Curve plot for load, d, and remaining insulation, ℓ.

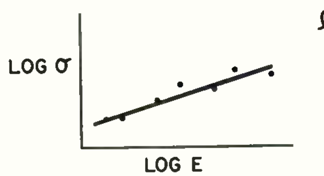


Fig. 10. Curve plot used to compute constants C and n.

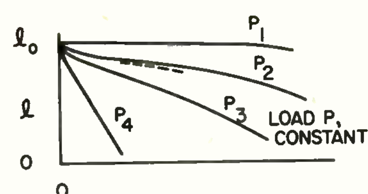


Fig. 11. Curve plotted from computer points. Load P4 > P3 > P2 > P1. Dash line indicates experimental data check.

Wire Number	Hours To Failure* — Temperature °F 78 — Applied Load — Grams			Temperature °F 150			Experimental Constants — Temperature °F 78		Temperature °F 150	
	50	500	1400	50	500	1400	n	c	n	c
0	∞	∞	14	∞	1 x 10	7×10^{-3}	19	7.94×10^{-12}	11.5	1.41×10^{-37}
2	∞	4.6×10^{4}	1.6 x 10	∞	7.5×10^{3}	2×10^{-4}	12.21	1.78×10^{-43}	16.8	7.94×10^{-57}
6	∞	∞	∞	∞	2.4×10^{5}	2.3×10^{-5}	18.18	2.82×10^{-75}	26.85	3.98×10^{-89}
7	∞	∞	7×10^{4}	∞	∞	11×10^{-16}	11.75	7.92×10^{-49}	35.1	7.94×10^{-114}
11	∞	∞	1.9×10^{4}	∞	1.5×10^{4}	35×10^{-7}	8.7	8.51×10^{-36}	21.55	2.51×10^{-73}
12	∞	∞	4.5×10^{4}	∞	∞	47×10^{-11}	11.15	7.94×10^{-45}	38.5	1.26×10^{-129}
14	∞	∞	∞	∞	∞	46×10^{-14}	22.15	6.31×10^{-86}	42.6	6.31×10^{-137}
15	∞	∞	∞	∞	∞	2×10^{-4}	21.75	1.26×10^{-86}	28.15	6.3×10^{-96}
17	∞	∞	∞	∞	∞	1.11 x 10	25.75	10^{-106}	26.1	7.94×10^{-98}
21	∞	∞	∞	∞	∞	9.9×10^{-5}	20.85	2.515×10^{-83}	66	3.16×10^{-239}
23	∞	∞	∞	∞	∞	∞	38.6	6.31×10^{-168}	10.52	5.01×10^{-37}
26	∞	∞	∞	∞	∞	4.9×10^{3}	18.06	1.12×10^{-84}	28.55	5.62×10^{-107}

* Insulations with predicted life greater than 219,000 Hrs. (25 years) are shown as ∞ hours to failure.

Fig. 12 - Experimental constants developed for twelve wire insulations.

FLOW GRAPH TECHNIQUES FOR RELIABILITY ENGINEERING

J. L. Burroughs and W. W. Happ
Microsystems Electronics Department
Lockheed Missiles and Space Company
Sunnyvale, California

Summary

Essentials of flow graph techniques are concisely summarized emphasizing those aspects directly useful in reliability analysis. Illustrative examples of applying flow graphs to reliability problems include:

1. Reliability of systems with many types of variables.
2. Error propagation in systems with interacting variables.
3. Optimization of multistage decision processes.

A survey of over 100 cited references pertinent to the application of flow graphs to reliability engineering leads to the conclusion that a large number of reliability problems can be reduced to a few basic patterns. Flow graphs clearly reveal these patterns, thereby providing a systematic approach to solve a wide range of problems by reverting to existing patterns, if possible, or by adapting or combining known solutions.

Basic Flow Graph Techniques

Introduction

The idea of a signal flow diagram was developed out of the concept of block diagrams which have been used so extensively in the representation and analysis of servomechanisms. In a signal flow graph, simple directed branches instead of blocks are used to denote functional relationships; and nodes are used to represent both summation and signal pickoff points. General techniques and formulas for evaluating signal flow graph relationships have been developed and applied to the analysis of deterministic, probabilistic, and logical systems.

Transmittances and Nodes

A signal flow graph (SFG) represents the relationships between a set of variables in such a way that these relationships can be evaluated by visual inspection. Nodes represent variables, and transmittances represent relationships between variables. For example, the equation

$$y = 10x \tag{1}$$

is represented in signal flow graph form as SFG(2).

x 10 y (2)

This signal flow graph consists of a single transmittance which is directed from node x to node y. This transmittance multiplies node x by 10 and transmits this product to node y. The value of node y is the value of the signal 10x transmitted to it by the transmittance. Thus, SFG(2) represents Eq. (1).

Transmittances may be classified as parallel, series, and loop transmittances; and nodes may be classified as departure nodes and target nodes. The node from which a transmittance is directed is a departure node; and a node to which a transmittance is directed is called a target node.

Parallel transmittances originate at the same departure node and terminate at the same target node, as shown in SFG(3).

x a y b (3)

The value of a target node is the algebraic sum of the signals transmitted into it.

Let transmittances be designated by giving their departure node, value, and target node in that order. The transmittance x, a, y multiplies x by a and transmits the product ax to node y. The transmittance x, b, y multiplies x by b and transmits the product bx to node y. Node y in SFG(3) is the algebraic sum of the two signals ax and bx. Thus, this graph may be written as SFG(4).

x a + b y (4)

In other words two (or more) parallel transmittances are equivalent to one transmittance with a value equal to the sum of the values of the parallel transmittances. A loop transmittance, or self-loop, as it is more commonly called, originates and terminates at the same node as, for example, the transmittance y, c, y does in SFG(5).

x a y b z c (5)

This SFG represents Eq. (6).

$$y = cy + ax + bz \tag{6}$$

Solving Eq. (6) gives Eq. (7) which may be represented in signal flow graph form as SFG(8).

$$y = \frac{ax + bz}{1 - c} \tag{7}$$

x $a(1-c)^{-1}$ y $b(1-c)^{-1}$ z (8)

From SFG(8) it follows that the presence of a self-loop on a node changes the value of the node by a factor of $(1-c)^{-1}$. Thus, each transmittance coming into that node must be multiplied by $(1-c)^{-1}$ when the self-loop is eliminated.

If the departure node of one transmittance is the target node of another transmittance, as in SFG(9), the transmittances are series transmittances.

x a y b z (9)

The equations of SFG(9) are Eq. (10).

$$y = ax \qquad z = by \tag{10}$$

Eliminating y from these equations gives Eq. (11)

$$z = abx \tag{11}$$

which in SFG form is SFG (12).

x ab z (12)

Thus, two (or more) series transmittances are equivalent to one transmittance with a value equal to the product of the values of the series transmittances.

In summary, a transmittance multiplies the value of its departure node by the value of its transmittance and transmits the resulting product to its target node. The value of a target node is the sum of all the signals transmitted into that node. Transmittances may be classified as series, parallel, or loop transmittances. The basic transformations that apply to these three classes of transmittances are summarized in Table 1.

Elimination Techniques

Elimination is the process of reducing the number of nodes in a flow graph without disturbing the mathematical exactness of the graph.

Consider Eq. (13) which are represented by SFG (14).

$$a\,w = x \quad bx = y \quad cy = z \tag{13}$$

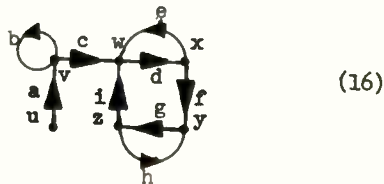
(14)

Eliminating the series transmittances

w,a,x; x,b,y; and y,c,z reduces SFG (14) to SFG (15).

w abc z (15)

In this manner, a whole series of nodes can be eliminated from a signal flow graph in one operation by visual inspection of the graph. However, if there is a self-loop present on a node, that self-loop must be eliminated before eliminating the node.



(16)

Values of transmittances which form continuous paths in a signal flow graph appear as products in the transmittances of a reduced signal flow graph. These paths are called signal flow paths. They originate at a departure node, form a continuous sequence of similarly directed transmittances, and terminate at a target node-without passing through any node more than once. A signal flow path is designated by giving the value of its departure node, its transmittances, and its target node, in the order that these quantities occur when the signal flow path is traced from its departure node to its target node. Some paths which occur in SFG (16) are:

u,a,c,d,f,g,z
v,b,v,
w,d,e,w
z,i,d,f,g,z
y,g,h,y

The value of a path is the product of the values of its transmittances.

There are two types of paths. A path whose target node is different from its departure node is an open flow path, and is called a path, for short. A path whose target node is the same as its departure node is closed flow path and is called a loop. The first signal flow path listed above is a path. The others are loops.

A multinode loop can be reduced to a self-loop by eliminating all of the nodes in the loop but one. Before the one remaining node can be eliminated, the self-loop on that node must be eliminated.

As long as at least one node from each loop is not eliminated, as many intermediate nodes, through which signal flow paths pass, can be eliminated in one step.

Basic Probability Theory

Basic Laws of Probability

Let (a) and ($\bar{a}$) denote events (a) and (not a) respectively; and let (a,b) denote the joint occurrence of events (a) and (b). Two fundamental laws of probability are:

(1) The probability of the joint occurrence of any event (a) and the contradictory event ($\bar{a}$) is zero.

$$p(a, \bar{a}) = 0 \tag{17}$$

(2) Given a set of n mutually exclusive events $\{b_1, b_2 \ldots, b_n\}$, one of which must occur, the total probability of any event (a) is the sum of the probabilities of the events (a,b_1), (a,b_2) $\ldots(a,b_n)$.

$$p(a) = \sum_{i}^{n} p(a,b_i) \tag{18}$$

where

$$p(b_i,b_j) = 0$$

for

$$i \neq j$$

and $\sum_{i}^{n} p(b_i) = 1$

Probability of an Event (a,b,c)

The conditional probability $p(b|a)$ of event (b) given event (a) is defined by SFG (19)

p(a) p(a|b) p(a,b) (19)

provided that p(a) is not zero. The use of conditional probabilities to evaluate probabilities of the joint occurrence of several events is illustrated by SFG (20).

p(a) p(b|a) p(a,b) p(c|a,b) p(a,b,c) (20)

If events (a), (b) and (c) are statistically independent, then

$$p(b|a) = p(b) \tag{21}$$

$$p(c|a,b) = p(c)$$

and SFG (20) simplifies to SFG (22).

p(a) p(b) p(a,b) p(c) p(a,b,c) (22)

Probability of an Event (a+b+c)

According to Eq. (18), the probability p(a+b) that at least one of the events (a) or (b) will occur is given by Eq. (23)

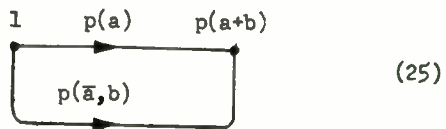
$$p(a+b) = p(a+b,a) + p(a+b, \bar{a}) \tag{23}$$

It can be shown from Eqs. (17) and (18) that

$$p(a+b,a) = p(a)$$

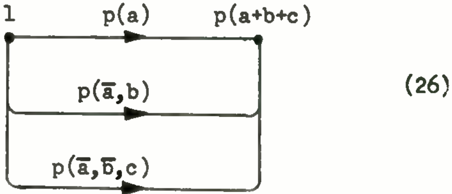
$$p(a+b,\bar{a}) = p(\bar{a},b) \tag{24}$$

Thus, the probability p(a+b) may be evaluated from SFG (25).



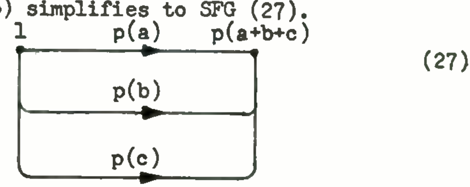
(25)

The extension of SFG (25) to the evaluation of probabilities of at least one event out of a set of more than two events is illustrated by SFG (26).



(26)

If the events a, b, and c are mutually exclusive, SFG (26) simplifies to SFG (27).



(27)

The use of conditional probabilities to evaluate the probabilities $p(\bar{a})$, $p(\bar{a},b)$ and $p(\bar{a},\bar{b},c)$ changes SFG (26) to SFG (28).

1 p(a) p(a+b+c)
p(ā) p(b|ā)
p(b̄|ā) p(c|ā,b̄)
(28)

Other combinations of probabilities exist, but the combinations illustrated above are adequate for the solution of most reliability problems.

Redundant Systems

Time Dependent Failures

In many systems, the failure of one active component changes the failure rates of the other components. For example, consider the power supply system in Fig. 1.

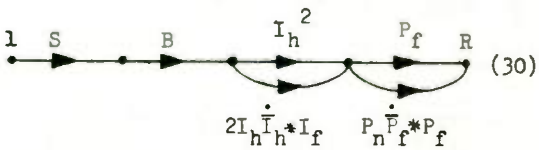
The inverters in the above diagram are in active redundancy because they share the load, whereas the power amplifiers each carry the entire load independently. The standby power amplifier carries no load until the working power amplifier fails; and then the standby amplifier carries the entire load.

The problem is to determine the reliability of the power supply system from the reliabilities of the individual components. The failure rates of the components are given in Table 2.

Assuming that each component reliability $R_c(t)$ is distributed exponentially as a function of time, $R_c(t)$ is given by Equation (29).

$$R_c(t) = e^{-\lambda_c t} \tag{29}$$

The reliability flow graph of the system diagrammed in Figure 1 is SFG (30).



(30)

The letters S, B, I, P, and R denote the reliabilities of the solar cell array, the battery, the inverters, the power amplifiers, and the entire power supply system, respectively. The subscripts f, h, and n distinguish reliabilities at full, half, and no load. The letters with bars over them denote probabilities of failure; e.g.,

$$\bar{I}_h = 1 - I_h(t) \tag{31}$$

The dot over a letter indicates differentiation with respect to time; e.g.,

$$\dot{\bar{I}} = \frac{d}{dt}\left[1 - I_h(t)\right] \tag{32}$$

is the probability that an inverter will fail at time t under half-load conditions.

Time-dependent failure rates give rise to convolution integrals. Convolution is indicated by the convolution integral operation * which is defined by the following identity.

$$f(t) * g(t) \equiv \int_0^t f(t_1)\, g(t-t_1)dt_1 \tag{33}$$

For example, consider the power amplifier sub-system, which consists of two power amplifiers in standby redundancy. This subsystem can operate until a time t in either one of two ways: (1) by operation of the active amplifier for the entire time t; or (2) by operation of the active amplifier until time t_1, and operation of the standby amplifier from time t_1 until time t. The probability of operating in the first way is simply $P_f(t)$. The probability of operating in the second way is $\left[P_n(t)\ \dot{\bar{P}}_f(t)\right] * \left[P_f(t)\right]$, that is:

$$P_n\dot{\bar{P}}_f * P_f = \int_0^t P_n(t_1)\, \frac{d}{dt_1}\left[1-P_f(t_1)\right] P_f(t-t_1)dt_1 \tag{34}$$

The probability that the standby amplifier will not fail with no load on it at or before time t_1 is $P_n(t_1)$. The probability that the active amplifier will fail at time t_1 with full load on it is $d\left[1-P_f(t_1)\right]/dt_1$. The probability that the standby amplifier will then operate at full load from time t_1 until time t is $P_f(t-t_1)$. Integrating the product of these three probabilities over all values of t_1 from 0 to t gives the total probability of operating in the second way.

In the case of the two inverters which are operating in active redundancy, there is one primary mode of operation and two secondary modes. The primary mode is operation of both inverters at half rated load for a period of time t. The probability of operating in the primary mode is $\left[I_h(t)\right]^2$.

The secondary modes each consist of operation of both inverters at half load until one inverter fails at time t_1, and operation of the other inverter at full load from time t_1 until time t. The system can get into a secondary mode by failure of either one of the two inverters. Thus, the total probability of operating in either secondary mode is $2\left[I_h(t)\ \dot{\bar{I}}_h(t)\right] * \left[I_f(t)\right]$, that is:

$$2I_h\dot{\bar{I}}_h * I_f = \int_0^t 2I_h(t_1)\, \frac{d}{dt_1}\left[1-I_h(t_1)\right] I_f(t-t_1)dt_1 \tag{35}$$

The values of the transmittances in SFG (30) can be evaluated by substituting data from Table 2 into Eq. (29).

$$S = e^{-1.6t} \tag{36}$$

$$B = e^{-1.8t}$$

$$I_h^2 = e^{-0.8t}$$

$$2\, I_h\dot{\bar{I}}_h * If = 2\int_0^t \left(e^{-0.4t_1}\right)\left(0.4e^{-0.4t_1}\right)e^{-0.6(t-t_1)}dt_1$$

$$2\, I_h\dot{\bar{I}}_h * I_f = 4\left(e^{-0.6t} - e^{-0.8t}\right)$$

$$P_f = e^{-0.5t}$$

$$P_n\dot{\bar{P}}_f * P_f = \int_0^t \left(e^{-0.1t_1}\right)\left(0.5e^{-0.5t_1}\right)\left(e^{-0.5(t-t_1)}\ dt_1\right)$$

$$P_n\dot{P}_f * P_f = 5\,(e^{-0.5t} - e^{-0.6t})$$

Substituting from Eqs. (36) into the expression for R given by SFG (30) gives

$$R = (SB)(I_h^2 + 2I_h\dot{I}_h * I_f)(P_f + P_n\dot{P}_f * P_f)$$

$$R = (e^{-3.4t})(4e^{-0.6t} - 3e^{-0.8t})(6e^{-0.5t} - 5e^{-0.6t}) \quad (37)$$

The example presented here demonstrates a method of formulating reliability flow graphs of systems in which the failure of one component affects the failure rates of other components. The method consists of using convolution integrals to "sum up" probabilities over an interval of time. Using convolution integrals and mutually exclusive events which cover all relevant possibilities converts reliability problems into ordinary algebraic problems which can be readily solved by flow graph techniques.

The extension of this method of formulating reliability flow graphs from the simple example illustrated in this paper to more complicated reliability problems which occur in practical engineering work is straightforward. The extensions involve a greater number of probabilistic relations between a larger number of events, but the basic method of formulation remains unchanged.

