DEPARTMENT OF COMMERCE · NATIONAL BUREAU OF STANDARDS

NBS SPECIAL PUBLICATION 487

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

Engineering Design

MFPG
25th Meeting

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Institute for Computer Sciences and Technology, the Office for Information Programs, and the Office of Experimental Technology Incentives Program.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of the Office of Measurement Services, and the following center and divisions:

Applied Mathematics — Electricity — Mechanics — Heat — Optical Physics — Center for Radiation Research — Laboratory Astrophysics[2] — Cryogenics[2] — Electromagnetics[2] — Time and Frequency[2].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials, the Office of Air and Water Measurement, and the following divisions:

Analytical Chemistry — Polymers — Metallurgy — Inorganic Materials — Reactor Radiation — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services developing and promoting the use of available technology; cooperates with public and private organizations in developing technological standards, codes, and test methods; and provides technical advice services, and information to Government agencies and the public. The Institute consists of the following divisions and centers:

Standards Application and Analysis — Electronic Technology — Center for Consumer Product Technology: Product Systems Analysis; Product Engineering — Center for Building Technology: Structures, Materials, and Safety; Building Environment; Technical Evaluation and Application — Center for Fire Research: Fire Science; Fire Safety Engineering.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus wthin the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Institute consist of the following divisions:

Computer Services — Systems and Software — Computer Systems Engineering — Information Technology.

THE OFFICE OF EXPERIMENTAL TECHNOLOGY INCENTIVES PROGRAM seeks to affect public policy and process to facilitate technological change in the private sector by examining and experimenting with Government policies and practices in order to identify and remove Government-related barriers and to correct inherent market imperfections that impede the innovation process.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data — Office of Information Activities — Office of Technical Publications — Library — Office of International Standards — Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.

[2] Located at Boulder, Colorado 80302.

MFPG
Engineering Design

Proceedings of the 25th Meeting of the
Mechanical Failures Prevention Group,
held at the National Bureau of Standards
Gaithersburg, Maryland, November 3-5, 1976

Edited by

T.R. Shives and W.A. Willard

Metallurgy Division
Institute for Materials Research
National Bureau of Standards
Washington, D.C. 20234

The 25th meeting of MFPG and these proceedings were sponsored by the Institute for Materials Research of the National Bureau of Standards, Washington, DC 20234; the Office of Naval Research, Department of the Navy, Arlington, VA 22217; the Naval Air Development Center, Department of the Navy, Warminster, PA 18974; the Frankford Arsenal, Philadelphia, PA 19137; the Federal Aviation Administration, Department of Transportation, Washington, DC 20591; and the Energy Research and Development Administration, Washington, DC 20545.

U.S. DEPARTMENT OF COMMERCE, Juanita M. Kreps, Secretary

Dr. Sidney Harman, Under Secretary
Jordan J. Baruch, Assistant Secretary for Science and Technology

U.S. NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Acting Director

Issued August 1977

Library of Congress Cataloging in Publication Data

Mechanical Failures Prevention Group.
MFPG—engineering design.
(National Bureau of Standards special publication ' 487)
"Sponsored by the Institute for Materials Research of the National Bureau of Standards ... [et al.]."
Supt. of Docs. no.: C13.10:487
1. Engineering design–Congresses. 2. Fracture mechanics–Congresses. I. Shives, T. R. II. Willard, William A. III. United States. National Bureau of Standards. Institute of Materials Research. IV. Title. V. Title: Engineering design.. VI. Series: United States. National Bureau of Standards. Special publications ; 487.
QC100.U57 no. 487 [TA174] 602'.1s [620'.004'2] 77-21810

National Bureau of Standards Special Publication 487
Nat. Bur. Stand. (U.S.), Spec. Publ. 487, 367 pages (Aug. 1977)
CODEN: XNBSAV

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1977

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $5.25

Stock No. 003-003-01829-5

FOREWORD

The 25th meeting of the Mechanical Failures Prevention Group was held November 3-5, 1976, at the National Bureau of Standards in Gaithersburg, Maryland. The program was organized by the MFPG committee on Design under the chairmanship of Jesse E. Stern of NASA/Goddard Space Flight Center. The Design committee, the session chairmen, and especially the speakers are to be commended for the excellent program.

With a few exceptions, the papers in these proceedings are presented as submitted by the authors on camera ready copy, except for some minor editorial changes.

Appreciation is extended to the following members of the NBS Metallurgy Division: T. Robert Shives and William A. Willard for their editing, organization, and preparation of the proceedings, Paul M. Fleming for handling financial matters, Todd Eudy for photographic work, Larry W. Ketron for drafting work, and to Marian L. Slusser for typing. Appreciation is also extended to Sara R. Torrence of the NBS Office of Information Activities for the meeting arrangements and to many members of the staffs of the NBS Metallurgy Division and the NBS Institute for Materials Research for their assistance in many ways.

HARRY C. BURNETT
Executive Secretary, MFPG

Metallurgy Division
National Bureau of Standards

TABLE OF CONTENTS

*Indicates speaker when a paper had more than one author.

ABSTRACT

These Proceedings consist of a group of twenty-three submitted papers from the 25th meeting of the Mechanical Failures Prevention Group which was held at the National Bureau of Standards in Gaithersburg, Maryland on November 3-5, 1976. The central theme of the Proceedings is the application of design to failure prevention, with emphasis on the techniques of design, materials of design, design validation, and the goals, purposes, and requirements of design as related to product liability.

Key Words: Coating materials; composite materials; design; failure avoidance; failure prevention; lubricants; materials of design; product liability; reliability of design; techniques of design; validation of design

UNITS AND SYMBOLS

Customary United States units and symbols appear in many of the papers in these Proceedings. The participants in the 25th meeting of the Mechanical Failures Prevention Group have used the established units and symbols commonly employed in their professional fields. However, as an aid to the reader in increasing familiarity with and usage of the metric system of units (SI), the following references are given:

NBS Special Publication, SP330, 1974 Edition, "The International System of Units."

ISO International Standard 1000 (1973 Edition), "SI Units and Recommendations for Use of Their Multiples."

E380-76 ASTM/IEEE Standard Metric Practice (Institute of Electrical and Electronics Engineers, Inc. Standard 268-1976).

Disclaimer:

Certain trade names and company products are identified in order to adequately specify the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Bureau of Standards, nor does it imply that the products are necessarily the best available for the purpose.

THE GROWING WEB AROUND THE DESIGNER

Charles O. Smith
Professor of Engineering
University of Nebraska at Omaha
P. O. Box 688
Omaha, Nebraska 68101

Abstract: The task of the designer has never been easy since he has had to consider a number of factors, not all of which are necessarily compatible. He has always had to make iterative tradeoffs among various requirements, including the very important one of cost. In the context of today's technical and societal climate, the designer's task has become even more difficult, if not almost impossible.

This paper summarizes the interrelated factors the designer must consider. There is further complexity as not all of the factors apply with equal weight in all designs, and some requirements conflict with others, yet all must be considered. None should be discarded without full justification, much less be ignored. The paper is essentially a checklist for the complex, confusing, and (sometimes) frustrating task of designing a product, not a "how to design" paper.

Key words: Design; checklist for design; factors in design; requirements in design.

Introduction

The task of the designer has never been easy since he has had to consider a number of factors, not all of which are necessarily compatible. He has always had to make iterative tradeoffs among various requirements, including the very important one of cost. In the context of today's technical and societal climate, the designer's task has become even more difficult, if not almost impossible.

Today's designer is required to account for so many factors and considerations that it is essentially impossible for one individual to be thoroughly conversant with all of these. At the same time, however, he can not afford to overlook or slight even one of these factors. Thus he is, in effect, forced to play a role something like that of an orchestral conductor. He must have some special competence of his own and a reasonable knowledge of the other "instruments". But above all, he must know how all these considerations fit together, what interactions are possible, and what sort of tradeoffs can and/or must be made. The various design considerations have been (somewhat arbitrarily) grouped into three categories as indicated in Table 1. It is recognized that these categories intersect and overlap. This hard fact is the central core of the difficulty in developing truly acceptable design in various products.

Functional Requirements

It is obvious that any design must necessarily be able to meet performance specifications. Thus this item is at the top of the list in Table 2. Performance specifications must fully reflect a carefully and completely defined need. One must be careful about specifications, however, since these are often rather specific about what shall be in the product rather than the functions required. The latter is much more important and should be so emphasized. Design can be improved as a result of clarifying the functional requirements of a product. For example, if a surface must only resist corrosion, coating or cladding may cost less -- and may require less material -- than using corrosion resistant material throughout (1)*, and still adequately meet performance requirements.

Awarding contracts to the lowest bidder encourages minimum material input (consistent with specifications). As a result there is frequently a shift from first cost to maintenance cost (2), not necessarily good design from the viewpoint of total life-cycle performance and cost.

---

*Numbers in parentheses refer to items in the bibliography.

Any product must have some sort of design configuration, i.e., it must take some physical, geometrical form. In all cases, there will be some demand function which the product must have the capacity to meet. Most efforts to balance that equation, while providing some margin of safety, are placed in a deterministic mode, using average values of demand and capacity. This, in effect, ignores the statistical nature of design parameters. To avoid unrealistic results, magnitude and frequency relationships, for both demand and capacity, should be considered. This suggests a probabilistic approach as portrayed by Haugen (3) and others. In the context of mechanical design, coupling of average stress from applied loading with the average available strength, while accounting for the variations in both, permits designing to any desired degree of failure probability (4).

In making mechanical stress calculations, the designer must give full attention to notches since notches render hardware highly vulnerable to failure (5). At the same time, it must also be recognized that the developing field of fracture mechanics provides evidence that some notches are benign and acceptable to a degree in some design applications. The designer must very carefully assess this aspect.

Increasing emphasis is being placed on reliability in equipment. In this context, reliability is defined as the probability of a device performing adequately for the period of time intended under the operating conditions encountered. There is no general definition of adequate performance, the criteria for which must be carefully and exactly detailed in advance of generating a design. In addition to reliability, there is also increasing emphasis on its probabilistic relatives such as maintainability, availability, and repairability.

The designer must anticipate failure in his product. This requires use of such tools as fault tree analysis, failure modes and effects analysis, and hazards analysis.

In the process of generating a design configuration, one should apply value analysis or value engineering, ie., appraising a design from the viewpoints of both function and cost to assure maximum value, or achieving the required function (without unnecessary functional capacity) at the lowest cost. This is neither simple cost-cutting nor does it lead to performance trade-offs.

Substantial improvement in function or cost can often be achieved by redesign. Formal design review procedures have proven very useful in enhancing product performance. In this context, techniques such as failure-modes-and effects-analyses are very valuable and, in many cases, essentially imperative.

Intensive and extensive standardization and simplification of design can lead to substantial savings without loss of performance. Retaining only the most efficient of the variety of sizes, types, grades and models of a given product will lead to considerable economies. During World War II, for example, the U. S. reduced the number of types of steel, brass, and bronze valves from 4080 to 2500 with an annual savings of 5000 tons of steel and 1900 tons of copper (6).

Functional substitution offers great opportunity for improvement through redesign. The aim is not simply to replace but to find a new or different way to do a job. Joining two pieces of metal can be done not only by using stronger bolts and nuts but also by welding or by adhesives. Jet engines replace piston engines and propellors. Telephones replace mail for transmitting some information (7).

Total Life Cycle

Materials, energy, and the environment form the physical basis of the economy (8). They also have a complex interrelationship (9). The designer thus has the problem of using the optimum combination of materials, energy, input, and environmental impact consonant with the functional requirements of a design. While many recognize that changes in one area start changes in the others, probably no one fully understands the interactions. The designer must recognize this and be prepared to cope as well as possible.

Designers often select a material for two incorrect reasons: (1) "It's always been used for this application," and (2) "Its properties." With regard to the first incorrect reason, the time when each application had its preferred material and each material had its secure markets is gone. In other words, tradition and good design do not necessarily mix. The second reason is incorrect because it implies that properties are something mysteriously, almost metaphysically, connected with the material, something a material has without reference to the specific use. On the contrary, a property should be regarded as the response of the material to a given set of imposed conditions.

In selecting materials for a given application, the first step is evaluation of service conditions. The characteristics (responses) of various materials are next evaluated to select the most promising materials. One common question is that of adequate strength. While strength may well be the primary selection criterion, it may also be ductility, toughness, corrosion resistance, specific gravity strength/weight ratio, and/or one or more of several other characteristics.

Stability of material in service is intimately related to temperature, fluctuations in temperature, and length of time at temperature. All of these directly affect strength and often lead to creep, but they can also produce changes in the microstructure of the material, e.g., tempering of martensitic steels and overaging in precipitation-hardening alloys.

Availability and fabricability are intimately related. A given material may show excellent promise for a given application on the basis of laboratory tests. If it is not available in the desired quantity and form at an acceptable cost, or if it is not reasonably possible to develop processes to get the desired form, then it is obviously unreasonable to base a design on it. Fabrication processes should be used which give minimum waste and/or maximum recovery of material.

In general, the designer is basically responsible for selecting materials. He should not, however, rely exclusively or even heavily on handbook data but should consult with materials specialists. At the same time, to best profit from such consultation, the designer must give the specialist the complete story of intended use, possible sources of supply; and fabricating facilities. To be most effective, there should be continual collaboration throughout the design process.

The designer must be fully aware of the option of alternative materials. In some cases this will involve substitution for preferred, but unavailable material. In other cases the alternative material will be selected because of superiority in the specific situation. In one case a tool manufacturer had been purchasing AISI 1078 and 1086 carbon bar with overlapping analyses. The decision was made to use only AISI 1078. This gives one less grade to inventory. Even more importantly, 1078 is rolled more frequently than 1086, thus effectively making it more available and easier to schedule (10). One aspect of alternative materials is the possibility of using composites which might also be considered as an aspect of value analysis and/or redesign.

The designer must be concerned with producibility. While this is related to fabricability, it is much broader as it necessitates consideration of manufacturing processes, assembly processes, etc.

Durability is related to, but differs from, reliability. Reliability is concerned with successful operation for a given period. Durability is concerned with how many cycles of such operation are possible. In other words, durability is a question of the length of life of a product.

In the context of total life cycle, the designer must think of the material he selects from its starting point (e.g., ore) to its disposition. While he is properly concerned with the durability of the product over the projected life, he must also be concerned with possible recovery or reclamation. The ultimate goal should be to combine materials selection, product design and manufacture, and recycle processing in a systems approach to optimizing product development (11). Emphasis must be placed on conservation of nonrenewable resources, improved material effectiveness, and much more intensive recyling of the discards of industry and consumers (12). With more attention to recyling by the designer, it becomes easier to disassemble a discarded product and feed components into appropriate material streams which return them to the secondary industry for reworking (13). With recycling, the same material may be reused many times over a period of years. With each recycle, an equivalent amount of virgin raw material is saved. Introduction of recycling, in effect, closes the loop on an otherwise open system of extraction, consumption and disposal (14).

Recycling post-consumer waste materials is not a panacea since it will not supply the majority of the Nation's raw material demands. On the other hand, substitution possibilities, with regard to both total consumption and domestic virgin material supply, are not insignificant (15).

In obtaining materials from basic sources (e.g., mining ore and extracting metal from ore), in fabricating into basic forms (e.g., sheet, plate, rod, etc), in further processing into semi-finished and finished products, etc, there is a continual requirement for energy. In fact, processing of materials is highly energy-intensive. There are further energy requirements in operating or using most products. If the product is ultimately discarded, there is loss of non-renewable material and energy plus some adverse impact on the environment. If the product is reclaimed and/or recycled, there is still some energy requirement although this is usually significantly less than in extracting virgin raw material. The designer must consciously seek to minimize the energy requirements of his product over the total life cycle.

The designer must consider the effect of his product on the environment. The uses of energy and materials may have adverse environmental effects associated with generation (oil drilling, water shortage, etc), delivery (oil spills, exhaust fumes, etc), fabrication, application and disposal (glare, noise, air pollution, water pollution, thermal pollution, debris, etc) (16). Consideration of the effects of the product on the environment is in itself a complex task. Nevertheless, it is a task which must be performed.

The designer must also consider the effects of the environment on his product. It may be argued that this is done when considering stability, corrosion resistance, wear, cavitation, etc. No matter when it is considered, it is most important.

One area often given too little attention is handling of the product after completion. This includes such aspects as inspection, testing, packaging, shipping, and storage. Inspection and testing are important in ensuring that the product is manufactured and functions as intended. Proper attention to such considerations during design can make inspection and testing easier and more effective. Packaging, shipping and storage can be a major cause of product malfunction or failure. For example, major home appliances such as washing machines, dryers, etc, often are subjected to stresses and environments during shipping and storage which are significantly more harmful than the normal operating environment.

Further Major Factors in Design

The complex economic machinery of the Nation requires that materials and products in commerce conform to prearranged standards and specifications, which usually represent voluntary agreement among the elements of industry (17). Although use of such standards (e.g., AISI, ASTM, etc) is appropriate, the designer should also recognize that since there is voluntary agreement, many standards are, in effect, least common denominators or minimum standards. Good design may well require more than this. At the same time, specification by standard sometimes requires a producer to meet some requirement which is actually non-essential (18). Emphasis in specifications should be on performance requirements.

A different area of standards relates to matters of safety and health. Requirements of the Occupational Safety and Health Administration (OSHA) and the Consumer Product Safety Commission (CPSC) are of prime interest to the designer. There are at least three aspects related to the CPSC which often receive too little attention. The intended purpose of a product is normally quite clear. Unintended uses of the same product are not always as obvious. Obvious or not, however, the designer has an obligation to foresee as many unintended uses as possible and to develop his design in such a way as to eliminate hazards to one who operates in an unintended (from the designer's viewpoint), but foreseeable manner. The designer must be aware that warnings are often required on products. These warnings must be clear, complete, unambiguous, and permanently attached. The designer must also be cognizant of all advertising material, owner's manuals, operating instructions, etc, which relate to the product since such information put out by the manufacturer has been held to constitute a warranty. There have been situations in which such information essentially promised the purchaser or user some characteristic which the product neither had nor was intended to have.

Although OSHA and CPSC impose safety considerations, the designer must recognize that there may be additional safety aspects which must be included in designing his product.

Consideration of human-engineering principles must proceed concurrently with each step of the design process. Each iteration implies a change in some aspect of configuration. Each configuration, in turn, may imply a decision which is crucial to operability and maintainability, two major aspects of the man-machine efficiency can be made more easily by modifying the machine rather than the man.

While the overall appearance of the finished product may have no influence on the capability of the device to perform the required functions, appearance must be considered. It is reasonably well established that if two machines are "identical" in performance characteristics, the one which has the greater aesthetic appeal will outsell the other. This aspect is especially difficult since there is no uniform pattern of aesthetic appeal.

Last, but certainly not least, is consideration of cost. In every case, the final configuration of a design depends on a number of compromises. One perpetual major determinant in these compromises is the matter of economic considerations. Industry has tended to optimize only that segment of the total life that deals with the incoming material through to the outgoing product (19, 20). There is growing recognition that the designer must assess the total economic consideration over the whole life cycle. This means assessing and including the cost of each factor, especially those relating to materials, energy use, and environmental impact.

Summary

This paper has presented a brief discussion of the factors which a designer must consider, explore, and use (or reject) in reaching a final design configuration for his product to enter the marketplace. These factors do not all have equal weight, nor do all of them apply in every design. All must be considered carefully, however, before being omitted in a given design. More satisfaction of some of these factors will mean less satisfaction of others. The designer, as in the past, must make tradeoff decisions. This task becomes increasingly complex. This paper tries to point out all the problems to be considered in developing a design. It makes no attempt to tell how to design. Good designing!

REFERENCES

1. Materials and Man's Needs, Summary Report of the Committee on the Survey of Materials Science and Engineering, National Academy of Sciences, 1973 (COSMAT) p. 57.

2. Material Needs and the Environment Today and Tomorrow, Final Report of the National Commission on Materials Policy, June 1973, (NCMP), pp. 4C-14.

3. Haugen, E. B., Probabilistic Approaches to Design, New York, John Wiley and Sons, 1968.

4. Smith, C. O., "Design Relationships and Failure Theories in Probabilistic Form," Nuclear Engineering Design, (1974), pp. 286-292.

5. Smith, C. O., "Never Overlook Notches," Proceedings, 21st Meeting, MFPG, NBS, Gaithersburg, MD, 7-8 November 1974.

6. NCMP, pp. 4C-16.

7. COSMAT, p. 56.

8. NCMP, pp. 3-11.

9. Azaroff, L. V., "Materials, Environment, Energy - One Vast System," Research/Development, March 1976, pp. 32-34, 36.

10. Metal Progress, September 1974, p. 50.

11. COSMAT, p. 147.

12. NCMP, pp. 4C-4.

13. NCMP, pp. 4C-15.

14. Resource Recovery and Source Reduction (SW-122), U. S. Environmental Protection Agency, 1974, p. 13.

15. SW-122, p. 15.

16. NCMP, pp. 5-3.

17. NCMP, pp. 10-13.

18. Metal Progress, September 1974, p. 45.

19. COSMAT, p. 78.

20. SW-122, pp. 109-112.

TABLE 1

General Categories of Factors in Design

Functional requirements
Total Life Cycle
Further Major Factors

TABLE 2

Functional Factors in Design

Performance specifications

- definition of need

Design Configuration

- probabilistic or deterministic approach
- stress considerations
- size/weight/volume
- reliability, maintainability, availability, and repairability
- failure anticipation
- value analysis

Redesign

- design review
- simplification and/or standardization
- functional substitution

TABLE 3

Total Life Cycle in Design

Material selection

- strength and/or ductility
- stability
- availability
- fabricability
- corrosion resistance
- properties of unique interest
- alternative materials

Producibility

Durability and/or recyclability

Energy requirements

- for production
- during use
- for reclamation

Environmental compatibility

- effect of product on environment
- effect of environment on product

Handling

- inspection and quality assurance
- testing
- packaging
- shipping and storage

TABLE 4

Further Major Factors in Design

State of the Art

- prior knowledge
- possible patent infringement

Conformance to Standards

- Occupational Safety and Health Administration
- Consumer Products Safety Commission
  - unintended uses
  - warnings
  - advertising
- Environmental Protection Agency
- American National Standards Institute
- National Safety Council
- American Society for Testing Materials
- Industry standards

Safety

Human Factors

- Ease of operating
- Ease of maintenance

Aesthetics

COST

- total economic consideration

SESSION I

TECHNIQUES

OF

DESIGN

(I)

Chairman: R. E. Maringer

Battelle,Columbus Laboratories

A SYSTEMS APPROACH TO MATERIALS IN DESIGN

Bruce P. Bardes
Bimba Manufacturing Company
Monee, Illinois 60449

Abstract: A systems oriented approach to the problems of materials selection and choice of manufacturing processes in engineering design is presented. The goal of this approach is to select the best combination of shape, material and manufacturing process by optimization with respect to several previously established criteria, such as performance, cost, availability of materials and manufacturing facilities and personnel. A flow chart is used to describe the process of evaluating the interactions between shape, material and process to determine the best combination. The flow chart also illustrates the use of prototype testing and reports of performance in service to provide feedback which may indicate changes in the design.

Key words: Engineering design; manufacturing processes; materials selection; systems approach to design.

All too frequently an object is designed with careful attention to its mechanical function or its esthetic appeal, but with little concern for the material from which it will be made or the manufacturing process by which it's made. A designer may have in mind the generic material (e. g. steel) and a manufacturing process (machine to size and heat treat), but specific choice of material and sequence of manufacturing processes may be neglected until a very late stage of the design process. The resulting object may perform satisfactorily, but it probably isn't a very efficient design, either in terms of performance or cost. This approach to design and materials selection invites patchwork redesigns to cover up shortcomings, should the object not perform as intended. In this paper, an approach of optimizing the combination of shape, material and manufacturing processes is described. The approach is not novel and it is used by good designers, consciously or otherwise. The reasons for describing this approach are to reinforce the thought processes used by successful designers and to emphasize the interrelation between shape, material and manufacturing process.

What the engineer finally designs to fill a particular function can have many forms, depending largely upon whatever constraints are placed upon his design. Identifying and describing the intended function of the object is the basic constraint on the design. Governmental regulations, interaction with other objects and availability of potential materials and manufacturing facilities are other examples of such constraints. In general, a designer has very little control over these constraints; they are derived from natural laws, public policy or management directive.

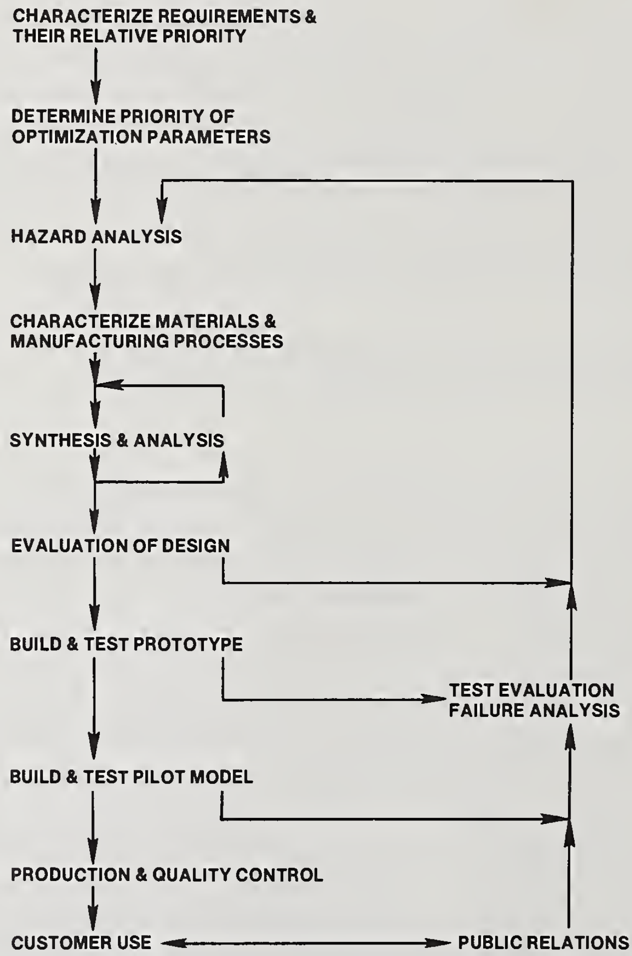


Figure 1. Flow chart depicting systems approach to design.

The engineer must order the priorities of the various factors which define the performance of the designed object. The most obvious performance criterion is the ability of the design to fulfill its basic function. In some cases the criterion is absolute; either the design works, or it doesn't. Particularly where service life is concerned, there may be degrees of performance acceptability. Cost, availability of materials and manufacturing facilities and esthetic appeal may be other criteria for evaluating the design; these criteria, in order of their relative importance, represent the parameters against which the design will be optimized. A flow chart, Figure 1, shows how identification of design constraints and performance parameters represent the first steps in engineering design.

Hazard analysis is the process in which a designer attempts to anticipate normal and abusive service conditions and the effect of these service conditions upon the object under consideration. Is fatigue likely to be a problem? Is impact loading likely? Will it be exposed to extreme temperatures or corrosive environments? What additional problems would be introduced by abusive service? The engineer must translate service hazards, as well as design constraints and performance criteria, into factors which are amenable to manipulation during design. This is often a difficult process, demanding careful engineering judgments.

Characterization of materials is another problem facing the designer. Unfortunately, readily measured and widely published data on tensile strength, percent elongation at fracture and chemical composition are not always useful in predicting whether or not a material can be successfully used or processed as desired. Careful engineering judgment is required to relate available test data to actual applications. Sometimes specialized testing which simulates the intended use or manufacturing process will provide more useful information than the usually published data.

Traditional usage of the word design is limited to the next step of the overall design process, formulating the shape of the object. Design may be based upon prior experience, or it may be done by synthesis or analysis. Relying upon the design of a similar object is convenient and often satisfactory, but there are cases where extrapolation from existing designs produces disastrous results. Synthesis is devising an object to produce a desired response to a given stimulus; analysis is predicting the response of a device to a given stimulus. Synthesis and analysis are closely related and complement each other. In practice design usually includes all of these techniques.

In principle, evaluation and optimization of a design could be accomplished by writing an equation which describes the performance as a function of all possible design, material and process variables, setting its partial derivative with respect to each optimization parameter equal to zero, and solving for the optimum conditions. In reality, that's impossible, so the designer must try to incorporate as many good features (relative to the various criteria for design) as possible. Evaluation of tentative designs begins during the previous design step; obviously inadequate designs are discarded, perhaps before they are even completed. Pencil and paper evaluation may suffice for simple objects, or objects whose design is thoroughly understood. For more complex designs, computer simulation or physical modeling may be useful. Changes can be made wherever necessary, and the evaluation repeated. Several iterations may be required. When the design appears to be reasonably close to optimum, a prototype is built. A frequent problem at this stage is that prototype materials and production methods are different from those planned for production. Similarly, prototype testing may not represent typical service conditions. However, the data obtained from prototype testing does provide the engineer with the first record of the performance of his design. This is an obvious opportunity to further improve the design, in retrospect, based on prototype performance. In the case of failure of the prototype, the engineer must identify the cause of the failure and take appropriate corrective action. The flow chart shows how evaluation and corrective action fit into the design process.

The manufacturing process must then be scaled up to pilot lots, then to full production. Early production units should be tested, both for quality control reasons and to compare their performance with that of the prototype. Any shortcomings in performance would suggest design changes.

Whether the product is used in its manufactured form or incorporated into some other product, the engineer probably loses contact with it. He must rely on service and sales personnel, who deal with the customer, to provide information concerning performance in service. In the case of a service failure, he must obtain from the user information about the circumstances of failure. Then he can attempt to isolate the cause of failure and determine what corrective action is warranted by the particular incident.

A small bore pneumatic cylinder used to punch holes in plastic film provides an example of how the systems approach to materials in design can function in practice. The perforated film is used in fabricating bags for fruits and vegetables. The piercing is accomplished while the film is moving by quickly extending and retracting a series of pneumatic cylinders.

When the piercing system was first constructed, stock cylinders were used. V-slots were milled into the stainless steel piston rods to act as cutters. Operation was acceptable, but there were several deficiencies. Locations of the several cylinders used in the system had to be changed frequently, depending upon the specified hole pattern, and the aluminum mounting threads would be worn out rather quickly. Because there are several cylinders operating simultaneously, the process was rather noisy. The piston rod was too soft to hold a cutting edge for more than a few hundred thousand operations, and the cutter sometimes became clogged with the circles punched out of the plastic film. The cutting edge could be reworked several times, but the useful life of the cylinders was limited to about a millon cycles.

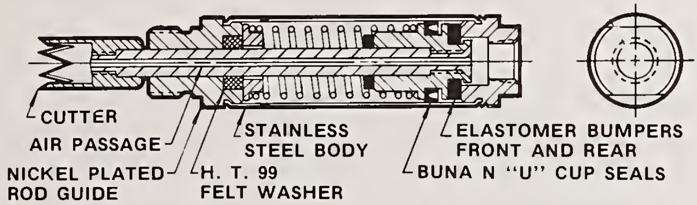


Figure 2. Cross-section showing internal construction of pneumatic cylinder designed for punching holes in plastic film.

Figure 3. Pneumatic cylinder designed for punching holes in plastic film. Cutter is part of piston rod.

The customer's report of deficiencies came to the cylinder manufacturer through a routine sales call. This information became the basis for a second cycle through the design process. See flow chart, Figure 1. The requirements imposed by the customer included sturdier mounting threads, some means of blowing the punched circles out of the cutter, quieter operation and a goal of 10 million operations without additional service. How these requirements were met is shown in Figures 2 and 3. A hard nickel plated steel piston rod guide provided corrosion resistance and a good bearing surface, but only when the bearing was well lubricated. Thus a lubricant reservoir became part of the design. To eject slugs from the cutter, the piston rod was made from small inside diameter tubing. For cutter diameters from 5/16" to 5/8", mild steel tubing, centerless ground and buffed to the required diameter and surface finish, was chosen for the piston rod . See Figure 2. The 1/4" diameter cutter, shown in Figure 3, was part of the piston rod; a free-machining medium carbon steel was gun drilled, then the end of the piston rod was hardened. Because all piston rods have the same diameter, regardless of cutter diameter, the jig used in grinding the cutters can be used for all cutter sizes. To reduce operating noise, elastomer bumpers were attached to both ends of the piston. The return spring was designed for the fastest possible cyclic operating rate, consistent with the largest possible space for the front bumper and long life without fatigue. A special lubricant was developed to reduce the need for a mist-lubricated air supply.

The cost of the redesigned cylinder is about twice that of the stock cylinder. But its service life has been more than five times that of the stock cylinder; no maintenance or resharpening of cutters has been necessary. The customer has been very pleased by the new cylinder. The manufacturer has found a substantial market for this product. By any standard, it has been successful. The success is due, at least in part, to the systematic approach to the redesign of the stock cylinder.

PROBABILISTIC METHODS IN DESIGN

Edward B. Haugen
The University of Arizona
Tucson, Arizona 85720

Abstract: It has sometimes been the case during the development of the various branches of science and engineering that, on looking back, initial solutions turned out to be special cases of more general problems. Such has been the case with mechanical design and analysis.

Until recently, many of the implications arising from the probabilistic nature of engineering phenomena were not fully realized, even though with the announcement of "Heisenberg's uncertainty principle," the whole of science (including engineering) was recognized as ultimately based philosophically on the concepts of experimental probability. A partial explanation is that some of the necessary supporting science and technology had not, until recently, sufficiently evolved to provide a base for a probabilistic design theory and methodology.

Beginning with a probability based design theory and reducing the variability (in the design variables) to the vanishing point, the residue amounts to a special case, i.e., a theory based on the deterministic assumption. It is essentially this special case into which classical deterministic design theory in its main features has evolved. Over the past several centuries an enormous methodology and supporting literature has been developed around the special case of assumed determinism.

Although in many classes of engineering problems the variables have been identified and the forms of their functional relationships correctly described the behavior of many individual variables has until recently remained imperfectly understood. In addition, probability mathematics, for coping with functions of random variables, has only recently been developed in forms lending themselves to the support of design synthesis and analysis.

In this paper, the focus of discussion is on a probabilistic design theory and methodology. Several points of theory in which the assumptions of determinism has led to untenable statements are examined. Other points are identified, whose existence under deterministic assumptions were not suspected.

The task undertaken here is to discuss points of theory and methodology such that the practicing engineer can readily appreciate its power. Design examples are given, to a reliability specification, in which both over and under designing is avoided and variability information is used not suppressed.

Key words: Probability; reliability; random variable; algebra of normal functions; correlation.

In a general sense the satisfactory performance of an engineered component or device in its use environment imposes the requirement that for a defined mission:

strength > stress = S > s .

Except in the simplest cases S and s are each functions of more than a single variable. Thus, usually:

$$S = f(x_1, x_2, \ldots, x_n) \tag{1}$$

and

$$s = g(y_1, y_2, \ldots, y_r) \tag{2}$$

Mathematical analysis and engineering science provide the vehicle for developing equations (1) and (2).

These models, however, are mute concerning the nature of the individual x's and y's. If the real conditions were such that each variable $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_r$ was single valued (as assumed in the past) and the relationships of these variables to the variables S and s were described by the functions f and g, then the transformation from $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_r$ to S and s would be single values. If, however, the variables $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_r$ were each multivalued, then the transformations to S and s would be distributions of probable values of S and s, the actual situation in engineering.

Since the solution of engineering problems will involve functions of random variables, mathematically tractable models of random variables and functions of random variables must be found. Certain familiar two parameter distributions (normal, log normal, gamma, Weibull, etc.) provide models that are mathematically simple and by means of which random variables are uniquely defined by their mean values (μ) and standard deviations (σ). The mean values and standard deviations of functions of such random variables and pairs of random variables (x and y) are also easily calculated. Given the independent random variables x and y, defined by statistics μ_x and σ_x and by μ_y and σ_y, the mean (μ_z) and standard deviation σ_z of the functions $z = h(x,y)$ are calculated as follows:

Function	Mean, μ_z	Standard Deviation, σ_z	
$a\ x$	$a\ \mu_x$	$a\ \sigma_x$	(1)
$a + x$	$a + \mu_x$	σ_x	(2)
x^2	$\mu_x^2 + \sigma_x^2$	$\sqrt{\alpha\mu_x^2\ \ \sigma_x^2 + 2\ \sigma_x^4}$	(3)
$x + y$	$\mu_x + \mu_y$	$\sqrt{\sigma_x^2 + \sigma_z^2}$	(4)
$x\ y$	$\mu_x\ \mu_x$	$\sqrt{\mu_x^2\ \sigma_z^2 + \mu_z^2\ \sigma_x^2 + \sigma_x^2\ \sigma_y^2}$	(5)
x/y	μ_x/μ_z	$\sqrt{\dfrac{\mu_x^2\ \sigma_z^2 + \mu_z^2\ \sigma_x^2}{\mu_z^4}}$	(6)

Functions of two independent random variables are often observed in engineering. Such functions are illustrated in examples 1, 2 and 3:

Example 1: A particular spur gear tooth is subject to a peak force, F, inducing at each revolution a tensile stress (34,000; 2,600) psi. Production processing has introduced a compressive residual stress (11,600; 1,500) psi, at the base of the tooth. What are the effective applied stress statistics, $\bar{s}$ and s_s?

Effective stress is the difference between the applied tensile and residual compressive stresses. Thus,

$$\bar{s} = 34{,}000 - 11{,}600 = 22{,}400 \text{ psi}$$

$$s_s = \sqrt{(2{,}600)^2 + (1{,}500)^2} \simeq 3{,}000 \text{ psi}$$

Example 2: If a bending equation were to be applied to make a rough estimation of the stress in a gear tooth, the moment at the weakest section would be given by $M = F_b \ell$. Given:

$$(\bar{F}_b,\ s_{F_b}) = (3{,}800;\ 460) \text{ lb.}$$

$$(\bar{\ell},\ s_\ell) = (0.730,\ 0.016) \text{ in.}$$

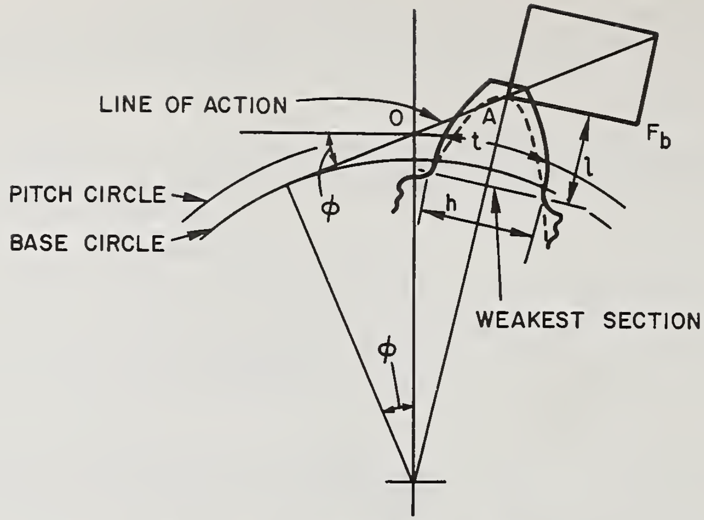


Figure 1. Beam Model of Gear Tooth

The statistics of M, assuming F_b and ℓ statistically independent, are:

$\bar{M}$ = 3,800   0.730 = 2774 in /lb.

$$s_M = [(3{,}700)^2 \cdot (0.016)^2 + (0.730)^2 \cdot (460)^2 + (460)^2 (0.016)^2]^{1/2}$$

$$= 346 \text{ in /lb.}$$

Example 3: $(\bar{s}, s_s)$ = (15,000; 1,600) psi are the statistics of stress in the material of a component having a modulus of elasticity (E, s_E) = $(29 \cdot 10^6, 3.3 \cdot 10^5)$ psi. If the model for elastic strain is

$\varepsilon = \frac{s}{E}$, estimate the statistics of ε.

$$\bar{\varepsilon} \simeq \frac{\bar{s}}{\bar{E}} = \frac{15{,}000}{29 \cdot 10^6} = 5.2 \quad 10^{-4}.$$

$$s_\varepsilon \simeq \sqrt{\frac{\bar{s}^2 s_E^{\,2} + \bar{E}^2 s_s^{\,2}}{\bar{E}^4}}$$

$$s_\varepsilon \simeq \sqrt{\frac{(1.5 \cdot 10^4)^2 (3.3 \cdot 10^5)^2 + (29 \cdot 10^6)^2 (1.6 \cdot 10^3)^2}{(29 \cdot 10^6)^4}} = 5.6 \cdot 10^{-5} .$$

The mean value and standard deviation of arbitrary functions of independent random variables can be estimated by a Taylor series expansion. Given the function $\psi = f(x_1, x_2, \ldots, x_m)$, expand ψ about $\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_m}$. Then

$$\mu_\psi \simeq f[\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_m}] \tag{7}$$

and

$$\sigma_\psi \simeq \left[\sum_{i=i}^{n} \left(\frac{\partial\psi}{\partial x_i}\right)^2 \sigma_{x_i}^{\;2}\right]^{1/2} \tag{8}$$

Equation (8) provides a valuable tool with which to draw sensitivity inferences, because the terms $(\partial\psi/\partial x_i)^2 \cdot \sigma x_i^{\;2}$ disclose the contributions of each x_i to the magnitude of σ_ψ. For instance, it is desirable to minimize the variability in applied stress, s. An examination of the expression for σ_s will disclose which variables can most profitably be closely controlled. Further, the tolerances on some variables (producing negligible effects) may be relaxed.