Open and Short Circuit Failures

Resistor Quad Analysis. The reliability of a circuit is frequently dependent on the failure modes as well as the probability of the occurrence of failures.

Each part of a circuit might, for example, operate normally, fail open, or fail short. Consider the four part resistor quad in Fig. 2. Such a quad of resistors might be used as the input resistor for the multivibrator circuit shown in Figure 3.

The resistor quad is equivalent to two parallel subcircuits connected in series as shown in Fig. 4.

Let $n(r_1)$, $o(r_1)$ and $s(r_1)$ denote the probabilities of normal operation, open circuit failure, and short circuit failure of component r_1. Normal operation is defined as operation without failing open or failing short. It follows that the events which correspond to the probabilities $n(r_1)$, $o(r_1)$, and $s(r_1)$ are mutually exclusive; and that one of these events must occur.

$$n(r_1) + o(r_1) + s(r_1) = 1 \quad (38)$$

The probability $\bar{s}(r_2)$ denotes the probability the component r_2 does not fail short.

$$\bar{s}(r_2) = 1 - s(r_2) \quad (39)$$

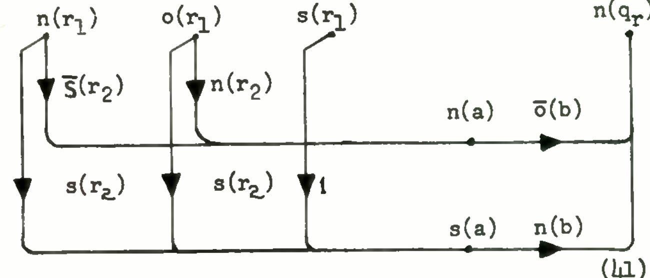
$$\bar{s}(r_2) = n(r_2) + o(r_2)$$

Similarly, $\bar{o}(b)$ denotes the probability that subcircuit b does not fail open.

$$\bar{o}(b) = 1 - o(b) \quad (40)$$

$$\bar{o}(b) = n(b) + s(b)$$

Using the notation described above, the reliability of the resistor quad is denoted as $n(q_r)$. The value of $n(q_r)$ is given by SFG (41).



(41)

The derivation of SFG (41) is as follows: Let the events which correspond to the probabilities $n(r_1)$, $o(r_1)$, and $s(r_1)$ be denoted as events $n(r_1)$, $o(r_1)$ and $s(r_1)$ respectively. Similarly, let $n(a)$ denote both the probability and the event of normal operation of subcircuit a. The probability of event $n(a)$ is given by Eq. (42).

$$n(a) = [n(a)|n(r_1)]\; n(r_1) + [n(a)|o(r_1)]\; o(r_1) + [n(a)|s(r_1)]\; s(r_1) \quad (42)$$

where $[n(a)|n(r_1)]$ denotes the probability of event $n(a)$ given event $n(r_1)$. Since subcircuit a consists of resistor r_1 parallel with resistor r_2, $[n(a)|n(r_1)]$ is the probability $\bar{s}(r_2)$ that resistor r_2 will not fail short.

$$[n(a)|n(r_1)] = \bar{s}(r_2) \quad (43)$$

Similarly

$$[n(a)|o(r_1)] = n(r_2) \quad (44)$$

$$[n(a)|s(r_1)] = 0$$

Combining Eqs. (42), (43), and (44) gives the signal inputs and transmittances associated with node $n(a)$ in SFG (41). The signal inputs and transmittances associated with the other nodes in SFG (41) can be obtained in a similar manner. Suppose that the probabilities $n(r_1)$, $o(r_1)$ and $s(r_1)$ have the following values:

$$n(r_1) = 1-u$$

$$o(r_1) = 0.959\ u \quad (45)$$

$$s(r_1) = 0.041\ u$$

If the probabilities associated with each resistor in the quad have the same values, the reliability of the resistor quad is given by Eq. (46). From SFG (41) and Eqs. (45) it follows that

$$n(a) = n(r_1)\ \bar{s}(r_2) + o(r_1)\ n(r_2)$$

$$n(a) = n(r_1)\left[1-s(r_1) + o(r_1)\right]$$

$$n(a) = (1-u)(1-0.918u)$$

$$n(a) = 1-1.918u + 0.918u^2$$

$$s(a) = n(r_1)\ s(r_2) + o(r_1)\ s(r_2) + s(r_1)$$

$$s(a) = s(r_1)\left[n(r_1) + o(r_1) + 1\right]$$

$$s(a) = 0.041u\ (2-0.041u)$$

$$s(a) = 0.082u - 0.001681u^2$$

$$n(q_r) = n(a)\ \bar{o}(b) + s(a)\ n(b) \quad (46)$$

$$n(q_r) = n(a)\left[1 - o(a) + s(a)\right]$$

$$n(q_r) = n(a)\ \left[n(a) + 2\ s(a)\right]$$

$$n(q_r) = (1 - 1.918u + 0.918u^2)x$$

$$(1 - 1.836u + 0.918u^2)$$

$$n(q_r) = 1 - 3.754u + 5.357u^2 + 0(u)$$

If the unreliability of the resistors u has a value of 0.00455, the reliability of the resistor quad turns out to be

$$n(q_r) = 0.98303 \quad (47)$$

The use of conditional probabilities to formulate reliability flow graphs of series and parallel combinations of components which are subject to open and short circuit failures has been demonstrated. The conditional probability method is relatively simple and straightforward, compared to other methods, such as multinomial expansions. The extension of the conditional probability method of formulation to components with more than two modes of failure is straightforward, and the method can certainly be applied to bridge circuits as well as series and parallel combinations of components. However, one bridge in a circuit is sufficient to change the problem of flow graph formulation from a relatively simple to a tedious task.

Reliability Flow Graphs of Multimode Systems

System Description

Figure 5 is a block diagram of a hypothetical satellite tracking system which has six different modes of operation. For searching, the system has search radar and optics; for tracking, it has track radar, optics, and a computer. Table 3 lists the six different modes in which the system can search for and track a target.

Operating Reliability

Let r_i denote the reliability of operating in the i th mode. Each r_i can be evaluated directly from Table 1. For example, the table shows that components A, B, C, and D operate in mode 1. Therefore, if A, B, C, and D denote the reliabilities of the corresponding components, r_1 is the product of these reliabilities.

$$r_1 = ABCD \quad (48)$$

Similarly

$$r_2 = ACDE \quad (49)$$

$$r_3 = ACE$$

$$r_4 = BCDE$$

$$r_5 = DE$$

Mode Effectiveness

The effectiveness E_i of mode i is the ratio of the quality of system performance in mode i to the quality of performance of the system in mode 1. The performance quality of each mode is estimated from an analysis of theoretical and actual operational data.

Mode Reliability

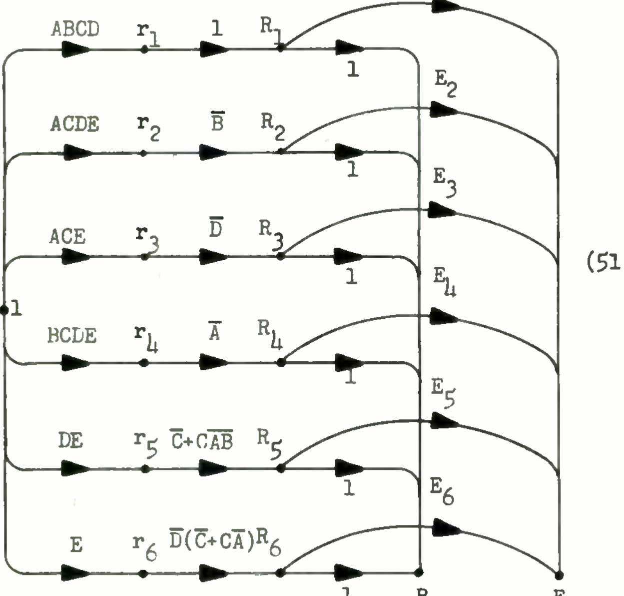
The probability of getting into the i th mode is denoted by p_i. The mode reliability r_i of getting into and operating in the i th mode is the product of p_i and r_i.

$$R_i = p_i\ r_i \quad (50)$$

Reliability Flow Graph

Signal flow graph (51) is the reliability flow graph of the system. The system reliability R is the probability that the system will operate in any of the six mutually exclusive modes. As indicated

by SFG (51), the system reliability R is the sum of the mode reliabilities.



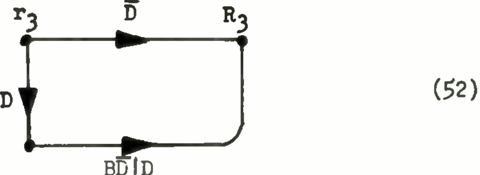
Similarly, SFG (51) shows that the system effectiveness E_s is the sum of the products R_1 E_1, R_2 E_2, etc.

The value of the transmittance r_1, 1, R_1 is value of p_1 , the probability getting into mode 1. This value must be unity, since mode 1 is the first mode in the mode sequence.

More complicated transmittances such as r_6, $\bar{D}(\bar{C} + \bar{C}\bar{A})$, R_6 can be evaluated from Table 3. The reasoning is as follows: Component E operates in mode 6. If component D were also operating, the system would be in mode 5. Therefore, component D must fail in order to eliminate the possibility of being in mode 5 instead of mode 6. With component E working and component D not working, the system could be in mode 3, because component E is used and component D is not used in that mode. To eliminate the possibility of being in mode 3, component A or component C must fail. This eliminates the possibility of being in any mode except mode 6. Thus, the value of the transmittance r_6, $\bar{D}(\bar{C} + \bar{C}\bar{A})$, R_6 is p_6, the probability of getting into mode 6.

Conditional Probability

The conditional nature of the probabilities associated with transmittances must be kept in mind when evaluating the probability of getting into a mode. For example, consider the transmittance r_3, $\bar{D}$, R_3. If components A, C, and E work and component D fails, mode 3 is the only possible mode. If component D does not fail, it would seem that it might be possible to get into mode 3 by having component B fail, but the system could also be in mode 2 if component B fails. To eliminate the possibility of being in mode 2, component D must also fail, but the probability $(\bar{D}|D)$ that component D fails given that component D does not fail is obviously zero. Signal flow graph (52) summarizes the above arguments.



The conditional probability $\bar{B}\bar{D}|D$ (that components B and D both fail, given that component D does not fail) is obviously zero. Therefore, SFG (52) reduces to the transmittance r_3, $\bar{D}$, R_3 which was used in SFG (51).

Generalization

The above example is probably too simple to demonstrate the effectiveness of signal flow graph techniques for analyzing the probabilities of obtaining the various modes of a multimode system. It is chiefly when systems become more complex that signal flow graph techniques reveal their power. Without the use of signal flow graphs it is difficult for one's mind to follow through all facets of a problem and to retain them in sufficient detail, so that a logically coherent solution can be obtained. On the other hand, the use of signal flow graphs systematizes problem formulation so that steps can be retraced and details can be scrutinized from a broad perspective viewpoint.

Special Flow Graph Techniques

Path Inversion

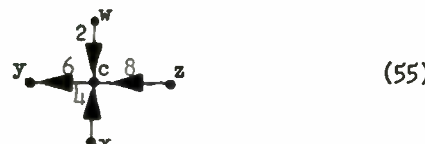
An operation that greatly enhances the flexibility of a signal flow graph as a mathematical model of a system is the operation of path inversion. A path represents a series of functional relationships which constitute a functional dependency of the target node of the path on the departure node of the path. Inversion is the operation of reversing the direction of a path so that its target node and departure node are interchanged, and the functional dependency denoted by the path is inverted. For example, consider SFG (53).

$$x \xrightarrow{a} y \qquad (53)$$

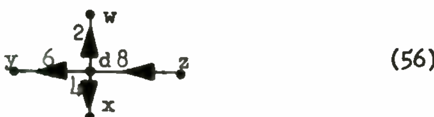
$$x \xleftarrow{\frac{1}{a}} y \qquad (54)$$

The inversion of the transmittance x, a, y in SFG (53) changes it to the transmittance y, $\frac{1}{a}$, x in SFG (54). This shows that inverting a transmittance reverses its direction and inverts its value.

In the process of inverting the transmittances of a path, it is convenient to distinguish three types of nodes: contributive nodes, distributive nodes, and unstretched nodes. In SFGs (55), (56), and (57), nodes c, d, and u are examples of contributive, distributive, and unstretched nodes respectively.



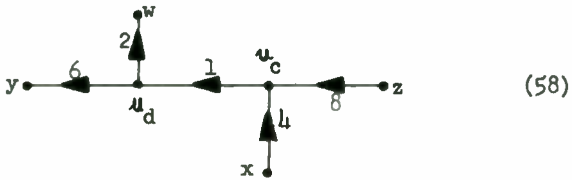
(55)



(56)

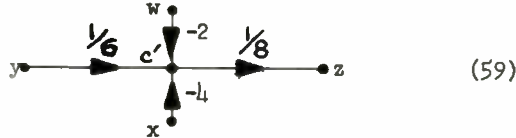
(57)

A contributive node is the target node of more than one transmittance and the departure node of only one transmittance. A distributive node is the target node of only one transmittance and the departure node of more than one transmittance. An unstretched node is the target node of more than one transmittance and also the departure node of more than one transmittance. An unstretched node can be split into a contributive node and a distributive node which are joined by a transmittance with a value of unity. This operation is known as stretching the node. The node u in SFG (57) can be stretched into a contributive node u_c and a distributive node u_d as shown in SFG (58).



(58)

Any transmittance not included in a signal flow path, but which contributes a signal to or distributes a signal from a node on the signal flow path is a tributary transmittance. Tributary transmittances which contribute signals to a node on a signal flow path are known as contributive transmittances. Tributary transmittances which distribute signals from a node on a path are known as distributive transmittances. To invert a path through a contributive node, the transmittances of the path must be inverted and the sign of contributive transmittances must be changed. For example, inverting the path z, 8, 6, y in SFG (54) changes it to the path y, 1/6, 1/8, z in SFG (59).



(59)

The value of the contributive node c' in SFG (59) is not the same as the value of the contributive node c in SFG (55). From SFG (55), the value of c is:

$$c = 8z + 2w + 4x \tag{60}$$

or

$$z = 1/8\,(c - 2w - 4x) \tag{61}$$

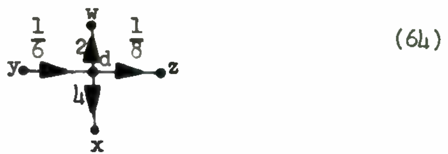
from SFG (59) the value of z is

$$z = 1/8\,c' \tag{62}$$

Comparing Eqs. (61) and (62) shows that

$$c' = c - 2w - 4x \tag{63}$$

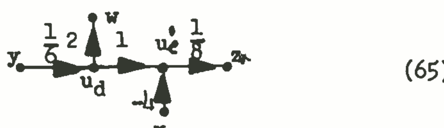
Substracting the contributed signals 2w and 4x from c gives a signal which when multiplied by the value of the inverted transmittance c', 1/8, z gives the correct value for z. The signal with the value of c is transmitted to node c' in SFG (59) by the transmittance y, 1/6, c' which is inverse of the transmittance c, 6, y in SFG (55). The inversion of a path through a distributive node is done by simply inverting the transmittances of the path. For example, inverting the path z, 8, 6, y in SFG (56) gives the path y, 1/6, 1/8, z in SFG (64).



(64)

Since the distributive transmittances d, 2, w and d, 4, x do not contribute anything to the signal flowing along the path, the signs of the distributive transmittances do not have to be changed when the path is inverted.

The inversion of a path through an unstretched node can be handled by the methods explained above, after the node has been stretched. For example, the node u in SFG (57) can be stretched as shown in SFG (58) and then the path z, i, 1, 6, y in that graph can be inverted as shown in SFG (65).



(65)

Obviously, if there are no tributary transmittances, there are no contributed signals to substract. Hence, a path through a node which has no tributary transmittances can be inverted by simply inverting the transmittances of the path.

Thus, the path through node y in SFG (66) can be inverted as shown in SFG (67).

x 6 y 8 z (66)

x 1/6 y 1/8 z (67)

In summary, the procedure for inverting a signal flow path can be summarized as follows:

(1) stretch all unstretched nodes and invert all of the transmittances in the signal flow path.

(2) change the sign of all contributive transmittances which contribute signals to the nodes on the path.

(3) indicate the change in value of all contributive nodes in the signal flow path.

By following this procedure, paths which start at a signal source may be inverted. A signal source is a node which is not the target node of any transmittance, but the departure node of one or more transmittances.

Signal Evaluation

At least as early as 1918, graphs similar to signal flow graphs were used to evaluate a determinants of systems of equations. A linear signal flow graph with n target nodes represents a set of n linear algebraic equations.

$$\sum_{i=1}^{n} a_{ij} x_i + y_j = x_j \qquad j = 1, 2, 3, \ldots n \tag{68}$$

Solving these equations for the y's in terms of the x's transforms Eqs. (68) to Eqs. (69).

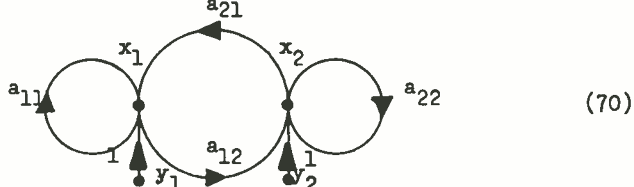
$$\sum_{i=1}^{n} (\delta_{ij} - a_{ij}) x_i = y_j \tag{69}$$

where

$$\delta_{ij} = 1 \text{ when } i = j$$
$$\delta_{ij} = 0 \text{ when } i \neq j$$

The determinant of the transformed equations is defined as the determinant Δ of the signal flow graph which represents Eqs. (68).

For example, consider SFG (70).



(70)

It represents Eqs. (71).

$$x_1 = a_{11} x_1 + a_{21} x_2 + y_1$$
$$x_2 = a_{21} x_1 + a_{22} x_2 + y_2 \tag{71}$$

Solving Eqs. (71) for the y's gives Eqs. (72).

$$y_1 = (1-a_{11}) x_1 - a_{21} x_2$$
$$y_2 = -a_{12} x_1 + (1-a_{22}) x_2 \tag{72}$$

The determinant of Eqs. (72) is .

$$\Delta = \begin{vmatrix} 1-a_{11} & -a_{21} \\ -a_{12} & 1-a_{22} \end{vmatrix} \tag{73}$$

$$\Delta = (1-a_{11})(1-a_{22})-a_{12} a_{21}$$

The determinant Δ is the determinant of SFG (60). The first term of Δ involves the values of the self-loops x_1, a_{11}, x_1 and x_2, a_{22}, x_2.