<u>Example 4</u>: Examine the sensitivity of the standard deviation on stress to the contributions due to applied force and the dimensional random variables. Consider the cantilever beam formula (rough approximation) sometimes used to model stress(es)in a spur gear tooth:

$$s = \frac{6W_t \cdot \ell}{Ft^2} \text{ lb/in.}^2 .$$

W_t = force = $(\overline{W}_t; \sigma_{W_T})$ = (5400; 360) lb.

ℓ = dimension = $(\overline{\ell}; \sigma_\ell)$ = (1.30; .040) in. (a)

F = face width = $(\overline{F}; \sigma_F)$ = (.625; .025) in.

t = root thickness = $(\overline{t}; \sigma_t)$ = (.75; .035) in.

The partial derivatives of s with respect to W_t, ℓ, F, and t evaluated at the mean values are:

$$\frac{\partial s}{\partial W_t} = \frac{6\overline{\ell}}{\overline{F}\overline{t}^2} = 22$$

$$\frac{\partial s}{\partial \ell} = \frac{6\overline{W}_t}{\overline{F}\overline{t}^2} = 92{,}160 \tag{b}$$

$$\frac{\partial s}{\partial F} = \frac{-6\overline{W}_t\overline{\ell}}{\overline{F}^2\overline{t}^2} = -\ 191,690$$

$$\frac{\partial s}{\partial t} = \frac{-12\overline{W}_t\overline{\ell}}{\overline{F}\overline{t}^3} = -\ 319,490$$

Utilizing equation (8):

$$\sigma_s \simeq \left[\left(\frac{\partial s}{\partial W_t}\right)^2 \cdot \sigma_{W_t}^{\ 2} + \left(\frac{\partial s}{\partial \ell}\right)^2 \cdot \sigma_\ell^{\ 2} + \left(\frac{\partial s}{\partial F}\right)^2 \cdot \sigma_F^{\ 2} + \left(\frac{\partial s}{\partial t}\right)^2 \cdot \sigma_t^{\ 2}\right]^{1/2}$$

$$\sigma_s \simeq \left[\left(\frac{6\overline{\ell}}{\overline{F}\overline{t}^2}\right)^2 \cdot \sigma_{W_t}^{\ 2} + \left(\frac{6\overline{W}_t}{\overline{F}\overline{t}^2}\right)^2 \cdot \sigma_\ell^2 + \left(\frac{-6\overline{W}_t\overline{\ell}}{\overline{F}^2\overline{t}^2}\right)^2 \cdot \sigma_F^{\ 2} + \left(\frac{-12\overline{W}_t\overline{\ell}}{\overline{F}\overline{t}^3}\right)^2 \cdot \sigma_t\right]^{1/2} . \qquad \text{(c)}$$

Substituting values (b) into (c):

$$\sigma_s \simeq [6.39 \times 10^7 + 1.36 \times 10^7 + 2.30 \times 10^7 + 12.5 \times 10^7]^{1/2}$$

$$= 15,020 \text{ psi.} \qquad \text{(d)}$$

$$\overline{s} = \frac{6\overline{W}_t \cdot \overline{\ell}}{\overline{F}\overline{t}^2} = 119,808 \text{ psi.}$$

Equation (d) discloses that the variability in t is the largest contributor to variability in s. Elimination of variability in t would result in a reduction in σ_s to approximately 10,000 psi.

It is frequently the case that two engineering random variables may be correlated, i.e., that the random value of one or more variables in a function is related to the random value of one or more other variables in the same function. The statistics of functions of correlated random variables may be very different than the statistics of identical but statistically independent random variables. Clearly, in design the requirements of a component or system will be dependent on the presence or absence of correlations among the relevant engineering variables. The following formulas are those required to estimate the statistics of correlated random variable pairs.

Function	Mean μ_z	Standard Deviation σ_z
$z = x + y$	$\mu_x + \mu_y$	$\sqrt{\sigma_x^2 + \sigma_z^2 + 2_\rho \sigma_x \sigma_y}$
$z = x - y$	$\mu_x - \mu_y$	$\sqrt{\sigma_x^2 + \sigma_y^2 - 2_\rho \sigma_x \sigma_y}$
$z = xy$	$\mu_x\mu_y + \rho\sigma_x\sigma_y$	$\sqrt{\mu_x^2\sigma_y^2+\mu_y^2\mu_x^2+\sigma_x^2\sigma_z^2+2\rho\sigma_x\sigma_z+\rho\sigma^2\sigma_x^2\sigma_y^2}$
$z = x/y$	$\frac{\mu_x}{\mu_y}[1+(\frac{\sigma_y}{\mu_y})][\frac{\sigma_y}{\mu_y} - \rho\frac{\sigma_x}{\mu_x}]$	$\frac{\mu_x}{\mu_y} \cdot [\frac{\sigma_x^2}{\mu_x^2} + \frac{\sigma_y^2}{\mu_y^2} - 2\rho\frac{\sigma_x\sigma_y}{\mu_x\mu_y}]^{1/2}$

Example 5: Difference of Correlated Random Variables. A necessary condition for equilibrium is that (assumed) deterministic forces acting in any direction must sum to zero:

$$\Sigma F = 0. \quad \text{(a)}$$

The probabilistic analogue of equation (a) is:

$$\Sigma(\overline{F}, \sigma_F) = (0, 0). \quad \text{(b)}$$

What are the necessary and sufficient conditions for equilibrium, that satisfy equation (b) for a random pair of interacting normally distributed force random variables?

The required conditions are:

1. $\overline{F}_1 - \overline{F}_2 = 0$

2. $\sigma_{F_1} = \sigma_{F_2}$

3. F_1 and F_2 correlated, with $\rho = 1$.

If F_1 and F_2 are dynamic loads, they must be in phase.

Example 6: Product of Correlated Random Variables. When a bar is loaded in tension by a force, F, the strain, δ, is induced. Strain energy, U, is:

$$\text{Work} = \text{total energy} = U = \frac{F\delta}{2}.$$

It is assumed that F and δ are linearly related, hence correlated, $\rho = +1$. Knowing the distribution of F, the distribution of δ is defined. The statistics of U are:

$$\bar{U} = (1/2)(\overline{F\delta} + \rho\,\sigma_F\,\sigma_\delta).$$

$$\sigma_U = (1/2)[\bar{F}^2\sigma_\delta^{\,2} + \bar{\delta}^2\,\sigma_F^{\,2} + \sigma_F^{\,2}\cdot\sigma_\delta^{\,2} + 2\rho\overline{F\delta}\,\sigma_F\sigma_\delta + \rho^2\sigma_F^{\,2}\,\sigma_\delta^{\,2}]^{1/2}.$$

Example 7: Quotient of Correlated Random Variables. The interaction between loading and deformation is important in mechanical design. Loading intensity, psi, called stress, s, is a random variable, $(\bar{s},\sigma_s)$. Also, deformation, in./in., called strain, ε, is a random variable $(\bar{\varepsilon},\sigma_\varepsilon)$.

A unique feature of truly elastic behavior is that deformation in the units above is a linear function of load intensity, since strain is a linear function of stress, they will be correlated, $\rho = +1$.

$E = s/\varepsilon$, the ratio of two correlated random variables, is a relationship called Hooke's Law, where:

$$\bar{E} \simeq \frac{\bar{s}}{\bar{\varepsilon}}\left[1 + \left(\frac{\sigma_\varepsilon}{\bar{\varepsilon}}\right)\cdot\left(\frac{\sigma_\varepsilon}{\bar{\varepsilon}}\right) - \rho\,\frac{\sigma_s}{\bar{s}})\right].$$

and

$$\sigma_E \simeq \frac{\bar{s}}{\bar{\varepsilon}}\left[\frac{\sigma_\varepsilon^{\,2}}{\bar{\varepsilon}^2} + \frac{\sigma_s^{\,2}}{\bar{s}^2} - 2\rho\,\frac{\sigma_\varepsilon\sigma_s}{\bar{\varepsilon}\cdot\bar{s}}\right]^{1/2}.$$

Example 8: Reliability Calculation. Given the normal random variables describing strength and stress in an eye-bar:

Strength = $(\bar{S}, \sigma_S)$ = 27,000; 3,200) psi.

Stress = $(\bar{s}, \sigma_s)$ = 18,400; 1,500) psi.

Compute: R and P_f.

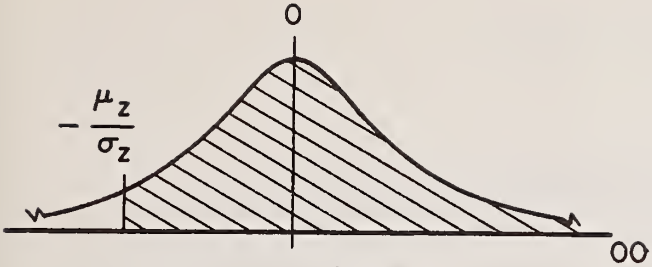


Figure 2

$$R = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\mu_z}{\sigma_z}}^{\infty} e^{-t^2/2}\, dt \;.$$

With S and s assumed to be normally distributed, reliability (R) is given by the area under the curve in the figure above. The integration is achieved by utilizing standard normal area tables.

(1) S and s statistically independent, with $\rho = 0$.

Write the difference expression $Z = S - s$.

$$\bar{z} = 27{,}000 - 18{,}400 = 8{,}600 \text{ psi.}$$

$$\sigma_z = \sqrt{(3200)^2 + (1500)^2} = 3{,}534 \text{ psi.}$$

The transformation of the random variable Z to the standardized normal R.V.t is:

$$t = \frac{z-\bar{z}}{\sigma_z} = \frac{0-\bar{z}}{\sigma_z} = \frac{-\bar{z}}{\sigma_z} \;.$$

Integration from 0 to ∞ is required, thus $t = \frac{-8600}{3534} = -2.43$.

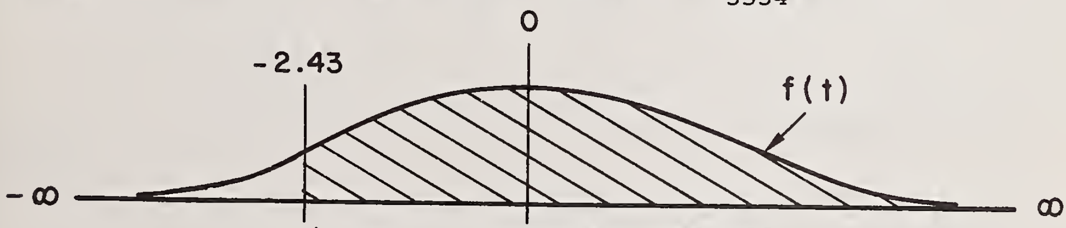


Figure 3. Standard Normal pdf

Utilizing standard normal area tables:

$$R = \int_{-2.43}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}\, dt = 0.9925, \text{ and } P_f = 0.0075.$$

(2) S and s correlated, with $\rho = +1$.

$\bar{z} = 8{,}600$ psi;

$$\sigma_z = \sqrt{(3200)^2 + (1500)^2 - 2(3200)(1500)} = 1{,}610 \text{ psi.}$$

With limits: $t = -8{,}600/1{,}610 = -5.34$ to ∞, integrating as in (1):

$R = 0.9999995$, and $\rho_f = 0.0000005$.

Recognition of the real nature of design variables has led to new theory, revisions to older theory and has opened a Pandora's box of incidental consequences in mechanical design. The end is not within sight.

The remaining pages of this paper will be devoted to mention of several non-trivial considerations arising from unique characteristics of random variables.

Engineering random variables do not constitute orderable sets. Thus, it is important in engineering to carefully scrutinize all inequality relationships.

The reason for the scrutiny is that with random variables it cannot be said that one random variable is larger or smaller than another.

It can be seen that these remarks apply in particular to the concept of principal stresses. These are usually designated as s_1 and s_2, acting in the directions of axes 1 and 2 on the planes perpendicular to which shear stresses are nominally zero. Referring to Figure 4 below the normal stress s_1 has been referred to as the maximum principal stress. The normal stress s_2 has been referred to as the minimum principal stress. Further, there is the suggestion that $s_1 > s_2$. The functional models for stresses s_1 and s_2 may be written:

$$s_1 = OA = OC + CD = \frac{s_x + s_y}{2} + \sqrt{\left(\frac{s_x - s_y}{2}\right)^2 + \tau^2}$$

$$s_2 = OB = OC - CD = \frac{s_x + s_y}{2} - \sqrt{\left(\frac{s_x - s_y}{2}\right)^2 + \tau^2}$$

$$|\tan 2\phi| = \frac{DE}{CE} .$$

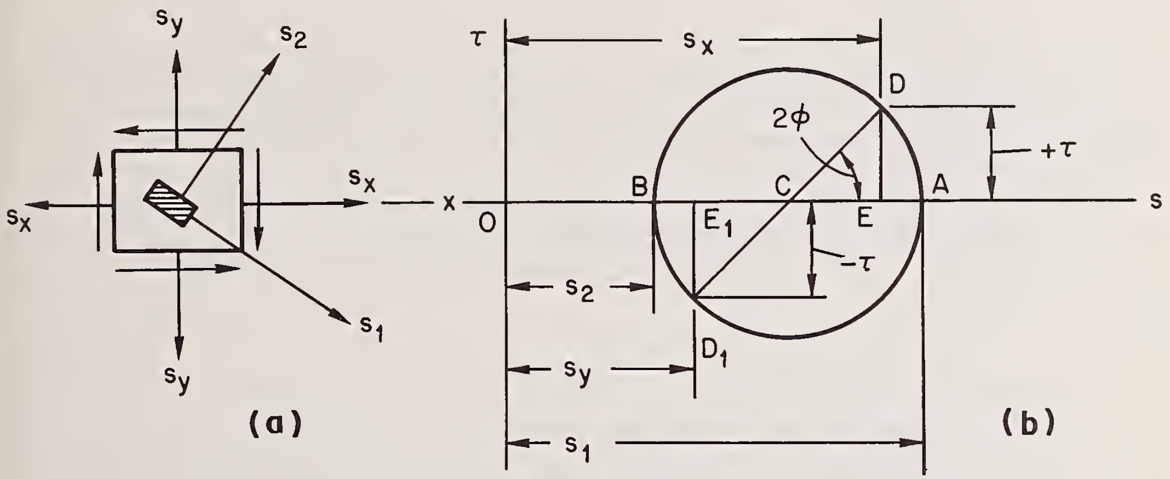


Figure 4. General Conditions of Two Dimensional Stress

The mean value estimators are $\bar{s}_1$ and $\bar{s}_2$, and the inequality, $\bar{s}_1 > \bar{s}_2$, is a meaningful statement. It is permissible to make statements concerning the maximum (or minimum) mean values of principal stress. The stress statistics, $\bar{s}_x$ and $\bar{s}_y$, cannot, however, be considered principal, because not only normal but also shearing stresses are acting on the planes perpendicular to the x and y axes.

$$\tau_{max} = \frac{s_1 - s_2}{2} = \sqrt{\left(\frac{s_x - s_y}{2}\right)^2 + \tau^2} .$$

The mean value and standard deviation of τ_{max} are calculated, using equations (7) and (8).

In the 19th century it was generally accepted in engineering that: (1) stresses were deterministic, and (2) principal stresses were orderable. These were unstated beliefs when Rankline (1802-1972) postulated the "Maximum Normal Stress" theory. However, the considerations above have shown such propositions to be untenable, in a strict sense.

Classical maximum normal stress theory postulated failure whenever $s_1 > S_y$ or $s_1 > S_u$, whichever was applicable. Loading was assumed deterministic static, and s_1 was the largest of the (deterministic) principal stresses.

In the uni-axial stress case, $s_2 = s_3 = 0$, and s_1 is a simple (random) variable. The maximum normal stress theory can be effectively employed in probabilistic design of many component types, such as connecting rods, frame members, valves, leaf springs, etc., involving ductile materials, subject to:

(1) A probabilistic requirement to avoid yield:

$$R(t) = p\cdot[S_y > s_1].$$

(2) A probabilistic requirement to avoid tensile fracture:

$$R(t) = p\cdot(S_u > s_1).$$

(3) A probabilistic requirement to avoid shear yielding:

$$R(t) = p(S_{sy} = 0.577\ S_y > s_1).$$

In the usual situation in design, s_1, s_2, and s_3 must be treated as random variables. Thus, probabilistically the inequality $s_1 > s_2 > s_3$ is usually, in a strict sense, without meaning. The following possible triaxial stress cases must, however, be considered:

1. s_1, s_2, and s_3 are statistically independent.

2. s_1, s_2, and s_3 are correlated.

If s_1, s_2, and s_3 are independent, $\rho = 0$, then since the ranges are $0 \leq s_i < \infty$, there will be a finite probability that $s_2 \geq s_1$ and even $s_3 \geq s_1$.

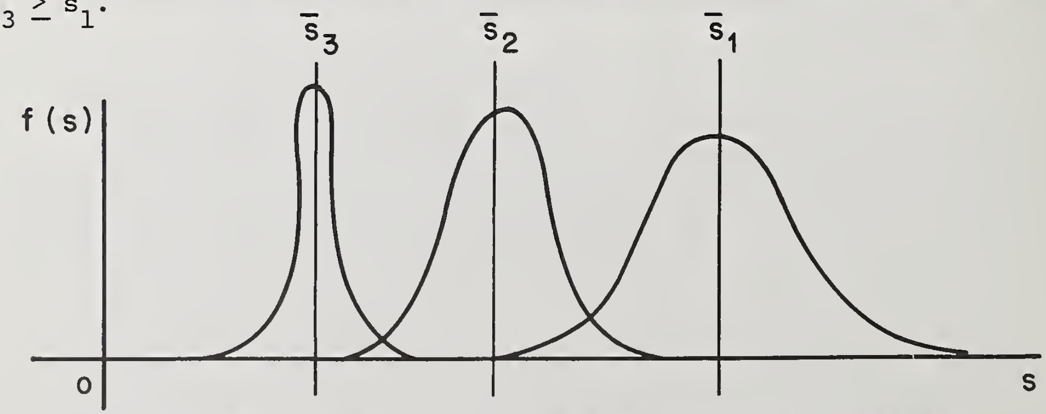


Figure 5. Ordering of Correlated Principal Stresses

If the principal stresses are independent, the only possible ordering is usually that of the mean values:

$\bar{s}_1 > \bar{s}_2 > \bar{s}_3$.

The scatter relationships not being ordered, component probability of failure must be estimated for each axial direction, and summed.

Considering biaxial applied stress, the strength of a material $\bar{S}_{u1}$ in the direction of mean principal stress $\bar{s}_1$ is very likely not the same as $\bar{S}_{u2}$ in the direction of mean principal stress $\bar{s}_2$. Assume that $\bar{S}_{u1} > \bar{S}_{u2}$ and that S_{u1} and S_{u2} are correlated, $\rho = + 1$, in the following example.

Example 9:

The figure above suggests that for high intensity stress excursions (larger than the mean values), the ordering $s_1 > s_2$ will be valid. For very low stress excursions, there will be a calculable probability that $s_1 < s_2$.

It is to be noted that the distortion energy theory of failure is not dependent upon an assumption of ordered principal stresses.

The importance of recognizing the presence of correlation versus statistical independence has previously been noted in this paper. The following example will reinforce this point.

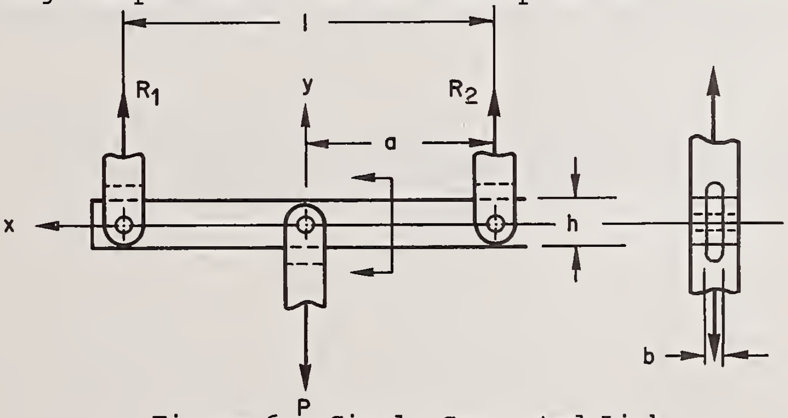


Figure 6. Simply Supported Link

Example 10:

If the reactions, R_1 and R_2, at the supports of a link (modeled as a simple beam) are to be calculated, considerable insight might be gained by considering correlation. In Figure 6, P, a, and ℓ are random variables. If $c = a/\ell$, then, $R_1 = cP$; $R_2 = (1 - c)P$.

If $\bar{c} >> s_c$ and $s_c/\bar{c} << 1$, c may be treated as an approximate constant $(\bar{c}, \sim 0)$. R_1 and R_2 can be considered correlated with P, with correlation coefficient $\rho \simeq 1$. Further R_1 and R_2 will have distributions of approximately the same general type as P, since the transformations from the distribution of P to the distributions of R_1 and R_2 would be nearly linear.

The unique considerations arising from the statistical and probabilistic nature of engineering phenomena, discussed in this paper, are far from exhaustive. Clearly probabilistic treatment will bring important changes to design practice. However, much important developmental work remains to be done.

References

1. Haugen, E. B., Probabilistic Approaches to Design, John Wiley & Sons, publ., New York, 1968.

2. Haugen, E. B., "Probabilistic Mechanical Design," Aerospace and Mechanical Engineering, The University of Arizona, 1976, to be published.

3. Haugen, E. B., and P. H. Wirsching, "Probabilistic Design Alternative to Miner's Cumulative Damage Rule," Proceedings of the Annual Reliability and Maintainability Symposium, Philadelphia, Pennsylvania, 1973.

4. Wirsching, P. H., and E. B. Haugen, "Probabilistic Design of Random Fatigue Loads," Journal of the Engineering Mechanics Division, ASEC, Vol. 99, No. EM6, December 1973.

5. Wirsching, P. H., and E. B. Haugen, "A General Statistical Model for Random Fatigue," Journal of Engineering Materials and Technology, ASME, Paper No. 73-Mat/EE, December 1973.

SESSION II

TECHNIQUES OF DESIGN (II)

Chairman: B. P. Bardes
Bimba Manufacturing Company

THE FINITE ELEMENT METHOD IN STRUCTURAL ANALYSIS

Harry G. Schaeffer
Associate Professor
University of Maryland
College Park, Maryland 20740

Introduction

The finite element method is a technique that allows the analyst to represent a complex structural system by means of an assemblage of basic building blocks such as those shown by Figure 1. These basic building blocks are termed finite elements and include:

- Truss-type elements that resist axial load and torsion
- Beam-type elements that resist bending, shear, axial load and torsion
- Membrane elements that resist plane loads
- Plate bending elements that resist bending moments
- Three-dimensional elements

By utilizing these elements the analyst can represent such complicated structures as an entire aircraft, a ship, or a nuclear containment vessel as shown by Figures 2, 3 and 4, respectively.

The technology represented by computer codes for finite element analysis was made possible by the almost simultaneous developments of:

- Large general purpose computers such as the IBM 704 and the UNIVAC 1102
- The concept of the finite element in the late 1950's
- Computer science advances including:
    - higher level languages
    - efficient computers
    - data base management

- Numerical analysis including:
    - efficient numerical algorithms for solving large sets of equations
    - efficient eigenvalue extraction techniques
    - efficient solution of ordinary differential equations

Today this technology is available in large computer codes which are called "general purpose finite element programs". These programs allow the analyst to model structural systems having thousands of degrees of freedom.

The purpose of this paper is to:

- Present an overview of the finite element method
- Discuss the technology incorporated in the current general purpose finite element programs
- Discuss the impact of current developments in computer hardware, computer science, numerical analysis and engineering mechanics on future programs
- Discuss the impact of the finite element method on automated design synthesis.
- Discuss the engineer's and the manager's responsibility in design certification using a general purpose program.

Overview of the Finite Element Method

The bases of the finite element method are:

- The replacement of the continuous structure by a set of finite elements which join each other only at a discrete number of points called nodes as shown by Figure 5.
- The representation of the displacement of the element in terms of the node point displacements

The element displacement is taken in the form

$$\underline{v} = \underline{N}\,\underline{\Delta} \tag{1}$$

where

$\underline{v}$ is the displacement vector having component u, v, w in the x, y and z-directions

$\underline{\Delta}$ is node point displacements

$\underline{N}$ is interpolation functions

Development of Element Stiffness Equations

Our goal is the development of a set of algebraic equations which relate the element node point displacements to the node point forces in the following form

$$\underline{k}_{(e)}\ \underline{\Delta}_{(e)} = \underline{P}_{(e)} \qquad (2)$$

where

$\underline{k}_{(e)}$ is the stiffness matrix for element (e)

$\underline{\Delta}_{(e)}$ is the set of all node point displacements for element (e)

$\underline{P}_{(e)}$ is the set of all nodal forces for element (e)

In order to derive the desired set of relations we start with the principle of virtual work for the element which states that if the virtual work is equal to zero for any possible virtual displacement in harmony with the constraints the system is in equilibrium. We can thus use the principle to obtain equilibrium equations provided we can express the virtual work in terms of virtual displacements of the element nodes.

The principle of virtual work for the element is expressed mathematically as follows

$$\delta U + \delta V = 0 \qquad (3)$$

where, δU is the virtual work of internal forces and δV is the virtual work of external forces.

The virtual work of the internal forces can be expressed in terms of stresses $\underline{\sigma}$ and strains $\underline{\varepsilon}$ as follows

$$\delta U_{(e)} = \int \underline{\sigma}^{T}_{(e)}\, \delta\underline{\varepsilon}_{(e)}\, dv_{(e)} \qquad (4)$$

And, the virtual work of the external forces can be expressed in terms of node point forces $\underline{P}_{(e)}$ and node point displacements, $\underline{\Delta}_{(e)}$ as

$$\delta V_{(e)} = -\ \underline{P}^{T}_{(e)}\ \delta\underline{\Delta}_{(e)} \qquad (5)$$

Representation of Strains in Terms of Nodal Displacements

The strains are related to displacements through the well known strain-displacement relations of the form

$$\varepsilon_{xx} = \frac{\partial u}{\partial x}; \qquad \varepsilon_{yy} = \frac{\partial v}{\partial y}; \qquad \varepsilon_{zz} = \frac{\partial w}{\partial z}$$

$$\varepsilon_{xy} = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}; \qquad \varepsilon_{yz} = \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}; \qquad \varepsilon_{zx} = \frac{\partial w}{\partial x} + \frac{\partial u}{\partial z} \tag{6}$$

If we define the set of strain components $\underline{\varepsilon}$, as

$$\underline{\varepsilon} = (\varepsilon_{xx} \quad \varepsilon_{yy} \quad \varepsilon_{zz} \quad \varepsilon_{xy} \quad \varepsilon_{yz} \quad \varepsilon_{zx}) \tag{7}$$

Then:

$$\underline{\varepsilon} = \underline{D}\ \underline{v} \tag{8}$$

where,

$$\underline{D} = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 \\ 0 & \frac{\partial}{\partial y} & 0 \\ 0 & 0 & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \end{bmatrix} \tag{9}$$

and where $\underline{v}$ is the displacement vector

Now we can relate the strains to the nodal displacements by using the approximate displacement functions given by equation (1)

$$\underline{v}_{(e)} = \underline{N}_{(e)}\ \underline{\Delta}_{(e)}$$

where we note that the shape functions, $\underline{N}_{(e)}$ are functions of position; and that the node point displacements are displacements at specific points in the structure.

The substitution of (1) into (8) then gives:

$$\underline{\varepsilon}_{(e)} = \underline{D}\ \underline{N}_{(e)}\ \underline{\Delta}_{(e)} \tag{10}$$

The virtual strains are then related to virtual displacements of the node points by

$$\delta\underline{\varepsilon}_{(e)} = \underline{b}_{(e)}\ \delta\underline{\Delta}_{(e)} \tag{11}$$

where

$$\underline{b}_{(e)} = \underline{D}\ \underline{N}_{(e)} \tag{12}$$

Constitutive Relations

The stresses, $\underline{\sigma}$, and the strains, $\underline{\varepsilon}$, are related by a generalized Hooke's law as follows

$$\underline{\sigma} = \underline{E}\ \underline{\varepsilon}$$

Since there are six strain and six stress quantities, $\underline{E}$ is seen to be a 6-by-6 matrix. The elasticity matrix, $\underline{E}$, is symmetric, and thus has at most 21 material constants. Most general purpose finite element programs assume this general relationship between stress and strain, but provide some convenient means for an analyst to define orthotropic or isotropic materials.

Element equations

We now have all the ingredients to define the element equilibrium equations.

- Principle of virtual work

$$\int \underline{\sigma}^T_{(e)}\ \delta\underline{\varepsilon}_{(e)}\ dv_{(e)} - \underline{P}^T_{(e)}\ \delta\underline{\Delta}_{(e)} = 0$$

with

$$\underline{\varepsilon} = \underline{D}\ \underline{v}$$

$$\underline{v} = \underline{N}\ \underline{\Delta}$$

$$\delta \underline{\varepsilon} = \underline{b} \quad \delta \underline{\Delta}$$

$$\underline{\sigma} = \underline{E} \; \underline{\varepsilon}$$

- $$\int \underline{\Delta}^T \; \underline{b}^T \; \underline{E} \; \underline{b} \quad \delta \underline{\Delta} \; dv_{(e)} - \underline{P}^T_{(e)} \; \delta \underline{\Delta}_{(e)} = 0$$

$$(\underline{\Delta}^T_{(e)} \; \underline{k}_{(e)} - \underline{P}^T_{(e)}) \; \delta \underline{\Delta}_{(e)} = 0$$

- $$\underline{k}_{(e)} \; \underline{\Delta}_{(e)} = \underline{P}_{(e)} \tag{13}$$

where

$$\underline{k}_{(e)} = \int \underline{b}^T \; \underline{E} \; \underline{b} \; dv_{(e)} \tag{14}$$

We see that the systematic application of principles of engineering mechanics allows us to derive an expression for the stiffness matrix for a general element. Looking back over this process we would note that we have:

- Made a specific approximation for displacement field through choice of $\underline{N}$
- Assumed linear strain-displacement relations
- Assumed linear stress-strain relation

Element Libraries

General purpose finite element programs contain libraries of element stiffness matrices which may be specified by the analyst to represent a complex structure. A library of structural elements that is contained in a typical general purpose program includes those elements shown by figure 1. These elements allow the analyst to model a structure as an assemblage of scalar (spring-type) elements, truss, beam, membrane plate or solid elements.

Modelling Structures Using General Purpose Codes

The general purpose finite element programs incorporate technology that allows the user to prepare a mathematical model in the following steps.

- Define node point locations at which displacement variables are to be determined

- Define the element connectivity; i.e. specify the type of element and the node points that it connects

- Define the element properties

- Define the material

- Prescribe the loads which may be:

    - Node point forces

    - Enforced element deformations

    - Enforced node point deformations

    - Thermal gradients

- Define constraint relationships

By using the user defined input data the general purpose program then performs the following operations which are transparent to the user.

I. Input Phase

- Reads input data and checks for syntax errors

- May resequence element numbers or node numbers to optimize bandwidth

II. Generation Phase

- Generate the elemental stiffness matrices

- Transform stiffness from local to global coordinates

- Add elemental stiffness to form the system matrix

- Incorporate displacement constraints

- Calculate loads

III. Solution Phase

- Perform matrix decomposition

- Perform forward sweep-backward substitution for each load case

IV. Output Phase

- Calculate derived behavioral variables
- Print results

The specification of input data together with any control parameters that are required for a specific program will cause the program to execute and produce results -- and therein lies a real problem. There is a tremendous amount of technology involved in the program requiring a knowledge of the following areas

- The theory of Finite Elements
- The theory of Matrix Structural Analysis
- Numerical Analysis
- Solid Mechanics

The use of the general purpose program is not a turnkey operation that may be intrusted to a junior engineer. Indeed, the validity of the results obtained from the program depend on the expertise of the user to prepare a valid model, and to interpret the results correctly. The results of any program should be viewed with skepticism, and the engineer must utilize all of the tools at his disposal to make sure that the results are reasonable . . . generally the "tools" are engineering judgement and a "feel" for the problem; tools that the junior engineer may not have developed as yet.

Technological Content of General Purpose Finite Element Programs

There is probably no such thing as a typical general purpose finite element program; therefore the technological content that we describe here is a union of the capabilities of several well known analysis systems. These capabilities may be characterized in general terms as follows:

- Matrix Abstraction Capability
- Numerical Analysis Capability
- Engineering Mechanics Capability
- Pre-and Post-Processors

Matrix Abstraction

There is an order-of-magnitude difference in the capability of programs built around a general matrix abstraction capability and those which are not. The matrix abstraction codes have primitive matrix operators such as

- Addition
- Multiplication
- Transposition
- Partitioning
- Matrix Decomposition

together with logical and executive functions and modelling modules. These operations are then prescribed in a logical sequence to perform a specific analysis type. NASTRAN[1] is an example of a matrix abstraction program. NASTRAN contains these primitive matrix operations, as well as executive control modules and modelling modules. These modules can then be linked together to specify a particular solution algorithm by means of the NASTRAN higher level language called DMAP, an acronym for "Direct Matrix Abstraction Programming". Specific sequences of operations which define analysis options which are of interest to a large segment of users are pre-written and stored within the program. These DMAP sequences are called rigid formats and may be specified by the user. Because of the inclusion of these preprogrammed rigid formats a matrix abstraction program such as NASTRAN may appear to be a general purpose finite element program.

Numerical Analysis

In the simplest terms a general purpose finite element program must solve the set of equations

$$\underline{k}\ \underline{u} = \underline{P} \qquad (19)$$

The program must, therefore, include an efficient algorithm for solving sets of linear equations of very large order. By taking advantage of the sparse matrix topology and either spill logic or virtual memory, current programs can solve systems represented by thousand of degrees of freedom at a reasonable cost.

In addition to static analyses there are other problems of interest to the structural analyst which include

- Static Stability Analysis
- Dynamic Response
- Nonlinear Analyses

In order to solve these classes of problems, a program must include the following numerical analysis capabilities

- Unsymmetric Matrix Decomposition
- Eigenvalue Extraction
- Solution of Ordinary Differential Equations
- Efficient Solution of Nonlinear Equations

The typical general purpose code incorporates the latest advancements in each of these areas in order to allow the analyst to solve a problem in the most cost-effective manner. If the program incorporates more than one method for a specific capability, the user may be required to specify the correct algorithm. It then becomes the user's responsibility to specify the most cost-effective algorithm that is appropriate for the analysis. A very general complex eigenvalue routine for finding all the eigenvalues and eigenfunctions could be incorporated in the program which would solve any eigenvalue problem defined by the user. On the other hand, if the problem is that of finding the lowest buckling load and if the system matrix is symmetric, then the user would be paying an enormous cost for the generality. Again, it comes down to my previous point that the technology embodied in these programs is so sophisticated that one should not intrust their use to the junior engineer. It is cost effective to retrain the mature engineer, if necessary, in order to bring his experience and knowledge to bear on the use of these high-technology tools to solve structural analysis problems.

Engineering Mechanics

The engineering mechanics technology may be categorized by the following:

- Type of Strain-Displacement Relations

  Linear

  Nonlinear

- Type of Stress-Strain Relations

  Elastic

  Plastic

  Time-dependent

Incompressible material

Temperature dependent

- Type of Analysis

Static

Dynamic

Stability

All these capabilities may not be available in any one program, but these capabilities do characterize the types of engineering mechanics representation that the analyst may request.

Pre- and Post-Processors

The amount of data to be prepared as input to the program, and the amount of data that are generated by the program may pose a significant problem in data definition and management. The engineer needs computerized assistance at both ends of the process. The design of these pre- and post-processors has been investigated by Herness and Tocher[2], who identify the tasks of pre- and post-processors as follows:

Pre-Processors

- Generate node point coordinates
- Generate element connectivity
- Generate nodal loads
- Check data
- Reorder node/element sequences to reduce matrix solution time
- Graphic display of input data
- Prepare connectivity data for substructuring
- Format data for target analysis program

Post-Processors

- Plot behavioral variables
- Compute principal stresses and strains

- Plot deflected structure
- Generate time histories of behavioral variables
- Prepare plots and tables for reports
- Combine solution cases
- Convert output to format suitable for input to another program

Impact of Current Developments

The technology that directly supports computational analysis and design has seen great recent advances. These include

- The development of mini-and micro-computers
- The delivery of 3rd generation large general purpose computers
- Low cost computer graphics
- Computer science developments including compiler design, new languages and data base management techniques
- Numerical analysis developments including more efficient standard algorithms together with adaptive algorithms appropriate to automated mesh refinement and nonlinear analysis
- Engineering Mechanics developments include higher order finite elements and nonlinear analysis

These developments could well have an impact on both the types of analysis problems that the analyst can solve, and the type of hardware/ software that is utilized.

The availability of relatively low cost mini-computer systems that have the capability of an IBM 7094 or a UNIVAC 1108 is now a reality. These mini-computer systems are not characterized by having sophisticated software support; but due to their relatively low price we will see a trend for large volume users of structural analysis software to purchase stand-alone mini-computer systems and to convert present software systems to operate on these mini-computer systems.

While this development will be cost-effective, it is the recent development and the availability of low-cost micro processors and memory that may significantly change the way that the user does business with the computer. One can envision the use of a separate micro-processor to represent each finite element, with all the micro-processors working in parallel. Considering the low cost (approximately $10/chip) and the small size of these devices, we may well see computing capability of the order presently embodied in our large general purpose programs in a desktop calculator -- at a fraction of the cost . . . and the day isn't too far away. Certainly, we will see the start of these developments within the next five years.

We are also seeing the introduction of low cost graphics display devices in the $2,000 - $4,000 range. At this price the hardware is cost effective and we can thus expect to see a typical structural designer sitting at his desk with a hand held NASTRAN analysis tool communicating by means of a graphic display unit. Again, the current technology will support such a fantasy; and its realization isn't too many years downstream.

Automated Design Synthesis

The objective of structure design is to define a structure system which is in equilibrium with the load and which satisfies all design constraints and which extremizes some objective function. Posed in this way, the problem of design may be considered to be a mathematical programming problem as follows:

Minimize

$$\phi\ (\underline{u};\ \underline{\alpha}) = 0 \tag{20}$$

Subject to:

$$\underline{k}\ \underline{u} = \underline{P} \tag{21}$$

$$\psi\ (\underline{u},\ \underline{\alpha})\ \underline{<}\ 0 \tag{22}$$

where

$\underline{u}$ is behavioral variables

$\underline{\alpha}$ is design variables

Great strides have been made in the development of efficient algorithms for obtaining the minimum weight structure of a specific topology that will satisfy all behavioral and design constraints for all loading cases. Generally these algorithms are costly and have not been included in general purpose finite element programs. There does appear to be a trend toward the incorporation of <u>fully</u> <u>stressed</u> design procedures in programs such as NASTRAN.

At the present time the reanalysis of the structure for all load cases is expensive. Current research is being directed at the development of efficient reanalysis algorithms that will significantly reduce the cost of computer-aided design synthesis and thus bring analysis into the design process rather than using analysis for design verification.