The second term of Δ involves the multinode loop x_1, a_{12}, a_{21}, x_1. It can be shown that the determinant of any signal flow graph with n target nodes is given by Eq. (74).

$$\sum_{q} \sum_{r} (-1)^q M_r^q S_r^q \tag{74}$$

where M_r^q is the product of the rth possible combination of q nontouching multinode loops; and S_r^q is a product of the form

$$(1-a_{11})(1-a_{22})(1-a_{33})\ldots(1-a_{nn})$$

in which the self-loops touched by the loops in

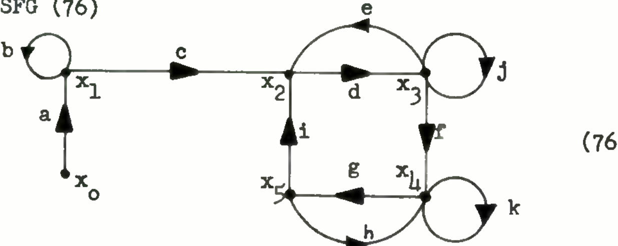
M_r^q are set equal to zero. The index q runs from zero to a maximum of n/2, where n is the number of target nodes in the graph. For q equal to zero

$$(-1)^0 M_r^0 S_r^0 = (1-a_{11})(1-a_{22})\dots(1-a_{nn}) \quad (75)$$

with each self-loop assigned the same value that it has in the signal flow graph.

Flow Graph Evaluation

To illustrate what M_r^q and S_r^q denote, consider the evaluation of the determinant of SFG (76)



(76)

The self-loops on nodes x_1, x_3, and x_4 have the values b, j, and k respectively.

$$a_{11} = b \qquad a_{33} = j \qquad a_{44} = k \quad (77)$$

The other nodes have no self-loops on them.

$$a_{22} = a_{55} = 0$$

Thus $\sum_r (-1)^0 M_r^0 S_r^0 = \prod_{i=1}^{5} (1-a_{ii})$ (78)

$$\sum_r (-1)^0 M_r^0 S_r^0 = (1-b)(1-j)(1-k)$$

There are three multinode loops in SFG (76): x_2, d, e, x_2; x_2 d, f, g, i, x_2; and x_4, g, h, x_4. Consequently, there are three possible ways of taking the loops one at a time. Thus, for q equal to one, r runs from 1 to 3.

$$\sum_r (-1)^1 M_r^1 S_r^1 = -deS_1^1 - dfgiS_2^1 - ghS_3^1 \quad (79)$$

The loop whose value (de) occurrs in M_1^1 touches nodes x_2 and x_3. Consequently

$$S_1^1 = (1-a_{11})(1-a_{44})(1-a_{55})$$
$$S_1^1 = (1-b)(1-k) \quad (80)$$

Similarly

$$S_2^1 = (1-b)$$
$$S_3^1 = (1-b)(1-j) \quad (81)$$

The only two multinode loops in SFG (76) that do not touch each other are x_2, d, e, x_2 and x_4, g, h, x_4.
Thus,

$$\sum_r (-1)^2 M_r^2 S_r^2 = deghS_1^2 \quad (82)$$

The two loops whose values (de) and (gh) appear in M_1^2 touch nodes x_2, x_3, x_4, and x_5. Therefore

$$S_1^2 = 1-a_{11} = 1-b \quad (83)$$

There are no combinations of three or more multinode loops which do not touch, so the complete expression for the determinant of SFG (76) is given by Eq. (84).

$$\Delta = \sum_q \sum_r (-1)^q M_r^q S_r^q$$

$$\Delta = (1-b)(1-j)(1-k)-de(1-b)(1-k)-(dfgi)(1-b)-(gh)(1-b)(1-j)+(degh)(1-b) \quad (84)$$

By the use of signal flow graph determinants, the following concise formula for the value of a target node x can be obtained.

$$x = \sum_i \sum_r S_i P_{ir} \Delta_{ir} / \Delta \quad (85)$$

where

S_i = the value of the i'th signal source

P_{ir} = the value of the r'th possible path from node S_i to node x.

Δ_{ir} = the determinant of the part of the signal flow graph which does not touch the path P_{ir}.

Δ = the determinant of the complete signal flow graph

For example, node x_3 in SFG (76) can be evaluated as follows. The only signal source is x_o. The only path from x_o to x_3 is x_o, a, c, d, x_3. The value of this path is acd. The determinant of SFG (76) is given by Eq. (84). The expression Δ_{ir} is called the cofactor of the path P_{ir}. Its value is the determinant of the signal flow graph evaluated with all the nodes in path P_{ir} ignored. This leaves only the loops through nodes x_4 and x_5 to be considered. The determinant of this part of the graph is Δ_{ir}

$$\Delta_{ir} = 1-k-gh \quad (86)$$

Using these values in Eq. (85) gives the value of x_3.

$$x_3 = x_o(acd)(1-b-gh)/\Delta \quad (87)$$

where

$$\Delta = (1-b)(1-j)(1-k)-de(1-b)(1-k)-(dfgi)(1-b)-(gh)(1-b)(1-j)+(degh)(1-b) \quad (88)$$

In summary, the formula for the value of a target node x is:

$$x = \sum_i \sum_r s_i P_{ir} \Delta_{ir} / \Delta \quad (89)$$

$$\Delta = \sum_q \sum_r (-1)^q M_r^q S_r^q \quad (90)$$

and S_r^q = the product $\prod_j (1-L_j)$ with values of the self-loops on the nodes in the loops contained in M_r^q set equal to zero.

The L_j denotes the value of the self loop on the j'th target node in the signal flow graph. The value of the zero'th term of the determinant is the product $\prod_j (1-L_j)$ where the values of all self-loops are the same as in the signal flow graph.

Markhov Processes

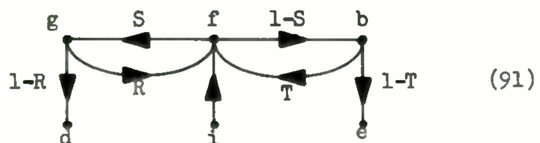
Definition and Example

A convenient way of approximating many statistical processes is to specify a discrete set of states which are related by a set of transition probabilities. If the transition probability of going from one state to another is a function of only the two states, the process is called a simple Markhov process.

As an example of a Markhov process, consider a recoverable missile system. The reliability of each missile is characterized by:

S = the probability of a successful flight
R = the probability of recovery after a successful flight
T = the probability of recovery after an unsuccessful flight

Assuming that the probabilities S, R, and T are constant, the reliability flow graph of such a missile system is SFG (91).

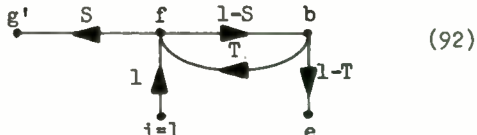


(91)

Each node corresponds to a specific event.

i ~ initiation of the first flight
f ~ a missile flight
g ~ a successful flight
b ~ an unsuccessful flight
d ~ non-recovery after a successful flight
e ~ non-recovery after an unsuccessful flight

The probability P_{ij} of the first occurrence of event j given event i is obtained by driving node i with a unit signal and evaluating node j as a strictly dependent node. For example, the probability of the first occurrence of event (g) given event (i) is obtained evaluating node g' in SFG (92).

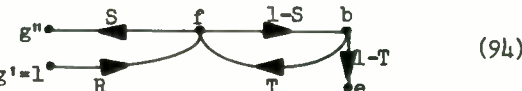


(92)

Node g' is made strictly dependent by removing the transmittances in SFG (91) which have node g as a departure node.

$$P_{ig} = \frac{S}{1-(1-S)T} \quad (93)$$

The probability P_{jj} of the first recurrence of an event (j) is evaluated in essentially the same way. For example, the probability of the first recurrence of event (g), given event (g), is obtained by evaluating node g" in SFG (94).



(94)

Node g in SFG (91) is split into an independent departure node g' and a strictly dependent target node g", and then node g" is evaluated with node g' driven by a unit signal. Thus, the following value is obtained for P_{gg}.

$$P_{gg} = \frac{RS}{1-(1-S)T} \quad (95)$$

The first occurrence probabilities of each node in a system flow graph, such as SFG (91), can usually be evaluated directly from the original system flow graph by mentally performing the simple flow graph transformations demonstrated above. The first occurrence probabilities associated with the nodes in SFG (91) are given in Table 4.

Probability of Simple Events

The probability distributions of the number of occurrences of each event associated with SFG (91) can be characterized in terms of the following functions.

$P(j=y|i)$ = the probability that event (j) will occur exactly y times, given event (i).
$E(j|i)$ = the expected number of occurrences of event (j) given event (i).
$V(j|i)$ = the variance of the number of occurrences of event (j) given event (i).
$G(j|i;t)$ = the generating function which gives the probability of the number of occurrences of event (j), given event (i), with t as a parameter.

If the event whose occurrence is given is implied by the conditions of the problem, the above notation may be simplified to $P(j=y)$, $E(j)$, $V(j)$, and $G(j;t)$ without causing any confusion. The expressions for these functions in terms of the first occurrence probabilities are:

$$P(j=y) = P_{ij}(P_{jj})^{y-1}(1-P_{jj}) \tag{96}$$

$$E(j) = P_{ij}(1-P_{jj})^{-1} \tag{97}$$

$$V(j) = P_{ij}(1-P_{jj}-P_{ij})(1-P_{jj})^{-2} \tag{98}$$

$$G(j;t) = 1-P_{ij}+(1-P_{jj})(P_{ij}t)(1-P_{jj}t)^{-1} \tag{99}$$

Equation (96) gives the probability of exactly y occurrences of event (j) as the product of three probabilities: (1) the probability of the first occurrence of event (j); (2) the probability of (y-1) recurrences of event (j); and (3) the probability that event (j) will not recur again.

Expanding equation (99) in a power series, for example by division, obviously generates the probabilities of successively greater numbers of occurrences of event (j). Equations (97) and (98), in turn, may be readily obtained from the generating function by differentiation. Thus, equations (97), (98) and (99) may be derived from equation (96). Table 5 lists the probabilities of exactly n occurrences and the expected number of occurrences of events (g), (b), and (f).

Probability of Compound Events

In the evaluation of missile effectiveness, the probability $P(g=n,b=k)$ of exactly n successful flights and k unsuccessful flights is a critical factor. This probability may be found by inspection of the system flow graph, SFG (91).

A state point is defined so that it moves from one node to another node as the state of the process moves from one event to another. The process of completing n successful flights and k unsuccessful flights starts with event (i) which corresponds to node i in SFG (91). The state point moves from node i to node f, as the missile goes into its first flight. If the flight turns out to be successful, the state point moves on to node g. In the event of an unsuccessful flight, the state point moves to node b instead of node g. Thus, the process can be visualized as moving from event to event as the state point moves from node to node.

The ability to visualize a process facilitates the evaluation of the probabilities of joint events. For example, consider the probability $P(g=n,b=k)$. The event $(g=n,b=k)$ must end with the non-recovery of a missile; that is, with either event (d) or event (e). It follows that the probability of event $(g=n,b=k)$ is given by Equation (100).

$$P(g=n,b=k) = P(g=n,b=k \text{ given } d)P(d) + P(g=n,b=k \text{ given } e)P(e) \tag{100}$$

The first occurrence of event (d) or event (e) is the last. Therefore

$$P(d) = P_{id} = \frac{S(1-R)}{1-SR-(1-S)T} \qquad P(e) = P_{ie} = \frac{(1-S)(1-T)}{1-SR-(1-S)T} \tag{101}$$

If the state point ends up at event (d), the point must also move completely around loop g, R, S, g (n-1) times, and completely around loop b, T, 1-S, b k times. These loop traversals are necessary for the state point to enter node (g) n times and node (b) k times. The probability of moving around a loop once is the value of the loop transmittance. Thus, the conditional probability of event $(g=n,b=k)$ given event (d) is $P(g=n,b=k|d)$.

$$P(g=n,b=k|d) = \binom{k+n-1}{k}(RS)^{n-1}T^k(1-S)^k \tag{102}$$

Similarly

$$P(g=n,b=k|e) = \binom{n+k-1}{n}(RS)^{n}T^{k-1}(1-S)^{k-1} \tag{103}$$

Substituting these probabilities in Equation (100) gives the value of $P(g=n,b=k)$.

$$P(g=n,b=k) = \left[(RS)^{n-1}T^k(1-S)^kS(1-R)\binom{k+n-1}{k} + (T)^{k-1}(1-S)^{k-1}(RS)^n(1-S)(1-T)\binom{n+k-1}{n}\right] \div \left[1-SR-(1-S)T\right] \tag{104}$$

The effectiveness of a recoverable missile system can be expressed in terms of the probability $p(g\geq n;m)$ that at least n successful flights are obtained using m missiles. The probability $p(g\geq n;m)$ can be expressed in terms of the probability $p(g=k;m,j)$ that exactly k successful flights will be achieved with m missiles when j missiles achieve no successful flights.

$$p(g\geq n;m) = 1-\sum_{k=0}^{n-1}\sum_{j=0}^{m} p(g=k;m,j) \tag{105}$$

The probability $p(g=k;m,j)$ can be expressed in terms of the first occurrence probabilities P_{ig} and P_{gg}.

$$p(g=k;m,j) = \binom{m}{j}\binom{n-1}{m-j-1}(P_{ig})^{m-j}(1-P_{ig})^{j} \times (1-P_{gg})^{m-j}(P_{gg})^{n-m+j} \qquad (106)$$

One possible use of these results would be to determine the minimum number of missiles which will give a specified value of the probability $p(g \geq n;m)$ of at least n successful flights.

Sensitivity

Another use of the above results would be to find a criterion for selecting the probability S, R or T to which the expected number of successful flights is most sensitive. In feedback theory, the sensitivity S_p^v of variable v to parameter p is defined as the ratio of $\partial v/v$ to $\partial p/p$.

The expected number of occurrences of successful flights is a bi-linear expression of each of the probabilities S, R and T.

$$E(g) = \frac{S}{1-SR-(S-1)T} \qquad (107)$$

This fact can be used to obtain a convenient formula for the related sensitivities. Let N denote the numerator of E(g) and D, the denominator.

$$E = \frac{N}{D} \qquad (108)$$

Since the numerator and denominator are each linear in S, they may be expressed as

$$N = N_1 S + N_0$$
$$D = D_1 S + D_0 \qquad (109)$$

The sensitivity S_S^E can now be obtained by differentiation.

$$E = \ln N - \ln D$$
$$\left(\frac{1}{E}\right)\frac{\partial E}{\partial S} = \left(\frac{1}{N}\right)\frac{\partial N}{\partial S} - \left(\frac{1}{D}\right)\frac{\partial D}{\partial S}$$
$$\frac{\partial N}{\partial S} = N_1 = \frac{N-N_o}{S}$$
$$\frac{\partial D}{\partial S} = D_1 = \frac{D-D_o}{S} \qquad (110)$$
$$S_S^E = \frac{S}{E}\frac{\partial E}{\partial S} = \frac{D_o}{D} - \frac{N_o}{N}$$

The quantities D_o and N_o are the values of D and N when S is zero.

$$D_o = 1-T$$

$$N_o = 0$$

$$S_S^E = \frac{D_o}{D} - \frac{N_o}{N} = \frac{1-T}{1-SR-(1-S)T} \qquad (111)$$

The sensitivities S_R^E and S_T^E can be calculated in the same manner.

$$S_R^E = \frac{1-(1-S)T}{1-SR-(1-S)T} - \frac{S}{S}$$
$$S_R^E = \frac{SR}{1-SR-(1-S)T} \qquad (112)$$

$$S_T^E = \frac{1-SR}{1-SR-(1-S)T} - \frac{S}{S}$$
$$S_T^E = \frac{(1-S)T}{1-SR-(1-S)T} \qquad (113)$$

If the resources available for investigating system reliability are limited, they obviously should be applied to the improvement of the probability to which the expected number of successful flights is the most sensitive.

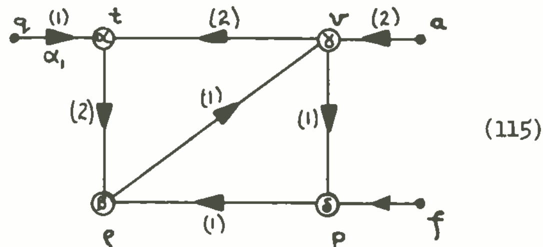
Error Propagation

Linear Perturbation Theory

Because flow graphs show the relationships between large numbers of variables so clearly, relationships between linear perturbations can be readily evaluated by using flow graph techniques. To establish a theoretical basis for the more detailed examples which follow, an aerodynamic system is analyzed. Suppose the following non-linear relationships hold between the static pressure p, the static temperature t, the velocity v, density ρ, the heat input q, the cross-sectional area a and the frictional head loss f of a gas flowing at a critical point in a propulsion system.

$$t = \alpha(q,v)$$
$$\rho = \beta(p,t)$$
$$v = \gamma(\rho,a) \qquad (114)$$
$$p = \delta(v,f)$$

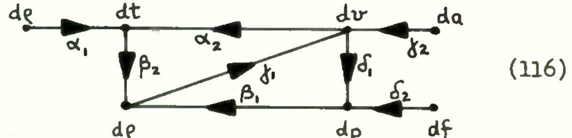
These non-linear equations may be flow graphed as in SFG (115).



(115)

The transmittances (1) and (2) indicate the order in which the variables are substituted into the function at the target node to give the value of the variable associated with that node.

Differentiating Equation (114) gives a set of differential equations which are flow graphed in SFG (116).



(116)

The transmittances of SFG (3) denote partial differentials in the following way: α_1 denotes the partial derivative of the function α with respect to its <u>first</u> argument q; and α_2 denotes the partial derivative of the function α with respect to its <u>second</u> argument v. The flow graph of these differential equations is a linear flow graph; so the transmittances and nodes have their usual significance and can be evaluated by means of the signal transmission formula. For example, the value of $d\rho$ is

$$d\rho = \alpha_1\beta_2 dq/\Delta + (\gamma_2\alpha_2\beta_2 + \gamma_2\delta_1\beta_1)da/\Delta + \delta_2\beta_1 df/\Delta \tag{117}$$

$$\Delta = 1 - \alpha_2\beta_2\gamma_1 - \delta_1\beta_1\gamma_1$$

If the system is subject to small deviations δq, δa, and δt of the input variables, the resulting deviation in ρ can be approximated by substituting the deviations δq, δa, and δt for the differentials in SFG (116).

$$\delta\rho = \alpha_1\beta_2\,\delta q/\Delta + (\gamma_2\alpha_2\beta_2 + \gamma_2\delta_1\beta_1)\,\delta a/\Delta + \delta_2\beta_1\,\delta f/\Delta + O(\delta q, \delta a, \delta f) \tag{118}$$

Thus, SFG (116) can be used to obtain first order approximations of small errors in the system outputs.

Propagation of Variance

If the errors in the system inputs are random variables taken from an ensemble of errors generated by a set of systems, the variances in the system outputs can be used to measure the precision of system performances.

Consider, for example, the system which is discussed in the previous section. That system has three degrees of freedom; that is, the state of the system is determined by the values of three input variables: q,a, and f.

Selecting p, t, and v as system output variables adequately defines the system performance, since the errors in ρ are determined by the errors in p, v, and t. The deviations δp, δv, and δt are linearly related to the deviations in the input variables, as shown in SFG (3). The values of the deviations δp, δv, and δt, which are obtained from SFG (116) may be expressed in the form of a matrix equation.

$$\delta X = \delta E\ A$$
$$\delta E = (\delta q, \delta a, \delta f) \tag{119}$$
$$\delta X = (\delta p, \delta v, \delta t)$$

where A is a 3x3 matrix obtained from SFG (116).

The covariance matrix of the three dimensional joint probability distribution function is $\langle \delta X^T \delta X \rangle$ where the superscript T indicates that δX^T is the transpose of δX; and $\langle\ \rangle$ denote the operation of taking the expected values of the matrix elements.

$$\langle \delta X^T \delta X \rangle = A^T \langle \delta E^T \delta E \rangle A \tag{120}$$

Written out in full, the result is:

$$\langle \delta X^T \delta X \rangle = \begin{Vmatrix} \langle(\delta p)^2\rangle & \langle \delta v \delta p\rangle & \langle \delta t \delta p\rangle \\ \langle \delta p \delta v\rangle & \langle(\delta v)^2\rangle & \langle \delta t \delta v\rangle \\ \langle \delta p \delta t\rangle & \langle \delta v \delta t\rangle & \langle(\delta t)^2\rangle \end{Vmatrix} \tag{121}$$

The expected value of $(\delta p)^2$ is the variance of the variable p; and the expected value of the product $\delta p\ \delta v$ is the covariance of the variables p and v. The probability that the deviations of the output variables lie in the intervals from

p- ⟨p⟩ to p+dp- ⟨p⟩
v- ⟨v⟩ to v+dv- ⟨v⟩
t- ⟨t⟩ to t+dt- ⟨t⟩

is $$\frac{\exp\left(-\frac{1}{2}\delta X^T M^{-1}\,\delta X\right)dp\,dv\,dt}{2\pi^3 |M|^{1/2}}$$

where $|M|$ is the determinant of the covariance matrix, M.

$$M \overset{\text{def}}{=} \langle \delta X^T \delta X \rangle \tag{122}$$

Even if the deviations of the input variables do not have Gaussian distributions, the deviations in the output variables tend to have joint Gaussian distributions by virtue of the central limit theorem.

In summary, flow graphs are a convenient means for evaluating the linear relations between perturbations of input and output variables. The perturbations in the output variables can be used to compute the covariance matrix. The variances which lie along the principal diagonal of the covariance matrix are the squares of the r.m.s. values of the deviations. These variances measure the dispersion of the output variables. However, the probabilities that the deviations will lie within given intervals of uncertainty cannot be determined from the variances alone, because the joint probability distribution of the deviations depends on all of the elements of the covariance matrix and not just the elements alongs its principal diagonal. The joint probability distribution of the output variables often tends to approximate a multi-dimensional Gaussian distribution. A multidimensional Gaussian distribution is completely determined by its mean and its covariance

matrix. Thus, both the intervals of uncertainty and the probability that the system performance variables are within those intervals can usually be adequately approximated by the procedure that has been outlined here. A convenient way to specify the resulting uncertainty of an output variable is as follows:

$$(19 \text{ to } 1) \quad p = (10 \pm 0.5) \text{ psi} \tag{123}$$

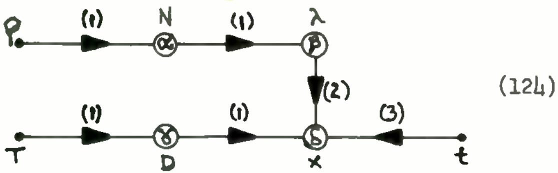
which indicates that the odds are 19 to 1 that the output variable p lies within ± 0.5 of 10 psi.

Diffusion Errors and Junction Depth

In the fabrication of semiconductor devices by diffusion of impurities into solids, the relationship between the process variables t, T, and ρ and the device variable X must be considered.

- t = the duration of the process (sec)
- T = the temperature of the process (°K)
- ρ = the resistivity of material (Ω-cm)
- X = junction depth (cm)

The relationships between these variables is shown by SFG (124).



(124)

- N = the impurity concentration at depth X and time t
- λ = a diffusion parameter which defines surfaces of constant N as a function of depth X and time t.
- D = a diffusion coefficient

The functions α, β, γ, and δ in SFG (124) are defined by the following equations:

$$N = (\rho q\mu)^{-1} = \alpha(\rho) \tag{125}$$

The constant $q\mu$ is the product of the electronic charge q and μ, the mobility of the majority carriers in the solid state material.

$$D = D_o \exp(-Q/RT) = \gamma(T) \tag{126}$$

The constant Q/R is the ratio of the activation energy Q to the universal gas constant R; and the constant D_o is a frequency factor associated with the entropy of activation.

$$\text{erfc}\, \lambda = \frac{N}{Ns} \tag{127}$$

The constant N_s is the concentration of the diffusant on the surface of the material. Equation (125) must be inverted to obtain the diffusion parameter λ as a function β of the impurity concentration N.

$$\lambda = \text{erfc}^{-1} \left(\frac{N}{Ns}\right) = \beta(N) \tag{128}$$

The function δ is defined by Equation (129):

$$X = 2\sqrt{Dt}\,\lambda = \delta(D, \lambda, t) \tag{129}$$

Differentiating these equations gives α_1, β_1, γ_1, δ_1, δ_2, and δ_3.

$$\begin{aligned}
\alpha_1 &= N\frac{d(\ln N)}{d\rho} = -N \\
\beta_1 &= \left(\frac{dN}{d\lambda}\right)^{-1} = \frac{\sqrt{\pi}\, \epsilon^{\lambda^2}}{-2N_s} \\
\gamma_1 &= D\frac{d(\ln D)}{dT} = \frac{DQ}{RT^2} \\
\delta_1 &= X\frac{\partial(\ln X)}{\partial D} = \frac{X}{2} \\
\delta_2 &= \frac{X\,\partial(\ln X)}{\partial\lambda} = X \\
\delta_3 &= X\frac{\partial(\ln X)}{\partial t} = \frac{X}{2}
\end{aligned} \tag{130}$$

From the SFG (124) it is clear that δx is given by Equation (131).

$$\begin{aligned}
\delta X &= (\gamma_1 \delta_1)\,\delta T + (\alpha_1\beta_1\delta_2)\,\delta\rho + (\delta_3)\,\delta t \\
\delta X &= \left(\frac{DQX}{RT^2}\right)\delta T + \left(\frac{N\sqrt{\pi}\,\epsilon^{\lambda^2} X}{2N_s}\right)\delta\rho + \left(\frac{X}{2}\right)\delta t
\end{aligned} \tag{131}$$

If δT, $\delta\rho$, and δt are uncorrelated random variables the variance of X is

$$\begin{aligned}
\langle(\delta X)^2\rangle = {} & \left(\frac{DQX}{RT^2}\right)^2 \langle(\delta T)^2\rangle + \\
& \left(\frac{N\sqrt{\pi}\,e^{\lambda^2} X}{2N_s}\right)^2 \langle(\delta\rho)^2\rangle + \\
& \left(\frac{X}{2}\right)^2 \langle(\delta t)^2\rangle
\end{aligned} \tag{132}$$

The square root of the variance of X is the standard deviation of X, $\sigma(X)$.

$$\sigma(X) = \sqrt{\langle(\delta X)^2\rangle} \tag{133}$$

If X has a Gaussian distribution, the probability that X lies within the interval $\langle X\rangle \pm 2\sigma(X)$ is approximately 95%. Thus, the uncertainty in X could be expressed as

$$(19 \text{ to } 1) \quad \langle X\rangle \pm 2\,\sigma(X) \text{ cm}$$

Replacing all of the one-σ deviations in Equation (132) by n-σ deviations does not change the validity of that equation, since it is linear in the variances.

$$[n\ \sigma(X)]^2 = \left(\frac{DQX}{RT^2}\right)^2 [n\ \sigma(T)]^2 + \left(\frac{N\ \sqrt{\pi}\ e^{\lambda^2}\ X}{2Ns}\right)^2 [n\sigma(\rho)]^2 + \left(\frac{X}{2}\right)^2 [n\sigma(t)]^2 \qquad (134)$$

Thus, the width of the 2σ interval for the junction depth X could be computed by setting n equal to 2 in Equation (134). If the process variables have Gaussian distributions, the odds that they will lie within their 2σ intervals will also be 19 to 1.

Information such as this can be used as a rule of thumb for (1) estimating the variances of process variables; or (2) selecting an interval of uncertainty such that the odds of being within that interval are acceptable for a given process.

Summary and Discussion

The method presented here for determining the variances of system (or process) outputs in terms of the variances of system inputs is only approximate. Even if linear perturbation theory gave the exact relations between deviations, only probability distributions which can be completely specified in terms of their second order moments can be treated with mathematical precision by the "propagation of variance" techniques which have been used in this example. However, if the system inputs are continuous random variables with continuous first order derivatives, the flow graph techniques used here can be used to find the Jacobian of the transformation which relates the distribution of the system outputs to the distribution of the system inputs. Once the Jacobian has been determined, the distribution of the system outputs can be computed by standard change of variable techniques of integral calculus.

If the distributions of the system inputs are discrete, the distributions of the system outputs can be calculated by basic probabilistic algebra, which can be flow graphed as shown in the first part of this paper.

Eccentricity Error for Circular Orbits

Limitations of Linear Perturbation Theory

As an example of a significant problem to which linear perturbation theory cannot be applied, consider the following problem. The eccentricity e of the Keplerian orbit illustrated in Figure 6 is related to the injection coordinates x and $\emptyset$ by Equation (135).

$$e^2 = x^2(1-\emptyset^2) + \emptyset^2 \qquad (135)$$

The coordinates x and $\emptyset$ are defined as follows:

$\emptyset$ = the flight path angle

$$x = \left(\frac{r_i}{r_c}\right)^2\left(\frac{v_i}{v_c}\right) - 1 \qquad (136)$$

The subscripts i and c refer to the injection and the circular orbit respectively. The r's denote orbital radii; and the v's denote orbital velocities.

The problem is to determine the r.m.s. value of the perturbations in the eccentricity e of nearly circular injection, assuming that the perturbations in the injection coordinates v_i, r_i, and $\emptyset$ are independently and normally distributed about the mean values v_c, r_c, and zero.

For such small values of $\emptyset$, Equation (135) is closely approximated by Equation (137).

$$e^2 = x^2 + \emptyset^2 \qquad (137)$$

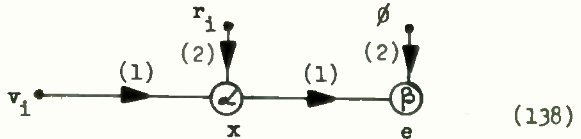
Linear perturbation theory cannot relate the perturbations in e to the perturbations in x and $\emptyset$ because any linear theory takes into account only the first two terms of the Taylor series of a variable and neglects all of the higher order terms. A Taylor expansion of x^2 at

$$x = 0$$

shows that the first two terms of the resulting series are zero. Hence, the third term of the series, namely x^2, cannot be neglected - in fact, it is the only non-zero term in the series.

Flow Graph Solution

The relationships between the eccesntricity e and the injection coordinates v,r,x, and $\emptyset$ are given by SFG (138).



(138)

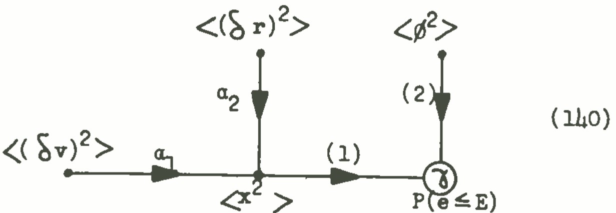
The definitions of the function α and β follow directly from Equations (135) and (136).

$$x = \left(\frac{v_i}{v_c}\right)^2\left(\frac{r_i}{r_c}\right) - 1 = \alpha(v_i, r_i)$$

$$e = (x^2 + \emptyset^2)^{1/2} = \beta(x, \emptyset) \qquad (139)$$

Although the perturbations in e are not linearly related to the perturbations in x and $\emptyset$, the perturbation in x is linearly related to the perturbations in v_i and r_i. Signal flow graph (140) gives the relationships between the var-

iances of the signals v_i, r_i, x and ϕ in SFG (138) to the probability $P(e \leq E)$ that the eccentricity e will be within a specified constant magnitude E.



(140)

The transmittances α_1 and α_2 may be computed as shown in Equation (141).

$$\alpha_1 = \frac{\partial x}{\partial v_i} = \left(\frac{2v_i}{v_c^2}\right)\left(\frac{r_i}{r_c}\right)\Bigg|_{\substack{v_i=v_c \\ r_i=r_c}} = \frac{2}{v_c}$$

$$\alpha_2 = \frac{\partial x}{\partial r_i} = \left(\frac{v_i}{v_c}\right)^2\left(\frac{1}{r_c}\right)\Bigg|_{\substack{v_i=v_c \\ r_i=r_c}} = \frac{1}{r_c} \tag{141}$$

The non-linear function γ is defined by Equation (142).

$$\gamma(\langle x^2\rangle, \langle \phi^2\rangle) = \iint_E \left(\frac{1}{2\pi}\right)^3 (\langle x^2\rangle \langle \phi^2\rangle)^{-1/2} \times \exp\left\{-\frac{x^2}{2\langle x^2\rangle} - \frac{\phi^2}{2\langle \phi^2\rangle}\right\} dx\, d\phi \tag{142}$$

The Equation (142) integral cannot be evaluated analytically by any elementary means, but has been calculated by computers. For convenience the data are plotted in Figure 7 using the following notation:

$$\begin{aligned} s^2(x) &= \langle x^2\rangle \\ s^2(r) &= \langle r^2\rangle \\ s^2(v) &= \langle v^2\rangle \\ s^2(\phi) &= \langle \phi^2\rangle \end{aligned} \tag{143}$$

Numerical Example

Given the data in Table 6 the probability that s(e) does not exceed 0.01 is obtained from SFG (140):

$$s^2(x) = s^2(r)/r_c^2 + 4s^2(v)/v_c^2 = (.0036)^2$$

$$\text{hence} \quad s(\phi)/s(x) = 0.5 \tag{144}$$

Figure 7 yields for E = 0.01 the desired result

$$P(e \leq E) = .997 \tag{145}$$

Dynamic Programming

Continuous Markhov Processes

Let $p(j|i)$ denote the conditional probability that a system will be in state j at time n+1 given that the system is in state (i) at time n. If the operation of a system is governed by transition probabilities such as $p(j|i)$, the operation of that system is a Markhov process. The equations of such a system are of the form

$$p(j;n+1) = \sum_i p(i;n)\, p(j|i) \tag{146}$$

The state probability $p(i;n)$ is the probability that the system is in state i at time n. If the state probabilities are continuous functions of time, Equation (146) becomes

$$p(j;t+dt) = \sum_i p(i;t)\, p(j|i) \tag{147}$$

The rate of change of the state probabilities in Equation (147) is given by Equation (148).

$$\frac{dp(j;t)}{dt} = \sum_i p(i;t) \lim_{dt \to 0} (p(j|i) - \delta_{ij}) \tag{148}$$

where $\delta_{ij} = 0$ for $i \neq j$

$\delta_{ij} = 1$ for $i = j$

Assuming that the limit in the above equation is a constant, say a_{ij}, Equation (148) reduces to Equation (149).

$$\frac{dp(j)}{dt} = \sum_i p(i)\, a_{ij} \tag{149}$$

where the t's have been omitted from $p(j;t)$ and $p(i;t)$ to simplify notation.

Markhov Processes with Rewards

Suppose that (1) the system earns a reward $q_i dt$ if it occupies state (i) for a duration of time dt; and (2) that $V_i(t)$ is the expected total reward at time t if the process starts in state (i). From these definitions it follows that

$$V_i(t+dt) = q_i dt + \sum_j p(j|i)\, V_j(t) \tag{150}$$

and that

$$\frac{dV_i}{dt} = q_i + \sum_j a_{ij} V_j \tag{151}$$

For large t, after the system has reached a steady state of operation, the expected total reward can be expressed as

$$V_i(t) = v_i + gt \tag{152}$$

where v_i is the expected total reward for the duration of transient operation, and g is average reward per unit time gained in steady state operation.

Substituting Equation (152) into Equation (151) gives

$$g = q_i + \sum_j a_{ij}v_j + gt\sum_j a_{ij} \tag{153}$$

The sum of the $p(j|i)$ over all j is equal to one, because the system must move into one of the possible states.

$$\sum_j a_{ij} = \sum_j [p(j|i) - \delta_{ij}] = 0 \tag{154}$$

This makes the last term of Equation (153) equal to zero, so that Equation (153) reduces to Equation (155).

$$g = q_i + \sum_j a_{ij}v_j \tag{155}$$

Policy Improvement

Suppose that one may select in each state (i) of a system S a number of alternatives which result in: (1) different earning rates q_i; and (2) different transition probability densities a_{ij}. A policy consists of a set of alternatives which contains one alternative for each state.