Manager's Responsibility

The technology represented by the general purpose finite element programs such as NASTRAN, ANSYS[3], MARC[4], and NONSAP[5] represent the cutting edge of technology in both structural mechanics and numerical analysis. While the technology is formidable, these programs respond to a set of correctly formatted input data by producing a set of output data which is only as good as the input data and the codified technology.

In my business of adult education, teaching graduate students and professionals to utilize large programs such as NASTRAN, I have become concerned by the lack of appreciation of the technology incorporated in the program. There is the danger that we will develop the undergraduate psychology of "don't try to teach me basics . . . just show me what formula to plug". With the public welfare riding on the correct interpretation of a computerized structural analysis of an aircraft component it makes sense to require something more of computer users than familiarity with the input data specification. Indeed the user must be well versed in structural mechanics, in numerical analysis and in computer systems. In short, he's a talented, high priced professional.

At the present time the engineering manager is the only certification mechanism that is available to guarantee that qualified professionals are utilizing computerized technology. It must be the manager's responsibility, therefore, to be aware of the technological content of programs utilized by his staff and to ascertain that the program users are qualified to do so.

Considering the qualifications that have been identified for certified users, it is clear that an organization cannot afford to hire a staff of them. The answer would seem to lie in providing the qualified users with pre- and post-processors and with a suitable staff to support their function . . . it's cost effective to do so.

Summary

The purpose of this paper has been to:

- Present an overview of the Finite Element Method

- Discuss the technology available in present codes

- Discuss present developments related to computer codes

- Discuss the use of the Finite Element Method in the design process

- Discuss the responsibility of engineering management in design certification

These objectives have been met by presenting a tutorial of the finite element method and by identifying the technological content of typical programs.

The development of low cost hardware such as mini-computers and micro-processors and low-cost graphics is considered and it is speculated that we will see a trend towards the development, first of stand-alone mini systems which utilize present software, and then, the development of low cost computing systems based on low cost, large scale integrated circuit technology which will provide the structural engineer with present analysis capability on a desk-top calculator . . . at a relatively low price.

Finally, management's role is seen to be that of protecting the public by insuring that qualified professionals are charged with utilizing computerized technology. It is concluded that it is cost effective for management to support qualified users with pre- and post-processors as well as adequate staff and computer graphics support.

References

1. NASTRAN Users Manual, NASA SP-222, Available from Cosmic, Univ. of Georgia, Athens, Georgia.

2. Herness, E. D., and Tocher, J. L., "Design of Pre- and Post-Processors", Structural Mechanics Computer Programs, ed. by Pilkey et al; Univ. of Va., 1974, pp. 887-898.

3. Swanson Analysis Systems, Inc.

4. MARC Analysis Corporation.

5. Dept. of Civil Engineering, University of California at Berkley.

6. Miller, R. E. and S. D. Hansen, "Large Scale Analysis of Current Aircraft," On General Purpose Finite Element Computer Programs, P.V. Marcal (ed.), ASME Special Publication, New York, N.Y., 1970.

7. Smith, C. S. and G. Mitchell, "Practical Considerations in the Application of Finite Element Techniques to Ship Structures," Proc. of Symposium on Finite Element Techniques, U. of Stuttgart, Stuttgart, Germany, June, 1969.

8. Corum, J. M. and J. E. Smith, "Use of Small Models in Design and Analysis of Prestressed-Concrete Reactor Vessels," Report ORNL-4346, Oak Ridge Nat. Lab., Oak Ridge, Tenn., May, 1970.

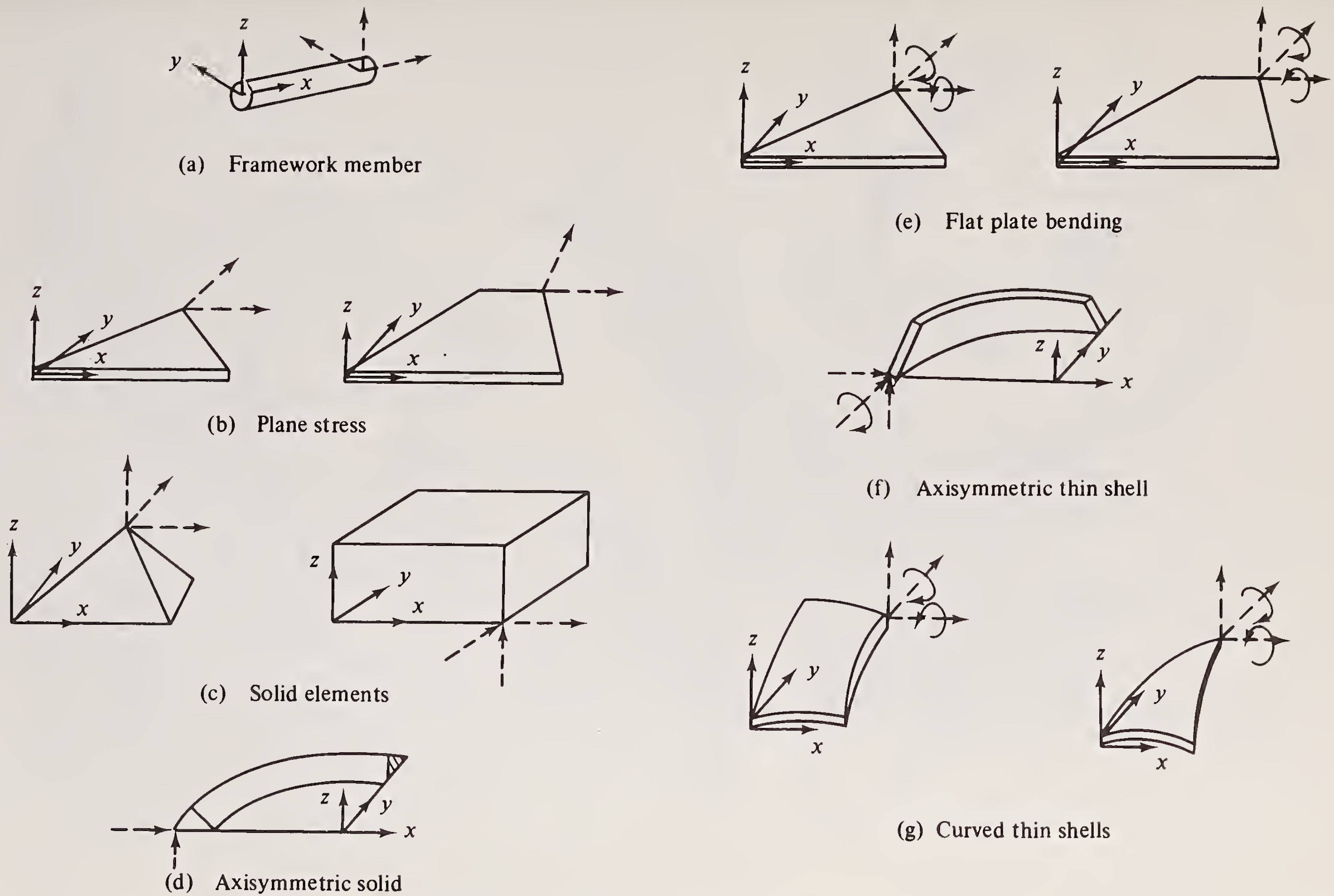
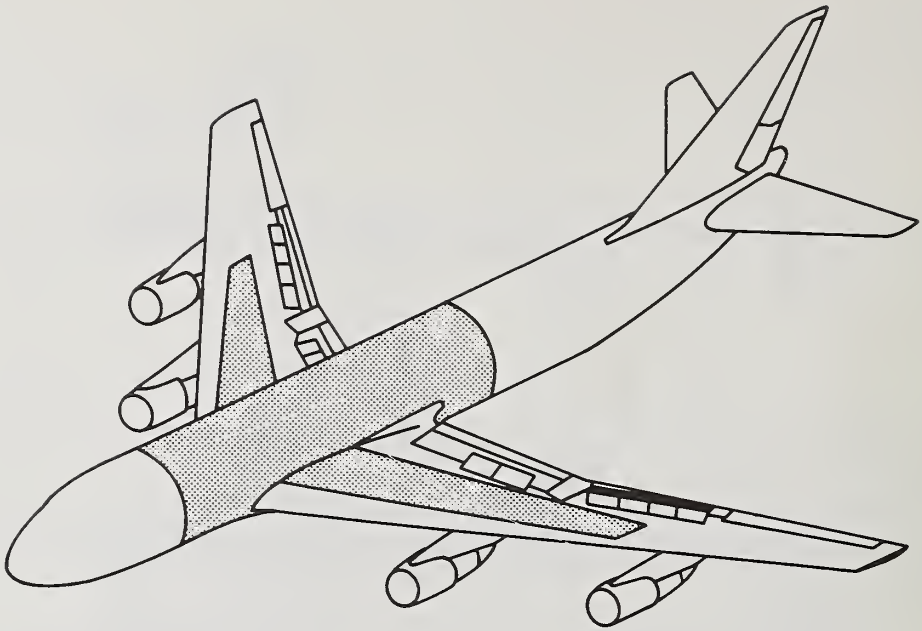
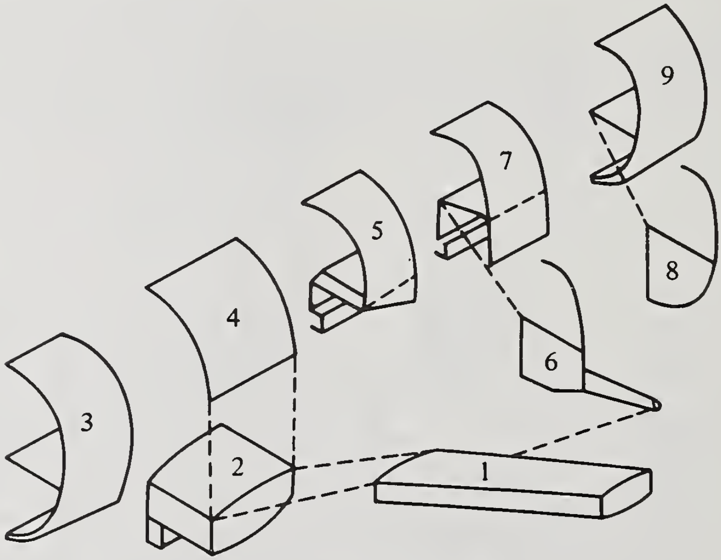


Figure 1 Types of Finite Elements



a. Boeing 747 Aircraft. (Cross-hatched area indicates portion of the airframe analyzed by finite element method.)



b. Substructures for finite element analysis of cross-hatched region.

Figure 2 Finite Element Analysis of Boeing 747 Aircraft (from Ref. 6)

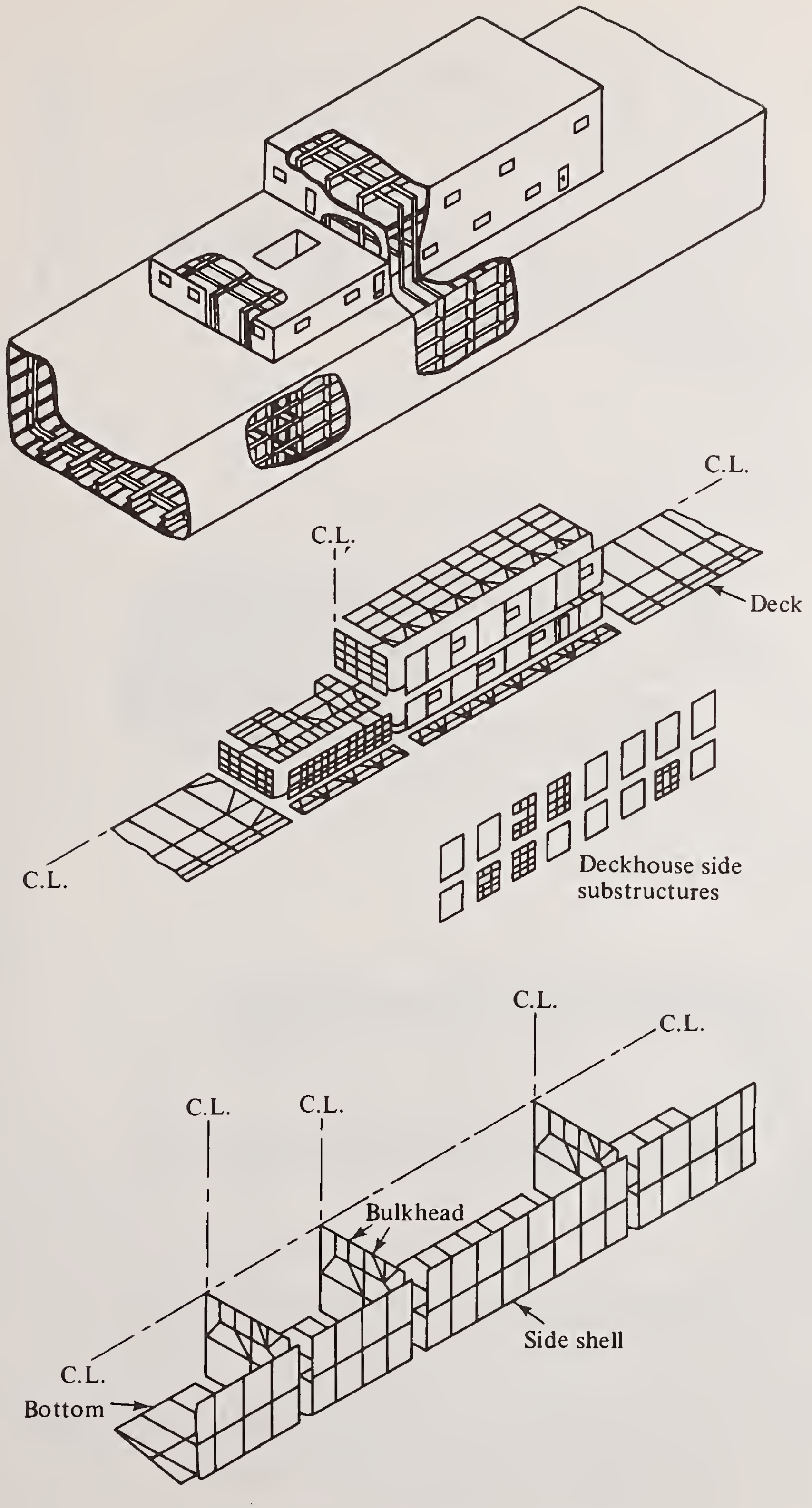
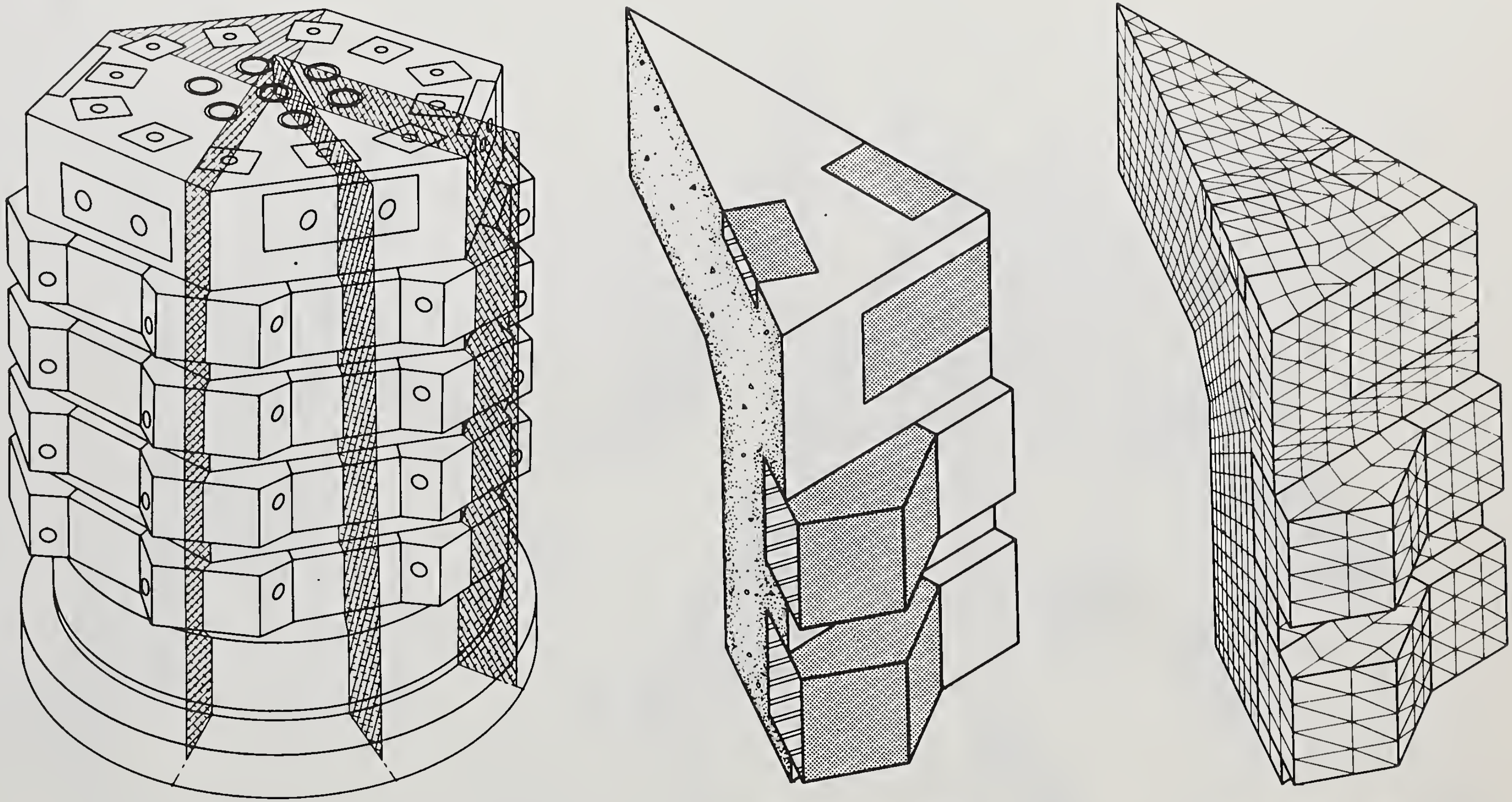


Figure 3   Finite Element Analysis of Ship Structure (from Ref. 7)



(a) Actual structure

(b) Octant of actual structure

(c) Finite element idealization (tetrahedronal elements)

Figure 4 Finite Element Analysis of Reactor Vessel (from Ref. 8)

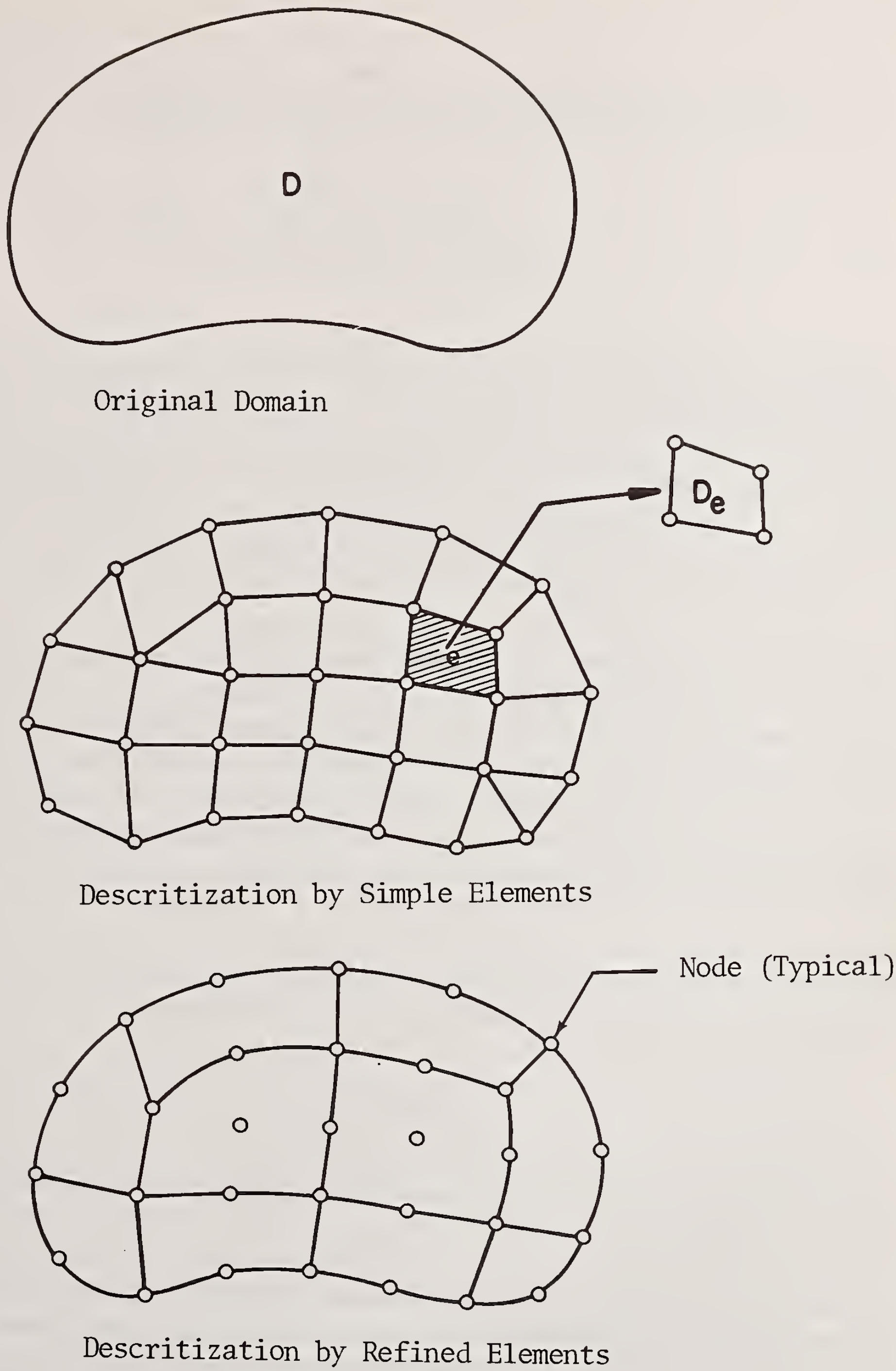


Figure 5 Finite Element Descritization

NASTRAN THERMAL ANALYZER IN INTERDISCIPLINARY THERMO-STRUCTURAL ANALYSES

Hwa-Ping Lee
NASA GODDARD Space Flight Center
Greenbelt, Maryland 20771

Abstract: The NASTRAN Thermal Analyzer ($\bar{N}$TA) is a unique, general purpose heat transfer computer program that is an integral part of the NASTRAN (NASA STRuctural ANalysis computer program) system. The $\bar{N}$TA is based on the finite-element method and was developed at GSFC but implemented by a contractor. This paper stresses the role played by $\bar{N}$TA in interdisciplinary thermo-structural analyses and illustrates the versatility and capability of this computer program. $\bar{N}$TA, together with its host program, NASTRAN, is best suited for analyzing large and complex thermo-structural problems such as those dealing with stresses, deflections or deformations, vibration, buckling, etc., that are induced by thermal effects. A comparison is made between (1) a unified approach to the thermo-structural problem, and (2) an approach that is a combination of using (A) a thermal analysis computer program employing a lumped-nodal network which is based on the finite-difference method, and (B) a structural analysis computer program which is based on the finite-element method. Solution accuracy, procedural efficiency and savings of engineering efforts are discussed.

Key words: Finite-element method; general purpose heat transfer computer program; interdisciplinary thermo-structural analyses.

INTRODUCTION

The evolution of the finite-element method has been one of the most notable advances in engineering practice over the past two decades. This method was originated by the aircraft industry in the structural mechanics field. It was a response to the need to provide precise and efficient solution methods that were used to calculate the strength and stiffness of air frame structures. The early analyses used an approach known as the force method, e.g. in [1]*, which was soon augmented by the displacement method, e.g. in [2], and various hybrid techniques, e.g. in [3]. The intrinsic nature of the finite-element method makes it suitable for computer automation; and the ease of varying material properties,

*Numbers in brackets designate References at end of paper.

model configuration, the refinement in element representation and boundary conditions are principal advantages possessed by this method. As a result, finite-element based structural computer programs proliferated, e.g. in [4-8], but NASTRAN has emerged to be the largest and the most versatile general purpose computer program for static and dynamic structural analyses using the displacement method [9, 10]. This NASTRAN was initially planned in 1964 with a view toward establishing a computer program to be used as a standard at NASA centers, and eventually to be accepted by the aerospace industry as well.

The NASTRAN Thermal Analyzer [11, 12], was originated and developed at NASA Goddard Space Flight Center as one of the software products resulting from a research and development program started in 1969. The software implementation was done by a contractor. The objective of this R & D program was to provide analytical analysis capabilities in the multiple disciplinary areas [13]. Special attention was paid to the interface problems impeding the reliable predictions of the thermo-stresses or deflections that might degrade the optical performance of a large space telescope system exposed to orbital thermal inputs.

The $\overline{\text{N}}$TA is a general purpose heat transfer computer program being integrated in the NASTRAN system. It is fully capable of rendering solutions of temperature and heat flow in structures subject to various boundary conditions that range from prescribed temperatures and specified thermal loads to convective and radiative exchanges at boundary surfaces, in both steady-state and transient cases. This heat transfer computer program was accomplished by the use of applicable functional modules, which had existed in the NASTRAN, and the addition of new modules including new elements and new solution algorithms. The feasibility of utilizing partial functional modules originally designed for structural analysis in thermal applications lies in the existence of a mathematical analogy between these two distinct physical systems. The $\overline{\text{N}}$TA is unique in that it is compatible with the structural counterpart of NASTRAN with regard to both problem size and the model representation. As a component of the unified thermo-structural analysis capabilities in the NASTRAN system, the $\overline{\text{N}}$TA is especially suited to compute temperatures for large and complex structures where thermo-structural problems under the influence of thermal effects are of concern.

The public announcement of the $\overline{\text{N}}$TA was made in September 1972 [11] after this computer program had been delivered. Efforts were then focussed on the verification of the program, debugging and maintaining, applications [14-16], and new developments [17-19]. The $\overline{\text{N}}$TA, having a full-fledged thermal analysis capability, was integrated into the NASTRAN system in its Level 15.5 version, which was made available for the general user through COSMIC in June 1973.

Experience with the $\overline{N}$TA started after the delivery of the IBM-360 computer program tape to GSFC in June 1972. Test problems were designed to verify program capabilities and to unearth programming defects. At the end of that year, the Colorado Experiment of the OSO-I was selected as the first flight experiment on which to test the developed analytical tools [14, 15].

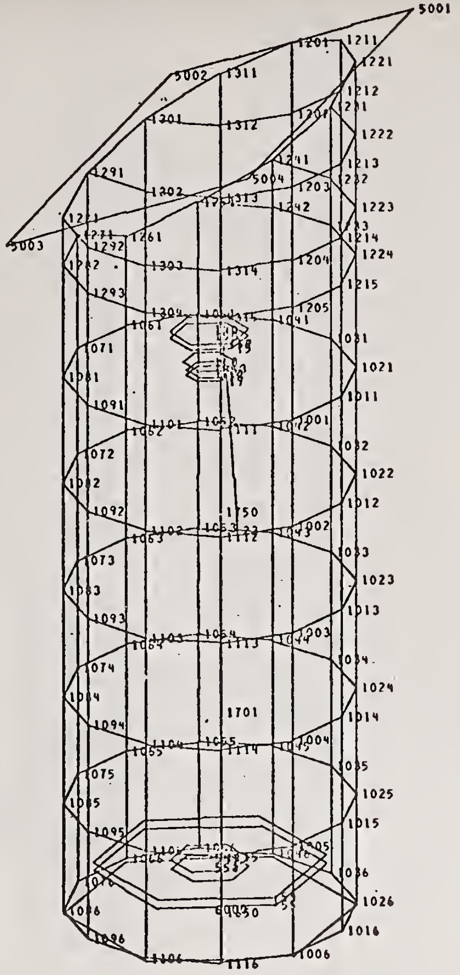
The $\overline{N}$TA has since been employed to support many scientific instrument packages for various flight programs at GSFC, such as IUE (International Ultraviolet Explorer Satellite) [16], SMM (Solar Maximum Mission), MMS (Multiple Mission Satellite), etc. Figure 1 shows some thermal models of the IUE. The $\overline{N}$TA has been employed by other users, who have contacted us for consultations, engaged in the areas of nuclear reactors, weaponry, helicopters, automobiles, railroad cars, oil refineries, computer hardware, and electronics.

With proven capabilities and the inherent flexibility of the finite-element method, the viable $\overline{N}$TA should be utilized to the maximum extent to benefit engineering designs in other applications. In view of the fact that thermal analysis using the finite-element method is a rather new innovation to thermal engineers, a comparison with the finite-element and finite-difference methods will be made to distinguish their fundamental differences. An overview of the $\overline{N}$TA organization together with its functional features will be presented. The procedure of model conversion in the interdisciplinary thermo-structural analyses will be given. Solution accuracy, procedural efficiency and savings of engineering efforts will be discussed.

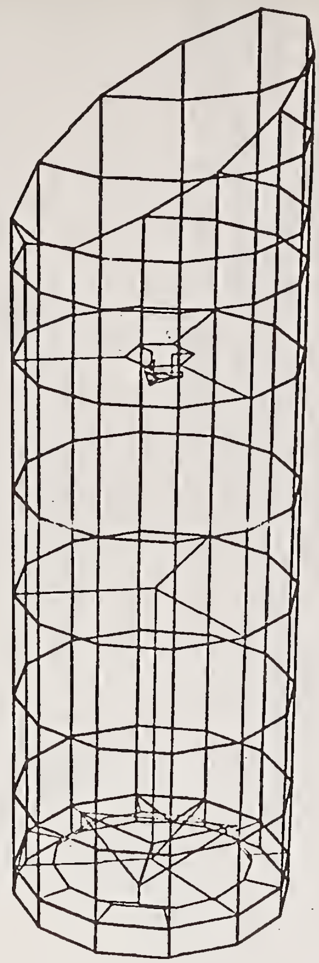
DISTINCTIONS BETWEEN THE TWO NUMERICAL METHODS

The finite-element method and the finite-difference method are two distinct numerical methods extremely valuable for solving complex engineering problems. While the latter has long been a standard numerical process being employed in thermal applications, e.g. [20, 21], the use of the finite-element method in computing temperatures is a rather recent endeavor. Since the uniqueness of the $\overline{N}$TA lies wholly in its underlying theoretical basis of the finite-element method, a comparison of these two methods is appropriate.

In the finite-element method, a continuum with infinite unknown variables can be approximated by an assemblage of finite number of discrete elements that reduce the problem to finite number of unknowns. The elements are interconnected at the grid points consisting of the end nodes (in 1-D rod element case) and vertices (in 2-D planar or 3-D solid element case). The conditions of equilibrium and continuity are established at those grid points. The finite-element



(a) THE BOUNDARY SURFACE ELEMENTS



(b) THE IUE TELESCOPE CONDUCTION ELEMENTS

Figure 1. The IUE telescope thermal model.

method deals directly with an approximate minimization of the functional and applies an approximation to the terms of a variational formulation. In applying the Rayleigh-Ritz procedure or the Galerkin weighted-residual procedure [22], a variational principle valid over the entire region is postulated, and the desired solution is the one minimizing the functional, which is defined by a suitable integration of the unknown quantities over the entire domain. Detailed treatments of using the finite-element method in thermal conduction and its extension to include the effect of radiative exchanges can be cited in [23, 24]. An essential part of the finite-element formulation is to obtain matrices for individual elements. The types of matrices are shown in the resulting general heat equation which is of the form:

$$[C]\{\dot{T}\} + [K]\{T\} = \{Q_{\ell}\} + \{Q_{r}\} \tag{1}$$

Where

- $\{T\}$ a vector of temperatures at grid points
- $[C]$ a symmetric matrix of heat capacitance
- $[K]$ a symmetric matrix of heat conductance
- $\{Q_{\ell}\}$ a vector of applied thermal loads
- $\{Q_{r}\}$ a vector of nonlinear thermal loads that depend on temperature

The general heat equation implies three classes of problems that require separate solution algorithms as will be discussed in the next section. Equation (1) directly represents a transient heat equation. $\{Q_{\ell}\}$ is permitted to be a vector of constant or time-dependent thermal loads and $\{Q_{r}\}$ is a nonlinear thermal load vector arising from radiative exchanges or other temperature-dependent variables such as the convective film coefficient. The steady-state equations obtained readily when $\{\dot{T}\} = \{\partial T/\partial t\} = 0$ is imposed. However, two separate solution algorithms are required for steady-state problems of the linear and nonlinear cases. The linear case allows only an array of constant conductivity coefficients and a vector of constant thermal loads. The nonlinear case allows a constant or a temperature-dependent conductance matrix and also a vector of nonlinear thermal loads including radiative exchanges.

The theoretical basis of the finite-difference method, in contrast, is to apply a direct approximation to the derivatives in a governing differential equation. The distinctions between these two numerical methods from a user's viewpoint are augmented by the following illustrations:

A simple 1-D rod, Figure 2, is employed. In the finite-element method, the rod is discretized by specifying nodal point locations along the length of the rod as

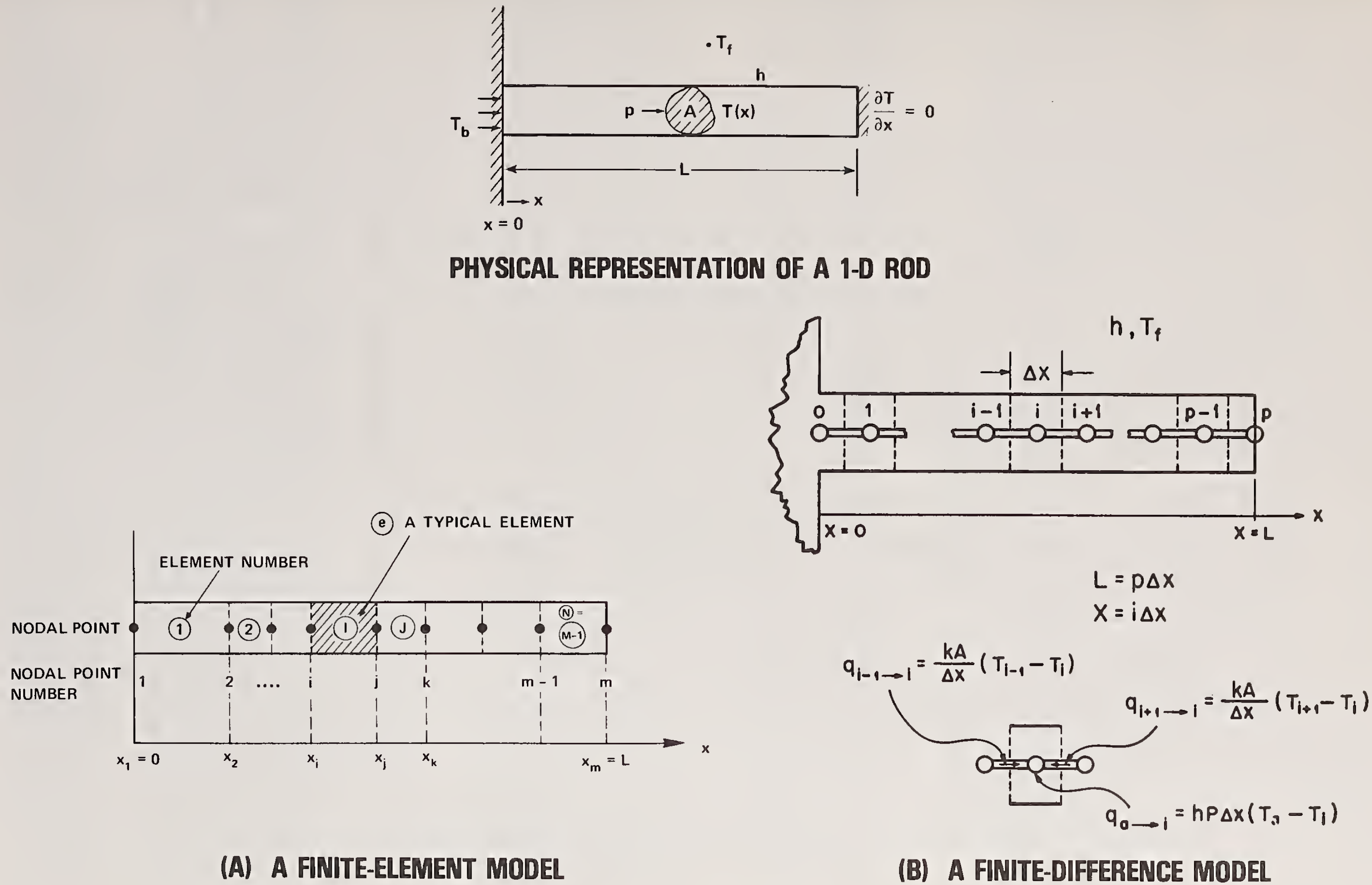


Figure 2. Model representations by the finite-element method and the finite-difference method.

shown in Figure (2-A). An element is defined as the subregion between two adjacent nodal points. These nodes are numbered from 1 to m, and these elements are numbered from 1 to N. e is a typical element bounded by the two end nodes i and j. A linear temperature variation is permitted within an element. Transferring heat, whether via conductive, convective or radiative mode, is automatically included in the derived element matrices. In the finite-difference method, the rod is also discretized but on a totally distinct basis. The rod is idealized into P elements, as shown in Figure (2-B), which are assumed to be isothermal segments individually. The temperature of an assumed isothermal element is represented by a lumped-node located at the center of that fictitious segment. Conductive couplings are simulated by connecting thermal conductors (the reciprocals of thermal resistance) between any two adjacent nodes, and radiative couplings are likewise simulated by nonlinear thermal resistances being connected between any two affected nodes.

Figure 3 shows the difference between a 2-D conducting medium being discretized by these two numerical methods. The continuum and the continuous independent variables are again replaced by a discretized system in each case. Temperature variables are represented at the vertices of each element in the case of the finite-element method, and each isothermal area is assumed and represented by a lumped-node at its center in the case of the finite-difference method.

In dealing with irregular shaped regions, the nodes near the boundary, in the finite-difference method, have to use separate equations other than those representing the interior points, but the finite-element method using triangular and quadrilateral elements provide a much better approximation of the same region than is provided by the other method. These finite-elements used at the boundary are no different than any interior ones. Consequently, no additional programming effort is required.

THE PROGRAM ORGANIZATION AND FUNCTIONS

An overview of the $\overline{\text{N}}$TA program organization together with an introduction of an $\overline{\text{N}}$TA input deck would be useful for realizing the ease of using this computer program and for appreciating its capability and versatility. Essential components in the $\overline{\text{N}}$TA consist of two main parts: (1) the elements which provide thermal capacitance and thermal conductance matrices and are able to accommodate thermal loading conditions, and (2) the methods of solution.