An optimum policy would be the one which results in the maximum steady state (rate of) gain g. As an illustration of how an optimum policy may be determined, consider the following example.

System S can be in either of two states: (w) working; or (d) down. The system can be assembled with cheap (less reliable) components, or expensive (more reliable) components. The alternative (c,w) of using cheap components results in

(1) an earning rate of \$4/hr while the system is in state (w); and
(2) a probability 2dt of failing in a time interval dt.

The alternative (e,w) of using expensive components results in :

(1) a lower earning rate of \$2/hr while the system is in state (w); and
(2) a probability of 1 dt of failing in a time interval dt.

These earning rates and probability densities apply to the working state (w). In the down state (d) two alternatives exist:

(d,c) the replacement of failed components with cheap components; and
(d,e) the replacement of failed components with expensive components.

Using cheap replacements results in:

(1) replacement costs of \$2 per unit of down time, and
(2) the probability 2dt of completing the replacement in a time interval dt.

Using expensive components results in:

(1) replacement costs of \$3 per unit of down time and
(2) the probability 3dt of completing the replacement in the time interval dt.

The problem is to decide which alternatives result in the maximum system effectiveness, measured in terms of g. Earning rates q_i and transition probability densities a_{ij} which are associated with each alternative are listed in Table 7.

To obtain alternatives included in the policy calculate g from Equation (155) and Table 6.

$$g = q_w + a_{ww}v_w + a_{wd}v_d$$
$$g = q_d - a_{dw}v_w + a_{dd}v_d \tag{156}$$

By virtue of Equation (154), adding a constant to the transient value of each state does not change g. For example, adding the negative of v_d to the transient value of each state in Equation (156) gives

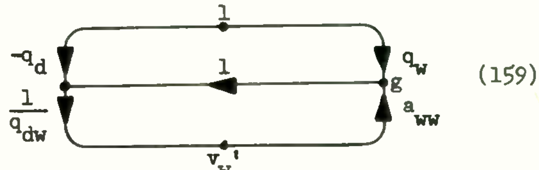
$$g = q_w + a_{ww}v_w'$$
$$g = q_d + a_{dw}v_w' \tag{157}$$

where v_w' is defined by

$$v_w' = v_w - v_d \tag{158}$$

The quantity v_w' denotes the transient value of state (w) relative to the reference state (d); and is referred to as the relative transient value of state (w).

Flow graphing Equations (157) gives SFG (159).



(159)

An optimal policy with a maximum gain can be found by the following procedure:

(1) Select an arbitrary policy and compute the relative transient values, i.e. the v_j', for that policy.

(2) Using the v_j' obtained in step (1) compute the relative gain g(i,p) for each alternative (i,p).

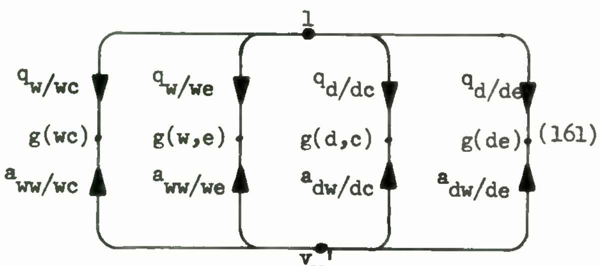
$$g(i,p) \overset{\text{def.}}{=} q_i + \sum_j a_{ij}v_j'$$

(3) Select as a policy a set of alternatives which maximize the relative gain $g(i,p)$ for each state (i).

(4) If the policy obtained in step (3) is the same as the policy in step (1), it is an optimal policy.

(5) If the policy obtained in step (3) is different from the policy used in step (1), repeat step (1) using the new policy.

Continue this procedure until an optimal procedure has been found. Step (1) corresponds to SFG (159). The values of the transmittances used in SFG (159) depend on the policy that is being used in step (1). The SFG corresponding to step (2) has the form of SFG (161).

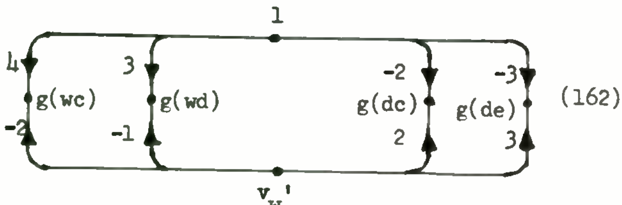


where

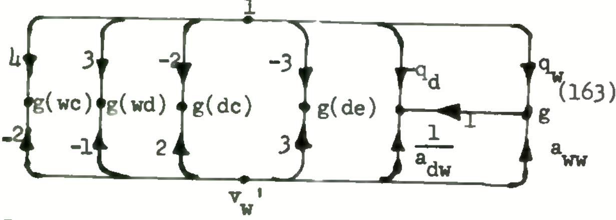
$q_{i/ip}$ = the earning rate of state (i) for alternative (i,p);

$a_{ij/ip}$ = the transition probability density from state (i) to state (j) for alternative (i,p)

The transmittances contained in SFG (161) pertain to every alternative. Consequently, the transmittances of SFG (161) always have the same values. Using the values from Table 7 for the transmittances of SFG (161) gives SFG (162).

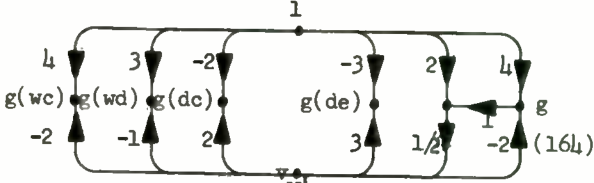


Steps (1) and (2) can be combined by joining SFG's (159) and (162) to obtain SFG (163).



Repetitive use of SFG (163) always yields an optimal policy.

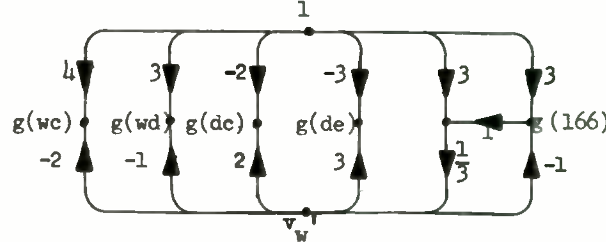
For example, consider as an initial policy the set of the alternatives (w,c) and (d,c). The values of the transmittances in SFG (163) which correspond to this policy may be otained from Table 7. Using these values in SFG (163) gives SFG (164).



Signal flow graph (164) gives the following results for the initial policy.

$$
\begin{aligned}
g &= 1 \\
v_w &= 1.5 \\
g(wc) &= 1 \\
g(we) &= 1.5 \\
g(dc) &= 1 \\
g(de) &= 1.5
\end{aligned}
\qquad (165)
$$

The greatest relative gain for states (w) and (d) respectively is $g'(we)$ and $g'(de)$. The policy consisting of alternatives (we) and (de) differs from the initial policy. Using the transmittances for this new policy in SFG (163) gives SFG (166).



For the new policy, SFG (166) gives

$$
\begin{aligned}
g &= 1.5 \\
v_w' &= 1.5
\end{aligned}
\qquad (167)
$$

Since this value of v_w' is the same as that obtained from SFG (164), the relative gains in SFG (166) will be the same as in SFG (164). In other words, the policy obtained from SFG (166) will be the same as the previous policy. Therefore, the policy which consists of alternatives (w,e) and (d,e) will yield the maximum effectiveness for this system.

Discussion

Dynamic programming is a method optimizing a sequence of decisions so as to maximize the "pay off" of a multistage decision process. The problem which was used as an example here involved a simple, continuous, non-deterministic process. The flow graph techniques which were used here can, however, be readily extended to discrete or deterministic processes.

Conclusions

This investigation was made to determine if the non-linear, probabilistic and decision making relationships which occur in reliability engineering problems could be specified in a form suited for signal flow graph representation and evaluation.

Flow graphs provide an effective technique to organize relationships between the variables of a system. An important task of reliability engineering is to establish and to evaluate mathematical models of processes and systems. These models must reflect realistically the functional relationship between the significant variables of the system. Therefore the first task of this investigation was to ascertain the feasibility of representing system models useful in reliability by flow graphs. The scope and limitations of applying flow graphs to reliability were examined using examples taken from current engineering practice.

No attempt was made to improve solution to existing reliability problems, rather to use known solutions to problems and present these by a visual, orderly, systematic and hence more effective display. Nevertheless, novel and interesting results were discovered.

As a result of formulating and solving a wide variety of reliability problems by flow graphs, it is concluded that advantages of flow graph techniques are reflected in the following facts:

(1) A technique for representing logical sums in terms of algebraic sums permits the basic laws of probability to be readily represented in flow graph form.

(2) Conditional probabilities were used to formulate relationships between operating and failure modes of multimode systems.

(3) A state point enables the movements of a Markhov system from state to state to be followed by tracing the path of the state point as it moves from node to node.

(4) Redundant systems were formulated by using convolution integrals to evaluate time-dependent failures. The use of convolution integrals is not limited to ensembles of components with exponentially distributed reliability functions.

(5) Flow graph techniques provide a systematic procedure for analyzing linear error propagation in systems with interdependent errors. Techniques for formulating both linear and non-linear perturbation problems were developed and illustrated by examples.

(6) Flow graph techniques applied to the optimization of multistage processes clarified the details of an iterative optimizing procedure. Furthermore, all of the steps involved in each iteration were combined into one signal flow graph, so that each iteration could be performed in just one step.

In summary, it was demonstrated that flow graphs provide a versatile and effective technique to the reliability engineer. Specific advantages are:

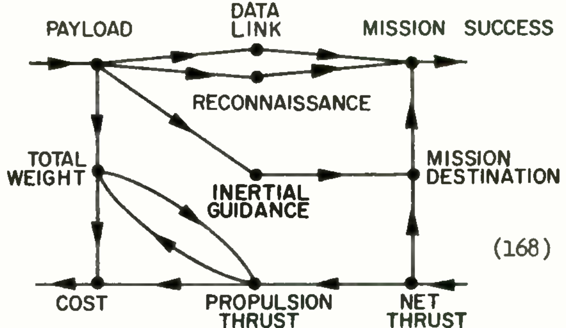
(1) The orderly formulation, analysis, and interpretation of system relationships is provided.

(2) The formulation of models of systems which involve linear, non-linear, probabilistic, and decision making features is simplified.

(3) Models of complex systems can be constructed by interconnecting signal flow graphs of subsystems.

Many aspects of the problem remain to be investigated, such as the following tasks:

(1) Investigate systems of increasing complexity, applying the models here presented as building blocks. SFG (168) indicates the scope of this avenue of approach.



(168)

(2) Develop a set of basic patterns of reliability models in addition to those here presented.

(3) Establish inter-disciplinary communication techniques.

(4) Provide numerical techniques for the evaluation of complex flow graphs by computer routines.

Literature Survey

The number of publications on flow graphs has increased from about 20 papers in 1955 to over 200 by 1960. It is convenient to examine this material from several points of view; namely, whether these flow graph investigations are of interest primarily (1) to the theoretical mathematician, (2) to the application-minded engineer, or (3) to the teacher. These categories are not mutually exclusive, but they are functional.

Source Material on Reliability Engineering

In applying new techniques to reliability engineering, it is desirable to illustrate solutions of known reliability problems. This survey, therefore, makes no claim to have solved new problems in the field of reliability. Examples from a large number of investigations were used to illustrate flow graph techniques. Hence grateful acknowledgement is made to the following sources for specific practical examples as well as basic theory:

Time dependent failures[1,2]

Open and short circuit failures[3,4]

Multimode systems[5]

Markhov Processes[6,7]

Error propagation and Jacobians[8-11, 108-110]

Dynamic programming[12,13]

Special flow graph techniques[6,14,15]

A bibliography on reliability engineering listing about 200 references gives a comprehensive survey of the field up to 1961[16].

Tutorial References

A bibliography[17] listing abstracts of 83 references reflects publications related to flow graphs up to 1960; this literature survey draws heavily from it. A monograph on flow graphs is in press at present[18]. The first engineering textbook to contain an account of flow graph theory appeared in 1955[19].

A number of engineering textbooks have since appeared which summarize flow graph theory. Several textbooks have developed important new material, for example, the application of flow graphs to the following fields:

Communication networks[14,107]

Operation research and system engineering[20]

Network analysis and synthesis[21,22]

System analysis[23,24,25,89,111]

There is also a steady flow of tutorial papers on flow graphs[26-33]. Standards on flow graph terminology have been proposed by the Institute of Radio Engineers[34,35].

Mathematical References

Publications concerned primarily with the mathematical aspects of flow graphs are grouped as follows:

Review of graph theory[36]

Flow graph theory based on linear equations[15,28,37-45]

Flow graph theory based on matrix methods[46-49]

Flow graph theory based on topological methods[50-53,106]

Probabilistic concepts interpreted by flow graphs[54-58]

Digital concepts evaluated by flow graphs[59,60]

Numerical techniques applied to flow graphs[37,61-65]

Non-linear systems approached by flow graphs[66]

References on Engineering Applications

Interest in applying flow graphs to engineering problems has developed primarily in the following fields:

Network analysis[32,67-71,103 105]

Electronic circuits[42,54,72-76,102]

Physical processes and physical systems[74-84]

Operational research[20,85-88]

Instrumentation[74,90-95]

Mechanical engineering[96-98]

Sampled data analysis[99-101]

References

1. J. Christian and K. W. Hollander, "Reliability of Switching Mechanisms," RADC TR 60-239, January 1961.

2. F. Lemus, "Reliability Evaluation of a Satellite Power Supply System," Missiles Systems Design, March-April 1961, pp. 7-10.

3. John R. Hanne, "Optimizing Simple Circuitry for Reliability and Performance by Failure Mode," presented at the Pacific Coast General Meeting of AIEE, August 8-12, 1960, San Diego, California.

4. R. H. Myers, "Predicting Reliability of Satellite Systems," presented at the Northeast Electronics Research and Engineering Meeting of the AIEE and IRE, November 14-16, 1961, Boston, Massachusetts.

5. R. I. Zagor, K. Curtin, and H. Greenberg, "Reliability of Multimoded Systems," Electronic Industries, April 1958.

6. C. S. Lorens, "Theory and Applications of Flow Graphs," Sc.D. Dissertation, Massachusetts Institute of Technology, July 1956.

7. G. G. Den Broeder, Jr. and W. E. Kane, "A Reliability Analysis of Recoverable Missiles," IRE Trans. on Reliability and Quality Control, June, 1959.

8. Harold Cramer, "Mathematical Methods of Statistics," Princeton University Press, 1958.

9. H. A. Thurston, "Partial Differentiation," Prentice Hall, 1961.

10. Joseph Edwards, "Treatise on the Integral Calculus," Chelsea Publishing Company, 1954.

11. J. E. Reynolds, "A Treatment of Diffusion Errors Affecting Junction Depth," IRE Trans., PGED-8, No. 5, pp. 377-380, September, 1961.

12. R. A. Howard, "Dynamic Programming and Markhov Processes," MIT Technology Press, 1960.

13. Operations Research Center, M.I.T., "Notes on Operations Research," M.I.T. Technology Press, 1959.

14. Samuel J. Mason and Henry J. Zimmermann, "Electronic Circuits, Signals, and Systems," John Wiley and Sons, Inc., 1960.

15. **J. L. Burroughs and W. W. Happ, "Signal Flow Graph Techniques," Lockheed Missiles & Space Company, Tech. Report No. 288959, Oct. 1959.**

16. H. E. Blanton and R. M. Jacobs, "A Survey of Techniques for Analysis and Prediction of Equipment Reliability," NEREM Convention Record, 1961.

17. **G. R. Evans and J. L. Burroughs, "The Theory and Applications of Signal Flow Graphs - An Annotated Bibliography," Lockheed Missiles & Space Company, SRB-60-10, 1959.**

18. Y. Chin and E. Cassignol, "Linear Signal Flow Graphs and Applications," Wiley, 1962, (in press).

19. J. G. Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill, 1955.

20. C. D. Flagle, W. H. Huggins, and R. H. Roy, "Operations Research and Systems Engineering," Johns Hopkins University Press, Baltimore, Maryland, Chapter 21: "Flow graph representation of systems," by W. H. Huggins, 1960.

21. S. Seshu and N. Balabanian, "Linear Network Analysis," Wiley, 1959.

22. S. Seshu and M. B. Reed, "Linear Graphs and Electrical Networks," Addison-Wesley, 1961.

23. D. K. Cheng, "Analysis of Linear Systems," Addison-Wesley, 1959.

24. P. E. Pfieffer, "Linear Systems Analysis," McGraw-Hill, 1961.

25. J. G. Truxal, Editor, "Control Engineers' Handbook," McGraw-Hill, 1958.

26. M. Boisvert, "Signal Flow Graphs," Ann. Telecommun, Vol. 13, No. 3,4, pp. 50-77, March-April 1958.

27. W. R. Abbott, "The Application of Systems Theory to Space Missions," Proceedings, Symposium on the Application of Switching Theory in Space Technology, Stanford University Press, 1962.

28. T. R. Nisbet and W. W. Happ, "Flow Graph Analysis - Visual Engineering Mathematics," Electronic Design, 9 and 23 December 1959 and 6 and 20 January 1960.

29. T. M. Stout, "Block Diagram Solution for Vacuum Tube Circuits," AIEE Trans., Paper 53-269, 1953.

30. **M. Tuero, "The Method of Signal Flow Graphs," Slaboproudy Obzor, v. 20, No. 10, pp. 608-15, 1959. (In Czech) (Electrical Engineering Abstracts, V. 63, No. 747, p. 146, Mar. 1960.)**

31. **S. Barabaschi and E. Gatti, "Modern Methods of Analysis of Electrical Linear Active Networks with Particular Regard to Feedback Systems," Energia Nucleare, V. 2, No. 12, pp. 105-19, 15 Dec. 1954. (In Italian) (Science Abstracts, Section B, Electrical Engineering, V. 58, P. 391, 1955.)**

32. C. F. Simone, "Equivalent Ladder Networks by the Use of Signal Flow Graphs," IRE Trans. on Circuit Theory, pp. 75-81, 1959.