In modeling, heat conducting structure is represented by a finite number of idealized heat conduction elements that are interconnected at grid points. Various forms of heat conduction elements are provided to represent common

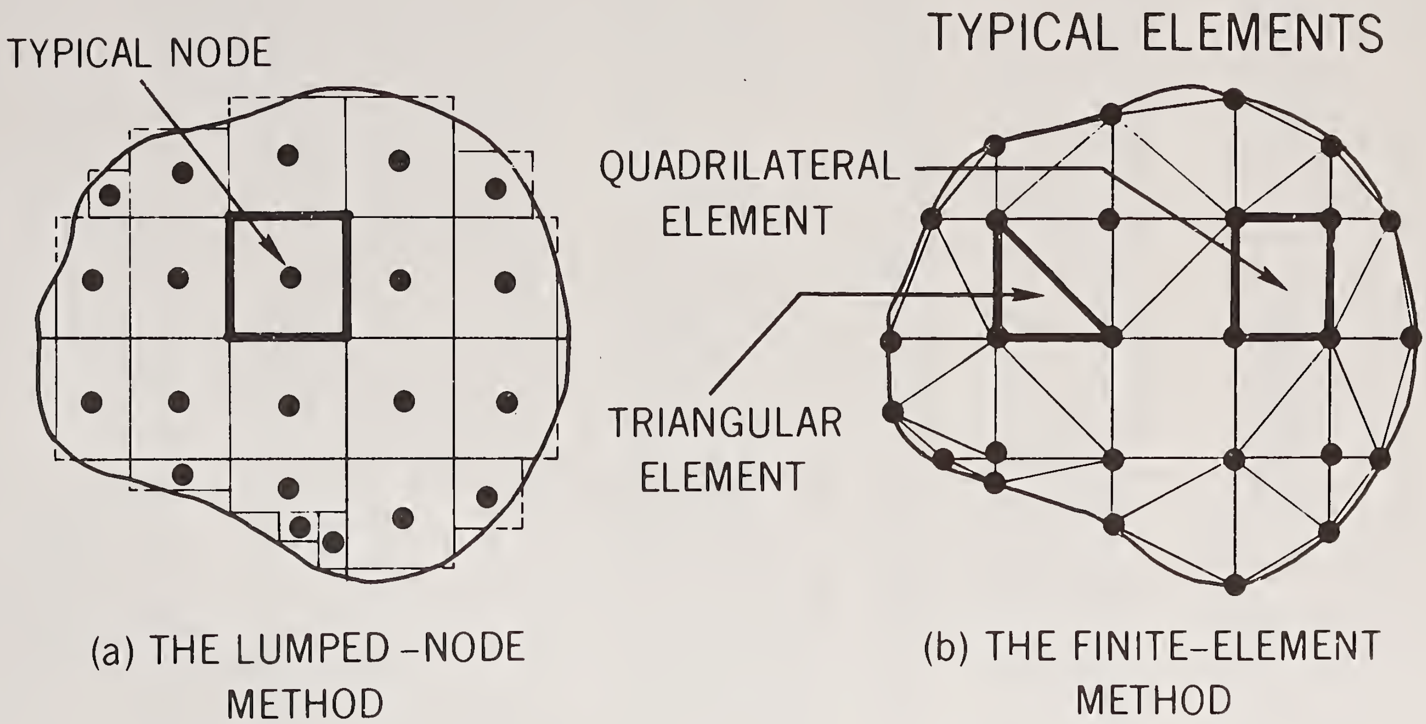


Figure 3. Distinction between thefinite-difference lumped-nodal method and the finite-element method.

structural members including 1-D rods, 2-D triangular and quadrilateral plates and axisymmetric rings of triangular and trapezoidal cross-sections, and 3-D solids such as wedges, tetrahedra and hexahedra, as shown in Figure 4. Also included are scalar heat conduction elements which can serve as linear thermal conductors connecting pairs of grid points with specified conductances. A special kind of boundary surface element is provided to serve as a medium for exchanging heat from exterior environment to the overlaid conduction element through the attached grid points.

Thermal loads in both static and dynamic modes may be applied directly to grid points, or indirectly via the boundary surface element. The primary types of thermal loads included are concentrated loads applied to the points, internally generated heat within an element volume, and uniform heat fluxes as well as directional radiant sources applied to the surface of an element.

The $\overline{\text{N}}$TA has been provided with three specialized solution algorithms that are able to yield accurate, efficient and stable solutions for the following analyses:

(1) Linear steady-state analysis: This solution is, in essence, a matrix inversion process.
(2) Non-linear steady-state analysis: The solution is an iterative process essentially based on the Newton-Raphson methods [25].
(3) Transient thermal analysis including both linear and nonlinear boundary conditions: The integration algorithm is a special form of the Newmark-β Method [26].

The $\overline{\text{N}}$TA computer program inherent in the program structure of the NASTRAN consists of a number of mathematical modules (subprograms) that are executed according to a sequence of macro-instructions. Such a permanently stored pre-arranged sequence for solving a specific type of problem is called Rigid Format. The described solution algorithms have formed three Rigid Formats. The $\overline{\text{N}}$TA also permits the creation of a special program by adding and editing the program stored matrix routines called DMAP (Direct Matrix Abstraction Program). This is a user-oriented programming language of macro-instructions. A similar modification, but to a lesser degree, to alter a Rigid Format is called ALTER.

A complete $\overline{\text{N}}$TA input deck begins with the required job control cards. The main body of the Data Deck consists of three sections:

(1) Executive Control Deck
(2) Case Control Deck
(3) Bulk Data Deck

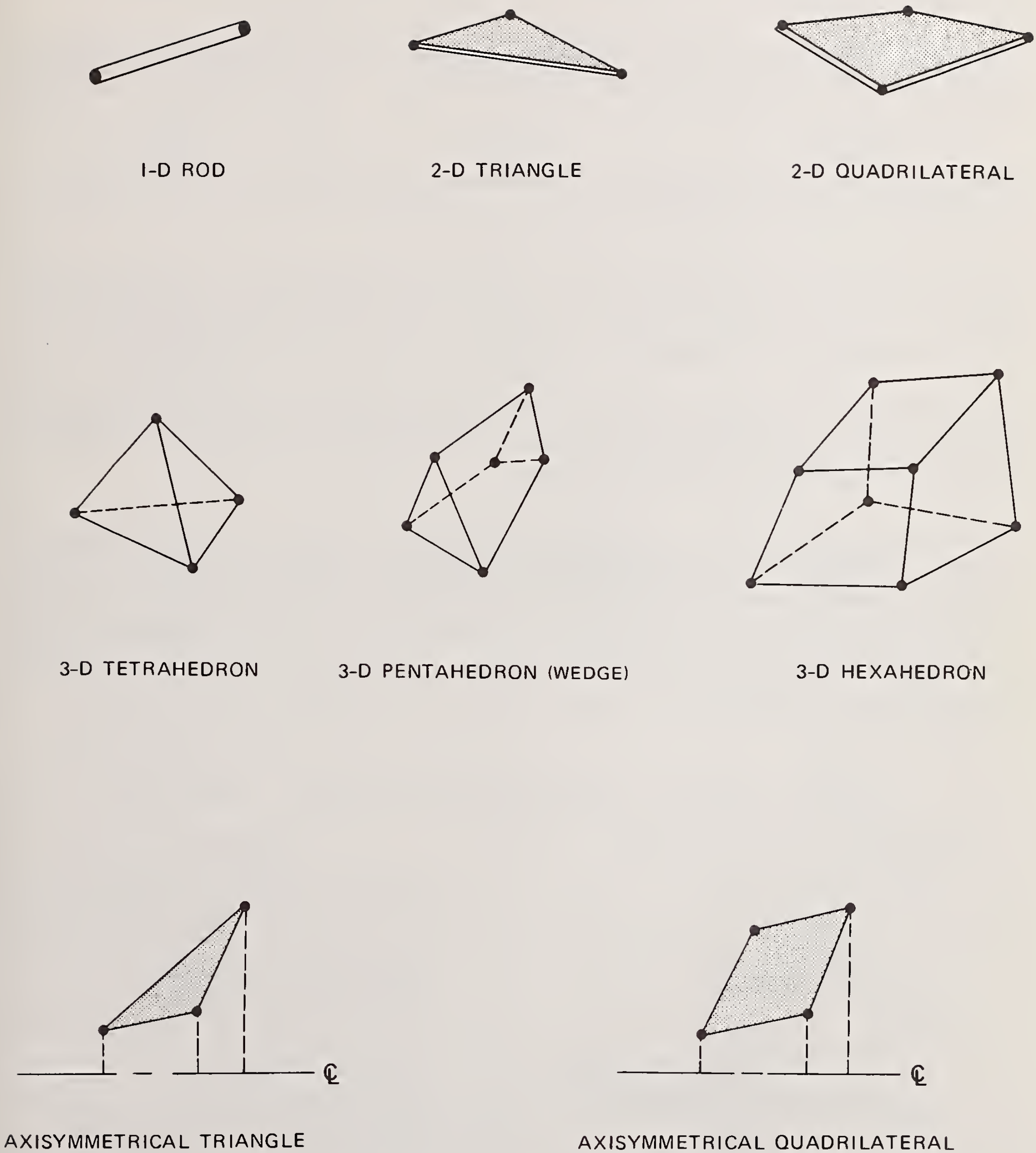


Figure 4. Representative heat conduction elements.

The Executive Control Deck identifies the job and the type of solution to be performed. It also declares the general conditions under which the job is to be executed, such as maximum time allowed, type of system diagnostics desired, restart conditions etc. The identification of the Rigid Format is declared along with any alterations to the Rigid Format that may be desired.

The Case Control Deck makes selections from the Bulk Data Deck, and makes output requests for printing, punching and plotting. It also defines the subcases, if any, for the problem.

The Bulk Data Deck contains all of the details of the idealized thermal model and initial and boundary conditions for the solution.

The functional flow of bulk data cards relative to the definition, constraints and thermal loadings of the finite-element thermal model is shown in Figure 5. This Bulk Data Deck constitutes the main portion of a complete $\overline{\text{N}}$TA input deck.

The grid point definition (GRID) forms the basic framework for the discretized thermal model. All other parts of the model structure are referenced either directly or indirectly to the grid points.

Heat conduction elements are defined on connection cards by referencing grid points. The library of heat conduction elements are listed in Table 1. In most cases the connection card refers to a property card, on which the cross-sectional properties of the element are given. The property card in turn refers to a material card which gives the material properties. If some of the material properties are temperature dependent, a further reference is made to tables for this information.

Table 1. Heat conduction elements.

Dimension	Type	Elements
1-D	Linear	BAR, RØD, CØNRØD, TUBE
2-D	Planar	TRMEM, TRIA1, TRIA2, QDMEM, QUAD1, QUAD2
	Solid of Revolution	TRIARG, TRAPRG
3-D	Solid	TETRA, WEDGE, HEXA1, HEXA2
–	Scalar	ELAS1, ELAS2, ELAS3, ELAS4

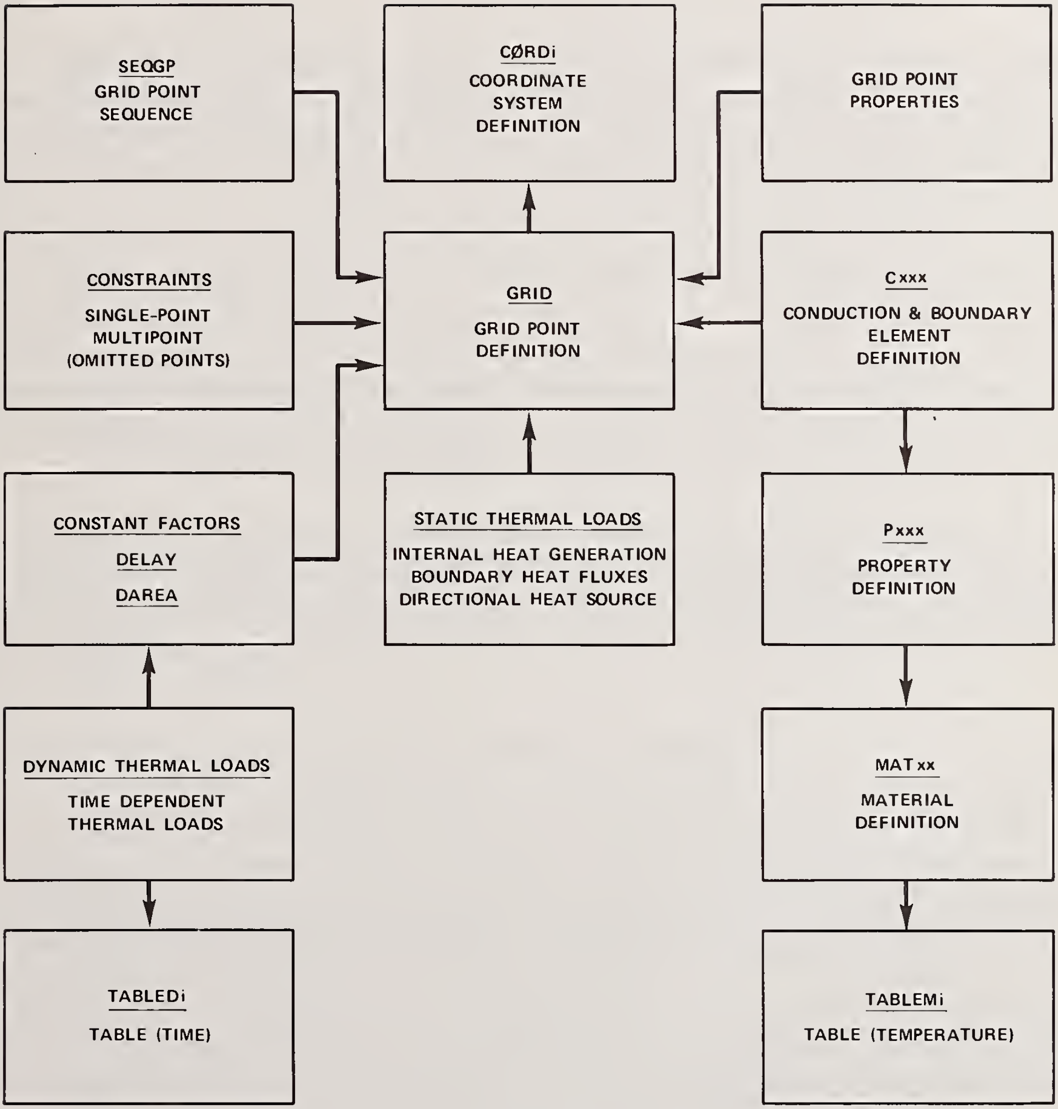


Figure 5. Functional diagram of the N̄TA thermal model.

The heat boundary element is provided to accept external heat fluxes. It is also used to form boundary surfaces of the heat conduction elements when thermal convection and radiative exchanges are present. The radiative surface properties, emissivity and absorptivity, can be entered on the associated property card. This card is also used together with the thermal material card (MAT4) if the convective film coefficient and the thermal capacity of the boundary film are simulated.

Various kinds of constraints can be applied to the grid point. Single-point constraints are used in steady-state thermal analyses to specify prescribed temperatures. Multi-point constraints are used to specify a linear relationship among selected grid point temperatures. Omitted points are used to reduce the number of unknown temperatures in transient analysis.

Static thermal loads consisting of the internally generated heat, boundary heat fluxes and heat input from directional radiant heat sources are provided to represent various types of heat inputs. The heat boundary element proportionates the total energy received by the boundary element surface to the vertices of that element according to its shape and size. The same loads may be modified by a time function to become time-dependent dynamic thermal loads.

The $\overline{\text{N}}$TA has been specifically designed to treat large problems with many unknown temperature variables. The only limitations on problem size are those imposed by practical considerations of running time and by the ultimate capacity of auxiliary storage devices. There are no dimension statements in the program.

MODEL TRANSFORMATION

In unified thermo-structural analyses, the grid point temperature data, as required by NASTRAN for analyzing thermally induced structural responses, are provided directly by the $\overline{\text{N}}$TA. It is achieved by using a pair of compatible models. Specifically, the input decks for these two physically distinct analyses share the same basic model in a finite-element discretization. Two separate input card decks, however, are still required. Either the NASTRAN structural model or the $\overline{\text{N}}$TA thermal model can be the first to become available in the back-to-back analyses. Alterations of cards in the input deck for the second model need be made only from the one first in existence to accommodate constraints, loadings, physical properties, parameters, etc. in accordance with the problem description. While remaining useable for those cards defining grid points and connection cards (finite-elements), they generally constitute the main body of a bulky input data deck, which would be the most time-consuming effort to model and prepare independently. When thermo-structural analyses are performed in tandem, the structural model, satisfying mechanical requirements and design criteria, is usually the first to be established. The transformation is, therefore, from a NASTRAN structural model to an $\overline{\text{N}}$TA thermal model. This conversion process can be done very efficiently, if steps are taken in a systematic manner. The detailed procedure of model transformation and modeling techniques are available, interested readers should consult with references [12, 16, 27]. Some essential points are given below to highlight the model transformation process:

(1) Pay special attention to areas with high power dissipations and crucial thermal contact couplings of the structure to be analyzed. It would often be the case that the structure model would be of insufficient representation in these areas, requiring additional modeling to be done.

(2) Select a proper Rigid Format to be used. It depends on considerations of required temperature results as input to the subsequent analysis and computer time, etc.

(3) Replace the Executive and Case Control Decks from a punched deck copy of the NASTRAN structural model by the new ones specifically designed for thermal analyses.

(4) Remove all Bulk Data Deck cards except those relating to the grid point definition (GRID, SPOINT), coordinate, connection, property, cards. In addition, applicable comment cards and common flag cards to the program code (e.g. BEGIN BULK and ENDDATA cards) are retained for thermal analysis.

(5) Remove any permanent constraints defined on the grid point definition cards of the structural model. Use "1" only in thermal analysis when degree-of-freedom information is requested by any input data cards.

(6) Examine any single-point-constraint, multi-point-constraint, or permanent constraints that have been removed from the structural model to affect the grid points properly coupled into the thermal model.

(7) Use the single-point-constraint to simulate specified temperatures at grid points for either case of the two steady-state solution algorithms. Specifically, SPC1 is used to list all grid points which are to be held at prescribed temperatures in the nonlinear steady-state case, and the temperature values are entered on TEMP cards. The simulation for specifying prescribed temperatures at grid points in transient analyses, however, are much more involved. Documentation containing eight detailed alternative modeling techniques for both constant-valued and time-dependent temperatures are available [27].

(8) Add material cards to define thermal conductivities for each material type referenced by the property cards in the model.

(9) Add the boundary surface elements (HBDY) to accommodate external heat inputs to grid points, and to allow radiative and convective thermal exchanges as well.

(10) Provide radiation matrix if gray-diffuse radiative couplings are present. RADLST and RADMTX cards are required to, respectively, define the radiatively active CHBDY cards, and the area-times-view factor coupling between the active CHBDY cards. A specially developed VIEW computer program [28, 29], using the same radiatively active CHBDY cards as input data cards, can automatically produce the required RADLST and RADMTX cards for $\overline{\text{N}}$TA runs.

(11) Add PARAM cards to define the convergence criterion, the maximum number of iterations, the datum temperature in the absolute temperature scale, and the Stefan-Boltzmann constant for the nonlinear steady-state case. Only the last two items are required for the transient thermal analysis when thermal radiation phenomenon is present.

(12) Define a "guess temperature vector" for all grid points in the model when the nonlinear steady-state solution algorithm is used. It is an option to use a "guess temperature vector" in the nonlinear transient thermal analysis. A special attention must be paid to those grid points listed on an SPC1 card, their temperature values be constrained consistently. In addition, to assure convergence, all guess temperatures should be no less than 80% of the final steady-state absolute temperature of the respective grid points.

(13) Select appropriate input data cards to simulate various thermal loads to be applied to any elements or grid points. In transient thermal analysis, all thermal loads must be referenced via TLOADi (i = 1, 2) cards, and TSTEP card must be added to specify the number of integration time steps, their length and the frequency of output.

(14) Supply an initial condition temperature set for a transient thermal analysis.

DISCUSSION AND CONCLUSIONS

For an accurate thermo-structural solution, the thermoelastically uncoupled analysis requires reliable temperature inputs to the NASTRAN structural model.

Prior to the existence of the $\overline{\text{N}}$TA, general purpose heat transfer computer programs were all of the lumped-nodal network type, which were based on the finite-difference method. They were not only limited in capacity but seriously handicapped by incompatibilities arising from the model representations inherent in the two distinct approaches. The intermodel transfer of temperature data led to extensive interpolation and extrapolation. The extra work proved not only a tedious and time-consuming process but also compromised accuracy. The $\overline{\text{N}}$TA eliminates the need to form two independent models with the concomitant requirement for intermodel interpolation of temperature data.

Table 2 compares the saving of modeling efforts with the two different approaches when a thermal analysis and a structural analysis are performed in tandem. With thermal radiation included in the analysis, a conventional approach requires that three independent models be prepared: A model defines discretized isothermal surfaces for generating view factors, another model gives discrete temperature nodes representing the isothermal surfaces for the lumped-nodal network thermal analysis, and a third finite-element structural model is for the NASTRAN structural analysis. In the unified finite-element approach, however, the view factors can be generated by a specially developed VIEW program, which uses the same boundary surface elements as required in the $\overline{\text{N}}$TA model for thermal analysis. The structural elements in the NASTRAN structural model are virtually identical to those heat conduction elements used in the $\overline{\text{N}}$TA. Only one model with modifications is required to be prepared.

Table 2. Number of models required for thermo-structural analysis including radiative exchanges.

Approach	Conventional Method	Finite-Element Method
View Factor Generation	(Surfaces)	
Thermal Analysis	(Nodal-Network)	(Finite Elements)
Structural Analysis (NASTRAN)	(Finite Elements)	
Total	3 Different Models	1 Unified Model

The economic advantage of using this computer program can be best expressed by examining the computer time expenditures required to complete a run. From our experience with the OSO-I Colorado telescope, for example, this model consisted of a total of 488 grid points. Each steady-state run required 10 min. CPU time and 12 min. I/O time, while a five-orbit transient solution with 120 time steps required 24 min. CPU time and 28 min. I/O time. It must be emphasized that structurally compatible temperature data cards were direct outputs included in these indicated computer time expenditures. The engineering-saving and cost-effectiveness is evident.

In conclusion, it is evident that the $\overline{\text{N}}$TA, a finite-element based general purpose heat transfer analysis capability in the NASTRAN system, is capable of providing valuable support to the engineering design processes. Principal advantages in employing this for the interdisciplinary thermo-structural analyses are summarized as:

(1) virtually the same discretized finite-element model can be used for thermal and structural analyses;

(2) temperature results determined by the program can be directly applied, with no error-inducing and time-consuming interpolation to the structural model. This permits the calculation of thermally induced structural effects more accurately and efficiently.

To equally benefit general engineering applications in fields other than the aerospace area, design engineers should become aware of the capability and availability so they also can make good use of this proven and viable analytical tool. The $\overline{\text{N}}$TA is especially suited to handle interdisciplinary thermo-structural analyses.

REFERENCES

1. De Veubeke, B. F.: Displacement and Equilibrium Models in the Finite Element Method, Chapter 9 of Stass Analysis, O. C. Zienkiewicz and G. S. Holister (eds.), John Wiley & Sons, London, 1965.

2. Turner, J. L., Clough, R. W., Martin, H. C., and Topp, L. J.: Stiffness and Deflection Analysis of Complex Structures, Journal of Aerospace Science, Vol. 23, No. 9, September 1956, pp. 805-825.

3. Pian, T. H. H.: Derivation of Element Stiffness Matrices by Assumed Stress Distributions, AIAA Journal, Vol. 2, 1964, pp. 1333-1336.

4. Melosh, R. J., Diether, P. A., and Brennan, M.: Structural Analysis and Matrix Interpretive Systems (SAMIS) - Program Report. TM33-307, Jet Propulsion Lab., December 1966.

5. Greene, B. E. et al.: ASTRA-Boeing's Advanced Structural Analyses, Proc. Int. Symp. on Structures Tech. for Large Radio and Radar Telescope Systems, MIT, October 1967.

6. von Fuchs, G., and Schrem, E.: ASKA - A Computer System for Structural Engineers, Proc. ISD/ISSC, Symp. on Finite Element Techniques, Institute fur Statik und Dynamite der Luft-und Raumfahrtkonstruktionen, University of Stuttgart, June 10-12, 1969.

7. Logcher, D., et al.: "ICES STRUDL II, the Structural Design Language User's Manual", Vol. 1 = Frame Analysis, MIT, Research Report R68-91, November 1968, Vol. II: Additional Design and Analysis Facilities, MIT, R68-92, June, 1969.

8. Mallett, R. H., and Jordan, S.: MAGIC - An Automated General Purpose System for Structural Analysis: Engineer's Manual, Vol. I, AFFDL TR-68-56. Air Force Flight Dynamics Laboratory, WPAF13, Ohio, January 1969.

9. MacNeal, R. H. (Editor): The NASTRAN Theoretical Manual, NASA SP-221, with Level 15.5 updates, December 1972, Washington, D. C.

10. McCormick, C. W. (Editor): The NASTRAN User's Manual (Level 15), NASA SP-222(01), with Level 15.5 updates, June 1972, Washington, D. C.

11. Lee, H. P., and Mason, J. B.: NASTRAN Thermal Analyzer - A General Purpose Finite-Element Heat Transfer Computer Program, The 2nd NASTRAN User's Colloquium, NASA TMX-2637, September 1972, pp. 443-454.

12. Lee, H. P.: NASTRAN Thermal Analyzer — Theory and Application Including a Guide to Modeling Engineering Problems, Volume 1: The NASTRAN Thermal Analyzer Manual, NASA Goddard Space Flight Center X-322-76-16, December 1975.

13. Lee, H. P.: Structural-Thermal-Optical Program (STOP), NASA-GSFC 1970 Science and Technology Review. NASA-Goddard Space Flight Center, January 13-14, 1971; also Significant Accomplishments in Technology: Goddard Space Flight Center, 1970, NASA SP-295, 1972, Washington, D. C., pp. 36-40.

14. Lee, H. P., and Jackson, C. E. Jr.: Thermal Analysis of the Telescope of the OSO-I Colorado Experiment, NASA-GSFC Test and Evaluation Division, Program STOP-Document, Greenbelt, Maryland, August 1973.

15. Lee, H. P., and Jackson, C. E. Jr.: Using the NASTRAN Thermal Analyzer to Simulate a Flight Scientific Instrument Package, Proceedings of the 20th Annual Meeting of the Institute of Environmental Sciences, Washington, D. C., April 1974, pp. 152-158.

16. Jackson, C. E. Jr: Thermal Analyses of the International Ultraviolet Explorer (IUE) Scientific Instrument Using the NASTRAN Thermal Analyzer ($\overline{N}$TA) - A General Purpose Summary, The Fifth NASTRAN User's Colloquium, NASA TMX-3428, Oct. 1976, pp. 99-118.

17. Lee, H. P., and Jackson, C. E. Jr.: Finite-Element Solution for a Combined Radiative-Conductive Analysis with Mixed Diffuse-Specular Surface Characteristics, AIAA Paper No. 75-682, AIAA 10th Thermophysics Conference, Denver, Colorado, May 27-29, 1975.

18. Lee, H. P.: NASTRAN Thermal Analyzer – Status, Experience and New Developments, The 4th NASTRAN User's Colloquium, NASA TMX-3278, Sept. 1975, pp. 523-528.

19. Lee, H. P., and Harder, R. L.: The GSFC NASTRAN Thermal Analyzer New Capabilities. The Fifth NASTRAN Users' Colloquium, NASA TMX-3428, Oct. 1976, pp. 119-126.

20. Dusinberre, G. M.: Heat Transfer Calculations by Finite Differences, International Text Book Co., Scranton, Pa., 1961.

21. Thermophysics Branch: Nodal Network Thermal Balance Program, NASA Goddard Space Flight Center, February 1968.

22. Hildebrand, F. B.: Methods of Applied Mathematics, Prentice-Hall, Inc. New York, 1954, pp. 120-226.

23. Zienkiewicz, O. C.: The Finite Element Method in Engineering Science, McGraw-Hill, London, 1971, pp. 116-120.

24. Lee, H. P.: Application of Finite-Element Method in the Computation of Temperature with Emphasis on Radiative Exchanges, Progress in Astronautics and Aeronautics, Vol. 31 – Thermal Control and Radiation, MIT Press, 1973, pp. 491-520.

25. Scarborough, J. B.: Numerical Mathematical Analysis, 4th Edition, The Johns Hopkins Press, Baltimore, 1958, pp. 192-211.

26. Newmark, N. M.: A Method of Computation for Structural Dynamics, Proc. ASCE, Journal of the Engineering Mechanics Div., EM-3, July 1959, pp. 67-94.

27. Lee, H. P.: On Modeling and the Use of the NASTRAN Thermal Analyzer, NASTRAN: Users' Experiences, Fifth Colloquium held at Ances Research Research Center, Moffett Field, California, Oct. 5-6, 1976, NASA TMX 3428, pp. 3-18.

28. Puccinelli, E. F.: View Factor Computer Program (Program VIEW) User's Manual, NASA-GSFC, Test and Evaluation Div., X-324-73-272, July, 1973.

29. Jackson, C. E. Jr.: Programmer's Manual for VIEW – A Modification of the RAVFAC View Factor Program for Use with the NASTRAN Thermal Analyzer on IBM-360 Series Computers, NASA-GSFC Test and Evaluation Div., X-322-73-120, March 1973.

THE USE OF CATHODE RAY TUBE GRAPHICS AS A COMPUTER AIDED DESIGN TOOL

Alyce Branum
Graphics Product Manager
Engineering Systems Group
Digital Equipment Corporation
200 Forest Street
Marlboro, Massachussetts 01752

Abstract: Considering how efficiently the human mind functions when communicating in pictorial terms, many graphics experts expected the use of CRT's to sweep through the engineering community much as the use of TV's had swept through the home entertainment marketplace. In addition to discussing this "boom that never happened" and why, the paper will provide a general background in display technology and its progress over the past decade. Particular emphasis will be placed on refresh calographic CRT's and their use in the design process.

Key words: Cathode ray tubes, graphics, design, computer aided design.

In going over some early papers to trace the history of Cathode Ray Tubes as design tools and to speculate on some of the reasons why they have not had as widespread use as predicted; I came across a 1968 paper by Mr. Carl Machover, a leader in the field and past president of the Society for Information Displays. He 'questimated' (and I would agree) that there were about 1000 graphics consoles installed in 1968. Mr. Machover went on to express the opinion then held by most graphics experts that there would be a billion dollars worth of graphics terminals purchased in 1973. Assuming an average terminal price of $50,000; twenty thousand of these graphics devices would have been sold in that year! Interesting, especially considering that best 'questimates' of the 1975 market placed the volume at just over $100 million. Today, these terminals are priced from as little as $5000 to as much as $100,000. Assuming the average to be $30,000 there were only about 3,300 terminals placed last year. Somewhat under the estimate for 1973! Aside from a depressed economy, how can we account for the over optimism - or under performance?

I propose the following:

1. Graphics and Design have been conceptually confused. Many of these consoles were predicted for the engineering, design and construction market - i.e. for use in CAD. Quite honestly CAD and computer graphics are not synonymous and have had parallel, not convergent development paths.

2. Design is far more than a picture. Our prediction authorities underestimated the level of computer science and software expertise necessary to tie the picture to a data base on one hand, and to a human engineered application interface on the other.

3. When Graphics devices were applied to the design process they were used to perform complex functions at a stage in the process where the proof of payoff was difficult at best; namely, the conceptual design phase and the drafting phase.

To illustrate these points let us first examine the complexity of the total design process. (This may in fact be a summary of all of the papers delivered here today on the various uses of computers in design). Secondly let us look at the technology involved in Cathode Ray Tube Graphics and how they are used in the picture creation and manipulation process. Lastly, we should explore historic landmarks in the use of Graphics in CAD - successful and unsuccessful. Perhaps then, we can make some educated guesses about the 1980's.

I am indebted to Dr. John Adams of our Engineering Systems Group for summarizing the very complex total design process in a recent tutorial called "The Anatomy of a Project". Actually, there are six distinct stages to any project.

Preliminary Investigation
Preliminary Specification
Preliminary Design
Final Design
Design Documentation
Commitment to Production

With each of these handled by a variety of people; and all of them in some way drawing upon computer resources to assist them in their particular project phase, we could come up with a computational nightmare. For example, the

Corporate Manager uses the computer to do engineering analysis and financial analysis and to ultimately prove feasibility. His computer use payoff comes from the ability to rapidly study alternatives and to provide quicker decisions on project feasibility. The Project Manager is the only person remaining with the project from start to finish. He uses the computer for feasibility studies, planning studies and preliminary scheduling. The Designer and Specialties Engineers use the computer heavily for computational analysis and occasionally for design investigation (PICTURES!) as well as for project control. Farther down the line, Draftsmen use the computer to help produce drawings. These will be final design drawings, working drawings and most important of all, the revisions. While drafting has been a traditional area for the use of graphics devices, I venture to say that it is probably one area in which the payoff is nominal, or at best difficult to measure. Specification Writers use computers for text processing and total documentation. Finally, we have the Fabricator using the computer to investigate alternatives and to keep track of scheduling, and of course money!

The conclusion I come to in looking at all this is that the total process is highly complex and that in fact 'Pictures' are not really that vital to the cost savings realized by using computers.

At this point it is probably appropriate to diverge and take a look at just what these picture generating devices are and how they are applied to the total process. Besides the CRT itself, there are several other devices of interest. First, digitizers which come in many forms, types and prices, but are all used to capture X,Y coordinate data in a non-interactive mode. On the output side, plotters are used to commit pictures to hard copy. Neither of these devices is interactive and neither actually represents any part of the problem other than the picture they capture or reproduce. The CRT, or Scope, is different in that it can be used in an interactive mode where the screen can be used to represent and to communicate with the heart of the problem. These Scopes come in many types and prices. They can be Storage or Refresh, Random-draw or Raster scan and in fact there are even many sub-varieties of these. The most common display system is simply a tube (or bottle) with a phosphorecent coating on the inside that can be activated by an electron beam. If the phosphor retains the image it is termed 'long persistance' and implies Storage graphics that must be erased before new pictures may be drawn. When the phosphor is low persistance, the image fades rapidly requiring regeneration. The scope is then categorized as Refresh. There is always some kind of instruction decoding that sets up either a point mode, a character mode or a vector mode. At Digital, because we are primarily an interactive computer company, we prefer the refresh displays because they allow for more dynamic changes and a higher rate of interactivity. All refresh displays require some type of memory from which the instruction decode unit picks up its commands. This is often called a Display Processing Unit. The Display Processing Unit operates like a CPU and is attached to the I/O bus, or the UNIBUS, of a minicomputer. In their simplest form the interpreters decode a command and place it in the appropriate circuitry for picture creation. More sophisticated systems add niceties such as Jump and Sub-routining capability and other features such as hardware windowing, scaling, and rotation. Some even have instruction look-ahead features. To complicate things even more, dual scope capability is easily available for those brave enough to tackle the software involved in providing two workstations. Too, often purchasers have underestimated the price of software development and the amount of extra effort in making the picture actually do something more than just be a picture. In the early days, few graphics systems were programmed in any other mode than assembly or macro code. Essentially 'Realtime' programming was needed to allow the operator immediate response to his attention getting interrupts. With faster Fortran, this is no longer a problem, and most manufacturers of refresh random-draw displays offer sets of Fortran Callable subroutines to simplify the process. To date there is no one standard interactive graphics language. Progress is being made and the Association for Computing Machinery does now have an active graphics standards committee...and that is 16 years worth of progress.

Just to get the terminology straight:

Display Files are the array of commands designated for building the picture. These are the commands that the Display Processing Unit decodes.

A Primitive is any basic graphic element such as a point, a line or some text.

A Sub-picture consists of a collection of Primitives and calls to other Sub-pictures. Each Sub-picture must be given a Name.

Finally, to interact and change any of these lines or pictures a Pointer is used to refer to specific Primitives.

Typical Fortran Subroutine groupings include those which Control the picture program. That is, they initialize or allow for its dimensioning. There are Drawing routines to allow for creation of vectors, circles, or dots. There are Sub-picture capabilities to allow for Manipulation, or the changing, deleting or inserting of different sub-pictures and Primitives into the Display File. In addition, almost all refresh displays provide registers for setting parameters such as blinking and light pen sensitivity. As if that were not enough, there must be routines for operator interaction through such devices as joy sticks, track balls, and writing tablets. The programming involved in attaching to each of the sub-pictures or picture elements other attributes such as component prices or nodal interconnections, or other geometrics can be complicated.

There are actually two major parts to the problem of inserting graphics into the design process. First, one needs to be a fairly sophisticated computer scientist to be able to understand how to develop data structures that really relate to the problem; how to not only draw the picture on the screen, but how to modify it and all its attributes. Secondly, there is the need for the human interaction to be closely tied to the application so that the person using the screen relates to his problem as he has known it in the past. Rarely does the computer scientist understand the human problems of the operator. Is his arm going to get tired from holding the light pen? Is the part he is designing presented so that he can understand it? Quite typically there is a great need for communication between the industry expert and the computer scientist.

Therefore we face:

The Design Process...A complex task

Computer Graphics...Complex solutions

And now how do we put them together?

Actually, the use of Cathode Ray Tube terminals dates back to the 1950's when the SAGE system used terminals not unlike those today in an Air Defense system. Incidentally, Digital Equipment Corporation supplied Cathode Ray Tubes with the PDP-1. In the early 1960's General Motors pioneered by developing a system still in use called DAC, or Design Augmented by Computer. By 1965 several large companies had developed experimental computer aided design systems using CRT's. Most of these were large and expensive. One of the most famous, the Lockheed CADAM system is reputed to have cost in the vicinity of $12 million to develope. I should add that this system is still in use and is far more than an experiment. Actually close to ten other aircraft companies are presently using the software.

In 1969 the use of computer graphics in CAD became more widespread when several small systems houses brought to market turnkey solutions to the drafting problem. They used storage tube CRT's digitizers and plotters under mini computer control to provide simplified systems for capturing the data, editing and producing final drawings or "photo plots" for creating Integrated Circuits. I view this as a major step in promoting the use of computer graphics devices as design tools. Despite the fact that the conceptual design phase remained manual, these systems were successful because they could be shown to save money in the manufacturing process and in the design verification, or editing stage.

A few Turnkey "Total" Design systems have also been successful. These systems are used in the Garment Industry for pattern layout, in Electronics for printed circuit layout, and in Manufacturing Engineering for geometric construction and for the specification of numerically controlled tool path generation. One consistent item stands out in looking at these systems. They all could be cost justified on $ savings realized in the production phase of the final product. The routing problem solved by the REDAC PC design system not only saved on people utilization, but improved the product by minimizing the use of copper and 'real estate' in the PC board.

The Garment Industry systems are quite popular because they can be justified on fabric savings from optimal pattern placement. They also produce a better product as the cut pieces, being more accurate, can be used in automated sewing machines.

One very successful package in the mechanical design area is that originated by Patrick Hanratty of Manufacturing and Consulting Services. He has developed a series of programs called ADAM 3 - dimensional geometric construction of parts to be machined using numerically controlled manufturing tools. One of the reasons for the widespread use of his programs is that they combine a high degree of computer science expertise with an indepth knowledge of the users needs. They also go beyond the drafting and conceptual design stage to relate that process to the associated data such as parts lists, bills of materials and the NC tapes for driving the tools. Once again payoff can be more easily proven the closer to the manufacturing process we direct the effort. On the computer science side, Mr. Hanratty describes all his geometries in the form of parametric equations so that a piece is defined as a surface generated by revolving an arc about a vector contrasted to the series of connected short vectors used by many automated drafting systems. Obviously the representation of these surfaces is not trivial, and requires considerable knowledge of the geometrics involved. From a human engineering standpoint, the ADAM system allows a part piece to be constructed on the screen by the designer complete with labels and dimensions and with the ability to be reconstructed altered if an error is detected. In other words, by modifying the point in error, all of the subsequent construction in the process is regenerated to reflect the change.

Most of these complex systems have taken many years and many iterations to reach the "successful" stage. Not many buyers are ready to make the man/money commitment that it takes to place computer graphics into their design process. All have gained by the efforts of these few "winners", and still their yield is low. I do not know of one system vendor grossing $20 million per year.

Alternatively, I believe that the use of graphics in the design process need not be complex if one capitalizes on modern software tools and applies the products at the proper place in the design process to justify purchase by proving immediate cost savings.

Allow me to illustrate this with a problem recently solved at Digital Equipment Corporation. We use some very expensive machine tools to punch out simple two dimensional sheet metal pieces that are incorporated into our computer cabinets and power supply boxes. The procedure for cutting such sheets was quite lengthy and required the preparation of a numerical control tape. To create this tape, we use a simple, non-graphical parts programming language called QUICKPOINT. Many such programs are commercially available. The primary problem was how to check the accuracy of the tape. Initially, this was done by running a first trial run on the machine. A very expensive solution since valuable machine time and precious metal were wasted. Several graphics techniques were considered including the use of plotters and turnkey drafting systems. When all commercially available systems were found lacking in some way it was decided to develope a system of our own. The solution was a simple, highly interactive technique based on a Digital Equipment GT46 Display system consisting of a Refresh Random-draw tube and a PDP 11/34 32K minicomputer using RT11, a standard operating system and the DECgraphic-11 Fortran extensions. The NC tape is loaded and the screen displays the overall pattern. The operator, a quality assurance inspector, checks this with the original drawing by zooming in on any particular part piece and requesting a dimension check. The inspector can accurately retrieve dimensions from anywhere on the sheet to a thousandth of an inch. This accuracy is not possible using plotter output. He can specify a local origin on the sheet and measure to or from it. He can place check marks on inspected areas to record his process. He can check shear lines for correct piece fallout. He can check the placement of multiple parts on the sheets for maximum metal usage. He can pick optimum tool paths for maximum production throughput. What was the payoff? Approximately 25 hours a week of expensive machine time formerly used for first run parts was eliminated, not to speak of the material savings. How long did it take to develop and who did it? The system was designed and developed in a little over three months by a mechanical engineer (not a programmer), Mr. Rich Simon, of Digital Equipment Corporation. Now, having easily recaptured his original investment he has justified the purchase of additional graphics and is investigating other phases of the design process to graphically automate. I would like to review an article I recently read in the APEC Journal in which a graphics questionaire sent to members was published. I was pleased to note that about 40% of the firms responding had been using graphics for design for about 5 years each. Expenditures amounted to 10% of their computer budget. However, it was their expectations and applications that concerned me. Few of the applications have been proven money savers. For example, the use of graphics for drafting was a 3 to 1 favorite. I contend, it is difficult to justify the purchase of $150,000 of capital equipment, especially in union shops, on the basis of people replacement, or even improved productivity.

Reading on, at the bottom of desired benefit list was "improved accuracy". I contend that more graphics systems would be installed if improved accuracy and improved products were used as the justification for the purchase of graphics in the design process.

With todays improved hardware, software and understanding of the products and problems, we may see a greatly increased use of Graphics in the CAD process, IF we keep our financial administrators in mind and direct the initial use of graphics to the simplest, most cost effective phases of the problem.

REFERENCE LIST

1. Carl Machover, "CRT Computer Graphics In The United States," Presented at SICOB, France, October, 1968.

2. Patrick J. Hanratty, "Computer Aided Design, Drafting and Fabrication On Low Cost Systems," Presented at SHARE XLII, Houston, Texas, March, 1974.

3. Leo M. Frawley, "APEC Graphics Questionaire Summary," Automated Procedures for Engineering Consultants, Inc., Summer of 1976.

SESSION III

MATERIALS

OF

DESIGN

(I)

-COMPOSITES-

Chairman: C. O. Smith
University of Nebraska at Omaha

REVIEW OF THE PROCESSING AND PROPERTIES OF PAN GRAPHITE-ALUMINUM COMPOSITES

Roger T. Pepper
H. Gigerenzer
T. A. Zack
Fiber Materials, Inc.
Biddeford Industrial Park
Biddeford, Maine 04005

Abstract: The development of a liquid metal infiltration process and successful fiber-matrix interface barrier coatings has led to an exciting new class of graphite fiber reinforced metals. Composite wire preforms are continuously cast on a pilot production basis using the liquid infiltration technique. Aluminum composites reinforced with PAN graphite fibers have shown strengths in excess of 150,000 psi and modulus of up to 20 million psi. Work is currently underway to develop low-cost fabrication processes for hardware, using the wire preform material, by the modification of conventional mechanical working techniques such as metal drawing. A review of the processes and mechanical behavior of graphite fiber reinforced aluminum is given in the paper below.

Key Words: Composite material; fabrication techniques; graphite fibers-aluminum; mechanical behavior; pultrusion/drawing; wire preform.

Introduction

The simultaneous advents of potentially low-cost, high-strength and stiffness multifiber graphite materials and new process technology to incorporate them in lightweight metals such as aluminum on a high volume production basis gives the potential for dramatic improvement in structural reinforcement without significant increase in cost.

High performance polyacrylonitrile (PAN) precursor graphite fibers of moderate cost have been available for the past several years and have been widely used to reinforce organic resin matrices. Past attempts to incorporate these PAN fibers into aluminum have been unsuccessful due to a lack of chemical stability of the fibers causing severe degradation of strength. Recently, significant progress has been made in the development of new interface barriers for PAN graphite fiber reinforced aluminum.

Graphite-Aluminum Composite Preparation

Aluminum does not readily wet graphite and attacks the carbon surface forming a metal carbide (Al_4C_3). Such carbide formation degrades carbon fibers severely, lowering their mechanical properties. Various techniques have been developed to overcome the wetting and degradation problems. Sara[1] shows that graphite-aluminum fiber composites can be formed by first coating the fibers with a tantalum film by electro-deposition from a fused salt bath, outgassing the fibers by pumping them down to a very low pressure and submerging the outgassed fibers into a pressurized molten aluminum bath to fill the interstices of the fibers. A similar process is also described in another U.S. Patent issued to R. V. Sara[2] in which the carbon fibers are first coated with silver or a silver-aluminum alloy by electrodeposition from a plating solution. The fibers are then contacted with aluminum foil and the combined foil-fiber is heated, while under pressure, to the solidus temperature of the foil. In both of these systems, it is also suggested that the metal coating can be applied by sputtering or by reduction of salts of the metal. Whether using silver or tantalum, it is difficult to obtain uniform thin coatings on the fibers and, in any event, the resulting composites are not desirable since they contain substantial amounts of expensive, heavy elements.

Electrodeposition and chemical deposition techniques have also been used to deposit the matrix material directly on graphite fibers, the aluminum coated fibers being subsequently hot pressed to form composites. The major disadvantage of forming composites with such deposition techniques is that, for the most part, the matrix material is usually limited to a pure metal which, for many purposes, has markedly inferior properties compared to alloys. A fourth process which has undergone development and laboratory demonstration is a liquid metal infiltration process[3] which uses nickel to promote the wetting of graphite fibers by liquid aluminum. The nickel coatings required for good wetting and protection of the fibers from attack by liquid aluminum are approximately 1000 angstrom units thick, giving rise to nickel contents of 5-10 weight percent in the aluminum alloy matrix at 30-50 volume percent fibers. Such high nickel impurity levels result in a drastic lowering of the ductility of the aluminum alloy matrix and the lowering of corrosion resistance.

A liquid metal infiltration process has been developed[4,5 and 6] which involves the use of very small concentrations of titanium and boron to promote the wetting of graphite fibers by commercial aluminum alloys. The graphite fibers are coated with a layer of Ti/B between 100 and 200 angstrom units thick to promote wetting and to protect them from attack by molten aluminum. This results in considerably less than 0.5 weight percent titanium and boron in the composite matrix. Similar concentrations of titanium and boron have been successfully used as grain refining additions to commercial aluminum alloys for the past 20 years and have beneficial effects by increasing alloy strength, ductility, and corrosion resistance. The titanium and boron are deposited on the fibers by the reduction of titanium tetrachloride and boron trichloride using zinc vapor, the zinc being continually extracted from the coating step of the process as zinc chloride gas. The coated fibers are infiltrated by drawing through a molten aluminum alloy bath producing a unidirectional graphite-aluminum composite wire. Significant advances have been made toward developing the Ti/B liquid infiltration process for future large-scale manufacturing (Figure 1). For example,

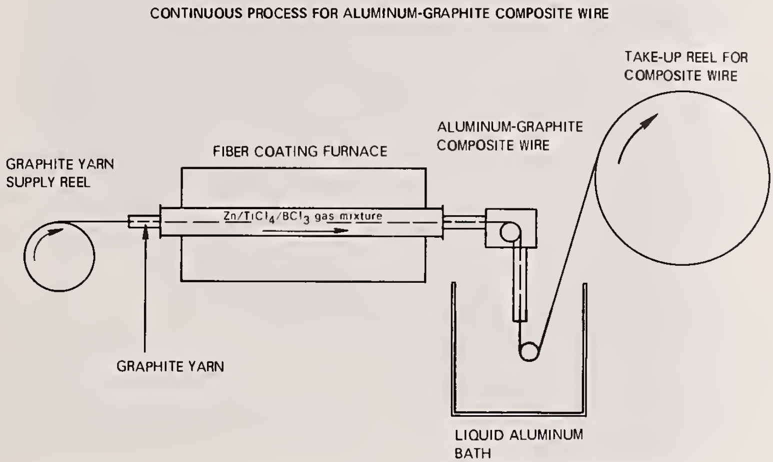


Figure 1

the introduction of textile techniques for feeding fiber into the Ti/B coating chamber and the use of foundry style aluminum open melt techniques for infiltration both facilitate composite wire handling and show the way for the future continuous casting of shapes directly from the melt.

Figure 2 shows a scanning electron micrograph of a polished end section of a graphite-aluminum composite wire containing approximately 11,000 filaments produced by the liquid infiltration unit. Good fiber distribution and excellent metal wetting characteristics are apparent. The graphite fiber in this composite is high modulus T50, a rayon precursor graphite fiber. Prior to 1976, rayon based graphite fibers were used almost exclusively in the development of metal matrix graphite fiber composites. Rule-of-mixture tensile strengths and elastic modulus have been attained utilizing a variety of aluminum alloys with these fibers. Typical fiber properties are shown in Table 1, and composite wire and fabricated bar properties are shown in Table 2.

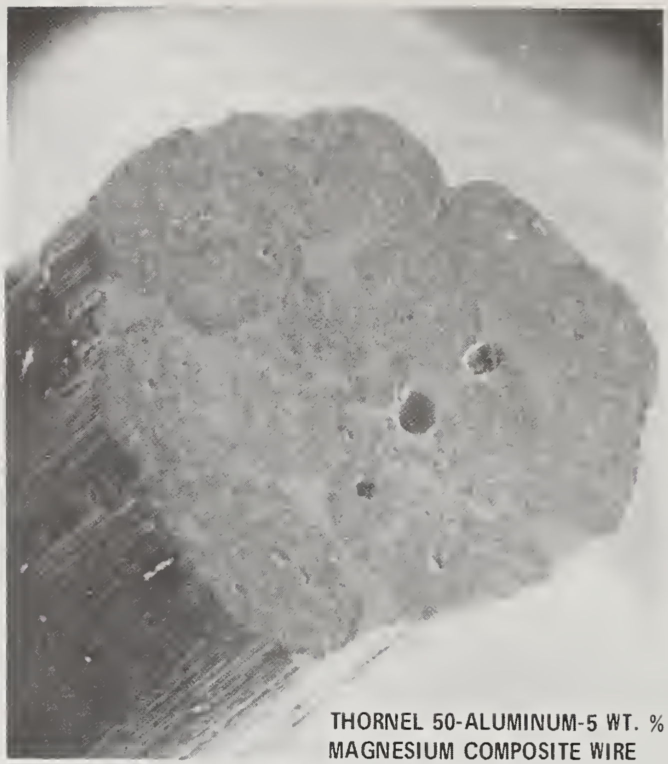


Figure 2

Table 1

TYPICAL FIBER PROPERTIES

Fiber Designation	Strength ($x10^3$ psi)		Elastic Modulus ($x10^6$ psi)	
T50 (rayon)	315	(2172 MPa)	57	(393 GPa)
T50 (PAN)	300	(2068 MPa)	50	(345 GPa)
T300 (PAN)	360	(2482 MPa)	34	(234 GPa)
HMS, HM/PVA (PAN)	340 min.	(2344 MPa)	50-55	(345-379 GPa)

Table 2

STATE-OF-THE-ART T50 GRAPHITE/ALUMINUM COMPOSITE MECHANICAL PROPERTIES

Property	35 v/o T50/A201 Wire	42 v/o T50/A201 Wire	30 v/o T50/A201 Test Bar
Tensile Strength	120 Ksi 828 MPa	155 Ksi 1069 MPa	90 Ksi 621 MPa
Tensile Modulus	27 Mpsi 193 GPa	30 Mpsi 207 GPa	20 Mpsi 138 GPa
Transverse Strength	5 Ksi 34.5 MPa	NOT TESTED	7.1 Ksi 49.0 MPa

Previous work using PAN based graphite fibers in metal matrices had shown PAN based graphite to be more reactive with liquid aluminum than rayon based graphite fibers. Therefore, poor translation of fiber strength and elastic modulus resulted in the composite (Table 3).

Table 3

MECHANICAL PROPERTY COMPARISON PAN BASED GRAPHITE-ALUMINUM COMPOSITES (PAST DATA vs. PRESENT STATE-OF-THE-ART)

Property	As Received T300 A201 Alloy (Wire)	Barrier Coated T300 A201 Alloy (Wire)
Tensile Strength	50.2 Ksi 346.4 MPa	180 Ksi 1242 MPa
Tensile Modulus	---	20 Mpsi 138 GPa
Fiber Content	45.9%	40.0%

During the past year, a significant breakthrough has been made in research and development on PAN graphite fiber reinforced aluminum composites. Modified fiber barrier coatings have been developed and demonstrated which prevent attack and degradation of the fibers by aluminum alloys during liquid metal infiltration and subsequent fabrication of composite wire into shapes (Table 3). Both moderate and high modulus PAN fibers have been processed with the new coatings. Typical mechanical properties for the newly developed PAN based graphite fiber reinforced aluminum matrices are shown in Table 4. Note the high translation of tensile strength and elastic modulus for fabricated test bars. These bars were produced from PAN precursor composite wire (approximately .035 in diameter) using a hot pressing technique.

Table 4

PAN BASED GRAPHITE FIBER REINFORCED ALUMINUM COMPOSITE PROPERTIES

Property	40 v/o T300 A201 (Wire)	30 v/o T300/A201 Test Bar	40 v/o T300/A201 Test Bar
Tensile Strength	180 Ksi 1242 MPa	125 Ksi 863 MPa	150 - 160 Ksi 1035 - 1104 MPa
Tensile Modulus	20 Mpsi 138 GPa	--- ---	21 Mpsi 145 GPa
Flexure Transverse	--- ---	16 Ksi 110 MPa	--- ---
Tensile Transverse	--- ---	--- ---	5 Ksi 34.5 MPa

Graphite-Aluminum Mechanical Behavior

The new fiber coating and bonding technology has resulted in the development of a ductile Thornel 300 (PAN based) graphite fiber reinforced aluminum composite. This achievement of ductility is extremely important since it implies that toughness or forgiveness necessary to prevent sudden catastrophic failures in highly stressed hardware can be developed in graphite-metal composites. To illustrate this composite ductility effect observed in the newly developed PAN based graphite-aluminum composites, a comparison of the mechanical behavior of rayon based (T50) and polyacrylonitrile based (T300) graphite fiber reinforced aluminum composites is made below. Figure 3 illustrates some typical stress/strain data for rayon (T50) and PAN (T300) precursor graphite fiber reinforced aluminum matrices in the as-cast condition[7]. See Table 1 and Table 5 for fiber and matrix properties, respectively. The T50/A201 Al wire composite at 42 volume percent fiber shows a completely elastic stress/strain relationship having a modulus of elasticity of 30 Msi (207 GPa) up to the ultimate tensile strength of 160 Ksi (1103 MPa). The total strain-to-failure for this composite is 0.53%, which is representative of the strain range 0.45 to 0.55% usually observed for this system. The stress/strain behavior is typically characteristic of a high modulus high strength composite material failing in a brittle manner. In contrast, the T300 (PAN based)/356 Al composites at 35 volume percent fiber show a pronounced increase in total strain-

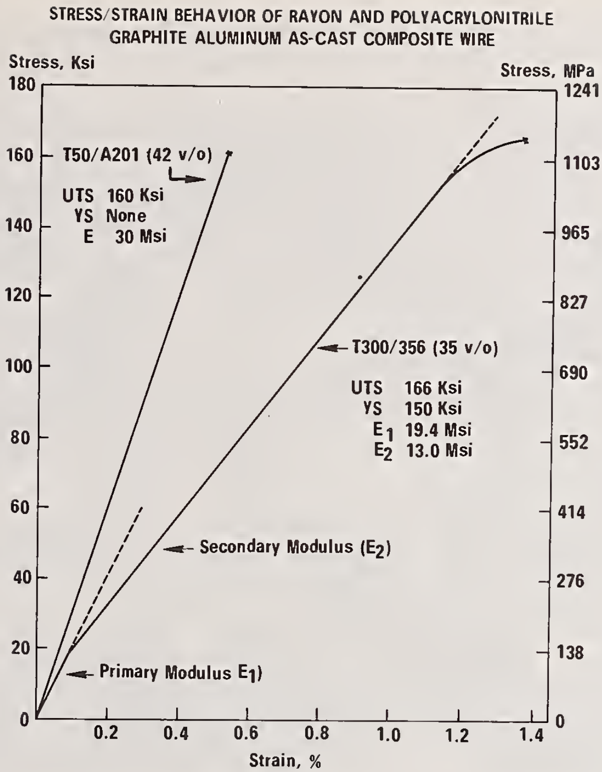


Figure 3

Table 5

TYPICAL MATRIX PROPERTIES

Matrix Designation	Nominal Composition (Weight Percent)						Strength ($x10^3$ psi)
	Cu	Ag	Si	Mg	Cr	Al	
A201	4.7	0.6	0.1	0.3	--	Bal	20[1] (138 MPa) 50[2] (345 MPa)
356	0.2	--	7.0	0.3	--	Bal	25[1] (172 MPa) 38[2] (262 MPa)
6061	0.25	--	0.6	1.0	0.20	Bal	18[1] (124 MPa) 45[2] (310 MPa)

NOTE:

(1) As Cast
(2) Heat Treated
(3) Elastic Modulus of All Alloys Approximately 10^6 psi (69 GPa)

to-failure, (1.4%) and show an initial high primary elastic modulus (typically 17-19 Msi, 117-131 GPa) with a transition at approximately 20 Ksi (138 MPa) stress level and 0.1% strain to a lower secondary modulus (typically 13-15 Msi, 90-103 GPa). The T300/356 Al composite exhibits a marked deviation from linear behavior prior to failure at the ultimate tensile strength of 166 Ksi (1145 MPa). Cyclic loading the T300/Al composites to their yield stress followed by unloading results in elimination of the primary modulus. The value of the resultant average elastic modulus obtained on cycling is typically 17 Msi (117 GPa). Similar ductile-like behavior is observed, but to a lesser degree for high modulus PAN based graphite fiber reinforced Al composites (Figure 4).

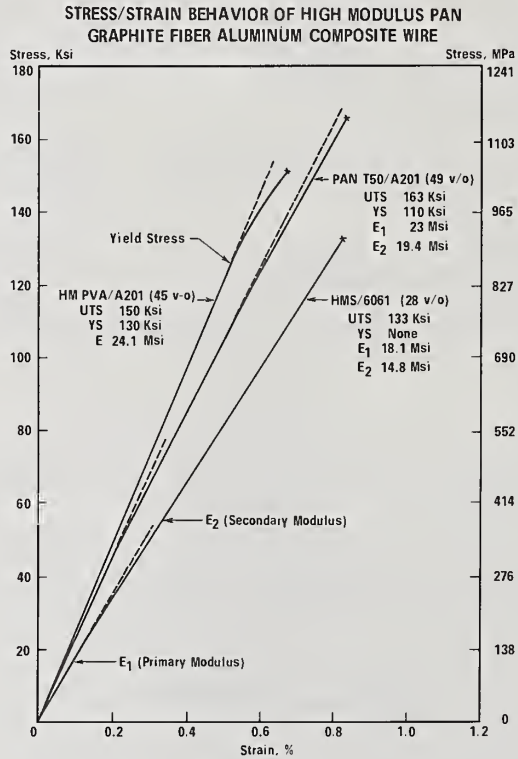


Figure 4

The linear stress/strain behavior of the T50 (rayon based)/A201 Al composite in the as-cast condition suggests that the fiber and matrix are essentially strained to failure elastically. There appears to be no plastic contribution by the matrix to the overall deformation behavior of the composite. This suggests that, in the rayon based graphite-aluminum composite, the matrix is not bonded to the fiber well enough for the matrix to contribute its plasticity to the composite when subjected to a tensile strain. The observed strength at which the composite fails, 160 Ksi (1103 MPa) and strain (0.53%), is probably due entirely to the fiber. That this is the case is supported by examining the tensile fracture surfaces of the as-cast T50/A201 Al composite (Figure 5). It is evident that extensive fiber pullout has occurred, which is indicative of a weak fiber to matrix bond.

T50 (RAYON BASED)/A201 TENSILE FRACTURE SHOWING FIBER PULLOUT

Figure 5

The stress/strain behavior of the PAN precursor graphite fiber aluminum composite (T300/356 Al), however, indicates that both fiber and matrix are strained elastically only up to a stress level of approximately 20 Ksi. At this point, the elastic modulus of the composite changes from the primary E_1 (fiber and matrix elastic) to the secondary E_2 stage (fiber elastic, matrix plastic), and this change is attributed to the onset of microplastic deformation in the matrix. In addition, the T300/356 Al composite also shows a definite yield stress of 150 Ksi (1035 MPa) at 1.1% strain, at which non-linear behavior (macro-plasticity) is evident. No plastic deformation occurs in T50 or T300 graphite fibers; thus, the non-linearity prior to failure in T300/Al is attributed to progressive failure of the fibers into discontinuous segments. This has been observed on tensile fracture edges and on chemically extracted fibers from tested tensile specimens. The total strain-to-failure of the composite is greater than the maximum strain-to-failure (1.1%) of the T300 fiber because the matrix continues to deform after fiber breakage but is still reinforced by the discontinuous broken fiber segments. For this effect to occur, an improved fiber/matrix bond is a prerequisite and must be maintained during progressive fiber failure. Examination of tensile fracture surfaces (Figure 6) of as-cast composite wire specimens of T300/356 Al reveals little fiber pullout and

T300 (PAN BASED)/A201 Al TENSILE FRACTURE SHOWING GOOD FIBER/MATRIX BOND.

Figure 6

pronounced matrix yielding (cellular structure) around the fibers. These observations indicate that an improved bond has been achieved between the matrix and fiber which effectively transfers tensile load from the matrix to the fiber by matrix shear.

Additional evidence of composite plastic behavior is shown by examining progressive stress cycling data (Figure 7). Loading the composite to below 20 Ksi (138 MPa) and unloading (cycle 1) shows only composite elastic behavior. Loading/unloading above 20 Ksi (138 MPa) shows E_1/E_2 modulus transition and retainment of permanent set (cycle 2 through 7) on unloading. The E_1/E_2 transition point increases rapidly with small increments of strain indicating that a high rate of strain hardening is occurring in the matrix (cycle 2 through 4). The abrupt E_1/E_2 transition disappears at cycle 5 through 7. True composite plastic behavior is evidenced when one considers the loading and unloading slopes of cycles 5 through 7. On unloading, the initial slope is parallel to the initial loading slope indicating that the first contraction on unloading is elastic. The remainder of the slope on unloading to zero load is parallel to the secondary E_2 slope on loading. This indicates that the remaining contraction on unloading is plastic. The elastic contractive force exerted by the fiber on the matrix results in plastic compression of the matrix. In conventional and completely elastic materials, all the contraction would be elastic when the load is

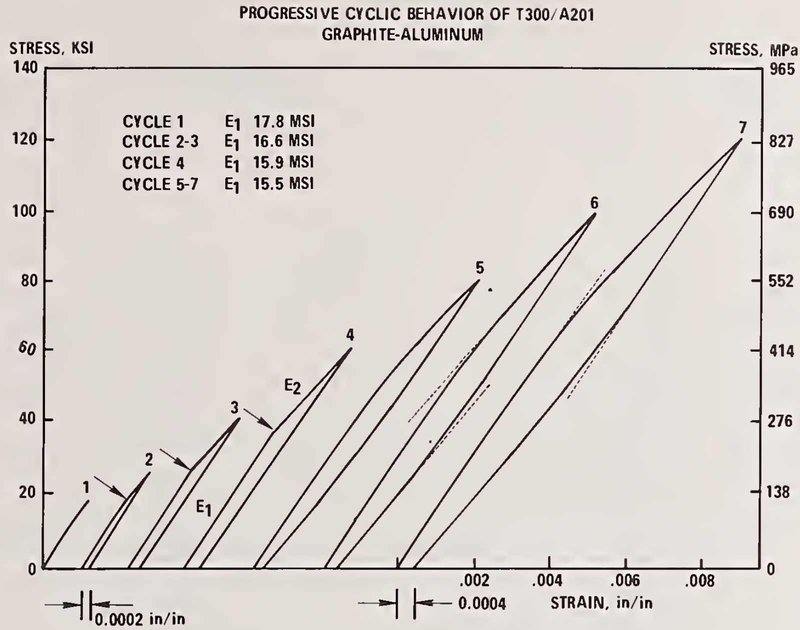


Figure 7

released. Cyclic loading between 20 and 120 Ksi (138-827 MPa) shows a relatively small hysteresis loss indicating potential high fatigue resistance of the composite.

Little fiber degradation has been observed in processed graphite-aluminum composites on heating in the solid state at times of up to three hours, Figure 8. Some reaction does occur, however, when solidus temperature (typically 500°C) is exceeded. Good compatibility exists between graphite fibers and aluminum matrix up to 500°C and, thus, the material is a prime candidate for applications requiring high strength and stiffness to density at elevated temperatures.

Graphite-Metal Fabrication Development

The anticipated large volume utilization of metal-graphite composites to satisfy the ever-increasing demands of present and future specialized structural applications will critically depend upon the successful development of high rate fabrication processes to effectively produce useful structural engineering shapes. As lower cost composite preform materials become available (for instance, the utilization of PAN based T300 fibers and the development of future pitch based fibers), high rate fabrication processes must be developed which are not only

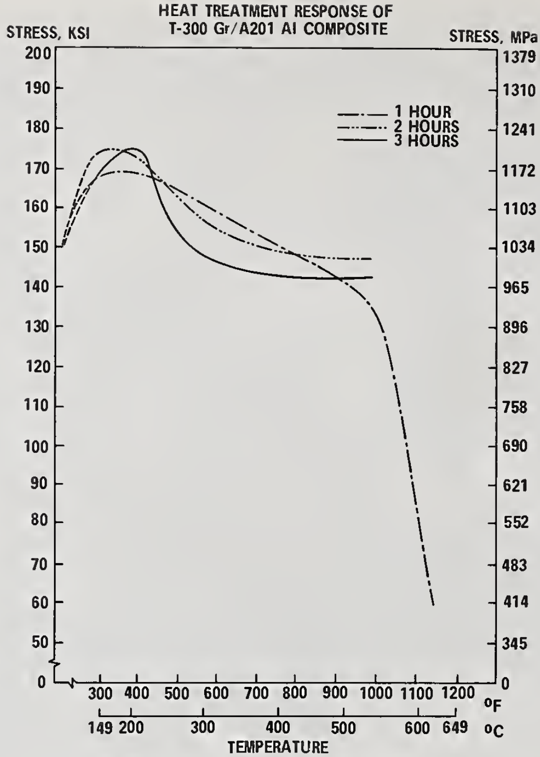


Figure 8

capable of translating composite wire preform properties into bulk composite engineering shapes but, in addition, must also be capable to achieve this end on a cost-effective basis. A development approach being taken for the fabrication of premium quality graphite-metal structural engineering shapes is to modify conventional high rate metalworking processes for adoption to metal-graphite processing. One such process having received much development and preliminary success is controlled temperature pultrusion/drawing for the fabrication of uniaxial graphite fiber reinforced tubes and long cylinders. Other composite fabrication processes under development include hot isostatic pressing (HIP) of complex structural shapes, mechanical hot pressing of unidirectional and cross-ply structural panels, rolling of graphite-aluminum tape preforms and strip, and casting of prewoven fiber shapes.

A large amount of matrix deformation occurs during the mechanical working of composites causing buckling stresses at the fiber/matrix interfaces which can lead to interface and fiber surface damage and, in extreme cases, severe fiber fragmentation. The low strengths observed in past attempts at hot working composites (extrusion) are thought to occur primarily because the fiber/matrix bond of the composite has been damaged to the point where only limited load transfer from the matrix to the fiber can be achieved or because the fibers are degraded by notching of their surfaces. A pultrusion process is under development to fabricate high strength metal-graphite structural composites. This process differs from other hot working techniques in that consolidation of the fibers and matrix occurs with considerably less overall matrix deformation. During pultrusion, the fibers are mainly loaded in tension[8,9], the direction of maximum fiber strength. Therefore, little fiber breakage and interface debonding occurs during the consolidation process since buckling stresses along the fiber/matrix interface are minimized.

Continuous T50 graphite fiber/A201 aluminum composite reinforced tubes 1" OD x 20" long and cylinders 2" OD x 30" long have been successfully fabricated to date[10]. Development of the process is being conducted on a modified 10 ton draw bench. Graphite-aluminum wire preform is sealed in an evacuated stainless steel can and drawn through a silicon carbide die at a constant temperature below the solidus or in the lower part of the liquid + solid range of the composite matrix alloy (see Figure 9 for schematic illustration of process). Standard graphite-base extrusion lubricants are used to minimize friction between the outer surface of the can and the die wall.

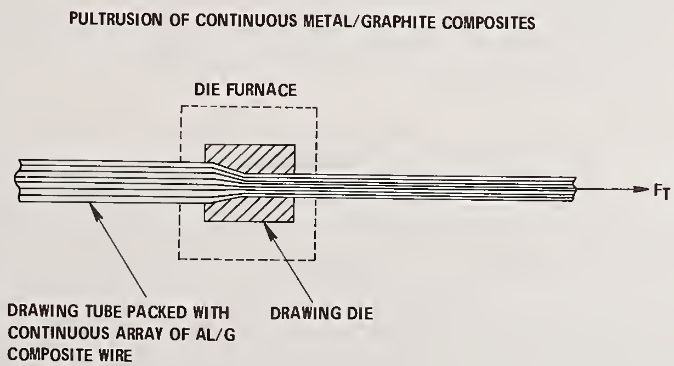


Figure 9

Thin wall tube configurations are a necessity for structural efficiency. Thus, a major goal of the graphite-aluminum pultrusion development is development of fabrication technology for thin wall tubes and cylinders. To achieve desired wall thicknesses in the range 1-60 mm, hot working of composites is necessary. Current efforts are concentrating on the T300 class of graphite-aluminum composites which have shown evidence of ductility at room temperature. Floating mandrel tube drawing techniques are being used with the major deformation mode being 'draw' (change in wall thickness) with minimum 'sink' (change in diameter). Extensive studies of the effects of hot working graphite-aluminum composites by pultrusion/drawing are in progress with particular attention being given to the effects of working on mechanical properties and fiber or fiber-matrix interface damage which may occur during deformation. Wrought aluminum alloys such as the 5000 and 6000 series, which are not susceptible to hot shortness, are being used. To achieve low-cost metal composites in fabricated forms such as tubes and other structural sections, the adaptation of conventional metal working processes to composite fabrication will be necessary.

Historically, the most used fabrication technique for metal composites has been solid state hot pressing. Sample bars, panels, and small parts have been prepared in this way by laying-up composite preform wire or tape in a closed die or can and pressing at just below the solidus of the matrix alloy (typically $500^{o}C$) for about 30 minutes. Such a fabrication method is not foreseen to have potential in production due to the long set-up and bonding times required, but modified versions of the process may develop for the closed die pressing/forging of complex shapes such as jet engine fan blades. Hot isostatic pressing is another technique that may have application to the fabrication of complex shapes, where the complexity of the part justifies the use of a slow batch-type of process.

Summary

High strength aluminum and titanium alloys have been developed over the past fifty years which are superior to steel on a strength to density basis and are now in widespread use. The development of premium high strength and stiffness graphite fibers over the past ten years surpasses the above developments of conventional metals in terms of potential improvement in the strength and particularly the stiffness of structural materials. For example, graphite fiber reinforced aluminum is two and three times better than commercial high strength aluminum alloys on a strength and stiffness to density basis, respectively. An R & D data base has been obtained over the past three years with graphite fiber reinforced aluminum and excellent mechanical properties have been attained. New technology is being developed for the production of high strength, high modulus graphite fibers from low-cost pitch, and it is now anticipated that, within the next five years, performance equal to that achieved with PAN graphite fibers will be realized. The development of low-cost pitch precursor graphite fibers will have a tremendous impact on the industry since it will facilitate the production of lightweight graphite-aluminum composites of substantially increased strength and stiffness in comparison to conventional metals at low increase in cost.

References

1. Sara, R. V., U.S. Patent No. 3,553,820.

2. Sara, R. V., U.S. Patent No. 3,571,901.

3. Perry, A. J., U.S. Patent No. 3,720,257.

4. Lachman, W. L. et al, Graphite Composite, U.S. Patent No. 3,860,443, 1975.

5. Lachman, W. L. et al, Graphite Composite, U.S. Patent No. 3,894,863, 1975.

6. Amateau, M. F., Progress in the Development of Graphite-Aluminum Composites by Liquid Infiltration Technology, Aerospace Corporation Report No. ATR-76 (8162)-3, 30 August 76.

7. Gigerenzer, H., Zack, T. A. and Pepper, R. T., Observations on the Mechanical Behavior of Rayon and Polyacrylonitrile Graphite-Aluminum Composites, Second International Conference on Mechanical Behavior of Materials, Boston, 1976.

8. Spencer, R. A. P., Advances in Pultrusion of Carbon Fibre Composites, Paper No. 21, Proceedings of the Second International Conference on Carbon Fibers, Their Place in Modern Technology, Plastics Institute, London, 1974.

9. Paton, W., Lockhart, A. H. and Montgomery, I. W., The Use of Drawing Methods in C.F.R.P. Fabrication, Paper No. 22, Proceedings of the Second International Conference on Carbon Fibers, Their Place in Modern Technology, Plastics Institute, London, 1974.

10. Pepper, R. T., Gigerenzer, H. and Zack, T. A., Development of Graphite Fiber Reinforced Aluminum Tubes, Proceedings of the 8th National SAMPE Conference, October 12-14, 1976, Seattle.

COMPOSITE INLAYS FOR JET ENGINES

W. Troha
USAF Aero Propulsion Laboratory
Wright-Patterson AFB, Ohio 45433

K. Swain
Detroit Diesel Allison Division
General Motors Corporation
Indianapolis, Indiana 46206

Abstract: A method of using composite material to improve turbine engine blade vibration stability has been demonstrated on a TF41 fan rig and engine. The procedure uses the high strength and stiffness properties of B/SiC - titanium composite material, bonded into the leading edge trip section of an all titanium fan blade. Comparisons are made between the blade response characteristics with and without the composite inlay. These comparisons are made for blade natural frequency, untwist due to rotational speed, and shift in flutter boundary. Engine performance improvements due to blade shroud removal are also presented.

Key words: Blade flutter control; composite materials; jet engines.

Introduction: The objective of this paper is to present the results of a recent demonstration for improving turbine engine fan blade flutter resistance by the placement of high modulus composite material in the outer span portion of a flutter sensitive TF41-A100 first stage fan blade. The TF41 is a two-spool low bypass turbofan engine in the 15,000-lb (6.7-kN) thrust class. The TF41 Fan-IPC (fan-intermediate pressure compressor) consists of a three-stage fan with no inlet guide vanes followed by a two-stage IPC on the same spool. The flow entering the fan is split into a bypass stream and a primary stream by a flow splitter located at the inlet to the IPC compressor.

The A100 compressor first stage rotor contains 25 monolithic, unshrouded titanium blades. Each airfoil is approximately 11 in. (28 cm) long with a 5-in. (12-cm) tip chord. Tip speed is 1500 fps (457 m/sec) at design. This blade experienced torsional flutter at a reduced frequency parameter ($b\omega_T/V_{rel}$) of 0.53 and at supersonic inlet relative Mach number.

This paper was first submitted by the authors at the ASME Gas Turbine Conference in March 1976, under publication No. 76-GT-29.

Design and Analysis: A parametric analysis was performed to define the optimum geometry, location, and fiber orientation of the B/SiC-Ti (boron fibers coated with silicon carbide diffusion bonded on a volume ratio basis of one to one in a 6AL-4V titanium matrix) composite inlay. The geometry of the inlay area (Fig. 1) was selected so that flutter stability could be significantly improved, and to allow adherence to customary blade vibration-stress design criteria and practical machining practices. Fiber orientations considered were radial, unidirectional at 45 deg, ± 45-deg cross ply, and transverse. The lower boundary of the inlay was restricted radially to an area of reasonably low stress in recognition of the potentially high stress concentrations occurring at the boundaries. The lower boundary was also restricted on the basis of total inlay length. A desirable objective is the limitation of the inlay length to the point where the overall amount of inlay twist and camber does not induce excessive "oil canning" in the composite plies during the lay-up operation. The upper and leading-edge boundaries, as well as the inlay cross-section, were established to obtain maximum inlay volume while preserving a high confidence of fabricability. Fiber orientation was found to have an insignificant effect on torsional stiffness of the inlay airfoil. Hence, a radial orientation was selected on the basis of its inherent fabrication advantages by being in tape form. In addition, various inlay cross-sections were analyzed, as shown in Fig. 2. As expected, the predicted dynamic responses of the core and no-core configurations were essentially the same. The constant thickness core cross-section was selected since the core provided ply locating advantages during fabrication. Selection of the chordwise location of the inlay was based on consideration for its effect on the position of the torsional mode twist axis. A more forward position (toward the leading edge of the airfoil) was considered desirable. The aeroelastic analyses methods available at the time predicted that the forward location of the inlay would produce a forward shift in the twist axis of approximately 2 percent C to a 46 percent C position. Actual blade test, however, showed the twist axis was moved aft by about 2 percent C. Analyses using test results and supersonic unsteady theory have since indicated that this aft shaft had little effect on stability of the inlay blade.

The calculated vibration characteristics of the inlay airfoil design are compared to those of the baseline monolithic titanium LP1 fan blade in Fig. 3. Finite-element techniques similar to NASTRAN were used in the analysis. No engine order coincidence problem was predicted.

The results of the finite-element stress analysis of the inlay airfoil showed a maximum stress of 98 ksi (676 mPa) at the airfoil hub. This compares with 104 ksi (717 mPa) for the all titanium airfoil. The slight blade weight reduction associated with the lower-density composite material accounts for lower overall stress levels. A stress of 57 ksi (393 mPa) was calculated for the area near the lower inlay boundary on the suction side of the airfoil. This region was further analyzed to assess the stress concentration associated with inlay scarf angle (Fig. 4). The results of this analysis are presented in Fig. 5 as a plot of maximum joint stress versus scarf angle. It was concluded that joint tensile stress is not greatly influenced by scarf angle. Fig. 6 is a Goodman diagram which shows that with a stress concentration of 1.54 and a mean stress of 57 ksi (393 mPa), an alternating stress of ± 28 ksi (193 mPa) can be tolerated. This is well above the usual blade design requirement.

Fabrication: A contoured graphite electrode was used to electrical discharge machine (EDM) the recesses in the blade airfoil. The electrode contained a series of coolant holes in the contour surface to flush out spark-eroded titanium particles between the tool and the work piece to obtain good surface finish and control recess geometry. These holes leave small pedestals on the thin titanium core which must be removed by a hand operation. A typical airfoil recess after EDM is shown in Fig. 7.

The first step in the blade lay-up was the cutting of composite monotape ply shapes. The computer-calculated outline was scribed on the monotape panel which was then rough cut with a straight-edge shear to the approximate contour, and hand ground on a rotating diamond disk to the scribe line contour. A typical set of finish cut plies is shown in Fig. 8. The monotape plies were etched to final thickness after rough cutting. After final edge grinding, the plies were given a light dip etch in the same solution, rinsed in water, dried, and stored for assembly. The blade recess and adjacent areas were scrubbed, acid etched, rinsed, and dried prior to lay-up. The clean blade was then positioned in a resin filled fixture, contoured to the blade geometry, with welding ground wires clamped to the platform and blade tip for capacitor discharge tack welding of the composite plies. Optimum tack welding parameters were determined such that diffusion bond type rather than nugget type welds were formed to minimize stress concentration factor. The first composite ply was tack welded to the Ti-6Al-4V core, with each succeeding ply tack welded to the ply below. A 0.020-in. (0.51-mm) Ti-6Al-4V cover sheet was then welded to the blade airfoil around the periphery of the recess, on each side of the blade, in an evacuated chamber. The final closed recess must be evacuated to obtain successful diffusion bonds.

The diffusion bonding was done using a hot isostatic press (HIP). A typical blade after diffusion bonding is shown in Fig. 9. Figs. 10(a) and 10(b) show the good filament array and excellent bond quality characteristic of all examined sections. The final operation was to hand finish the airfoil surfaces back to original dimension as shown in Fig. 11.