33. G. V. Woodley, "The Flow Graph - A Short Cut to Network Simulation," Electronic Design, January 1959.

34. Institute of Radio Engineers, Technical Committee 4, Proposed Standard Definitions of Terms for Linear Signal Flow Graphs, April 1958.

35. Institute of Radio Engineers, "IRE Standards on Circuits: Definitions of Terms for Linear Signal-Flow Graphs, 1960," Proc. IRE, pp. 1611-1612, September 1960.

36. D. Konig, "Theorie der Endlichen und Unendlichen Graphen," Leipzig, 1936; reprinted New York 1950.

37. C. E. Shannon, "The Theory and Design of Linear Differential Equation Machines," Privately printed, January 1942, under the auspices of Section D-2 (Fire Control) of the National Defense Research Committee.

38. S. J. Mason, "Feedback Theory - Further Properties of Signal Flow Graphs," Proc. IRE, Vol. 44, No. 7, July 1956 and Technical Report 303, Research Laboratory of Electronics, M.I.T., 20 July 1955.

39. S. J. Mason, "Feedback Theory - Some Properties of Signal Flow Graphs," Proc. IRE, Vol. 41, No. 9, p. 1144, September 1953 and Technical Report 153, Research Laboratory of Electronics, MIT, 2 February 1953.

40. S. J. Mason, "On the Logic of Feedback," Sc.D. Thesis, MIT, 1952.

41. C. A. Desoer, "The Optimum Formula for the Gain of a Flow Graph or a Simple Derivation of Coates' Formula," IRE Proc., p. 883, 1960.

42. W. W. Happ, "Dynamic Characteristics of Four-Terminal Networks," IRE Convention Record, Part 2, Circuit Theory, pp. 60-76, Mar. 1954.

43. C. S. Lorens, "Basic Theory of Flow Graphs - Inversion," Quarterly Progress Report, Research Laboratory of Electronics, MIT, p. 64, 15 April 1956.

44. C. S. Lorens, "A Proof of the Nonintersecting Loop Rule for the Solution of Linear Equations by Flow Graphs," Quarterly Progress Report, Research Laboratory of Electronics, MIT, p. 97, 15 January 1956.

45. E. B. Stear and A. R. Stubberud, "Nonstationary Signal Flow Graph Theory," Report #60-64, November 1960, University of California, Los Angeles, California.

46. E. W. Barankin, "Procedence Matrices," Univ. of Calif, Management Sciences Research Proj., Report No. 26, December 1953.

47. F. Harary, "A Graph Theoretic Method for the Complete Reduction of a Matrix with a View Toward Finding its Eigenvalues," J. Math. Phys., Vol. 38, P. 104-111, 1959.

48. F. Harary, "On the Consistency of Precedence Matrices," J. Assoc. Comp. Mach., Vol. 7, No. 3, pp. 255-259, July 1960.

49. R. B. Marimont, "A New Method of Checking the Consistency of Precedence Matrices," J. Assoc. Comp. Mach., Vol. 6, pp. 164-171, 1959.

50. R. Ash, "Topology and the Solution of Linear Systems," Electronics Research Labs., Columbia Univer. Report No. CU48-58-AF-677-EE, AFOSR TN-58-726, 15 June 1958, 18 pp., ASTIA Report No. AD-162 261.

51. R. B. Ash, "Topology and the Solution of Linear Systems, "Franklin Inst. Journal, pp. 453-463, December 1959.

52. W. W. Happ, "Signal Flow Graphs," IRE Proc., Vol. 45, No. 9, pp. 1293, 1957.

53. F. Harary, "The Number of Functional Digraphs," Math. Ann., Vol. 138, pp. 203-11, 1959.

54. R. F. Hoskins, "Signal Flow Graphs. Application to Linear Circuit Analysis," Electronic & Radio Engineer," Aug. 1959.

55. W. H. Huggins, "Signal Theory," IRE Transactions on Circuit Theory," pp. 210-16, Dec. 1956.

56. W. H. Huggins, "Representation and Analysis of Signals. Part V. Flow Graph Representation of Systems and System Dynamics," Johns Hopkins Univ., AFCRC TN 59-971, 15 Nov. 1959, pp. 95.

57. C. S. Lorens, "Statistical Application of Flow Graphs, Mean Occurrence of a Markov State," Quarterly Progress Rept., Research Laboratory of Electronics, MIT, 15 July 1956.

58. C. S. Lorens, "Stationary State Distribution of Discrete Markov Systems," Quarterly Progress Rept., Research Laboratory of Electronics, MIT, pp. 60, 15 April 1956.

59. D. D. Aufenkamp, "Analysis of Sequential Machines," IRE Trans., PGEC-7, 4, pp. 299-306, Dec. 1958.

60. F. E. Hohn, S. Seshu and D. D. Aufenkamp, "The Theory of Nets," IRE Trans., pp 154-161, Sept. 1957.

61. G. S. Axelby and R. H. Plath, "Practical Application of Signal Flow Graphs Using an Array of Non-Touching Loops," Technical Memo, Westinghouse Electrical Corp., Baltimore, Md., 1960.

62. A. Gill, "Analysis of Nets by Numerical Methods," J. Assoc. Comp. Mach., Vol. 7, 3, pp 251-54, July 1960.

63. W. W. Happ and J. L. Burroughs, "Formulation of the Signal Flow Graph of a Complex Servo-System," Pacific General Meeting, AIEE, San Diego, Calif., 8-12 Aug. 1960.

64. S. Seshu and N. Balabanian, "Linear Network Analysis," N.Y., John Wiley & Sons, Inc., London.

65. W. K. Linvill, "System Theory as an extension of Circuit Theory," IRE Transactions on Circuit Theory, pp. 217-222, Dec. 1956.

66. T. A. Bickart, "Flowgraphs for Nonlinear Systems," Radiation Lab., Johns Hopkins Univ., Technical Report No. AF-76, May 1960.

67. C. S. Lorens, "Formulation of Cascade Circuits in Terms of Flow Graphs," Quarterly Progress Rept., Research Laboratory of Electronics, MIT, 15 July 1956.

68. W. A. Lynch, "A Formulation of the Sensitivity Function," Institute of Radio Engineers, Transactions, CT-4, 3, pp. 289, Sept. 1957.

69. L. P. A. Robichaud, Jr., "Flow Graph Theory of Quadripoles," Universite Laval, Dept. Genie Elect., Rapport de Rech. #8, 1956.

70. O. Wing, "Ladder Network Analysis by Signal-flow Graph--Application to Analog Computer Programming. IRE Trans. on Circuit Theory, pp. 289-294, Dec. 1956.

71. T. M. Stout, "A Block Diagram Approach to Network Analysis," AIEE Trans., Vol. 71, Part II, pp. 255-60, 1952.

72. S. R. Hawkins and W. W. Happ, "Radiation Stabilization of Transistor Circuits by Active Feedback," Lockheed Aircraft Corp., Missiles & Space Div., LMSD No. 5031, 1958.

73. P. Kaufmann and J. J. Klein, "Flow Graph Analysis of Transistor Feedback Networks," Semiconductor Products, pp. 37-40, Oct. 1959.

74. W. M. Locke, V. W. Moore Jr., and W. W. Happ, "The Function of the Transistor in a Phase-Shift Oscillator," Lockheed Aircraft Corp., Missiles & Space Div., LMSD-288124, May 1960.

75. C. S. Lorens, "Applications of Flow Graphs to Transistor Circuits," General Electric Technical Information Series No. DF55ELP168, 20 Dec. 1955.

76. A. B. Vanrennes, "Signal-flow Graphs and Stability Analysis of Nuclear Reactors." IRE Trans. on Nuclear Science, pp. 1-5, Mar. 1960.

77. J. L. Burroughs, "Signal Flow Graph Analysis of Physical Systems." Proceedings, Montana Academy of Sciences, May 1958.

78. W. W. Happ, "Analysis of Ferromagnetic Resonance in Thin Films by Topological Methods," presented at Joint Meeting of AIEE and IRE Professional Group on Microwave Theory and Techniques, Palo Alto, Calif., 15 Jan. 1960.

79. W. W. Happ, "A New Teaching Aid: Flow Graphs," Lockheed Aircraft Corp., Missiles & Space Div., LMSD 2418, 1958.

80. C. S. Lorens, "Theory and Applications of Flow Graphs," Research Laboratory of Electronics, MIT, Report No. TR-317, 15 July 1956.

81. L.P.A. Robichaud, Jr., "A Study of Physical Systems Through Signal Flow Graphs. Part II: Direct Analysis of Electrical Networks." Canadian Armament Research & Development Establishment. CARDE Technical Memo No. 167/57, 1 volume, part 2, Jan. 1959. (Encl. to Air Attache Ottawa, Rept. No. TL-78-59).

82. L. P. A. Robichaud, Jr., "A Study of Physical Systems through Signal Flow Graphs, Part 4, Direct Simulation on Analog Computers," Canadian Armament Research & Development Establishment. CARDE Tech. Memo No. 167-57, Dec. 1957, 32 pp. (Encl. to Air Attache Ottawa, Rept. No. TL-24-58.

83. W. W. Happ, "A Systematic Treatment of Error Propagation for Design Evaluation and Process Control of Electronic Devices," Technical Report, Lockheed Missiles & Space Company, LMSC 3-38-61-7, 1961.

84. W. W. Happ, "Flow Graphs - A Topological Approach to Physics Problems," Technical Report, Lockheed Missiles & Space Company, LMSC 288058, 1959.

85. F. Harary, "Graph Theoretical Methods in the Management Sciences," Management Science, Vol. 5, pp. 387-403, 1959.

86. F. Harary and Z. Norman, "Graph Theory as a Mathematical Model in Social Science. Ann Arbor, Institute for Social Research, 1953.

87. W. W. Happ, "Mission Success and Cost Analysis by Flow Graphs," Technical Report, Lockheed Missiles & Space Company, 1961, LMSC 3-38-61-7

88. J. L. Burroughs, "Signal Flow Graph Techniques for Analyzing Satellite Mission Effectiveness," Technical Report, Lockheed Missiles & Space Company, 1959, LMSC 3-38-61-1.

89. E. Mishkin and L. Brown, Adaptive Control Systems, McGray-Hill, 1961.

90. J.M.L. Janssen, L. Ensing, J. B. Van Erp, "A Constant-temperature-operation Hot-wire Anemometer," Proc. of the IRE, pp. 555-67, 1959.

91. S. B. Dunn and G. F. Ross, "Signal Flow and Scattering Techniques in Microwave Network Analysis," Sperry Engineering Review, pp. 10, Dec. 1959.

92. J. K. Hunton, "Analysis of Microwave Measurement Techniques by Means of Signal Flow Graphs," Applications Note #38, Hewlett-Packard Co., Palo Alto, Calif., 1960, also as "Signal Flow Graphs Simplify Microwave Measurements," Electronic Design, Aug. 1960.

93. S. B. Dunn and G. F. Ross, "Signal Flow and Scattering Techniques in Microwave Network Analysis," Sperry Engineering Review, V. 12, #6, 10-22, Dec. 1959.

94. Alfred Krausz and H. A. Kahle, "Flow Graphs Applied," - Voltage Regulation in Modern Aircraft Electric Power Systems, AIEE Trans., Applications and Industry, July 1960.

95. P. K. C. Wang, "Mathematical Models for Time Domain Design of Electrohydraulic Servomechanisms," AIEE Trans. Applic. and Industry, Nov. 1961, pp. 252-259.

96. D. Broadbent and H. R. Chinn, "The Application of Signal Flow Diagrams to Interconnected Electrical Power Systems," Aus. J. Appl. Sc., 9, pp. 353-359.

97. H. Putterman, "Signal Flow Graph Techniques and its Applications," Sperry Engineering Review, Vol. 10, #6, p. 8-15, Nov. 1957.

98. J. L. Burroughs and W. W. Happ, "Signal Flow Graph Analysis of Electromechanical Systems," Technical Report Lockheed Missiles & Space Company, 1960. LMSC 6-90-61-120.

99. R. Ash, W. H. Kim and G. M. Kranc, "A General Flow Graph Technique for the Solution of Multiloop Sampled Systems," Transactions of the ASME, p. 360-370, June 1960.

100. R. Ash, W. H. Kim and G. M. Kranc, "A General Flow Graph Technique for the Solution of Multiloop Sampled Systems," Electronics Research Labs., Columbia Univ. Report No. CU-57-59-677-EE, AFOSR TN-59-824, 31 July 1959. TR-T-37/B.

101. P. C. Watson, "Sampled-Data Feedback Systems and Iteration Procedures for Simultaneous Equations," SM Thesis, MIT, EE Dept., 23 May 1955.

102. F. N. Fitting, "Modular Flow Graphs," Electro-Technology, Oct. 1960, pp. 1-9.

103. C. S. Lorens, "A New Approach to Reflection and Image Transfer Loss and Phase," Bell Telephone Laboratories, New York. L3 Coaxial System, Case 35552-124, 24 August 1954.

104. R. W. Sittler, Analysis and Design of Non-Linear Noise Filters, Sc. D. Thesis, MIT, 1954.

105. T. P. Sylvan, "Flow Graph Analysis of Transistor Circuits," Semiconductor Products, pp. 38-42, Jan.-Feb. 1958.

106. O. Wing, "Signal Flow Graph and Network Topology--How to Avoid Them," IRE National Convention Record, Part 2, Circuit Theory, p. 48, 1958.

107. H. J. Zimmermann and S. J. Mason, "Electronic Circuit Theory," N.Y., John Wiley & Sons, Inc., London, Chapman and Hall, Ltd., 1959.

108. H. L. Harter, "Circular Error Probabilities," J. Am. Statistical Assoc., pp. 723-731, Dec. 1960.

109. E. L. Peterson, "Statistical Analysis and Optimization of Systems," Wiley, 1961.

110. S. B. Belknap, LMSC A09963/62 - 41, Lockheed Technical Report, "Uncertainty in Eccentricity and/or Apogee-Perigee Difference Relative to a Circular Orbit," Sept. 1961.

111. A. D. Hall, "A Methodology for Systems Engineering," D. Van Nostrand Company, Inc., 1962.

TABLE 1

Basic Signal Flow Graph Transformations		
Original SFG	Transformation	Transformed SFG
x, y; a, b	Adding Parallel Transmittances	x, a+b, y
x, a, y, b, z	Multiplying Series Transmittances	x, ab, z
x, a, y, b, z; c	Eliminating a Self-Loop	x, $\frac{a}{1-c}$, y, $\frac{b}{1-c}$, z

Table 2

Component Failure Rates	
Component	**λ = Failures/10^6 hours**
Solar Cell Array	1.6
Battery	1.8
Inverter	
Half Load	0.4
Full Load	0.6
Power Amplifier	
No Load	0.1
Full Load	0.5

TABLE 4 FIRST OCCURRENCE PROBABILITIES

Node	Probability of First Occurrence	Probability of First Recurrence
g	$\frac{S}{1-(1-S)T}$	$\frac{RS}{1-(1-S)T}$
f	1	$SR + (1-S)T$
b	$\frac{1-S}{1-SR}$	$\frac{(1-S)T}{1-SR}$
d	$\frac{S(1-R)}{1-SR-(1-S)T}$	0
e	$\frac{(1-S)(1-T)}{1-SR-(1-S)T}$	0

TABLE 3 SEQUENCE OF OPERATING MODES

MODE SEQUENCE	OPERATING COMPONENTS A	B	C	D	E
1	X	X	X	X	
2	X		X	X	X
3	X		X		X
4		X	X	X	X
5				X	X
6					X

TABLE 5 PROBABILITIES, EXPECTED VALUES

Event	Type of Flight	Probability of Exactly n Occurrences	Expected Number of Occurrences
(g)	Successful	$\frac{R^{n-1}S^n[1-(1-S)T-RS]}{[1-(1-S)T]^{n+1}}$	$\frac{S}{1-(1-S)T-SR}$
(b)	Unsuccessful	$\frac{T^{n-1}(1-S)^n[1-SR-(1-S)T]}{[1-SR]^{n+1}}$	$\frac{1-S}{1-(1-S)T-SR}$
(f)	Either	$[SR+(1-S)T]^{n-1}[1-SR-(1-S)T]$	$\frac{1}{1-(1-S)T-SR}$

TABLE 6 Eccentricity Error - Numerical Values					
	Symbols		Typical Numerical Values		
Keplerian Orbits	Injection Orbit	Circular Orbit	Circular Orbit	Deviation	Units
radius	r_i	r_c	3699	s(r)=10	Nautical Miles
speed	v_i	v_c	2.5×10^4	s(v)=30	ft/sec
flight path angle	∅	0	0	s(∅)=.0018	radian
eccentricity	e	0	0	s(e)=.0045	-

TABLE 7

Alternative Earning Rates and Probability Densities

State	Alternative	Probability Density a(.w)	Probability Density a(.d)	Earning Rate q(.)
w	(w,c)	-2	2	4
w	(w,e)	-1	1	3
d	(d,c)	2	-2	-2
d	(d,e)	3	-3	-3

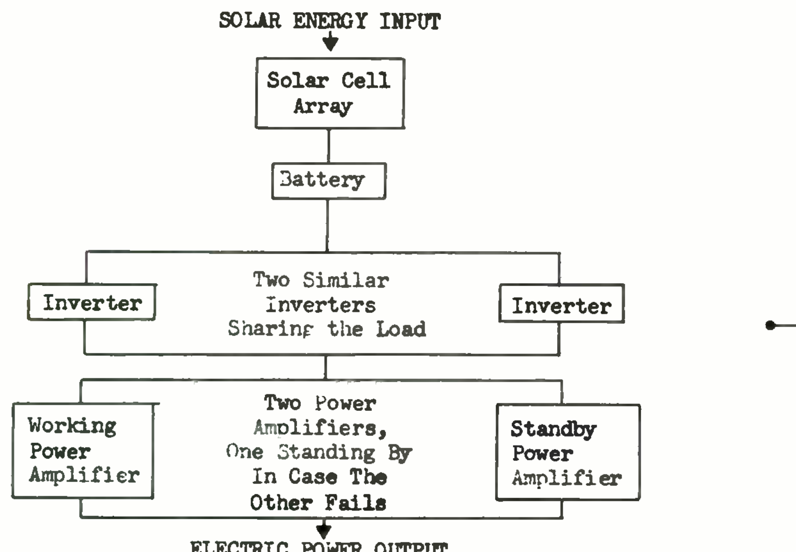


Fig. 1. Solar energy conversion power supply.