Test and Evaluation: The first three natural frequencies of 27 blades were determined with the aid of an acoustical siren and a 30,000-lb

(13.3-kN) base load on the dovetail. Table 1 is a listing of the average frequencies of 27 blades for three modes, and a comparison with predicted values. The blade-blade scatter of the composite inlay blades was approximately the same as that observed for a production fan blade. Holographically determined mode shapes showed that the torsional mode twist axis was shifted approximately 2 percent C, toward the trailing edge in the tip region, to a 50 percent C position.

The fatigue or endurance strength of the inlay airfoils were determined by a stress increment test. In this type of fatigue test, the airfoil is vibrated in a particular mode for 5×10^6 cycles at successively higher alternating stress levels until failure occurs. The initial blade fatigue test was conducted in the first torsional vibration mode. First torsion was selected because the high speed flutter encountered during rig testing of the standard all-titanium fan blades was a first torsional mode response. The blade failed at an endurance strength of $\pm$ 42 ksi (290 mPa) in the parent titanium material about 1 in. (2.54 cm) up from the platform on the trailing edge. The $\pm$ 42 ksi (290-mPa) test value is comparable with that expected at the trailing edge radius of an all-titanium blade. This would indicate that the fabrication of the inlay blade does not degrade the fatigue properties of the titanium portion of the airfoil and that the fatigue properties in the inlay region are adequate if the blade were excited in first torsion during operation. However, the stress survey indicated that the relative stresses in the inlay area are very low when the blade is excited in first torsion and that near-maximum alternating stress in the airfoil does occur near the lower boundary of the inlay when the blade is excited in the second bending mode. Subsequent blade fatigue tests were conducted in the second bending mode. Alternating stress levels of $\pm$ 13 ksi (90 mPa) and $\pm$ 23 ksi (159 mPa) in the parent airfoil material adjacent to the inlay interface were run for 5×10^6 cycles each without failure. Failure was eventually induced at the dovetail-stalk radius after 3.6×10^6 cycles into the third stress level step of $\pm$ 33 ksi (228 mPa). Although the fatigue properties at the interface were not defined, the test demonstrated adequate endurance strength for the planned rig and engine testing. In addition, it indicated that the residual stresses and local stress concentrations associated with diffusion bonding and tack welding were not of catastrophic magnitude. An inlay interface fatigue failure was produced when a test specimen was vibrated in a free-free mode. A fatigue strength of $\pm$ 37 ksi (255 mPa) was demonstrated. Based on these tests and analytical predictions, the inlay interface produces a notch factor of approximately 1.5. Subsequent specimen tests have shown a fatigue strength of 50 ksi at the composite/titanium interface. A limited amount of static or bench type bird impact testing of one composite inlay blade was done at the University of Dayton Test Facilities, Dayton, Ohio. Preliminary results indicate that the inlay blade withstood an RTV 560 bird-strike (3/4 span) at a velocity of 1436 fps (438 m/sec). A 2.4-oz (0.068-kg) slice weight was obtained which (at the given velocity) produced an impact approximately equivalent to a 3-lb bird-strike in an engine. The maximum blade tip deflection at impact as determined from high speed photography was approximately 2 in. (5.1 cm). After testing, the blade was visually and X-ray inspected and was found to be undamaged.

An uninstrumented spin proof test was run to demonstrate the structural integrity of the B/SiC-Ti inlay blades. A first stage fan wheel was assembled with 21 "dummy" blades, two standard titanium blades, and two composite material inlay blades. The "dummy" blades were sized to produce centrifugal loads comparable with those of actual blades. The assembly was accelerated to 115 percent design speed and maintained at that speed for 3 min. without failure. The planned rig/engine running was not to exceed 105 percent design speed. A second spin test was conducted for measuring airfoil tip untwist characteristics. High speed photography was used to record tip rotation of composite inlay and standard blades at various rotor speeds. Fig. 12 is a comparison of the untwist characteristics for the two-blade types with predicted values.

An instrumented first stage fan stage with B/SiC-Ti inlay blades was installed in a TF41 Fan-IPC rig. Eight of the 25 inlay blades were instrumented with a total of 15 dynamic strain gages. Nine additional dynamic strain gages were installed on the other stages of the rig. To determine if the original flutter boundary of the all-titanium blades had been eliminated, a compressor performance map was generated by slowly accelerating from 90 to 105 percent corrected speed at successively higher bypass pressure ratio loadings. The high speed flutter phenomenon was encountered at a corrected speed of approximately 102 percent and a relative bypass pressure ratio of 1.17. These values are in contrast to a 90 percent corrected speed and a relative bypass pressure ratio of 0.92 at which high speed flutter was experienced during rig testing of the all-titanium unclappered fan blades at comparable inlet temperatures and pressures. The frequency at which flutter occurred for both blade sets was characteristic of the first torsional mode as identified in the bench test programs. This indicates little or no modal coupling existed, i.e., pure torsional flutter. This significant expansion of the flutter-free operating regime of the compressor is best illustrated by the comparison in Fig. 13. Note that the inlay blades moved the unstable boundary from within the operating envelope to a region well above the estimated operating line of the engine. Thus, on the basis of the rig test results, the B/SiC-Ti composite material insert showed sufficiently improved stability over the TF41 unshrouded first stage fan rotor airfoils to be acceptable for operating in the engine. This improvement was later verified by actual engine testing. A total of six hours of rig running time was accumulated. Analysis has isolated the stability improvement demonstrated by the B/SiC-Ti inlay blade over that of the baseline configuration as primarily the result of the 5 to 6 percent higher torsional frequency of the inlay blade. Therefore, it is estimated that placing the composite over the airfoil full-chord would improve high speed flutter resistance 12 to 14 percent. The rig aerodynamic performance characteristics show a 2.5 percent airflow increase and a 2.4 percent higher efficiency at takeoff thrust conditions using inlay first fan rotor airfoils as compared to shrouded airfoil blades. These improvements are attributed to elimination of the shroud aerodynamic losses and improved matching of the fan stages at high rotor speeds as a consequence of increased first rotor airfoil untwist.

As stated previously, the results of the rig test indicated that the B/SiC-Ti inlay improved the stability of TF41 unshrouded first fan rotor airfoils to a level outside normal engine operating requirements. To demonstrate the engine operating characteristics using the inlay blades for adverse operating conditions, both increased inlet temperature and pressure running was conducted on a TF41 engine. A test was conducted at 78 F (298.7 K), 28.25 in.-Hg (95.4 kPa) abs, and 103.9 percent of design corrected fan rotor speed. Additional tests were conducted at ram pressures of 43 in. (145 kPa) to 49 in.-Hg (165 kPa) abs with inlet temperatures of 84 F (302 K), 146 F (336 K), 200 F (366.5 K). No development of high speed flutter was indicated at any time during a total accumulation of four engine test hours. The maximum speed pressure point run for the engine test is plotted in Fig. 13. Demonstrated engine performance improvements due to replacement of the shrouded first fan rotor airfoils with inlay airfoils verified the rig test results. At rated thrust reductions in specific fuel consumption and turbine inlet temperature of approximately 1.1 percent and 12 F (6.7 K), respectively, were demonstrated.

Conclusions: The fabrication of viable B/SiC-Ti composite inlay fan blades has been successfully demonstrated. This offers the designer a tool for controling flutter without degradation in aerodynamic performance, as compared to the more conventional approach which employs part span shrouds and/or increased airfoil thickness.

Nomenclature:

- b = blade chord length/2
- C = blade chord length
- R = radius to engine centerline
- rpm = revolutions per minute
- T/C = blade thickness to chord ratio
- V_{REL} = air velocity relative to blade
- v/o = volume fraction, i.e., the ratio, based on volume, that the fiber portion of a composite material is to the whole
- T = torsional frequency

Abbreviations:

- B/SiC-Ti = metal matrix composite material consisting of silicon carbide coated boron fibers diffusion bonded in a titanium matrix
- HIP = hot isostatic press (diffusion bond process)
- RTV 560 = room temperature vulcanizing silicon rubber. A product of General Electric Co.

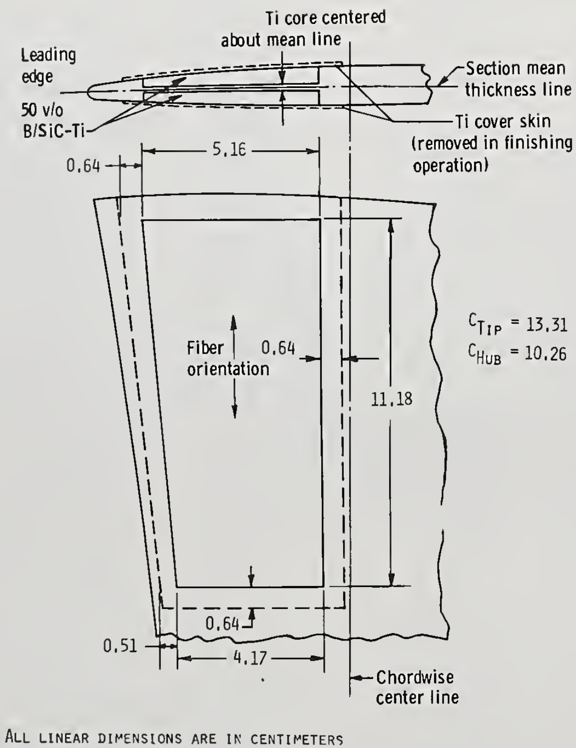


Fig. 1 Detail Schematic of Composite Inlay Airfoil

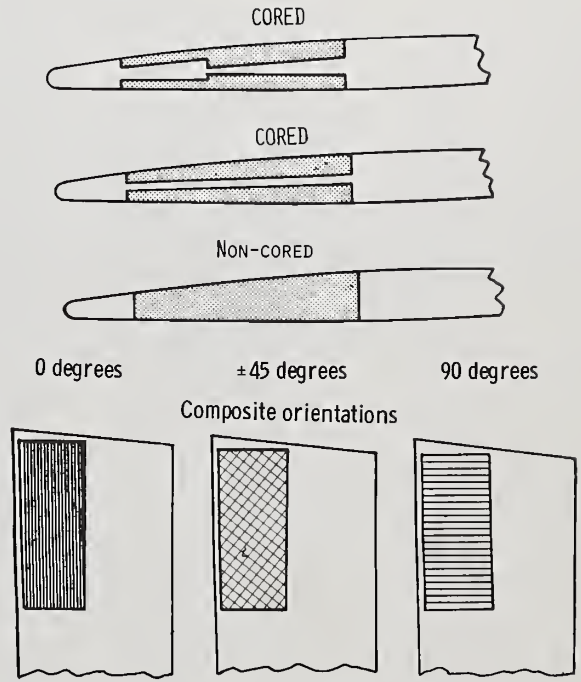


Fig. 2 Composite Inlay Configurations Analyzed

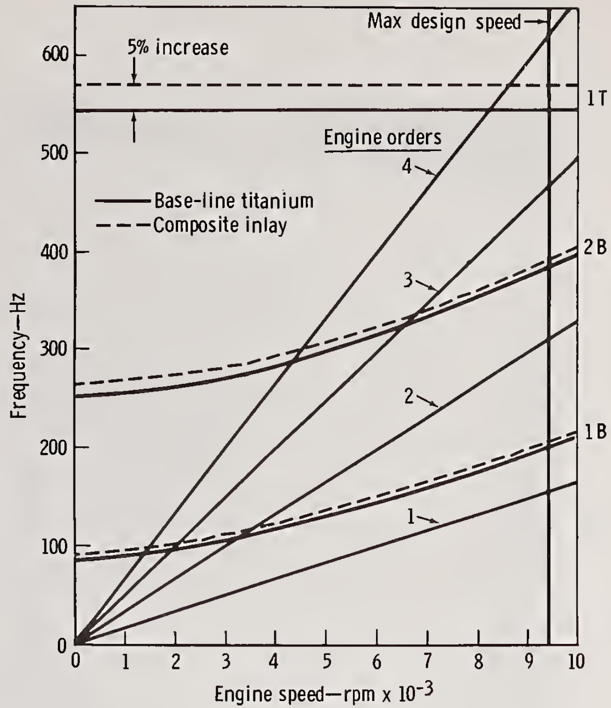


Fig. 3 Comparison of Calculated Frequency Characteristics

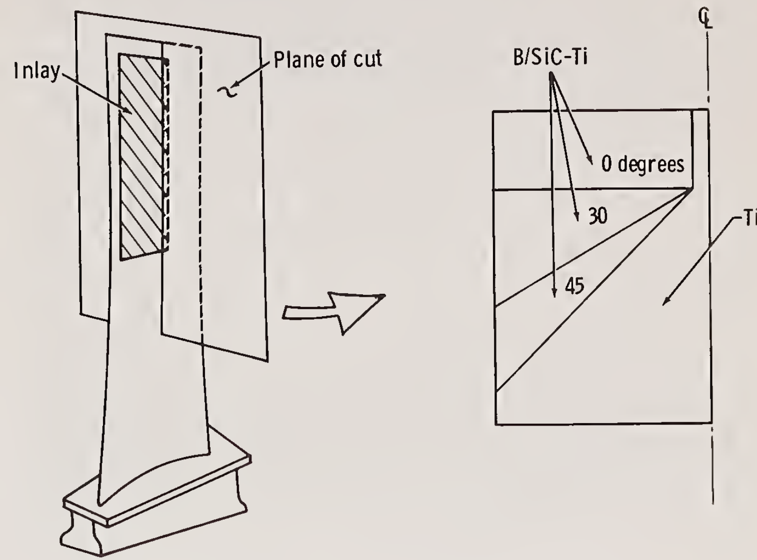


Fig. 4 Model Used to Evaluate Scarf Angle Stress Concentration

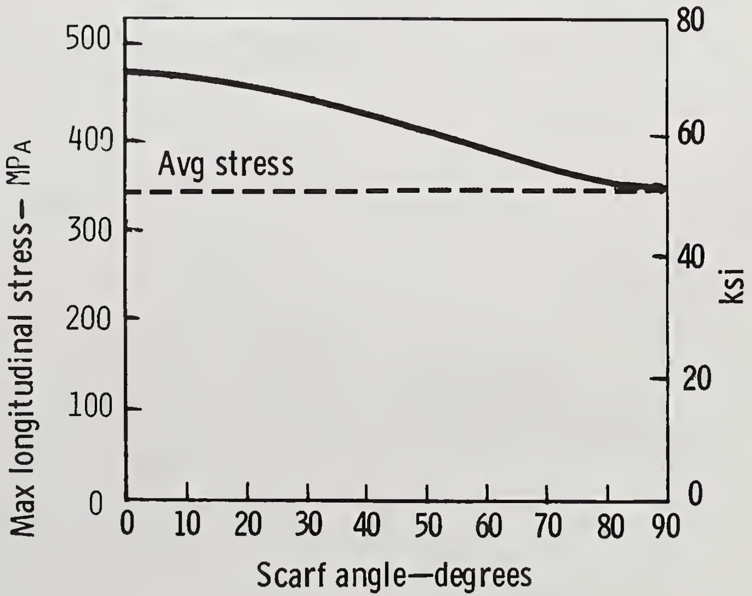


Fig. 5 Influence of Scarf Angle on Composite-Titanium Interface Stress

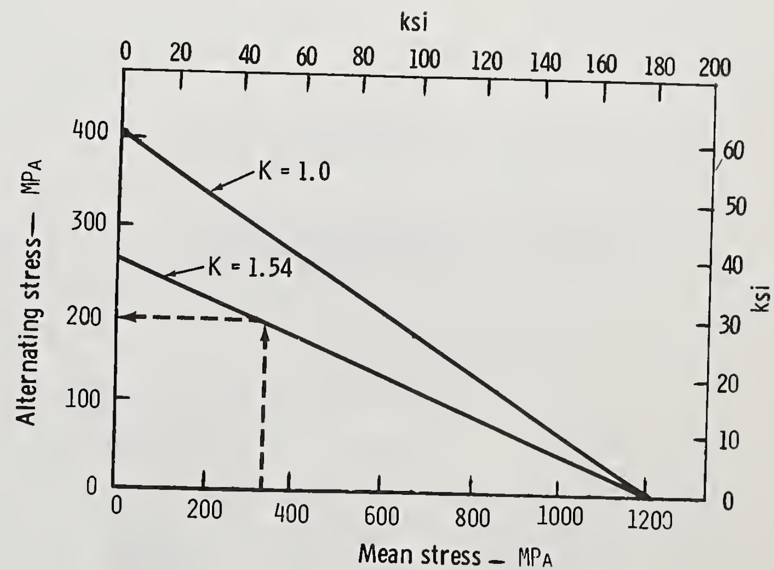


Fig. 6 Goodman diagram for 50 v/o B/Sic-Ti

Fig. 7 EDM Airfoil Recess

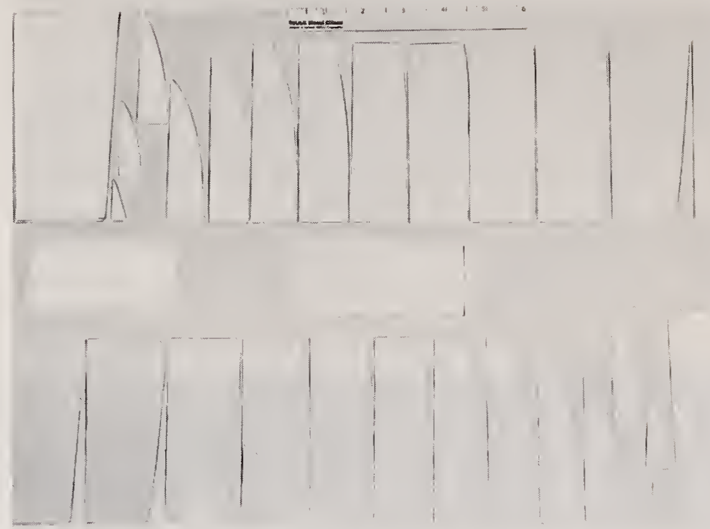


Fig. 8 One Set of Finish Cut Plies

Fig. 9 Typical Blade after HIP Bonding and Vapor Blasting

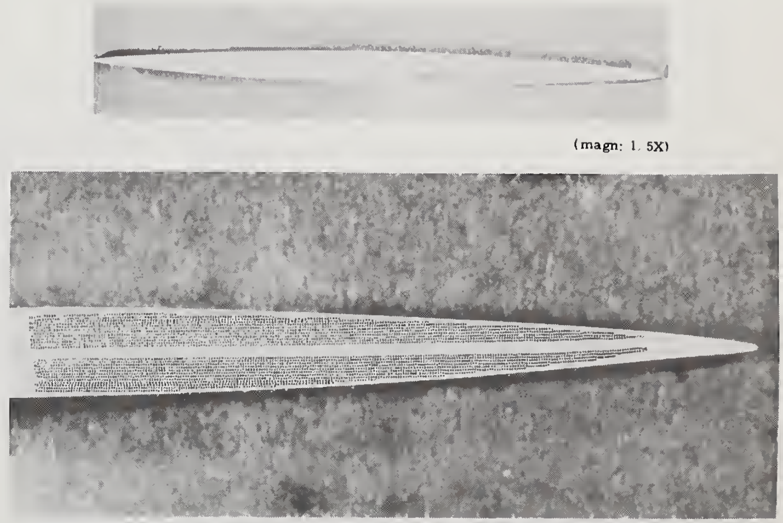


Fig. 10 Cross-Section Through Base of Composite Inlay Blade (S/N 204)

Fig. 11 Typical Blade after Final Blend

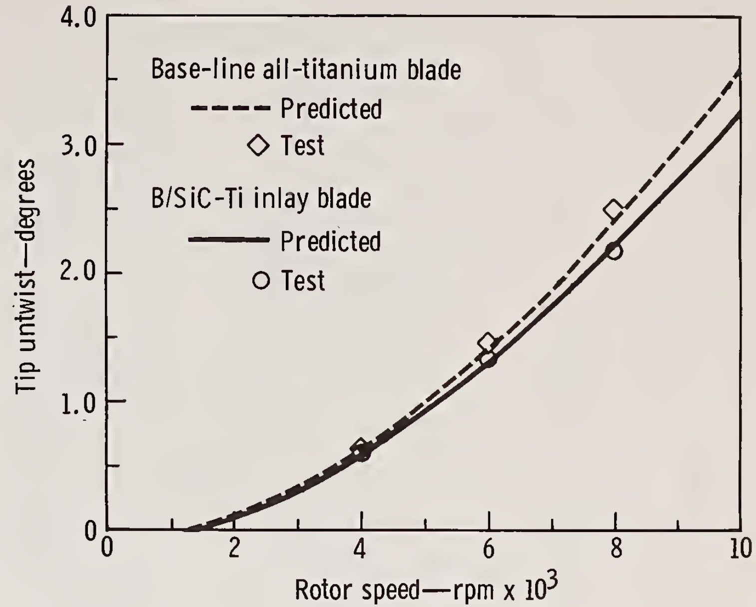


Fig. 12 Comparison of Airfoil Tip Untwist Characteristics

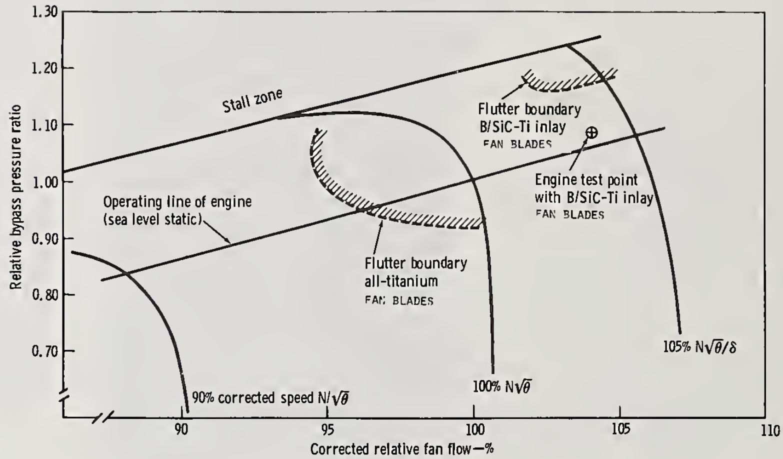


Fig. 13 TF41 Fan-IPC Compressor Rig Performance Map

STRESS CONCENTRATIONS IN COMPOSITE MATERIALS

G. Kardos
Carleton University
Ottawa, Canada
K1S 5B6

Abstract: The use of high strength laminated composites has added a new dimension to the structural designers arsenal. The designer may now tailor the properties of the material to match the load requirements of his structure. But the use of laminated composites is a mixed blessing; many design concepts have to be revised when applying these composites. The most important of these is the significance of stress and strain concentrations.

The analysis of simple laminated composite panels containing a hole demonstrate that neither the stress nor the strain concentration has the same significance as in isotropic homogeneous materials. Comparison with published test data shows that proceeding on the basis of older concepts can lead to erroneous results.

Key Words: Anisotropy; composite laminates; failure; strain concentration; stress concentration.

Introduction

The use of high strength laminated composites holds great attraction for the structural designer. The ability to tailor the material to fit the load promises lighter, stronger structures. Analytical techniques are available whereby the designer can predict, with some assurance, the performance of the composite he selects. Unfortunately, this new freedom has made the situation more complex. The designer can no longer simply analyse his stresses and then simply choose a material that matches his requirements. Now he must concern himself with the internal structure of the composite before he can undertake his analysis.

In spite of all the advances made in analysis and theory, design must start conceptually within the designer. The nature of this conceptual design is dependent on his experience and his qualitative understanding and prediction of the performance of his design. If his qualitative understanding or experience falls short of reality, the best that can be expected is that later detail analysis will reveal the shortfalls. They can then either be corrected or the designer can move back to the beginning armed with the new information. But much more costly is the case where misinformation results in the critical analysis never being made, in this case the shortcomings may not show up until the design is tested.

To take full advantage of laminated composites the designer should select his composite so that the properties are anisotropic. When this happens many of the responses we take for granted no longer hold true. Without the proper precautions simple tensile loads can cause bending, failure may first take place in unexpected locations and directions; and the maximum stress at a discontinuity may not be responsible for failure.

One of the principal tenents of designing with isotropic homogeneous materials is that where there are geometric discontinuities in a loaded structure the stress concentration will be the limiting factor in the design. It follows that if the maximum stress (usually evaluated as the nominal stress times the stress concentration) is less than the working stress of the material the structure is satisfactory; and if the maximum stress is greater than the working stress failure will occur at the point of maximum stress concentration.

With laminated composites the elastic constants are orthotropic, therefore the stress distribution about a discontinuity is different than with isotropic materials, the magnitude and location of the maximum stress are dependent on the elastic orthotropy. An additional complication is introduced because the strength properties are also orthotropic, which means the point at which failure takes place may not be at the point of maximum stress. Finally with laminated composites the strength of each lamina is different and failure may take place in different lamina at different locations. Because of these properties of laminated composites the designer must take great care in the assumptions he makes when using composites having discontinuities.

To acquire some appreciation of these factors a computer study was undertaken at Carleton University to explore the characteristics of a simple laminated composite panel with a hole subjected to tension or shear. The results are presented here to demonstrate what can be expected due to stress concentrations in composites.

Composite Design

The details of analysing laminated composites will not be given here, several excellent publications are available (Ref. 1 and 2) that detail such analysis. We will limit ourselves to simply outlining procedures that are necessary to design a satisfactory composite structure.

The necessary design steps are illustrated in Figure 1. The process is iterative, a composite configuration is chosen, analysed, and if found not to be optimum a new composite configuration is chosen and the process repeated.

A laminate composite design starts with the choice of the basic constituents, the reinforcing fibre and a matrix binder. In our analysis we used boron fibres and an epoxy matrix. From the known properties of the fibre and the matrix the properties of the individual lamina is determined. The properties of the lamina will be dependent on the volume fraction of fibre and can be computed using micromechanics theory. The elastic and strength properties of the lamina are orthotropic and are defined by a 3 x 3 symmetric matrix.

The lamina are assembled into a laminate. The elastic properties of the laminate are determined considering the properties of the individual lamina, the number of layers, the orientations of the lamina and the sequencing of the lamina. Lamination theory is used to establish the constituative behavior of the finished laminate. The result is based on the assumption that the strain will be continuous through the thickness of the laminate. The resultant orthotropic laminate properties will be given by a 6 x 6 matrix. These relate the applied stresses and couples to the deflections. The preferred practice is to make the composite symmetric about the mid-plane which decouples the bending and direct stress deflections, this is called a "specially orthotropic" laminate.

Once the constitutive behavior of the proposed laminate is established, the structural analysis can proceed. The analysis must be in terms of the orthotropy of the material. A number of closed form solutions for structural elements are available (Ref. 3), and if closed solutions are not available, finite element techniques can be used to give specific solutions. Closed solutions are available for a panel with a hole. The analysis is to give the stress and strain distribution and in some cases the bending couples and deflections on any element in the structure.

The local strain must then be resolved into the direction of the principal directions of each lamina to find the local stresses and strains in each lamina. The principal stresses and strains are then checked against the orthotropic strength properties of the specific lamina.

A number of theories of failure for composites have been proposed. In the analysis discussed here the maximum strain theory has proved to be as satisfactory as any other. For our purposes we assumed that the structure would be considered to be overloaded if any single lamina fails. In practice this would be considered a conservative design procedure, the redundant nature of composite laminates permits considerable overloads to be carried even after a single lamina has failed.

If the failure analysis shows that the laminate selection is unsatisfactory, either because a lamina has failed or because all lamina are under-stressed, the designer must return to the beginning. On the basis of his experience with the analysis he selects a new laminate construction and proceeds through the analysis again.

Analysis Results

Greszczuk (Ref. 4) investigated the stress concentration in a single ply composite, Fig. 2. He shows the tangential stress around a circular hole for different orientations of the principal material direction with respect to the loading direction. In this case the value and location of the maximum stress shifts with orientation but the failure load and location remain very close to the same value. Stress concentration for a two lamina cross-ply composite, is given in Fig. 3, but failure is not indicated.

In the Carleton studies (Ref. 6) four-ply symmetric laminates were considered with loading in the principal material direction. The individual lamina directions α were symmetric with respect to the principal material direction, i.e. $[+\alpha/-\alpha/-\alpha/+\alpha]$.

Figure 4 shows the tangential stress at selected points on the hole periphery with changes in lay angle α. The figure also shows the absolute value of the maximum tangential stress with respect to lay angle. This indicates that in this case the stress concentration would be a minimum with a lay of $\alpha = 65^{o}$.

Figure 5 shows the variation in tangential stress around the periphery for three different constructions $\alpha = 0^{o}$, $\alpha = 65^{o}$, and $\alpha = 90^{o}$. The point where the laminate will first fail is also indicated. It should be noted that in no case does failure take place at the location of the maximum tangential stress.

Figure 6 shows the tangential strain for selected points with changes in lay angle α. The strain is always maximum at $\theta = 90^{o}$. The smallest maximum strain is at a lay angle of 15^{o}.

Figure 7 shows the tangential strain distribution around the periphery of the hole for selected lay angles. The points of failure do not always correspond to the maximum strain locations.

Figure 8 compares the load capacity of a panel without a hole to the load capacity of the same panel with a hole, for different lay angles. The strength reduction ratio is shown. The optimum lay angle without a hole is $\alpha = 10^0$. The optimum lay angle with a hole is $\alpha = 19^0$ and the minimum strength reduction ratio is at $\alpha = 8^0$.

Figure 9 is an expanded version of the load capacity vs. lay angle data for a panel with a hole. It shows that the optimum panel design ($\alpha = 19^0$) would have a capacity of approximately 16,000 lbs/in. width. Whereas a panel with a minimum stress concentration construction ($\alpha = 65^0$) would only have a capacity of 2500 lbs/in. width.

Figure 10 shows the stress around the periphery of the hole for three different composites when the specimen is subjected to pure shear. With shear loading as with tensile loading the location of the first failure is not at the point of maximum stress. Although in the expected ideal lay (45^0) it is nearly so.

Figure 11 shows the load capacity of a shear panel with a hole with changing lay angles. It can be seen that the maximum capacity is not at the expected ideal lay angle of $\alpha = 45^0$ but at lay angle of $\alpha = 35^0$ and $\alpha = 55^0$. Although in this case the load capacity is near constant over the range $\alpha = 30^0$ to $\alpha = 60^0$.

Significance

It can be seen from these analysis that neither stress concentrations nor strain concentrations are a reliable criteria for load capacity of a laminate composite structure. Nor can they be used to identify the location of the probable point of failure. In fact it can be said in general that stress and strain concentration as normally used or understood in isotropic design are not a meaningful index in the design of anisotropic composites.

Furthermore since the objective of structural design is to optimize the load capacity of the structure, the use of stress concentration as a design criteria would lead to a highly unsatisfactory sub-optimum design (in this example $\alpha = 65^0$ instead of $\alpha = 19^0$).

In addition, the introduction of a hole in a composite panel does not simply reduce the load capacity, as it does in isotropic materials. It requires re-evaluating the internal structure of the composite itself. The introduction of the hole requires changing the direction of the lamina lay to achieve optimum strength.

Comparison with Experiments

The results presented are based upon computer analysis. But how close do these reflect what happens in a real composite? To determine if the analysis could predict the performance of a real composite an analysis was made of composite tests by Daniel and Rowlands (Ref. 5).

Tests had been carried out on boron-epoxy laminated composite panels 10 in. wide by 26 in. long, with a central hole. The laminate was a nine ply specially orthotropic laminate with lay angles:
[$90^o/0^o/+45^o/-45^o/0^o/-45^o/+45^o/0^o/90^o$].

Strain gages were implanted at $\theta = 90^o$ measuring the strain in the tangential direction. The local strain was plotted against the nominal applied stress.

An analysis was carried out on this laminate construction using the elastic constants for the laminate quoted by Daniel and Rowlands, the lamina allowable stresses and strains had to be taken from typical values quoted elsewhere. The test results and the results of the analysis are shown in Figure 12.

The analysis indicated the first failure should take place at 20,000 psi at $\theta = 90^o$ in the outer layers ($\alpha = 90^o$). Further increase in load indicates the next failure should take place at 29,000 psi at $\theta = 70^o$ in the third, fourth, sixth and seventh layers ($\alpha = \pm 40^o$). Further loading should cause final failure at 34,500 psi at $\bar{\theta} = 90^o$ in the remaining lamina ($\alpha = 0^o$). In the analysis it was assumed that there would be no change in elastic constants because of local failures.

The first failure predicted by analysis corresponds almost exactly with test results, probably because the location of the failure corresponds to the placement of the test gages. The second failure predicted by the analysis does not have its counterpart in the tests. This is probably because the failure is predicted to take place at $\theta = 70^o$ where there is no test gages. As a check of this, an analysis was made to determine when the $\alpha = 45^o$ lamina would fail at $\theta = 90^o$. This should take place at a load of 40,000 psi. This value corresponds closely to the final breaking strength on the tests. The predicted failure of the $\alpha = 0^o$ lamina at 34,500 psi corresponds quite closely to the second discontinuity in the test results.

If we consider the test results to only reflect failure at $\theta = 90^o$ location the sequence of failures as predicted by the analysis is as follows. First failure took place in the 90^o lamina at 20,000 psi, next the 0^o lamina failed at 35,000 psi, finally the 45^o lamina failed at 40,000 psi. This corresponds quite closely with discontinuities in the stress strain curve from the tests.

Thus we see that analytical methods used give reasonable predictions of failure in laminated composites. Consequently the expectations predicted for the effect of holes in composites can be considered highly reasonable.

Conclusion

The designer when first faced with the use of composites must be cautious about making any assumption based upon experience with other materials. Geometric discontinuities do cause stress concentration but their significance is not the same.

References

1. "Primer on Composite Materials: Analysis" J.E. Ashton, J.C. Halpin & P.H. Pettit; Technomic Publications, Stanford, Conn., 1969.

2. "Mechanics of Composite Materials" R.M. Jones: McGraw-Hill; 1975.

3. "Theory of Elasticity of an Anisotropic Elastic Body", S.G. Lekhnitskii; Holden-Day, San Francisco; 1963.

4. "Stress Concentration and Failure Criteria for Orthotropic and Anisotropic Plates with Circular Openings", L.B. Greszczuk; ASTM STP 497: American Society for Testing and Material; 1972.

5. "Experimental Stress Analysis of Composite Materials", I.M. Daniel and R.E. Rowlands; ASME 72-DE-6; American Society of Mechanical Engineers; 1972.

6. "Analysis and Design of Fibre Reinforced Composite Structural Elements"; S.A. Allidina; M.E. Thesis, Carleton University; 1974.

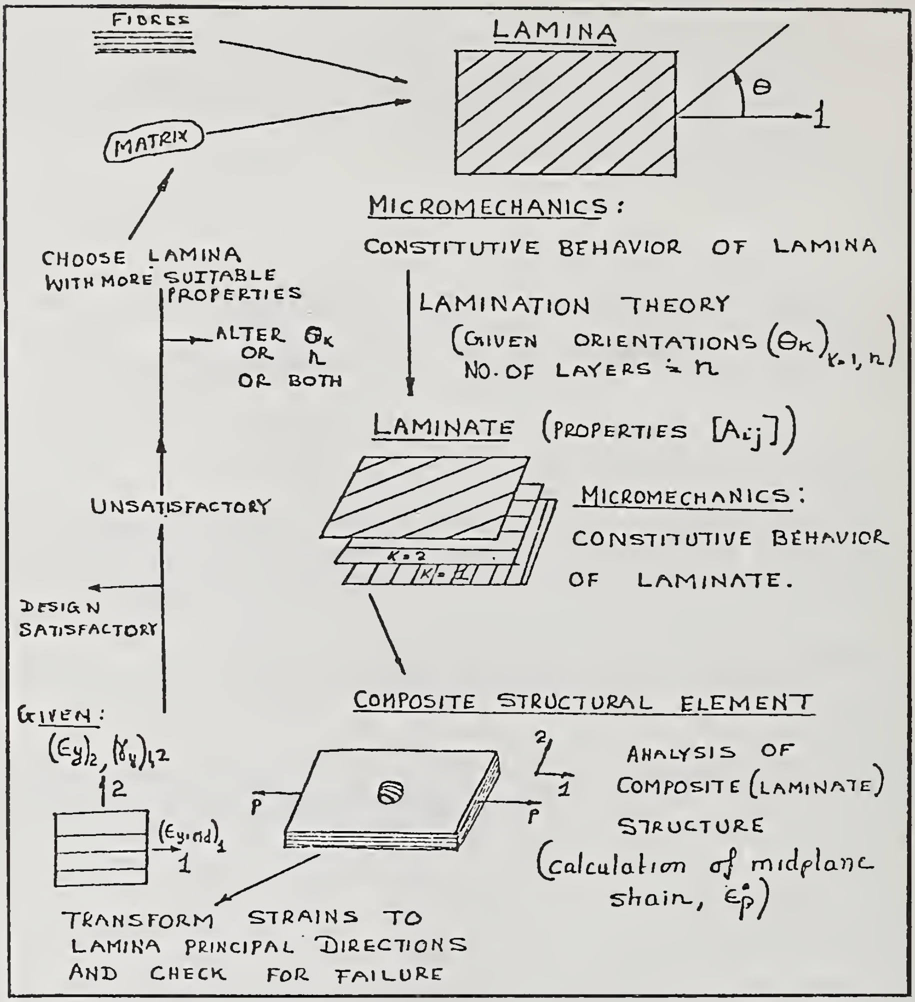


FIG. 1 DESIGN PROCEDURE FOR LAMINATED COMPOSITES

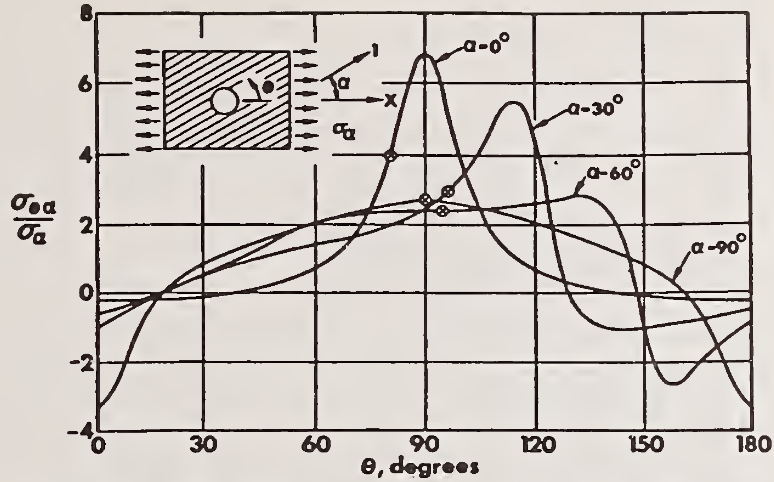


FIG. 2 Stress concentration at edge of a circular hole in an anisotropic plate subjected to stress at angle α to principal material direction.

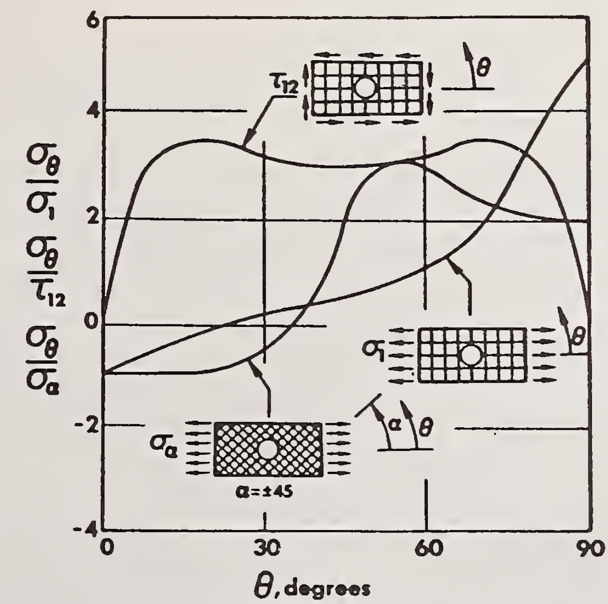


FIG. 3 Stress concentrations at the edge of a circular hole in a cross-ply laminate.