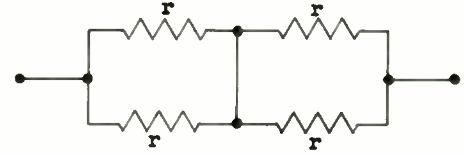


Fig. 2. Redundant elements.

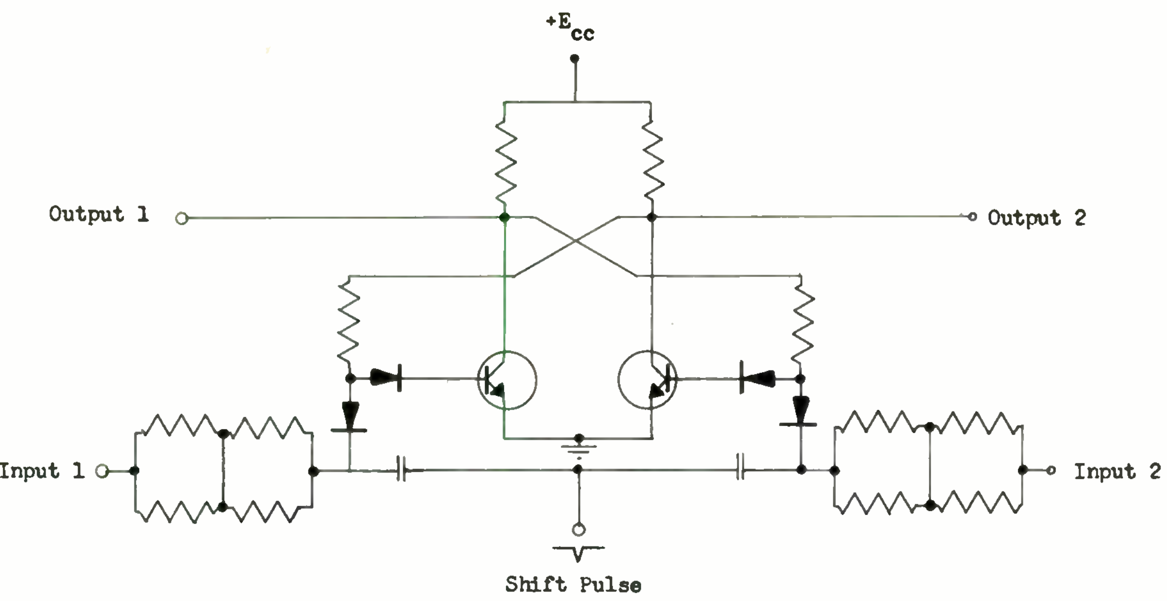


Fig. 3. Redundant circuit.

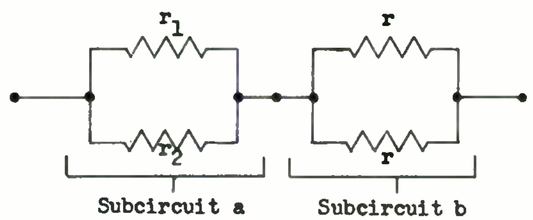


Fig. 4. Redundant subcircuits.

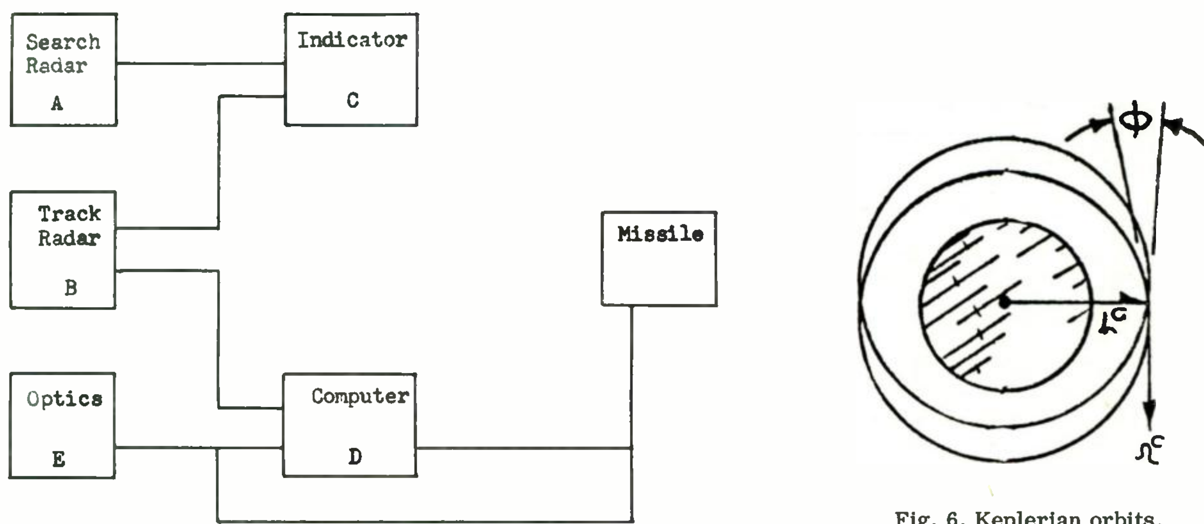


Fig. 5. Multimode failure of tracking system.

Fig. 6. Keplerian orbits.

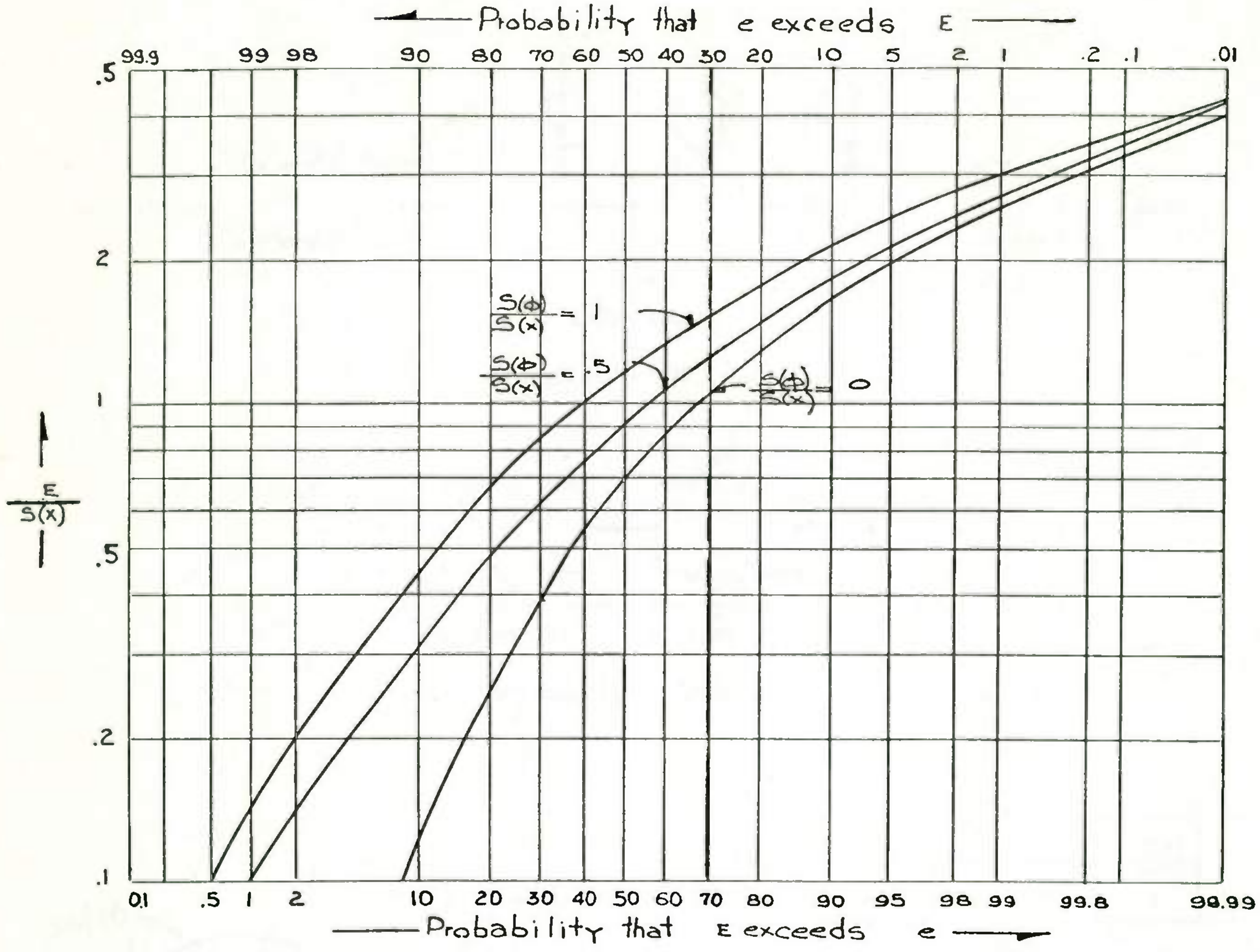


Fig. 7. Probability chart for quadratic uncertainties.

TEST FIXTURE VARIANCE AND RELIABILITY

L. G. Reynolds
Orlando Aerospace Division
Martin Marietta Corporation
Orlando, Florida

Summary

The problem of designing reliable production test fixtures which can be duplicated as production requirements increase has been of concern to engineers for many years. Fixtures which are designed for the manufacture of electronic, electromechanical and mechanical assemblies frequently fail to agree with presumably identical fixtures when the same test specimen is subjected to each fixture. Attention has been focused particularly on production and rework fixtures which have a high rate of use, which test according to specification, and which are required to be more accurate than the assembly itself which is under test. The effects of wear, aging, handling, abuse, and improper maintenance are nowhere more apparent than in production test fixtures. The accuracy of these fixtures constantly degenerates with use. Problems are compounded as additional fixtures are required for rework and test in other parts of the world. For purposes of this paper, reliability is defined as the probability that, using the same test specimen, the same set of test readings will be obtained on a group of test fixtures of the same type. Factors which affect the accuracy of the fixtures, such as drift, wear, human factors, calibration, parts and environment, are discussed.

A solution to some of these problems, as they affect the reliability of the test fixture, is offered. Techniques will be discussed which can eliminate the effects of switch and connector wear, simplify calibration, and reduce the effects of environment.

Test Fixture Variance and Reliability

To design a functional and accurate production test fixture is a substantial engineering task; to design a test fixture which can be duplicated reliably is a real challenge. Test fixture design entails many considerations which are unique. Attention will be focused on test fixtures which have the following general characteristics:

a) Designed to test electronic, electromechanical and mechanical assemblies during manufacture.
b) Used repetitively at a relatively high use frequency.
c) Designed to test an assembly according to a specification.
d) Are considerably more accurate than the equipment under test. (Accuracy is not usually required for preliminary aging and run-in fixtures.)
e) Have a design life of five (5) years or two hundred thousand (200,000) cycles.

The effects of wear, aging, drift, and severe handling are nowhere more apparent than in production fixtures. The accuracy and precision of these fixtures constantly degrades with use. When two or more of the same type test fixtures are required, disagreements in readings often result. When fixtures are shipped to other areas of the world, accuracy and reliability assume a new order of importance. It is essential that fixtures be dependable regardless of location and environment. The question of fixture agreement is complicated by the fact that fixtures are rarely mass produced. Typically, they are custom made as production demands increase, using available materials, parts and labor.

Frequently, depending on urgency and workload, second and third generation test fixtures are vendorized. If the vendor is not provided with an original model from which to copy, he rarely can build an identical unit. Ironically, it is during this period that the original model is usually in maximum demand.

There are two definitions, neither of which is usually applied, that will be useful in understanding this paper.

a) Reliability is defined here as the probability that, using the same test specimen, an identical result will be obtained when the test specimen is tested on a group of test fixtures of the same type.
b) Assembly is defined here (as a matter of convenience) to include sub-assembly, assembly, equipment or system.

In many companies manufacturing test fixtures used to test electronic, electro-mechanical and mechanical assemblies are called tools. In this paper, the term test fixtures will be used throughout.

Environmental considerations are vitally important in planning for the design of test fixtures. The effects of temperature, altitude, electro-magnetic fields, power sources, humidity, vermin incidence, shock, vibration, and fungus must be evaluated.

Some examples of the effects of different localities are:

a) In humid areas, and especially near the seacoast where the atmosphere is saltladen, corrosion and dampness can effect tolerances on mechanical and electrical parts.
b) In a sand and dust environment, these micro-particles will eventually permeate and clog all exposed mechanisms.
c) The effects of temperature are universally

recognized. Tolerance, coefficient of friction and expansion and contraction are factors that must be compensated for if temperature stability is to be achieved.

d) Altitude effects the accuracy of pressure instruments, bellows, pressure transducers, bourdon tubes and other pressure sensitive devices.

e) Vermin affect fixtures in some areas by eating the insulation off the wires, and by being electrocuted, causing partial and intermittent shorts.

f) A high incidence of magnetic metallic particles in some manufacturing areas can accumulate conductive material on electronic assemblies giving rise to leakage paths, and can find their way into D'Arsonval instrument movements adding to the friction and reducing the accuracy.

g) Power sources may differ in frequency, voltage, harmonic content, transients and interruptions. When digital or inertial navigation assemblies are being tested, these factors can be crucial, causing loss of stored information or tumbling of the gyros. Accuracy is frequently affected by these variables.

Temperature extremes should be defined and used in establishing the derating factors for the capacitors, resistors, transformers and other parts. It should be emphasized that the ambient temperature is not the room temperature or even the compartment temperature but is the local temperature around the part. The difference between this temperature and the room temperature must be added to the highest ambient room temperature to be encountered, and the part should be derated accordingly.

For stability and long life the following general rules are established:

a) Resistors should not be operated at more than 1/3 of their normal wattage rating. If the immediate ambient temperature is very hot, further derating may be needed.

b) Capacitors should not be operated at more than 1/2 of their normal voltage rating.

c) Inductive devices should be limited to a 20°C internal temperature rise.

d) Tubes, transistors and diodes should be voltage and wattage derated by 50%.

e) Gears and moving mechanical parts should be designed to operate at between 1/4 and 1/2 yield strength depending on the frequency of use and the length of life expected. Kinematic extremes must be used as a basis for introducing these safety margins.

f) Hydraulics should only be used where accurate positioning is necessary. Pneumatic equipment is more reliable.

Shock, vibration, and acceleration characteristics of the test fixture must be considered after the transportation and ultimate use environments have been evaluated carefully.

Test fixtures are usually not completely documented with a full complement of part specifications, layout drawings, assembly drawings, harness and wiring drawings, and photographs. To fabricate a twin of the first fixture is difficult; a subcontractor or vendor can hardly be expected to duplicate another company's product. For one thing, each company may use unique materials, processes and soldering methods. They may be totally unaware of the methods and processes employed by another company.

These pivotal factors may affect the reliability of the fixture in many ways, some examples of which are indicated:

a) The vendor may have an internal "standard design manual" spelling out a comprehensive grounding procedure. His grounding plan may eliminate the chance of quadrature, cross-talk and noise.

b) The vendor may automatically use harnessing boards for production wiring. If the original fixture was fabricated using point to point wiring, the capacitive, electrostatic and electro-magnetic coupling will vary.

c) The vendor may use various sprays, epoxys, silastics and special finishes on all products as a matter of policy. These finishes and treatments may prevent corrosion, contamination, dampness absorption, fungus and hot spots. In addition, they may influence the problems of modification, repair and calibration, rendering them anywhere from much easier to impossible.

These factors, taken individually, may have little or no effect on the reliability of the fixture. In the long run it is a thousand little things that will cause critical one and two percent differences, differences which will pass all test specimens on one test fixture and fail a large percentage when tested on another.

When several fixtures are in existence and are being used at different locations there is a strong tendancy for them to disagree when testing the same test specimen. Some reasons for this are listed.

1) The environments, though not necessarily hostile, are non-uniform.

2) The fixtures were never identical in the first place. These differences are accentuated by aging and environmental factors.

3) Operators frequently obtain discrepant results when running a single test specimen on a single test fixture. These differences are due to operator parallax, interpretation, and interpolation. These variances are much greater when two or more operators and fixtures are involved at separate locations. Electro-magnetic fields, local power sources, altitude and temperature may, and frequently do, have an effect.

4) Part drift and wearout constitute an important contributing factor to variance. Tubes, transistors, diodes, and resistors; wear, change in friction, backlash, variation in spring constant, and magnet aging; together with many other factors contribute to change.

5) Maintenance practices, both preventative and corrective, together with modifications, may effect the reliability of the fixture. The frequency and thoroughness of preventative maintenance, including lubrication, filter replacement, dust and vermin removal, worn parts replacement,

tube, transistor and potentiometer replacement; these and many other factors will affect the degradation rate. Modifications, although supposedly made using the same parts and methods, may differ in many details. These minor differences may affect the accuracy, reliability, stability, quadrature, rfi, electro-magnetic coupling, drift rate, gain, etc.

6) Calibration practices can affect the reliability of fixtures in many ways. The standards available in one area may differ in many respects from those in another. The fundamental accuracy, impedance, readability, and resolution may disagree. The accuracy of the standards always must be suspect. The calibration frequency of the test fixture and of the standards is usually not on a strict schedule. In some areas certain standards may be non-existent. In many instances calibration charts prepared for the standards and fixture instruments are being misinterpreted. Calibration is frequently difficult owing to the location of the meter terminals. To calibrate a voltmeter, for example, one terminal would have to be removed and a parallel standard and variable power source introduced. Where several instruments are involved, the opportunity to lose lockwashers and flatwashers in the assembly and disassembly operation is ever present. Repeated calibration practically guarantees loose hardware. These bits of hardware never fail to lodge in the worst possible places, and eventually may cause a power short circuit, an intermittent signal short, or some equally troublesome occurrence. The human element must be considered, as the military often requires a high frequency of calibration. There is a natural tendency to "overlook" some of the calibration work, simply by dating and signing cards. This tendency is increased when the terminals of the instruments are nearly inaccessible.