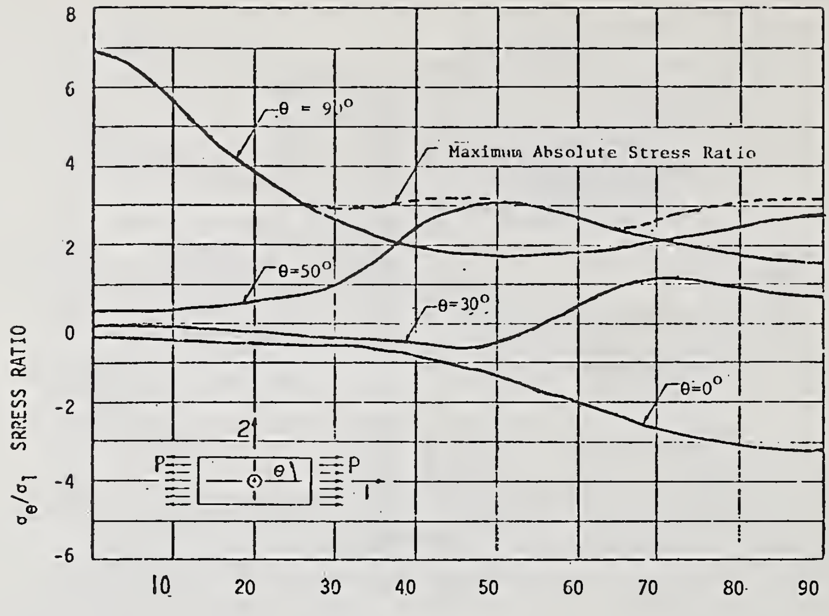


FIG. 4 MAXIMUM STRESS VS. APPLIED STRESS, BORON EPOXY
[+α/-α/-α/+α]

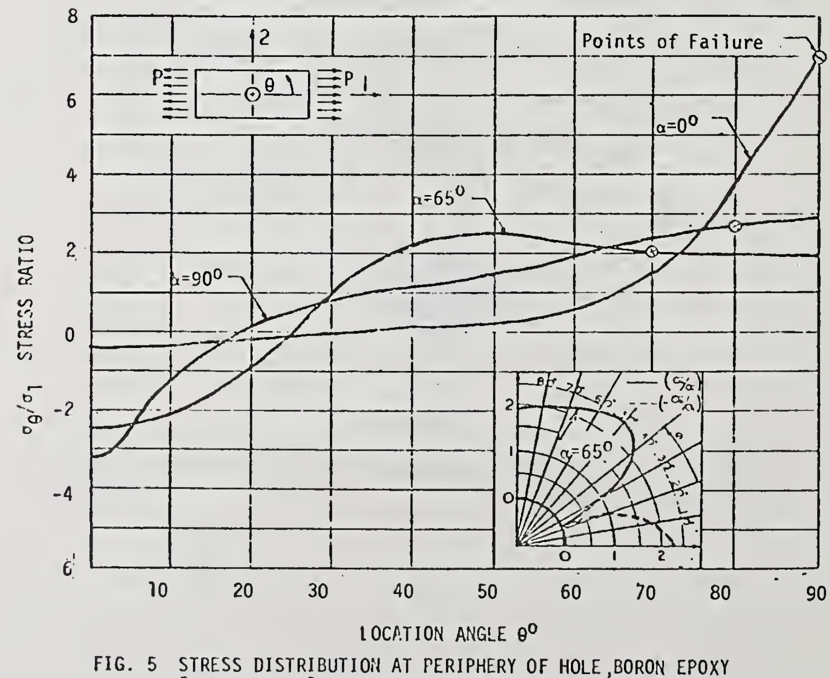


FIG. 5 STRESS DISTRIBUTION AT PERIPHERY OF HOLE, BORON EPOXY
[+α/-α/-α/+α]

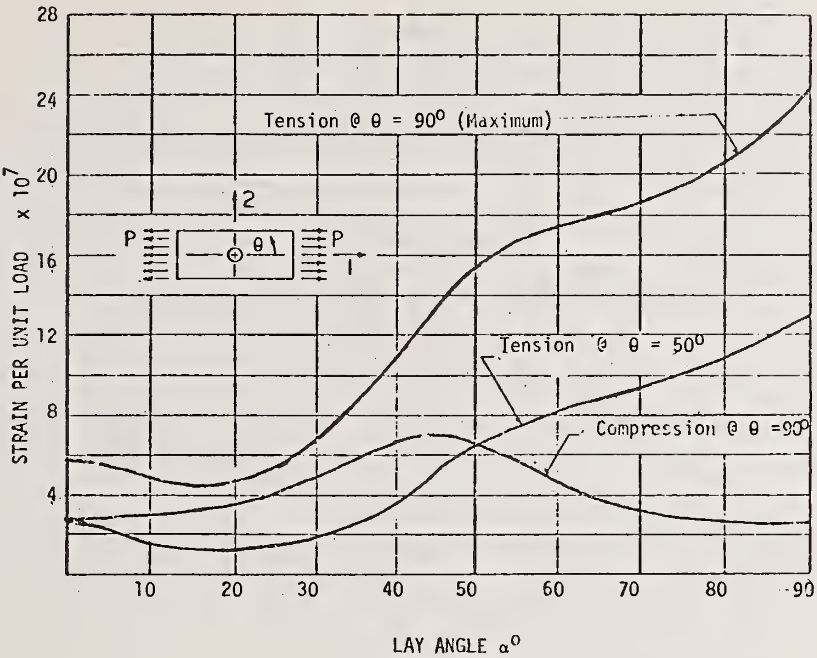


FIG. 6 TANGENTIAL STRAIN VS. LAY ANGLE, BORON EPOXY
[+α/-α/-α/+α]

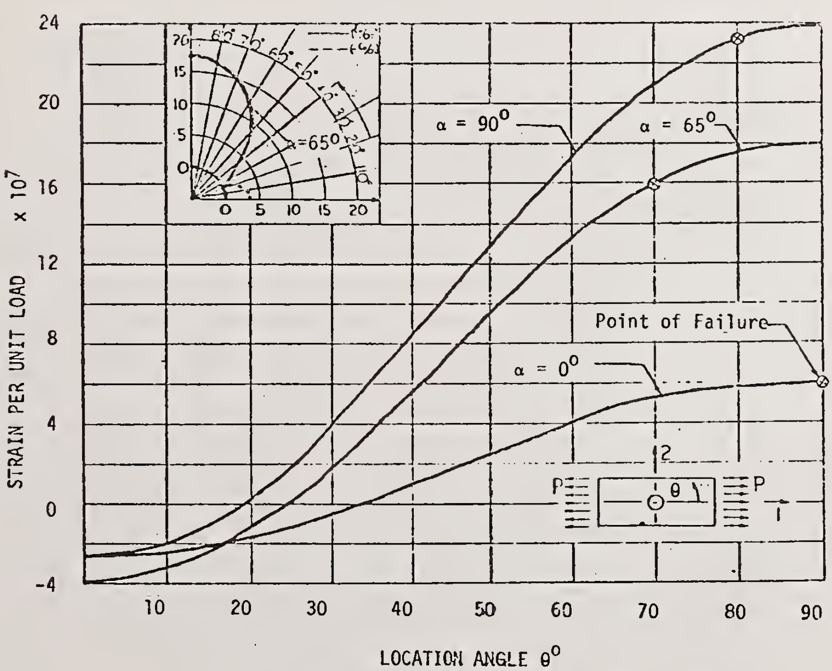


FIG. 7 STRAIN DISTRIBUTION AT PERIPHERY OF HOLE, BORON EPOXY
[+α/-α/-α/+α]

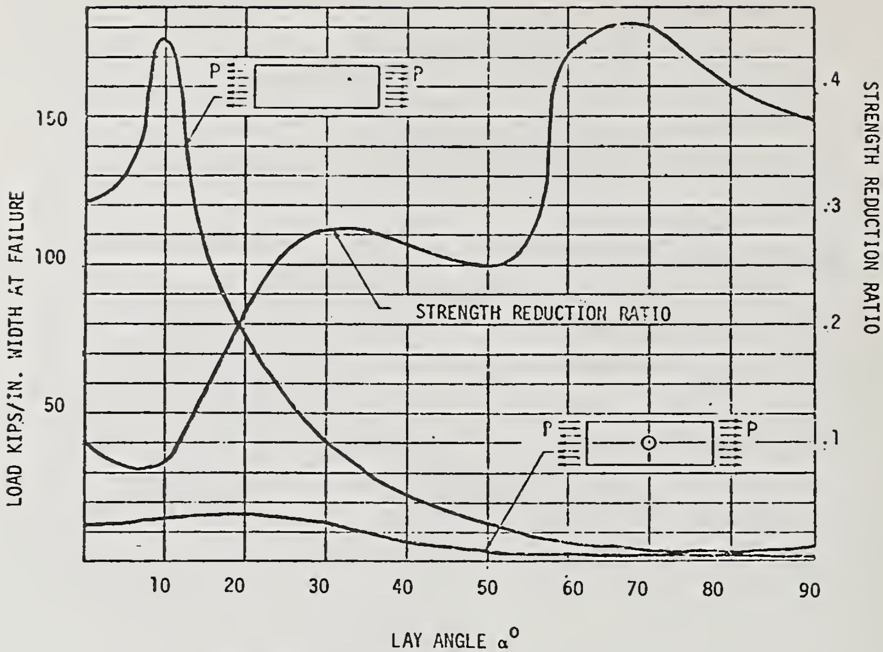


FIG. 8 LOAD CAPACITY OF PANELS IN TENSION, BORON EPOXY [+α/-α/-α/+α]

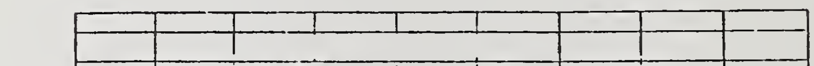


FIG. 9 LOAD CAPACITY OF COMPOSITE PANEL WITH HOLE, IN TENSION, BORON EPOXY [+α/ α/-α/+α]

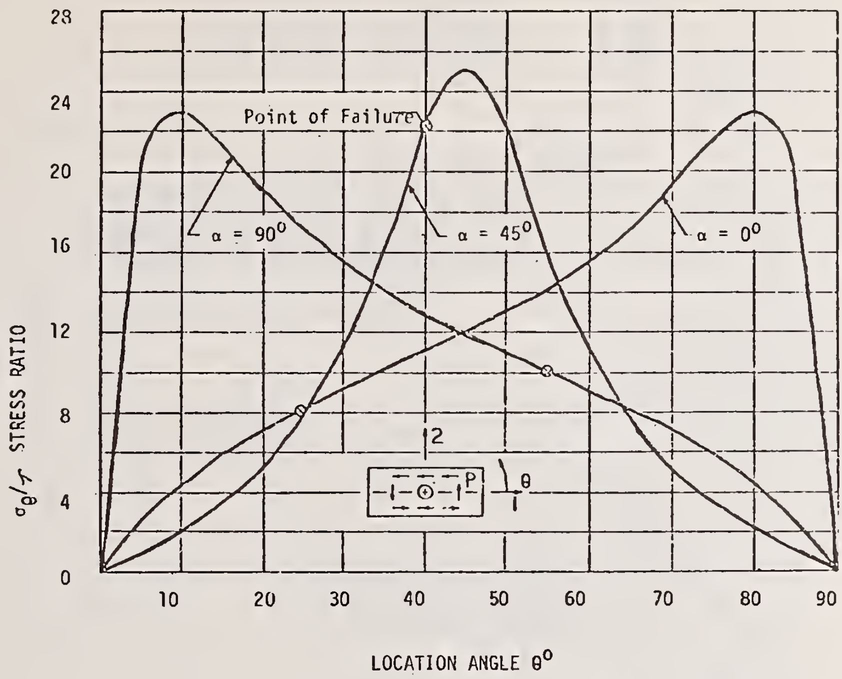


FIG. 10 STRESS DISTRIBUTION AT PERIPHERY OF HOLE IN SHEAR PANEL BORON EPOXY [+α/-α/-α/+α]

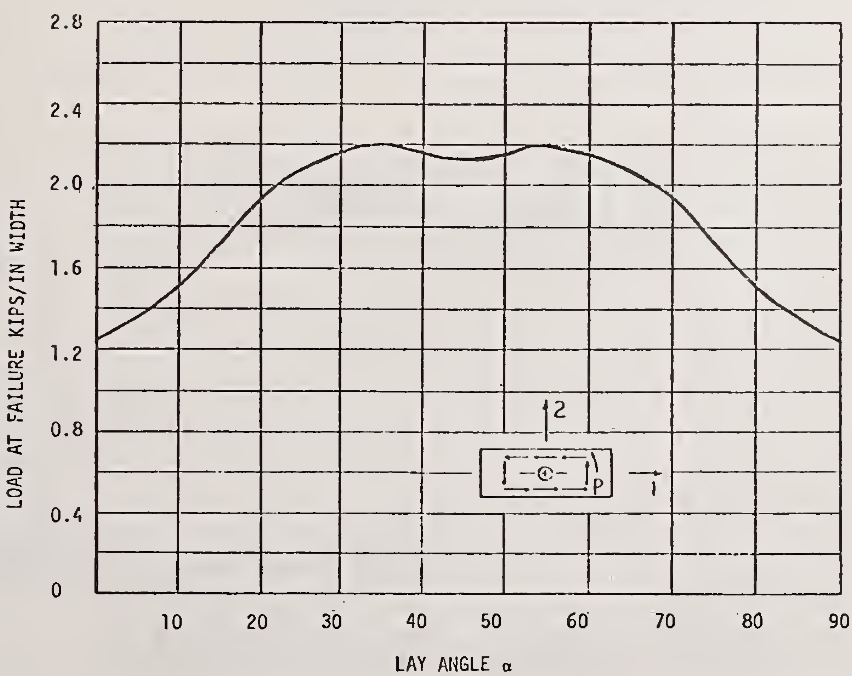


FIG. 11 LOAD CAPACITY OF PANELS IN SHEAR, BORON EPOXY [+α/-α/-α/+α]

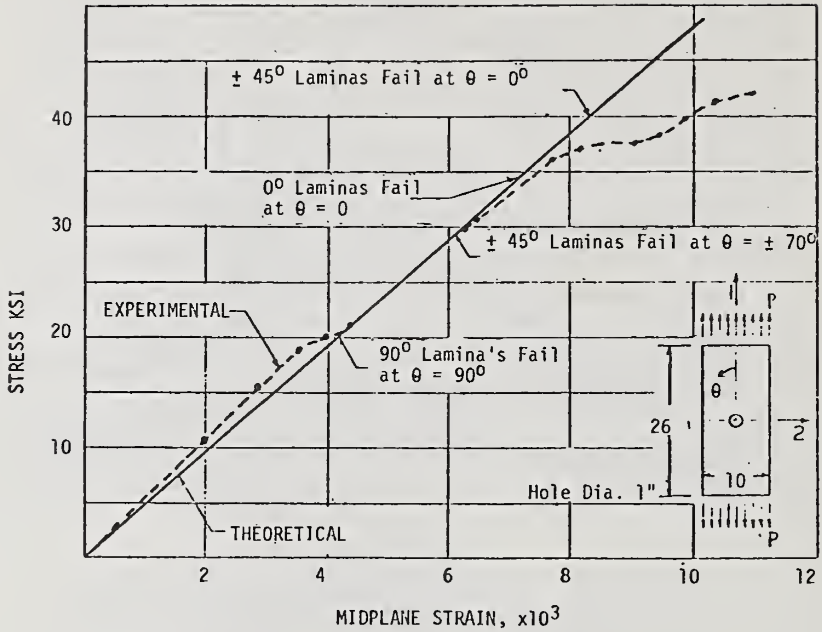


FIG. 12 Experimental and Theoretical Stress - Strain Relationships For Nine Ply Boron Epoxy Laminate
[90°/0°/± 45°/0°/±45°/0°/90°] Compared

SESSION IV

MATERIALS

OF

DESIGN

(II)

-LUBRICANTS-

Chairman: J. Scialdone
NASA, Goddard Space Flight Center

ELASTOHYDRODYNAMIC LUBRICATION

L. B. Sibley
SKF Technology Center
King of Prussia, PA 19406

Abstract: The basic mechanism of elastohydrodynamic (EHD) lubrication is described, including the historical development of the mathematical formulations and their experimental verification. The results of recent studies of the thermal, chemical and rheological conditions existing at the EHD contacts in machine elements are reviewed, as well as new developments in EHD starvation theory and experiments. Critical effects of EHD phenomena as they influence the performance of rolling bearings are illustrated, and their implications are given for the performance of other machine elements having concentrated contacts, such as gears, cams, seals, foil bearings and articular cartilage. The use of computers is described both for obtaining theoretical EHD solutions and for EHD calculations in rolling bearing design and application engineering. Typical computer computations of EHD film thickness, traction, power loss and starvation effects are presented, showing how EHD analysis can be used to design ball bearing lubricants and lubrication systems for minimum friction torque with maximum lubrication effectiveness and reliability.

Key words: Bearings; computer programs; elastohydrodynamics; failure analysis; film thickness; gears; Hertz contacts; lubrication; rheology; starvation; traction.

Elastohydrodynamic (EHD) lubrication concepts and analyses play an important role in the application engineering and lubrication failure prediction of mechanical components. Basically, EHD provides insight into the fundamental lubrication phenomena operating in a wide variety of mechanisms containing concentrated contacts, some of which are listed as follows:

- -the Hertz contacts between the balls or rollers and the ring raceways in rolling bearings,
- -the contacts between gear teeth,
- -cam and follower contact surfaces,
- -foil bearing surfaces, such as in tape heads,

-lip-type elastomeric seal contacts,
-animal joints of articular cartilage surfaces.

In all these applications, the critical load-carrying contacts consist of "bearing" surfaces which deform elastically under the action of the "lubricant" pressures generated hydrodynamically by the relative motion of these deformed surfaces. EHD lubrication prevails when the surface deformation becomes large compared to the lubricant film thickness in the contact. Theoretical solutions of the EHD problem date back only 10 to 20 years(1)* and consist of the simultaneous solution of both the hydrodynamic and elasticity equations with appropriate material property variations. A typical example of the hydrodynamic Reynold's equation for two-dimensional steady-state lubrication with a Newtonian fluid, including the compressibility effect, is

$$\frac{dp}{dx} = 6\eta(u_1 + u_2)\frac{h - h_o \rho_o/\rho}{h^3}$$

where p = lubricant film pressure
u_1, u_2 = surface velocities
h = film thickness
h_o = film thickness where dp/dx = 0
η = absolute viscosity of the lubricant
ρ = density of the lubricant
ρ_o = ambient value of lubricant density
x = distance along the surface in the direction of motion

The boundary conditions are

$p = 0$ at $x = -\infty$

$p = 0$ and $h = h_o$ at $x = x_o$

and the two-dimensional elasticity equation is

$$h = h_o + \frac{x^2 - x_o^2}{2R} - \frac{4}{\pi E'}\int_{\xi_1}^{\xi_2} \ln\left|\frac{\xi - x}{\xi - x_1}\right| p(\xi)\, d\xi$$

where R = the effective radius of curvature of the contacting surfaces in the direction of motion = $1/(1/R_1+1/R_2)$

*Numbers in parentheses refer to List of References at the end of this paper.

E' = the reduced Young's modulus of the contacting surfaces = $2/[(1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2]$

ζ = a dummy variable

x_1 = a reference point where the elastic normal deformation of the surfaces is zero.

For applications where metal contacting surfaces are used, such as in rolling bearings and gears, the EHD lubricant film pressures in the contacts are so high that the following lubricant property variations with pressure must be included in the analysis:

$$\eta = \eta_o e^{\alpha p}; \qquad \rho = \rho_o e^{p/p_s}$$

where α = the pressure-viscosity coefficient

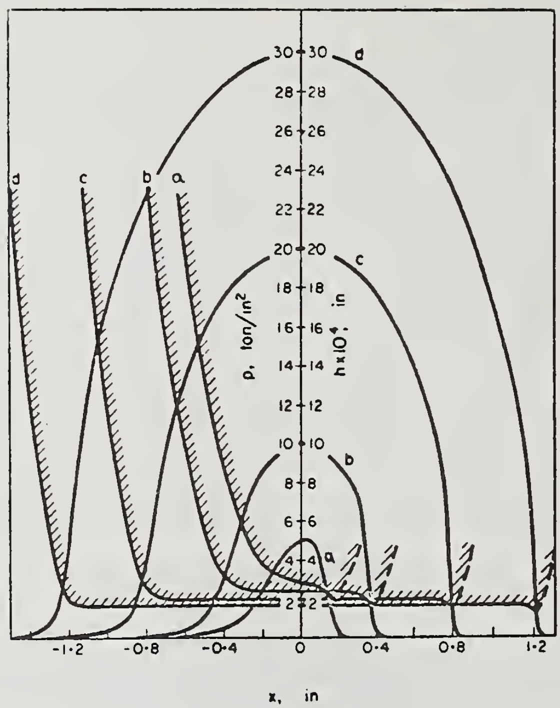
η_o = absolute viscosity at atmospheric pressure

p_s = the bulk modulus of compressibility.

Iterative solutions of the above equations using high-speed digital computers are best represented by the diagrams given in Figures 1 and 2 taken from the work by Dowson and Higginson (1,2).

Figure 1 shows how with increasing contact load the contact area increases rapidly, but the lubricant film thickness decreases very slowly, which is a basic characteristic of EHD lubrication quite different from the ordinary hydrodynamic lubrication of plain sliding bearings. This means that once EHD lubrication is established at the load-carrying contacts in a mechanical device, very little effect on the lubrication condition is caused by rapid large changes in load such as from vibration or shock loading.

In Fig.1 under high load the EHD pressure distribution in the lubricant film is almost identical to the semi-elliptical Hertz elastic surface pressure distribution, so that using Hertz formulas for computing lubricant contact film pressures does not introduce significant errors under many practical EHD lubrication conditions. Figure 2 shows the variation in film shape and pressure over a wide range of lubricant viscosities and surface velocities at a constant load. Here the film thickness varies rapidly with the viscosity-velocity product. The deformation and pressures are very similar to the Hertzian except at the highest levels of viscosity-velocity product where there is so



a, b, c, d = curves of increasing load on cylinders

x = distance from center of contact in rolling direction

p = oil film pressure

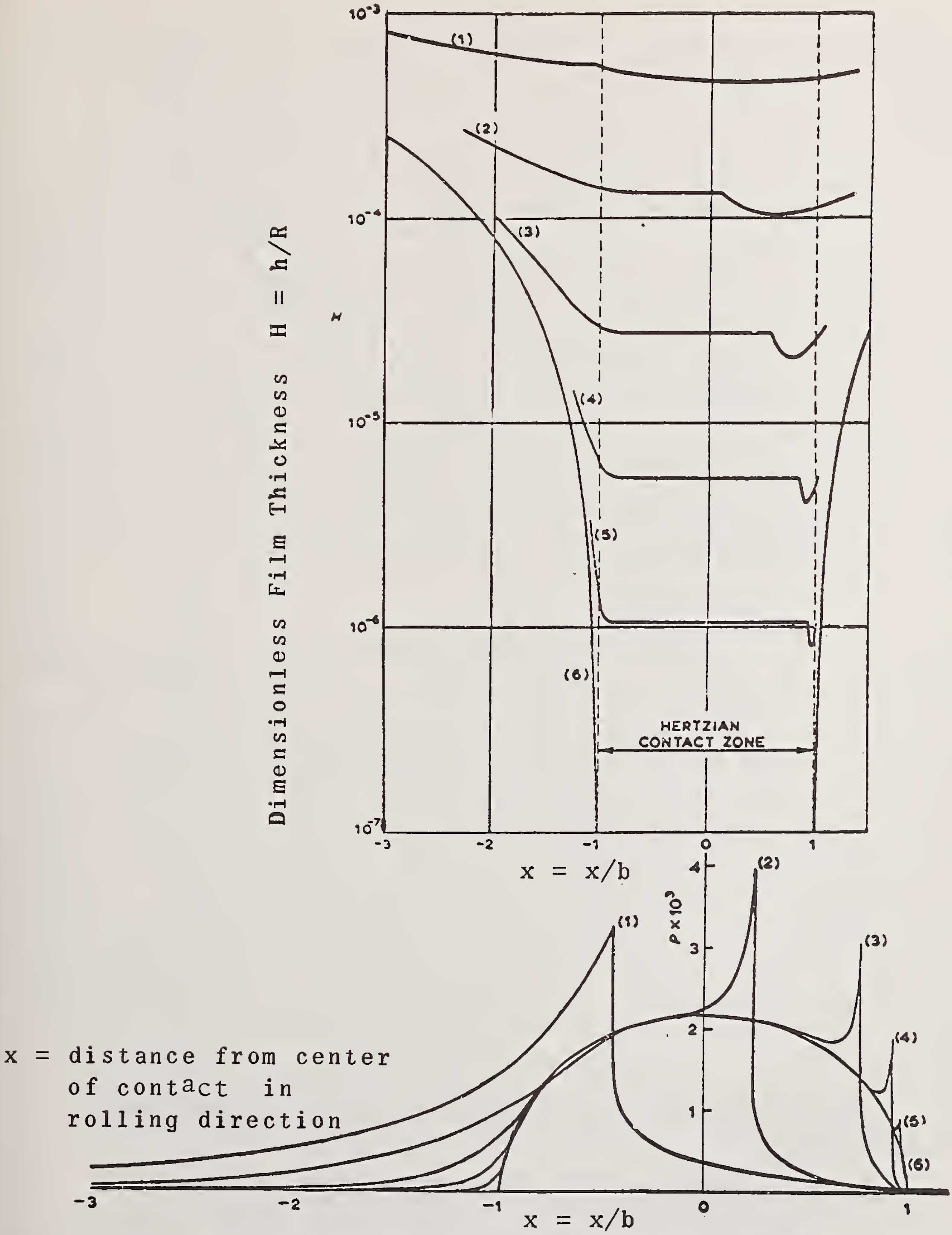
h = oil film thickness

From Dowson and Higginson(2)

Fig. 1 Pressure distributions and film shapes with pressure-dependent viscosity and elastic cylinders for a wide range of loads.

little surface deformation that the lubrication mode there is more hydrodynamic than elastohydrodynamic.

It is important that any new theoretical development be supported by experimental evidence, and a variety of experimental EHD techniques have been developed (3). Probably the most striking experimental verification of the existence



U = (1) 10^{-9}, (2) 10^{-10}, (3) 10^{-11}, (4) 10^{-12}, (5) 10^{-13}, (6) 0 (dry contact).
$W = 3 \times 10^{-5}$, $G = 5000$.

Fig. 2. Pressure distributions and film shapes for a compressible lubricant as a function of the product of viscosity times velocity.

From Dowson and Higginson (1)

of EHD films at machine element contacts is the results of radiotracer wear tests (4) shown in Figure 3. In this test, the wear rate of two steel balls in lubricated rolling/spinning contact with each other in a rolling four-ball tester was measured by making one of the balls radioactive in a nuclear reactor and measuring the radioactive wear debris in the lubricating oil using high-sensitivity radiotracer techniques. It can be seen that the wear rate was indeed quite small, only a small fraction of a microgram per 1000 ball revolutions, requiring the highly sensitive radiotracer method for any practical test program. This small wear rate fell even lower as the test speed was increased, all other conditions remaining constant, dropping off at high speeds to what seems to be an asymptotically negligible wear rate. This behavior is consistent with a hydrodynamic lubrication mechanism in which the film thickness increases with speed, and it is not consistent with a boundary lubrication mechanism in which the chemical reaction of the lubricant with the steel surfaces to form lubricating films would be expected to cause greater metal removal with increases in speed.

Studies using several new experimental EHD techniques presented at the recent American Chemical Society Symposium on Lubricant Properties in Thin Lubricating Films show that the shear conditions at EHD contacts are apparently severe enough to cause structural changes in some lubricants. Figure 4 from Lauer and Peterkin (5) in this symposium shows how the infra red spectrum for a high stability polyphenylether lubricant taken directly on the EHD film in the contact between a ball and a flat surface changes significantly with an increase in load on the contact. Sanborn and Winer (6), however, could show no significant molecular changes for lubricants having molecular weights below 600. The effect of such lubricant structural changes is mostly a modification of the contact traction characteristics which results in different heat generation rates and friction losses of mechanical components. The EHD lubricant film thickness is not usually affected significantly, since such a small quantity of lubricant is entrained in the contact.

An important EHD phenomenon that does have significant practical effects on both heat generation and oil film thickness in mechanical components such as rolling bearings and gears is the recently discovered EHD lubricant film starvation (7). Figure 5 shows plane views of two 3-dimensional EHD contacts which are identical except for the

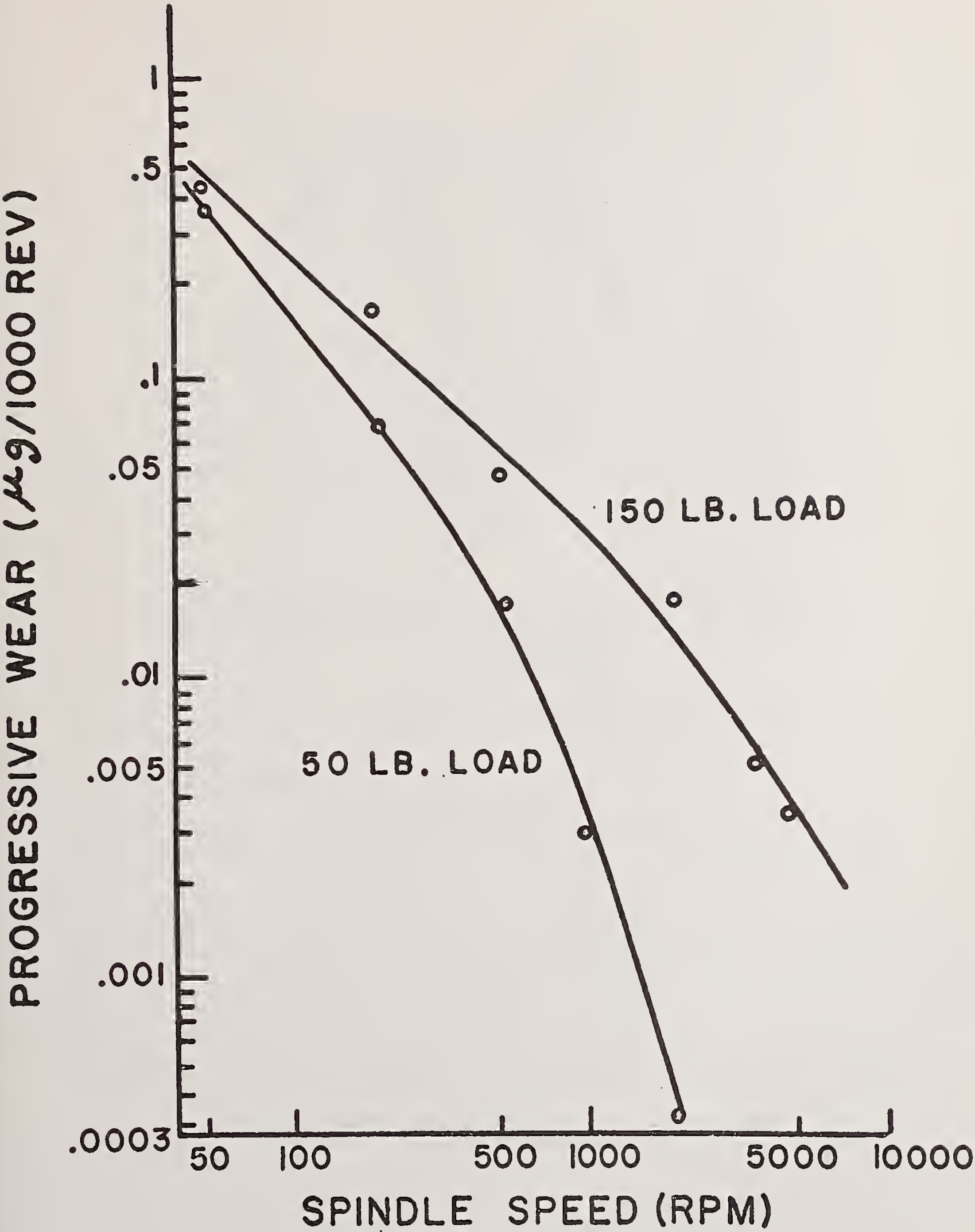
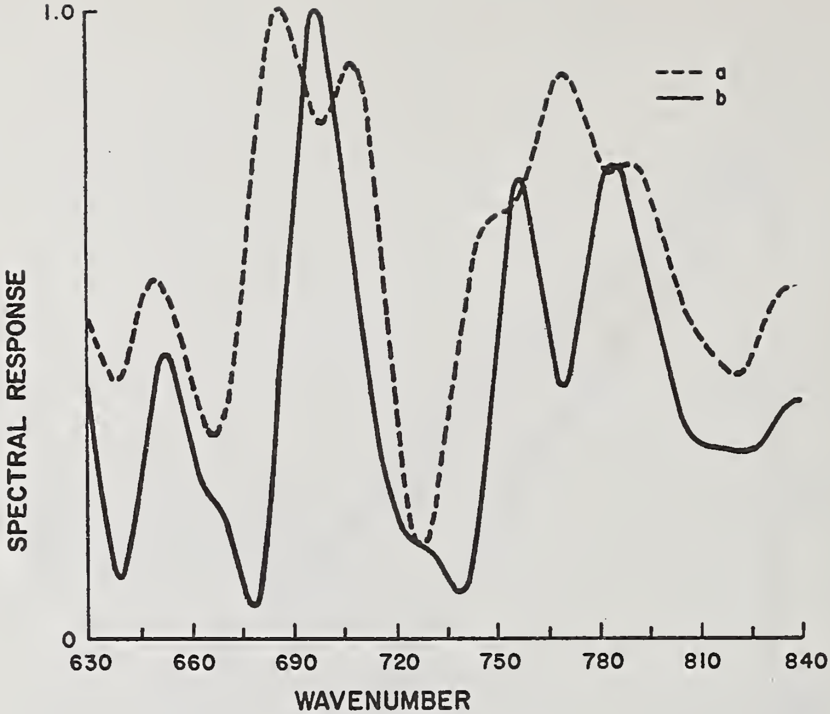


Fig. 3. Progressive wear rate as a function of rolling speed at a rolling/spinning ball-to-ball contact measured by radiotracer technique.



a. 28.5 kg load, 480 rpm
b. 11.0 kg load, 480 rpm

From Lauer and Peterkin (5)

Fig. 4. Emission spectra of polyphenylether 5P4E in an EHD contact.

degree of lubricant supply or starvation. The overall circular shape of these contacts results from the Hertzian elastic deformation of the contact surfaces, in this case between a ball and a flat. The lubricant films sticking to the rolling ball and flat surfaces sweep into the contact region from right to left on Figure 5, the direction of the surface motion with respect to the contact. These surface films merge together into a continuous oil film between the converging surfaces at the meniscus line, which separates the layer of air between the oil-wetted surfaces on the right side of the line from the continuous oil film on the left. As the amount of oil replenishment into the track between successive ball contacts decreases, the meniscus line moves closer to the Hertz contact region, as shown in the progression from Sketch (A) to Sketch (B) in Figure 5. This partial EHD starvation results in a decrease in the film thickness (from a plateau film thickness of 18.5

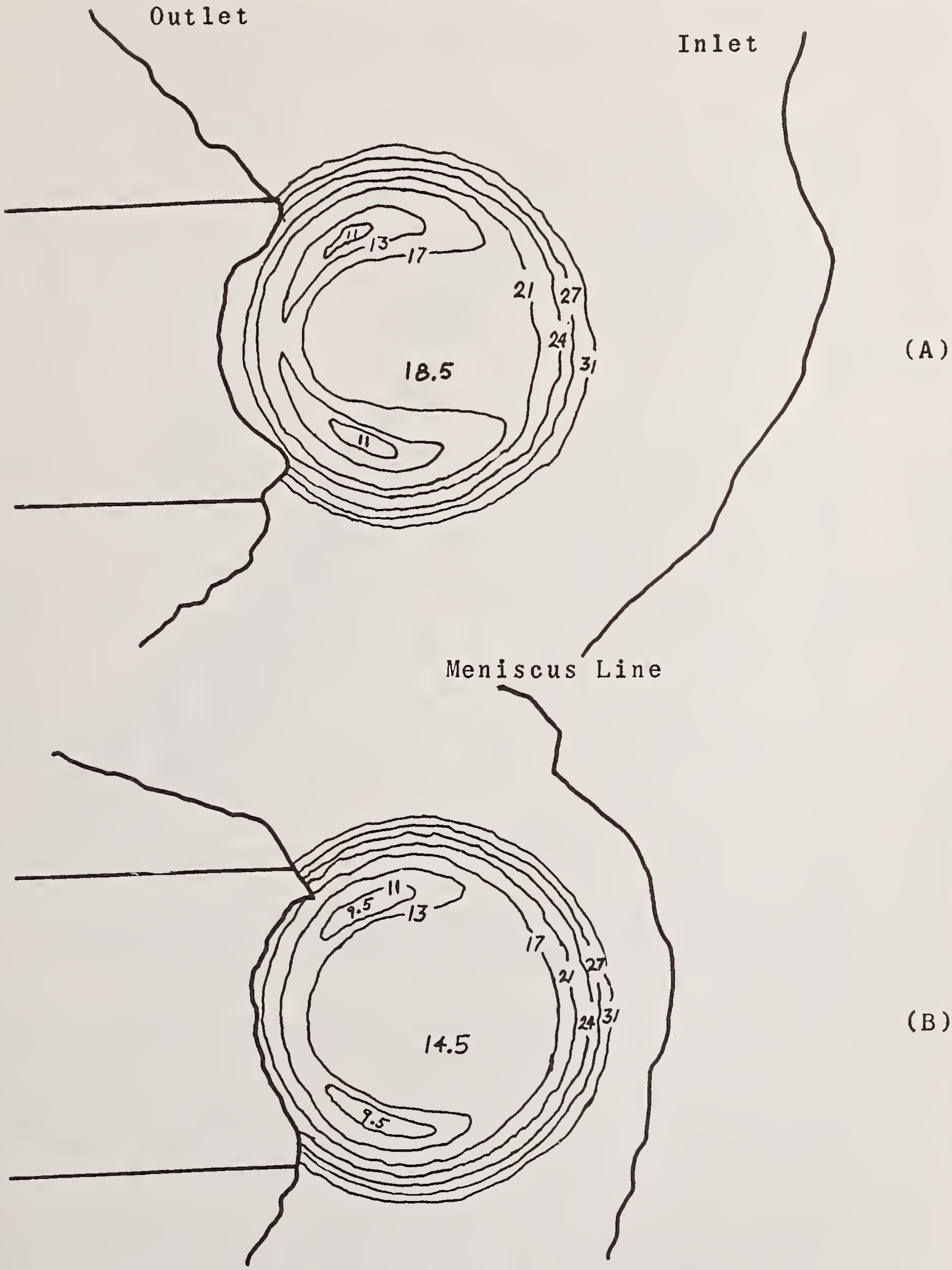


Fig. 5. Contour lines of equal lubricant film thickness in microinches at two identical ball/flat EHD contacts except for location of meniscus lines where separate oil films on converging inlet surfaces merge into a continuous film of oil between the surfaces.

microinches to 14.5 microinches in the example shown in Figure 5) and in a broadening of the plateau region of the contact making it less hydrodynamic and more Hertzian.

The reduction of EHD film thickness at critical lubricated contacts in mechanical components from whatever source, either EHD starvation or simply too little oil viscosity or velocity of the surfaces, is not necessarily harmful to component performance. A useful parameter for judging the lubrication condition at EHD contacts is the ratio of the minimum film thickness to the composite roughness of the two surfaces in contact, defined as

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$$

where σ_1 and σ_2 are the rms roughnesses of the two contacting surfaces. Figure 6 shows the percentage of time, t/t_0, that typical EHD contacts in a bearing or gear of practical surface roughness are separated by a complete lubricant film plotted against this ratio of minimum film thickness to composite surface roughness. For typical high speed bearings and gears, film thickness/roughness ratios less than about 1.5 result in the generation of surface distress in the tracks manifested by destruction of all traces of the original surface finish and formation of a glazed appearance. Continued operation under surface distress conditions leads to the generation of surface microcracking and a multiplicity of micropits that eventually result in surface initiated fatigue spalling at a much lower fatigue life than would occur with the same surfaces operating under more favorable lubrication conditions. The critical film thickness/roughness ratio needed to avoid this lubrication-related surface distress is lower for slower speed bearings and gears, down to about 0.4 for very slow, medium size, heavily loaded roller bearings. As shown in Figure 5, at very high film thickness/roughness ratios there is evidence that L_{10} fatigue life, the life at which 90% of a population of bearings or gears will survive without fatigue spalling, can be expected to increase by two or three times over the accepted "rating life" at intermediate film conditions (8).