7) Working standards are assumed to be those which are used in calibrating the test fixtures. Working standards are calibrated against transfer standards which, in turn, are calibrated against master reference standards. The variety of working standards is a practical guarantee that every area will have a unique set available. These will vary in impedance, accuracy, readability, availability and in the frequency of calibration. The transfer standards share these diverse characteristics with the working standards. Detective work will often uncover substantial differences in these standards or in the loading effects of these several devices on the unit under test.

8) There is a prevailing tendency for those testers who daily use test fixtures, to generate an unfounded confidence in their ability to modify, adjust, realign, recalibrate, trim, and redesign the fixture. These energetic and well meaning operators frequently make "improvements" to the fixture which causes it to yield test results in discord with other results on all other fixtures. In some instances the working standards have been diligently and deviously altered. A common tendency among testers is an unyielding motivation to reset the zero point on all panel instruments.

9) A well defined system of grounding is generally another missing link in the design of fixtures. Signal, power, shield and structural grounds are often unspecified. Each fixture may be variously grounded, or not grounded at all, and the ground returns may be run in several different ways. These dissimilarities often lead to quadrature, pickup, noise, spurious signals, interference, inter-modulation and other undesirable effects.

10) When an assembly is first interconnected to the test fixture the wires are not necessarily on the correct pins. Frequently power is applied to a circuit which was not designed to accept power. Before the operator realizes the difficulty severe damage may occur to either the test fixture or to the assembly under test, or both. To avoid this costly type failure each fixture should be equipped with a semi-automatic impedance measuring device which gives audible warning and a red light when an impedance is off by more than a pre-established limit. The particular circuit involved should also be indicated. A stepping switch device is often applicable.

A wise man does not dive into a pool until he knows the depth of the water. An engineer should not plunge into the design of a test fixture until he has complete knowledge of its purpose, reliability, maintainability, and human factor requirements. Test fixtures are subject to continuous use and abuse, and should be designed to withstand severe and unusual handling and wear. Fixtures are frequently moved, dropped, crated and shipped, and otherwise subject to shock and vibration. Fixtures may topple onto concrete when they have high centers of gravity, when someone trips over one of their cables, or when someone gets shocked electrically and knocks one over in the excitement.

Governing considerations in designing a test fixture are:

a) Can it do the job according to the specification?

b) Will it be reliable and maintainable?

c) Can it be calibrated easily and accurately?

d) Can it be duplicated at any reasonable time in the future, by another company, using the prints and schematics available?

e) Can an operator use the fixture easily and without unusual fatigue, using the prescribed test procedure?

f) Can the fixture be readjusted and otherwise tampered with by an operator without knowledge of his supervisors?

g) Does the fixture have self checking capability?

h) Can a left-handed color-blind person use the fixture as well as a right-handed person with normal color perception?

Attentiveness to these vital aspects of fixture design will pay important dividends in reduced maintenance and calibration time, but more importantly, the critical problem of achieving agreement between fixtures will be minimized.

Environment, even when not hostile, will have an effect on the reliability of test fixtures. Test fixtures are frequently untested and unqualified under the customary environmental conditions;

hence little is known of the effects of minor changes in temperature, altitude, humidity, shock, vibration, power source variations, radio frequency interference and other factors. It is recommended that the test fixture be subjected to a moderate and non-destructive series of environmental tests. Following, or during each test, a small control lot of the test specimens, which are permanently set aside, be tested and the data recorded and stored for future reference. If changes occur during any of the tests, indicating unusual variance, this condition should be investigated and corrected. A change of reading of twenty percent of the total acceptance tolerance is, in general, cause for an investigation.

In a high production operation, and particularly where fixtures are used on a two and three shift basis, wear, tear and degradation must be expected. In a typical case a rotary switch may have fifty or more wires connected to it, and may be cycled two thousand (2,000) times per day. Normally, a switch of this type will fail about once every four months with this switching frequency. To change a switch of this kind several things must be considered. The leads may not reach the new switch after they have been cut off for the second or third time. The leads may not be colorcoded and may get out of order while working in a cramped and poorly lit space. A soldering iron may damage adjacent wires because of their unusual closeness. Solder drippings may lead to other difficulties. The time involved in changing a switch of this type is prohibitive.

Customarily the test specimen is connected to the fixture through a connector. Here again the arguments used above apply. A solution is presented which should simplify the maintenance effort and reduce the skill level required for ordinary trouble shooting.

a) For all multi-connection devices which are subject to frequent use, a second interface should be provided consisting of a terminal strip or a connector.
b) When the switch or connector on the panel appears to be defective or worn it can be separated by removing three or four screws and a connector.
c) Two or three spares for all of these devices should be prefabricated and available. Spare cables also should be available for fixtures.

The time involved to get a fixture "back on the air" can be greatly reduced. If an intermittent condition develops it will no longer be necessary to make a major change to find out if it is in the switch, in the connector, or in the device under test. A simple change can pinpoint the trouble.

A test fixture is expected to be forever accurate and reliable. It must not drift, age, corrode or respond to ambient change; it must be stable and unaffected by power variations and rfi. These difficult and worthwhile objectives prohibit compromise with part quality, with material quality, and with design reliability. A test fixture must be more accurate and more dependable than the assembly under test.

Following are a regimen which must not be violated in selecting and applying parts.

1) Each part shall be selected for stability and durability.
2) A parts list shall be prepared which shall include a description of each part, a part number and a part specification. Not less than two vendors shall be selected for each part. The preferred vendor shall be starred and should always be first choice.
3) Tolerances shall be assigned to all critical circuit elements including resistors, capacitors, inductors, transformers, gears, motors, electron tubes, transistors, and diodes.
4) Adequate safety margins and derating factors: wear, drift, and stability allowances, should be made for all mechanical and electrical items.
5) It is important to purchase spares for the test fixtures from the same lot as the original parts were procured from. It is equally important to purchase enough parts to build all anticipated test fixtures from this lot.

An example of numerical resistor tolerances which must be designed for are provided in Table I. These data represent conservative limits, rather than optimistic claims, and will aid the fixture design engineer in avoiding serious design deficiencies. Note that original manufacturing limits do not reflect final use limits by relatively large factors.

Limits which far exceed original manufacturing limits must be expected with all electrical parts such as capacitors, tubes, transistors, etc. Experienced test fixture designers will compensate for wide variations in part values. Advantage may be taken of "Schmoo" plots, a systematized method of using a computer to establish total fixture variation as a function of extreme part values.

Postulating that complex equipment cannot be identical, and that test specifications must assume this inequality of test fixtures, effective techniques can minimize these inequalities. It is an exacting task; it cannot be accomplished by accident nor by wishful thinking. In addition to some of the general rules already given, some specific guidelines for fabrication of duplicate fixtures are provided.

Sufficient documentation must be generated to enable a competent outside company to manufacture a reliable* fixture without benefit of having seen the original fixture and without receiving extensive verbal instruction. The following documentation and/or hardware is considered to be minimum:

*See again the definition of reliability as used in this paper.

TABLE I

Resistor Type (1/2 watt)	Resistance Tolerance	Resistance Range	Temperature Coefficient (-55° to +125°C)	Voltage Coefficient	Ambient Underrated Operating Temperature (°C)	Maximum Rated Voltage	Overall Design Tolerance (%)
Composition	±5	10 ohms to 20 meg	±10%	350 PPM/v	70	350	±18
High Stability Film	±0.5	10 ohms to 5 meg	-200 to -500 PPM/°C	NIL	70	350	±4
Accurate Wirewound	±0.5	1 meg (max)	0 ± 5 PPM/°C	NIL	125	350	±0.2
Metal Film	±0.5	250 ohms to 1 meg	0 ± 25 PPM/°C	NIL	125	350	±1.5

1) A complete set of drawings is essential, including layouts, details, assemblies, subassemblies, cable assemblies, schematics, wiring diagrams and harnessing diagrams.

2) A complete set of specifications and/or procedures must be available including purchase, test, part, material, calibration, environmental, process, procedure, and manufacturing specifications.

3) A complete set of photographs showing wiring layouts, color coding, test set-ups, panels and other important details.

4) Special written procedures for adjusting, aligning, calibrating, trimming and other special operations are essential. Each switch, knob, fuse, connector and adjustment must be identified to permit positive reference to it in procedural and alignment instructions. Each position of each switch must be identified by number.

5) As indicated previously, parts must be ordered and checked for all anticipated fixtures from a single lot, and must be supplied to the fabrication agency.

6) Where working standards are used as ancillary equipment, these standards must be provided to the test group for fixture qualification and acceptance tests.

7) Several test specimens together with their individual test data must be provided to the test group to serve as a standard for the original qualification and acceptance of the fixture. These test specimens should be kept in a bonded area, and never sold just as patterns and dies are never sold.

8) A complete grounding procedure should be provided. Frequently, structural, panel, motor case, transformer case and other grounds are not indicated in drawings. An ultimate power ground together with where and how this ground is to be connected should be defined. Signal and shield grounding should be covered. Electrical input line connections should be polarized to be grounded in the same manner, and this ground should be carried through the fixture. For safety, an electrical power ground must be carried to all panels, relay racks and principle assemblies. This ground must not be fused. This same power ground return must be reliably carried to the assembly under test.

The engineer who is assigned development responsibility for the test fixture must retain this as an undelegatable responsibility throughout the life of the fixture. Whenever reliability responsibility is transferred to a separate organization and problems are encountered, the problem can be solved better by the design engineer.

When a fixture is fabricated it customarily enters the de-bugging or shakedown operation. Highly trained technicians correct and alter equipment to make it work. Filters are inserted, grounding is modified, parts are reworked to fit, resistors are bridged, quadrature is removed, pickup is eliminated, and frequently not one mark is made on the original set of prints to indicate where or how these alterations were made. The next fixture may be debugged by an entirely new group of technicians. It is improbable that these fixtures will ever fully agree. A set of prints must be marked up as changes are made to ensure that no information is lost. When these changes have been fully authenticated, the original prints should be changed accordingly and all changes must be fully documented.

Following shakedown, several specimens should be tested and the data recorded permanently. Test fixture variance should be plotted and the test specimens retained.

Methods have been developed which tend to prevent tampering with test fixtures by unauthor-

WRH

ized personnel. Some companies lockwire the fixture panels and drawers closed and use a lead embossed seal. Others use a special type cup around one of the holding screws, filling the cup with a resinous material. These are frequently embossed with a letter or symbol. To be meaningful, company policy must impose severe penalties for unauthorized personnel who open these seals. Ancillary equipment, working standards, and transfer standards should be likewise sealed.

One of the more intriguing aspects of reducing test fixture wear has nothing to do with their design. It is a production policy consideration. In a typical case an equipment is rejected at the final test station and returned for rework. The unit is passed at the rework station and forwarded to the final station where it is again rejected. An indefinite number of tests are conducted on units which fall into this category with concomitant wear on the fixture. Adequate records should be kept to flag an equipment which is cycling on rework. At this point, the unit usually fails the final test marginally and barely passes the rework test. A decision should be made, based on the fact that no two fixtures can ever agree, to rework the equipment towards a centralized specification value, or to sell the unit under a slight variation order.

Human factors, to an extent which is greater than generally realized, will bear on the reliability of fixtures. In setting dials and gages, and in reading instruments, there are interpolations in analog devices which can vary with different operators.

An interesting method of evaluating operator homogeneity consists of having trained operators test the same equipment, using the same test fixture, without telling them that the several recorded results will be compared. Repeat the experiment telling the operators that their results will later be compared with other test results. The results of this simple experiment are frequently astonishing.

A surprisingly cheap method of maintenance and calibration simplification through application of human factor information is available. Consider the ordinary problem of generating enough light to see clearly while working inside a relay rack under a chassis. If the inside surface of the relay rack and the entire chassis, except the front side of the front panel, were painted white, visibility would be increased for testing, troubleshooting and maintenance operations. A second advantage exists for application of a white finish to the inner surfaces of the panel and rack assemblies. Much of the radiated heat that causes operator loss of efficiency will be eliminated.

In a recently completed Martin study* sponsored by the USAF, future trends in test fixture design were examined. Pertinent facts were concluded as follows:

1) Future requirements show a continuing need (for test fixtures) even for the most reliable equipment.
2) The need for a universal checker in the broad sense is rare. However, there are applications for multi-system testers.
3) The feasibility of standardizing functional chassis was found to be 40% for digital and analog building blocks.

These findings should be valid for at least ten years in the future.

The following recommendation is made in the paper "The Payoff is the Standardization of Checkout Gear at the Building Blocks Level." . . . "standardization of equipment naturally leads to beneficial standardization of stocking, repair, availability and other logistics factors. Such standardization is feasible." To this it might be added that a much higher reliability of test fixture could be achieved by standardizing on certain building blocks and then concentrating on the improvement in dependability of these units under continuously severe use factors.

The best opportunity for application of the multi-system tester exists in rear area depot type checkout and maintenance areas.

Many of the precepts which have been discussed are immutable and an experienced designer will not violate them. An untold fortune is being spent in unnecessary product rework because these rules of design are either not known or are disregarded. Another fortune is spent in attempting to obtain agreement between test fixtures which should but do not agree. Many of the ordinary design pitfalls have been exposed in this paper.

In addition, through the application of calibration simplification techniques, herein described, the cost of periodic calibration can be appreciably lowered. Conversely, the likelihood of periodic and thorough fixture calibration will be substantially increased. The frequency and time per maintenance operation can be reduced through application of rapid replacement techniques, for known wearout items.

The composite effect of application of the methods presented will yield fixtures of high individual dependability and high related reliability. Down time will be decreased while accuracy is increased.

Production managers will move to a new level of confidence in making decisions affecting rework, accept-reject, and quality.

References

1. TR 59-416-1 "Reliability Stress Analysis for Electrical Engineer" by H. L. Wuerffel - RCA dated January 15, 1959.

---

*Cambell, W. O. "Design Criteria for Automatic Test and Checkout Systems," ARDC Study Requirement No. 17530

2. "Effects of Maintenance on System Reliability" ARINC Research Corp., September 15, 1959 - Publication #101-16-144.

3. "Mechanical Reliability" Product Engineering Magazine, November 16, 1959, p. 27.

4. "Reliability Training Text" published by IRE. Edited by Cryna, McAfee, Ryerson and F. Werling. Second Edition.

5. MIL-STD-441 "Reliability of Military Electronic Equipment."

6. AF Bulletin 506 "Reliability Monitoring Program for use in the Air Weapon System."

7. Exhibits RADC - 2629 "Reliability Requirements for Ground Electronic Equipment." 31 October 1958.

8. MIL-R-19610 (AER) Reliability of Production Electronic Equipment.

9. MIL-R-25717A. "Reliability Assurance Program for Electrical Equipment." 11 July 1957.

10. "Handbook Preferred Circuits" Navy Aeronautical Electronic Equipment prepared by National Bureau of Standards, NAVAER 16-1-519 - 1 September 1955.

11. MIL-R-26484 (USAF) 4 June 1958. "Reliability Requirements for Development of Electrical Sub-systems or Equipment."

12. MIL-M-26512. "Maintainability Requirements for Weapon Systems."

13. USAF Bulletin #510 "Guides for Reliability Organization" 30 June 1959.

14. MIL-R-26674 (USAF) "Reliability Requirements for Weapons Systems." 18 June 1959.

15. "Reliability of Military Electronic Equipment." 4 June 1957. For sale by Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. - Price $1.75

16. "NEL Reliability Design Handbook." U. S. Dept. of Commerce; Office of Technical Services. PB 121839, Price $3.00.

17. "Handbook of Instructions for Aircraft Designers" - HIAD ARDCM 80-1 USAF Volumes I, II & III Confidential.

18. "Production Design Handbook" ARDCTR 59-1 ASTIA Document No. AD 204 500 December 1958. Air Research Development Command.

19. "Human Engineering" by E. J. McCormick - McGraw-Hill Book Company - 1957.

20. "Interim Dictionary of Terms and Definitions Relating to Reliability" 7 December 1959. Aerospace Industries Assn. of America - 610 Shoreham Building, Washington 5, D. C. Price $1.25.

21. "Procedure and Data for Estimating Reliability and Maintainability." 9 July 1959. Martin-Denver Reliability Publication - M-M-P-59-21.

22. "The Prediction and Measurement of Air Force Ground Electronic Equipment Reliability" RCA. RADC-TN-58-307, AD-148977, 15 August 1958.

23. Dr. Dan Noble, Motorola, Incorporated: Supercolossal Gadgetry.

24. Robert Lusser: Production Environmental Testing.

25. Robert Lusser: Reliability Specifications for Guided Missiles - A Proposal (Superseded by "Reliability Through Safety Margins")

26. Robert Lusser: Testing to Specified Limits versus Testing to Failure.

27. Robert Lusser: The Notorious Unreliability of Complex Equipment A Condensed Presentation for Management Use.

28. Reliability Through Safety Margins October 1958 by R. Lusser USA Ordnance Missile Command.

29. Hickey, John E. "What Price Reliability" Electronic Industries, September 1961.

30. Leake, Charles E. "Understanding Reliability, Test Engineering" August 1. Serialized.

31. A Report under Contract No. bsr-64508, "Analysis of Operator-Equipment Relationships," 19 Dec. 1958 (No. 101-7-131)

32. A report under Contract No. bsr-64508, "Effects of Maintenance on System Reliability," 30 June 1959 (No. 101-16-144).

33. "RADC Reliability Notebook," RADC-TR-58-111. ASTIA document No. AD-148868.

34. Philosophy and Guidelines for Reliability Prediction for Ground Electronic Equipment, RADC TN-58-20. ASTIA Document No. AD-148556.

35. "Reliability Factors for Ground Electronic Equipment" McGraw-Hill Book Co. ASTIA Document No. AD-114274.

36. "Minuteman, Catalyst for Reliability," Electronic Industries, December 1960.

37. Design Criteria for Automatic Test and Checkout Systems (Task 17530). Final Report. The Martin Company. SSD No. 7993-17530 dated 29 September 1959. Volume I Unclassified - Volume II Secret.

38. The Design of Test Equipment for High Reliability Product, L. J. Paddison, Sandia Corporation.

1245 895 231
103
WRH