There is every reason to believe that the above EHD concepts can be used to characterize the performance of concentrated contact mechanisms other than bearings and gears, for which their application so far has been most extensively developed. Even though the contact surface motion is essentially sliding and usually not rolling in such components as cams, elastomeric seal lips and foil bearings, there is still a

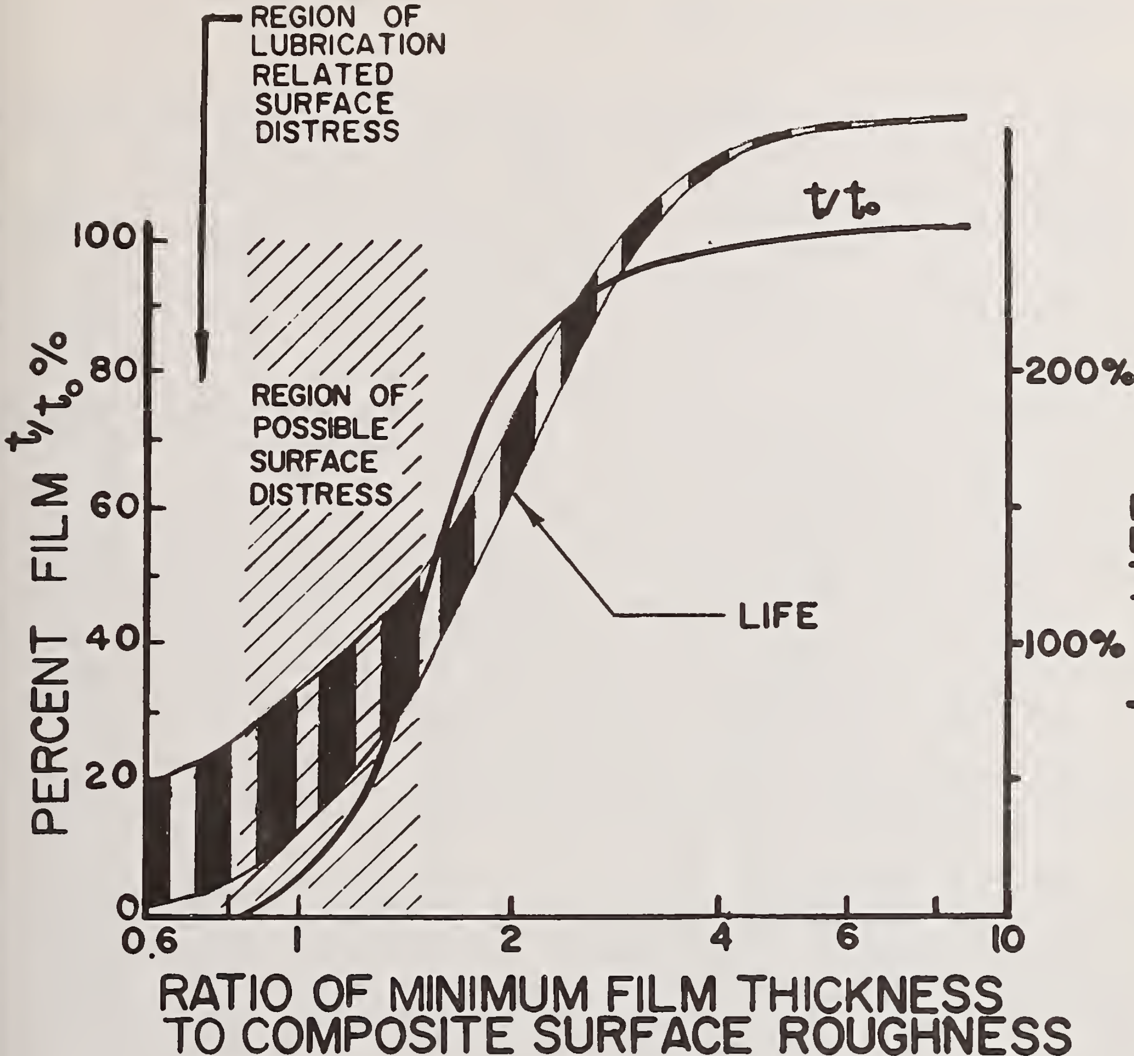


Fig. 6. EHD film condition and L_{10} fatigue life as a function of the film thickness/roughness ratio.

substantial "sweeping" motion of the surfaces carrying "lubricant" into the contacts to support EHD films. In the case of seals and foil bearings the "bearing" surfaces deform much more readily than rolling bearing or gear contact surfaces, so that EHD-type lubrication is realized without the need to account for the changes in lubricant viscosity and density under the high contact pressures found in bearings and gears. The porous, compliant, articular cartilage surfaces in animal joints, such as the human hip joint, are another example of "bearing" surfaces that deform easily under the hydrodynamic pressures generated in the synovial fluid, which acts as a highly non-Newtonian "lubricant" film.

The effective use of EHD lubrication concepts in component design entails the coordinated application of a wide variety of complex calculations that demand high-speed digital computer systems for their solution (9). A convenient application of computers for this purpose is a time-sharing conversational-mode system such as Program STARBRG illustrated in Figure 7. Program STARBRG obtains all required input data for the analysis of a ball bearing application in the answers to the six questions shown in the input data section of Figure 7. The output data section of Figure 7 shows how, in this case of a very high speed angular-contact ball bearing operating under a thrust load as in a typical turbine application, the ball contact force on the outer ring is much greater than on the inner ring, due to the high centrifugal forces on the balls, causing the inner and outer ring contact angles to be widely divergent. Under these conditions the program computes the rolling velocities, maximum film pressures and dimensions of the contact ellipses at the inner and outer ring contacts, as well as the standard rating capacity for fatigue life calculation, the spin/roll ratio as a measure of the amount of sliding in the contacts and the cage speed for vibration prediction and comparison with measured values to determine ball skid tendencies.

The EHD film thickness at each ring contact is computed by Program STARBRG for the partially starved condition resulting from lubricant replenishment into the ball and outer ring tracks only (due to the centrifugal forces on the oil films in the bearing), the replenishment being effected only by surface tension induced capillary flow from the track edges, as a conservative design base-line. The corresponding meniscus distance and starved oil-film thickness reduction factor are also computed for each ring contact based on these replenishment assumptions. If it should turn out that both the EHD film thickness and the starvation reduction factor are lower than needed to obtain the required performance in the application, the designer knows to provide some means of obtaining direct lubricant replenishment onto the bearing tracks, such as by the use of an under race oil supply system.

More complex computer programs have been developed for analysis of the effects of different amounts of lubricant replenished directly onto bearing tracks as well as for

```
                  STARBRG, SKF PROGRAM AT75Y001,  VERSION 1975 - 1 - 8
          CALCULATION OF BALL BEARING CONTACT SIZE,PRESSURE AND FILM THICKNESS
*************************************************************************

THE PROGRAM WILL RETURN TO 'WHICH QUESTION(S)' IF 999999 IS ENTERED
AS THE LAST VALUE OF QUESTIONS 4 - 6

  1.  ENTER TITLE? BALL BEARING FOR MAIN SHAFT

  2.  IS IT UNDER PURE THRUST LOAD ONLY? YES

  3.  DO YOU WANT AN EXPLANATION OF TERMINOLOGY? YES

          Z=NUMBER OF BALLS
          TR=THRUST LOAD,NEWTONS
          RPM=REVOLUTIONS PER MIN.
          D=BALL DIAMETER,MM
          FO=OUTER RACE CONFORMITY
          FI=INNER RACE CONFORMITY
          BET= NO LOAD CONTACT ANGLE,DEG.
          SRR= SPIN TO ROLL RATIO
          CAP= LOAD CAPACITY OF A CONTACT,NEWTONS
          DM=PITCH DIAMETER,MM
          CP=DYNAMIC VISCOSITY AT OPERATING TEMP.,CENTIPOISE
          ALFA=PRESSURE VISCOSITY INDEX,MM2/N
          SUT=OIL-AIR SURFACE TENSION,N/M
          CAP=LOAD CAPACITY PER RING,NEWTON
  4.  DO YOU WANT TO COMPUTE FILM THICKNESS? YES

      ENTER IN ORDER SHOWN:CP,ALFA,SUT? 2.5, 0.0145, 0.028

  5.  ENTER IN ORDER SHOWN:Z,D,DM,FO,FI,BET? 21, 20.6375, 159.4885, 0.521, 0.516, 26

  6.  ENTER TR, RPM? 9996, 20000
```
Fig. 7. Conversational - mode computer program STARBRG - Data input

BALL BEARING FOR MAIN SHAFT

I N P U T

NO. OF BALLS	21.		
BALL DIAMETER	20.64 MM	PITCH DIAMETER	159.49 MM
OUTER RING CONFORMITY	.5210	INNER RING CONFORMITY	.5160
INNER RING SPEED	20000. RPM		
THRUST LOAD	9996. N	FREE CONTACT ANGLE	26.00 DGR.
DYNAMIC VISCOSITY	2.5 C.P.		
PRESSURE VISCOSITY INDEX	.0145 MM2/N	SURFACE TENSION OF OIL	.0280 N/M

O U T P U T

	OUTER RING	INNER RING
CONTACT FORCE (N)	3742.4	670.8
CONTACT ANGLE (DGR.)	7.3	45.2
ROLLING VELOCITY (M/S)	95.7	74.7
MAX. PRESSURE (N/MM2)	1799.8	1101.8
SEMI-MAJOR AXIS (MM)	2.718	1.691
AXIS RATIO OF ELLIPSE	7.4	9.8
LOAD CAPACITY (N)	9360.8	6461.5
SPIN ROLL RATIO	0.002	0.699
CAGE SPEED	10156. RPM	
EHD FILM THICKNESS (MM)	0.00042	0.00037
MENISCUS DISTANCE (MM)	0.1223	0.1112
STARV. REDUCTION FACTOR	0.5479	0.6167

Fig. 7(cont'd). Program STARBRG - Output data calculation of EHD effects on bearing friction torque and power loss. A partial output from such a program called SHABERTH is shown in Figure 8. The computed EHD film thickness/roughness ratio, labeled H/SIGMA in Figure 8, is used to adjust the calculated L_{10} fatigue life of the five different bearings included in the bearing/shaft system analysis which comprises SHABERTH. The bearing frictional heat generation rates are computed by including the tractions at all EHD contacts in the bearings both at the ball or roller/raceways and all cage contacts. These tractions are summed up for both EHD fluid film and surface asperity frictions as well as for the viscous losses in the meniscus films around each contact and the drag forces of the ball or roller/cage assembly through the oil mist existing in the open bearing spaces.

It is clear that EHD analysis provides insight into the basic lubrication mechanisms at work in properly lubricated mechanical components that can be used to guide their design and to predict failure mode paths to be avoided when recognized early enough in the design and development cycle of new machines. Computer simulation of these EHD lubrication mechanisms has already proven to be a powerful tool for implementation of EHD analysis in advanced engineering design.

Acknowledgements

The author wishes to thank SKF Industries, Inc., for permission to publish this paper. The support of various government agencies in the development of some of the technologies and computer programs described is sincerely appreciated. The contributions of many members of the SKF staff, especially Y.P. Chiu, W. J. Crecelius, J. I. McCool and T. E. Tallian, to the technologies discussed in this paper are gratefully acknowledged.

** S H A B E R T H / A B R ** TECHNOLOGY DIVISION S K F INDUSTRIES INC. **

POWER INPUT MODULE *STEADY STATE LUBRICATED ANALYSIS* 1450 SHP WP 7/76

	FATIGUE LIFE (HOURS)			H/SIGMA		LUBE-LIFE FACTOR		MATERIAL FACTOR	
BRG.	O. RACE	I. RACE	BEARING	O. RACE	I. RACE	O. RACE	I. RACE	O. RACE	I. RACE
1	2.494+05	1.960+05	1.184+05	.888	.728	.437	.428	5.00	5.00
2	6.764+04	1.461+05	4.917+04	1.50	1.51	.536	.539	5.00	5.00
3	6.692+04	1.518+05	4.936+04	1.43	1.46	.520	.526	5.00	5.00
4	6.493+04	1.358+05	4.675+04	1.42	1.44	.518	.522	5.00	5.00
5	7.292+04	1.816+05	5.519+04	1.43	1.41	.520	.515	5.00	5.00

FRICTIONAL HEAT GENERATION RATE (WATTS) AND FRICTION TORQUE (N-MM)

BRG.	O. RACE	O. FLNGS.	I. RACE	I. FLNGS.	R.E.DRAG	R.E.-CAGE	CAGE-LAND	TOTAL	TORQUE
1	941.	0.000	380.	0.000	1.651+03	16.4	8.31	2.997+03	2.105+03
2	424.	0.000	301.	0.000	1.260+03	30.4	31.4	2.047+03	1.437+03
3	411.	0.000	294.	0.000	1.193+03	26.3	27.4	1.951+03	1.370+03
4	410.	0.000	294.	0.000	1.190+03	25.9	27.2	1.947+03	1.367+03
5	412.	0.000	277.	0.000	1.194+03	25.9	26.7	1.936+03	1.359+03

EHD FILM THICKNESS, FILM REDUCTION FACTORS AND HEAT CONDUCTIVITY DATA FOR THE OUTER AND INNER

BRG.	FILM (MICRONS)		STARVATION FACTOR		THERMAL FACTOR		MENISCUS DIST. (MM)		CONDUCT
1	.226	.185	1.000	1.000	.938	.949	8.59	4.64	21.1
2	.286	.289	1.000	1.000	.896	.889	7.07	3.89	8.42
3	.274	.279	1.000	1.000	.902	.894	7.07	3.89	8.68
4	.272	.276	1.000	1.000	.902	.895	7.07	3.89	8.74
5	.274	.270	1.000	1.000	.902	.898	7.07	3.89	8.49

Fig. 8. Selected output data from computer program SHABERTH

List of References

1. Dowson, D., and Higginson, G., "Elastohydrodynamic Lubrication, The Fundamentals of Roller and Gear Lubrication", Pergammon Press, 1966.
2. Dowson, D., and Higginson, G.R., "Numerical Solution to the Elastohydrodynamic Problem" Journal of Mechanical Engrg. Sc. Vol. 1 No. 1 p. 6 (1959).
3. Seth, B. B., and Willis, T., "Techniques for Film Thickness Measurements in Elastohydrodynamic Lubrication", ASME Paper No. 76-DET-79 (1976).
4. Tallian, T.E., Chiu, T.P., Huttenlocher, D.V., Kamenshine, J.A., Sibley, L.B., Sindlinger, N.E., "Lubricant Films in Rolling Contact of Rough Surfaces," ASLE Trans. 7, 109-126 (1964).
5. Lauer, J.L., and Peterkin, M.E.,"Analysis of Infrared Emission Spectra of Elastohydrodynamic Lubricant Films", Proceedings of Symposium on Lubricant Properties in Thin Lubricating Films, American Chemical Society, New York, pp. 48-56 (1976).
6. Sanborn, D.M., and Winer, W.O., "Lubricant Properties in Thin Lubricating Films," Proceedings of Symposium on Lubricant Properties in Thin Lubricating Films, American Chemical Society, New York, pp. 57-67 (1976).
7. Chiu, Y.P., "An Analysis and Prediction of Lubricant Film Starvation in Rolling Contact Systems," ASLE Trans. 17 pp. 22-35 (1974).
8. Liu, J.Y., Tallian, T.E., and McCool, J.I., "Dependence of Bearing Fatigue Life on Film Thickness to Surface Roughness Ratio," ASLE Trans. 18 (2), pp. 144-152 (1975).
9. Sibley, L.B., and Pirvics, J., "Computer Analysis of Rolling Bearings," Proceedings of Symposium on Computer-Aided Design of Bearings and Seals, ASME, New York, pp. 95-115 (1976).

POLYFLUOROALKYL-ALKYL POLYSILOXANE GREASE FOR INSTRUMENT LUBRICATION

John B. Christian
Air Force Materials Laboratory
Wright-Patterson Air Force Base, Ohio 45433

Abstract: A wide temperature range grease has been developed for instruments which operate under light to heavy loads and low to moderately high speeds. It is especially suited for use in applications involving rolling, sliding or oscillating motions and where very little wear can be tolerated. The grease is based on a fluorinated ethylene-propylene or polytetrafluoroethylene thickener in a polyfluoroalkyl-alkyl polysiloxane fluid. The grease's outstanding qualities are its wide temperature range (-100° to 450°F) capability, its extreme pressure and antiwear characteristics, its non-migratory nature, and its low foreign and/or opaque particle content.

The grease is suitable for use in aircraft actuators, gears, gimbal rings, oscillation bearings, antifriction bearings, and plain spherical bearings. It is especially suitable for use in applications employing miniature bearings, and has yielded thousands of hours of operation in some of these applications including blower motors, motor generators, plastic clutches and gears, servo motors, and others.

Key words: Polyfluoroalkyl-Alkyl Polysiloxane; antimony dialkyldithiocarbamate; polytetrafluoroethylene; fluorinated ethylenepropylene; extreme pressure; antiwear; grease; gear lubrication; bearing lubrication.

Advancing aerospace technology has placed ever increasing requirements on the lubricants industry. With the advent of high speed aircraft and aerospace vehicles and their support equipment, and the increase in their scope of operations under extreme pressure conditions, severe demands have been placed on grease lubricants. To meet these demands prior research and development efforts have achieved substantial advancements in the synthesis of a variety of grease lubricants which have many desirable lubricating characteristics and are capable of functioning under moderately high load conditions. Present demands and anticipated demands for grease lubrication in a variety of extreme environments have generated research and development efforts which have resulted in many advanced grease lubricants. While a grease in its simplest form consists of a lubricating fluid thickened by a solid to impart a semisolid consistency, the lubrication problems associated with specific applications have led to the evolution not only of advanced fluid and solid materials, but also third and fourth additive components. Various grease developments have selectively ensured improved antiwear, antibleed, extreme pressure, high speed, antirust, and thermal characteristics. Among these are greases based on the various types of esters and the chlorinated silicones which showed improved lubricity over the straight methylphenyl silicones. Prior efforts have been made to develop extreme pressure greases using these fluids individually and combined in blends. Some success was achieved in developing a series of extreme pressure greases which were usable over a relatively wide temperature range. Their extreme pressure characteristics were improved to yield a Load Wear Index value of about 60 by the use of additives such as pentachlorobenzene thiol and its thioether and heavy metal salt derivatives, and by the use of pentachlorophenyl mercapto acetic acid. Greases utilizing moderate concentrations of the latter extreme pressure additive possessed Load Wear Index values approaching 100. Anti-wear, however, was not one of the outstanding characteristics of this series of greases. In fact, their antiwear properties were very poor.

Today's requirements for grease lubricants to function at temperatures as low as -65 to -100 °F and as high as 450 °F preclude the use of many fluids as base materials. This wide temperature range requirement together with requirements for antiwear and extreme pressure preclude all but the most advanced fluids for use in this type of grease. Prominent among these base fluids are the polyalkyl-alkyl polysiloxanes.

This report discusses the development of greases based on this class of fluid. This class of greases incorporates three very important properties into single grease compositions, namely, wide temperature range, extreme pressure capability, and antiwear, and thus represents an important break-through in grease lubrication technology.

Research efforts were conducted on the development of extreme pressure (EP), antiwear greases for lubricating heavily loaded surfaces required to operate under rolling, sliding and oscillating motions at slow to high speeds, and over a wide temperature range. The materials comprising the grease developed under this effort are described:

1. Investigations showed that the Polyfluoroalkyl Polysiloxane class of fluids possessed most of the basic properties to render them prime candidates as base materials for use in research on extreme pressure, anti-wear greases. The polysiloxane fluid used in this effort has the general formula:

$$CF_3CH_2CH_2Si(CH_3)_2O\left[\begin{array}{c}Si(CH_3)O\\ |\\ CH_2CH_2CF_3\end{array}\right]_n Si(CH_3)_2CH_2CH_2CF_3$$

For the purposes of this research it was desired that the polysiloxane have a molecular weight and value of n which would yield a viscosity of about 50-100 centistokes, and preferably 65-85 centistokes at 100 °F. The polysiloxane covered by the above general formula found to possess the desired properties has the formula:

$$(CH_3)_3SiO\left[\begin{array}{c}Si(CH_3)O\\ |\\ CH_2CH_2CF_3\end{array}\right]_n\left[Si(CH_3)_2O\right]_{n'}Si(CH_3)_3$$

It was also desirable that the viscosity at various other temperatures be in the following ranges:

200 °F	15-25 cs
0 °F	650-725 cs
-65 °F	10,000-11,000 cs

Other desirable properties included:

Pour Point	below -100 °F
Flash Point	above 500 °F
Fire Point	above 600 °F

It was found that the polysiloxane conforming to the second formula above, a trifluoropropylmethyl-dimethyl polysiloxane, possessed the desired properties. These properties are given in Table I.

2. A group of grease thickeners found to be highly effective in producing grease based on the previously described polysiloxane fluid are the fluorinated polymer powders polytetrafluoroethylene

(PTFE) and fluorinated ethylenepropylene (FEP). The investigation of these thickeners was under-taken first by defining parameters that would allow the characterization of the thickeners. Tests were then conducted by appropriate methods to characterize the thickeners as to particle size, particle distribution, surface area, critical surface tension, particle configuration, coefficient of friction, and bulk density. Results suggested trends of efficiency that can be observed in grease formation with these thickeners.

PTFE powders are manufactured by a variety of techniques including grinding, cryogenic pulverizing, molecular chain scission, and pyrolysis, and the behavior of thickeners in grease compositions produced by the different above techniques varies. FEP thickeners present entirely different characteristics. The low surface energy of the PTFE and FEP thickeners means that little shear energy is required to form a grease. The chemical resistance of these materials is excellent and their critical surface tension is below the surface tension of most liquids. An unusual consequence of this is that most greases in which PTFE or FEP is used as a thickener exclude water from interface boundaries and are thus hydrolytically stable. The surface characteristics of an FEP and a PTFE thickener found to be especially compatible with the polysiloxanes to form greases in this effort are described in Table II. The FEP thickener particles are quite spherical. The PTFE, manufactured by molecular chain scission, is practically spherical, and is physically similar to the FEP. These particle configurations were determined by scanning electron microscopy (SEM).

Surface area is an important parameter when used in conjunction with particle size distribution. The nitrogen adsorption method was used in this investigation. A monolayer of nitrogen is absorbed on a known amount of thickener solid. From the volume of nitrogen required and the cross-sectional area of the nitrogen molecule ($15.8\ \mathring{A}^2$) the surface area of the thickener particles may be calculated.

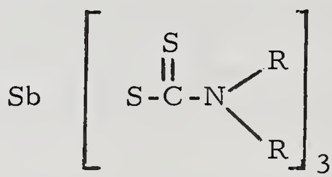
Critical surface tension is also related to thickening. It furnishes information on the behavior of thickeners in oils or liquids. Liquids whose surface tension is below the critical surface tension of the thickener solid will wet the surface and show a contact angle of 0-degree. The lower the surface tension of an oil with respect to the critical surface tension of thickener solid, the greater the thickening action of the solid.

Oil absorption number is a direct test for the relative absorption of thickeners. This standard method involves the titration of linseed oil onto a known amount of thickener until free oil appears.

3. Grease Formulations: A number of grease compositions were prepared by using the polyfluoroalkyl-alkyl polysiloxane base fluids and the previously described FEP and PTFE thickeners. The average proportions are given in Table III.

The greases were prepared by combining the base fluid and thickener in a suitable vessel and mixing with a high speed stirrer. When mixing was complete a thick semi-grease consistency had formed. The mixture was passed through a homogenizer set at 6,000 to 7,000 psi pressure for a sufficient number of times to produce a grease having consistencies representative of those given in Table III. Two to three passes through the homogenizer may be necessary to result in the desired consistencies. A three-roll mill equipped with floating or fixed rollers may also be used with equal success in the preparation of greases.

The use of antimony dialkyldithiocarbamate additive was used to enhance the extreme pressure (EP) and antiwear properties of the greases. The dithiocarbamate is characterized by the general formula



Antiwear and EP properties of the carbamate additives are functions of the alkyl group. Activity generally decreases as the alkyl group molecular weight increases. Activity is also a function of the additive concentration in the grease formula.

Thermal stability of the carbamates increases as molecular weight increases. Since EP activity decreases as molecular weight increases, it was necessary to establish the molecular weight range that would provide a good compromise between EP activity and thermal stability. The greater emphasis was placed on EP activity.

For these investigations the antimony diamyldithiocarbamate was selected. Typical grease formulations using this carbamate additive with trifluoropropylmethyl-dimethyl polysiloxane base fluid and PTFE or FEP thickeners are given in Table IV. The performance of the PTFE and FEP thickened greases were very similar in preliminary and subsequent performance investigations. Therefore, these greases will be discussed as a class employing both thickeners.

The trifluoropropylmethyl-dimethyl polysiloxane greases were tested by using Federal Test Method Standard 791 and ASTM procedures and methods. Table V show the typical physical properties and some of the performance characteristics of this class of greases.

1. Thermogravimetric Analysis (TGA): Thermogravimetric analysis (TGA) was used as a screening test to determine the thermal and oxidative stability of the base fluid, thickeners, additive, and grease. Samples of ten milligrams were tested in an air atmosphere at a programmed rate of 18 °F per minute. The heating rate was controllable from 7 to 32 °F per minute over a temperature range of 32 to 1650 °F.

2. Coefficient of Friction: Coefficient of friction is a parameter directly associated with lubricity. It is defined as the ratio of the force required to move one surface over another to the total force pressing the two surfaces together. The Four-Ball method was used in determining the coefficient of friction involving the grease in this investigation. A load of 15 kg was applied and the speed was adjusted to 1,200 rpm. The test lasted for one hour. Friction coefficient and temperature were measured at the beginning of the test and each 15 minutes during the one hour period. Temperature ranged from 77 °F at the start of the test to 101 °F at the end of the one-hour test. Friction coefficient is given in Table V.

3. Wear Preventive Characteristics: The antiwear characteristics of the grease were determined under sliding steel-on-steel conditions. The standard Four-Ball Wear Test was used to conduct these tests. Conditions were:

Speed	1,200 rpm
Load	40 kg
Duration	2 hours
Temperatures	167, 400, and 450 °F
Specimens	52100, M-10, and 440C steel balls

The average wear scars in duplicate test runs were:

52100 steel at 167 °F	1.18 mm
M-10 steel at 450 °F	1.00 mm
440C steel at 400 °F	1.28 mm

4. Load Wear Index: The extreme pressure characteristics or the ability of the grease to support heavy loads under sliding steel-on-steel conditions were measured by Load Wear Index and weld point determinations using the standard Four-Ball Extreme Pressure Tester. The test is conducted by one load steel ball rotating at 1770 rpm speed against three steel balls held stationary. Test temperature is 80 °F. The load wear index (an index of the ability of the grease to prevent wear at applied loads) and weld point (the lowest applied load in kilograms at which the rotating ball seizes and then welds to the three stationary balls) are given in Table V.

5. Antifriction Bearing Performance: The grease was evaluated for its ability to lubricate 20 mm anti-friction bearings. The tests were conducted under the following conditions:

Bearing	SAE No. 204, 20 mm
Speed	10,000 rpm
Temperature	450 °F
Thrust Load	5 pounds
Radial Load	5 pounds
Cycle	20 hours/day until failure

The average of six tests is shown in Table V.

6. Oscillation Bearing Testing: The oscillation bearing test is another tool to determine the ability of a grease to provide lubrication under heavy loads. This test differs from the fretting corrosion tests in which moving parts are subjected to oscillations of short amplitudes. In this procedure the oscillations were conducted under large amplitudes. The grease lubricated plain spherical bearings were mounted in the tester and a load of 3000 pounds was applied. The test was conducted at 300 °F and 450 °F ± 10 °F and 250 cycles per minute with a total angle of oscillation of ten degrees. The reported number of cycles given in Table V are the average of four runs with a maximum deviation between runs being not more than 20 percent from the average.

7. Starting and Running Torque: Grease torque characteristics were determined in accordance with the ASTM-D 1478 procedure for determining the low temperature torque data. Modifications were made to the test equipment to allow for the use of miniature bearings. The test bearing was an R-4, Grade ABEC 7, double shielded 440C stainless steel. The bearing was packed with 80 microliters of grease, a quantity which previously had proved to be the optimum for this size bearing.

Torque characteristics at the two higher temperatures were obtained on the Sunoco Bearing Grease Torque Tester, Model R-4. The same bearing and quantity of grease were used in all torque tests. The results of these determinations are shown in Table VI.

8. Dirt Content: The standard procedure outlined in Federal Test Method Standard No. 791 was used to determine the concentration of foreign or opaque particles in the grease, except determinations were made for particles in the 10-35 micron range. The following results were determined:

	No. particles/cc
10-35 microns	23
Over 35 microns	11

The majority of the particles observed were opaque and are considered to be agglomerations of the thickener rather than dirt or foreign particles. The grease is considered to be clean and essentially free of dirt or foreign particles. One of the most significant characteristics of this grease is its non-creep property. Creeping was not detected at any stage during the extensive testing and evaluation phases of the program.

Trifluoropropylmethyl-dimethyl polysiloxane grease is being implemented into numerous applications. One of its first uses involved the actuators which control the variable exhaust nozzle (VEN) on the engines of the F-5 aircraft. These actuators were previously lubricated by a proprietary dry film lubricant. The dry film lubrication technique was unsatisfactory because it lasted only about 200 hours, after which time the actuators were replaced. The implementation of the grease extended the actuator lives to 2400 hours with a 400 hour relubrication cycle. Relubrication is accomplished by the addition of lubricant, without removal of the actuator.

This new grease is currently being used in an increasing number of instrument bearing applications. These applications were the result of serious problems which had been assigned to the Rivet Gyro Project Office (currently PRAM Project Office) for timely solution. (The Rivet Gyro Project Office was established by a directive from the Pentagon to develop specific corrective actions to improve the reliability of certain Air Force systems).

1. Analysis of various data had indicated that fully one third of the failures occurring in LN-12 platforms were related to failure of the servo-motors. The primary mode of failure was excessive breakaway torque, which was related to lubricant failure in the bearings. A number of the motors were disassembled, their bearings cleaned, lubricated by the reported grease, and tested. It was demonstrated that the motors repaired in this manner performed better than new motors. It was also demonstrated that the new grease would extend the life of the motors and improve their mean-time-before-failure (MTBF).

2. Based on component failure data it was determined that a problem existed in the spin motor bearings and gimbal bearings of the 7900 D platform which was also related to lubricant failures. The new grease was evaluated for use in the gimbal and spin bearings with the goal of increasing life, reducing torques, and improving system MTBF. The torque reduction would make it possible to eliminate the need for a triple race bearing in the gimbal, thus eliminating the need for the "agitator". This reduction in complexity would result in an increase in MTBF. These evaluations proved to be successful.

3. Motor generators such as those in the Attitude Heading Reference Unit (AHRU) of the C-5 had shown a high failure rate. Examination of these units indicated no electrical failures, however, the bearings were completely dry. Failure was due to failure of the lubricant. A number of these units were reclaimed, cleaned, and relubricated with the new lubricant reported herein. At the end of 1500 hours the tests were terminated and the bearings were examined. It was found that the new grease had performed without failure. Implementation of the grease into this application is allowing for the reclamation of "failed" parts and greatly extending the MTBF of these units.

4. The stabilator actuator rod-end bearing for the F-4 aircraft had experienced excessive wear. The lubricant being used was showing indications of drying out and failing to carry the required load. Implementation of the new grease into this application combined with the resolution of a bearing outer race staking problem resulted in the elimination of the entire problem involving the stabilator.

5. Implementation of the new grease into AAR-34 pumps is resulting in extended MTBF and improved efficiency of the pump. In an efficiency test two pumps were lubricated, one with the prior lubricant, and the other with the new grease. Both pumps were set at 1.1 psi and the 100 hour test was begun. During the test it was noted that the pump lubricated with the new grease ran 20°F cooler. At the end of the 100 hour test, the pump lubricated with the prior lubricant was running at 0.5 psi while the pump lubricated with the new grease was running at 2.2 psi. This indicated an increase in efficiency.

6. A clutch for a radar recorder in the F-111 is now lubricated with the new grease. This clutch was considered to be one of the leading causes of camera failure due to lubricant deterioration. This clutch is now expected to last the life of the camera instead of a few hundred hours.

The grease is being implemented into several other applications including servomechanisms, gears, gearboxes, velocity potentiometers, elevation and azimuth drive motors, tachometer generators, micro switch assemblies, reduction gears, vacuum pumps, NOD drive assemblies, cryo-pumps, and camera drive mechanisms.

Conclusions: A trifluoropropylmethyl-dimethyl polysiloxane grease lubricant has been developed and has been found to be capable of functioning over the temperature range of -100 to 450 °F. Its most outstanding characteristics are its non-creep properties, which preclude the necessity of barrier films or coatings where anti-creep is an important lubrication factor; its extreme pressure capability or its ability to support extremely heavy loads without allowing seizure or welding of the moving parts to occur under dynamic conditions; and its ability to prevent excessive wear of heavily loaded sliding or rolling surfaces. These properties of the grease render it exceptionally effective as a lubricant in applications such as actuators, gear boxes, antifriction and plain spherical bearings, and gears.

The grease has been extensively tested in many components where failure of the component had occurred due to lubricant degradation. These applications and components include the actuators of the variable exhaust nozzle on the engines of the F-5 aircraft, and miniature bearings and gears in various blower motors, servomoters, gyro motors, vacuum pumps, spur gears and discs, micro switches, velocity potentiometers, motor generators, radar recorders, cameras, etc., in the inertial navigation and attack radar systems of aircraft such as the F-4, F-111 and C-5A. The grease is rapidly being implemented into these applications and is resulting in greatly improved performance, significantly increased mean-time-before-failure, and highly improved reliability where high rates of lubricant degradation and failures had previously occurred. Although it is impossible to cite the magnitude of the cost savings attributable to the implementation of the grease into the various systems it is conservatively estimated to be several million dollars.

It is anticipated that this grease will be implemented into several more applications and will make further contributions to the improvement of Air Force operational capabilities.

TABLE I

TRIFLUOROPROPYLMETHYL-DIMETHYL POLYSILOXANE FLUID PROPERTIES

PROPERTY	RESULTS
Viscosity at 200°F, cs	22
Viscosity at 100°F, cs	75
Viscosity at 0°F, cs	692
Viscosity at -65°F, cs	10,427
Pour Point, °F	Below -100
Flash Point, °F	555
Fire Point, °F	680
Density, g/cc	1.136
Surface Tension, dynes/cm	25.2
Evaporation, 22 hrs at 400°F, %	1.1
Evaporation, 22 hrs at 450°F, %	10.1
Four Ball Wear Scar at 1200 rpm, 40 kg load, 167°F, 52100 steel 2 hrs, mm	1.3

TABLE II

THICKENER PROPERTIES

PROPERTY	FEP	PTFE
Density, g/cc	2.1-2.2	2.1-2.3
Particle Size, microns Scanning Electron Microscope, microns	1-2	3-8
Surface Area, m^2/gram	10.1	7.6
Coefficient of Friction	0.07-0.10	0.06-0.08
Critical Surface Tension, dynes/cm	21.9	20.9
Bulk Density (Uncompacted) grams/liter	400	425
Oil Absorption No., lbs oil/lb thickener	47.3	40.9
Molecular Weight	140,000-190,000	25,000-40,000

TABLE III

TYPICAL POLYFLUOROALKYL-ALKYL POLYSILOXANE GREASES

COMPONENTS AND PROPERTY	RESULTS	
Base Fluid, %	65	65
PTFE, %	35	
FEP, %		35
Penetrations		
0 Stroke	330	315
60 Strokes	325	335
10,000 Strokes	300	340
Oil Separation		
30 hours @ 400 °F, %	18	20
30 hours @ 450 °F, %	23	26
Evaporation		
22 hours @ 400 °F	7.1	5.2
22 hours @ 450 °F	10.2	11.1
Coefficient of Friction	0.144	0.145
Load Wear Index	70	70

TABLE IV

FORMULATED TRIFLUOROPROPYLMETHYL-DIMETHYL POLYSILOXANE GREASES

COMPONENTS AND PROPERTY	RESULTS	
Base Fluid, %	62	62
PTFE, %	33	
FEP, %		33
Antimony Diamyldithio-Carbamate, %	5	5
Penetrations		
0 Stroke	298	284
60 Strokes	311	313
10,000 Strokes	310	315
Oil Separation		
30 hours @ 400 °F	10.6	13.4
30 hours @ 450 °F	15.8	18.6
Evaporation		
20 hours @ 400 °F	6.4	3.5
20 hours @ 450 °F	10.3	9.9
Coefficient of Friction	0.086	0.091
Load Wear Index	104	115

TABLE V

PERFORMANCE CHARACTERISTICS OF TRIFLUOROPROPYLMETHYL-DIMETHYL POLYSILOXANE GREASE (TYPICAL)

Penetration	
0 Strokes	290
60 Strokes	315
10,000 Strokes	315
Density, g/cc	1.34
Evaporation, 22 hrs at 400 °F, %	5.50
Evaporation, 22 hrs at 450 °F, %	10.00
Water Resistance, % loss	3.10
Coefficient of Friction	0.089
Load Wear Index	107
Weld Point, kg	500+
Oscillation Bearing Test, 450 °F, Cycles	140,000
Oscillation Bearing Test, 300 °F, Cycles	300,000
High Temperature Bearing Performance Test, 10,000 rpm, 5 lbs load, 450 °F, 20 mm bearings	2,200 hrs

TABLE VI

STARTING AND RUNNING TORQUE CHARACTERISTICS

TEMPERATURE	STARTING TORQUE (in-oz)	STARTING TORQUE (g-cm)	RUNNING TORQUE (in-oz)	RUNNING TORQUE (g-cm)
200 °F	0.04	2.80	0.02	1.40
100 °F	0.09	6.70	0.05	3.70
-65 °F	10.42	750	0.61	44
-100 °F	78.67	5,664	16.39	1,180

BIBLIOGRAPHY

1. Smith, R.K. "Development of High Temperature Heavy Load-Carrying Greases, " WADC TR 58-288, Part III (December 1960).

2. Christian, J.B., "Elasto-Hydrodynamic Lubrication with Fluorinated Polyorganosiloxane Colloidal Lubricants, " NLGI Spokesman, XXXV, 1, (1971).

3. Christian, J.B and Arkles, B., "Polytetrafluoroethylene and Fluorinated Ethylene-propylene Grease Lubricants," Lubrication Engineering, 30, 3, (1974).

4. Christian, J.B., "Aerospace Lubricants-Lubrication Demands and the State-of-the-Art," AFML-TR-68-180, (September 1968).

5. House, W.E., "An Investigation of the Volatility Characteristics of Greases by Thermogravimetric Analysis," AFML-TR-69-78, (1969).

6. Zisman, W.A., "Contact Angle, Wettability, and Adhesion," ACS Advances, Washington, D.C., (1964).

7. Zettlemoyer, A.C., "Spreading at the Teflon/Oil/Water/Air Interface," Journal of Colloidal and Interface Science," 34, 4, (1970).

8. Poster, A.R., "Handbook of Metal Powders," Reinhold Publishing, New York, (1966).

9. Schumaker and Garland, "Experiments in Physical Chemistry," McGraw-Hill, New York, (1967).

10. 1973 Annual Book of ASTM Standards, Part 17, American Society for Testing and Materials, Philadelphia, Pennsylvania, (1973).

11. Federal Test Method Standard 791a, General Services Administration, Washington, D.C., (December 1961).

12. Christian, J.B. (to the United States of America), U.S. Patent 3,642,626, (February 15, 1972).

ANTIWEAR AND LUBRICITY ADDITIVES FOR LUBRICANTS

S. M. Hsu
Amoco Chemicals Corporation
Naperville, Il. 60540

Abstract: The physical and chemical characteristics of the common antiwear additives such as zinc dithiophosphate, tricresyl phosphate are discussed. Effective region of these additives is defined by the load, speed, bulk and junction temperatures, and oxygen concentration. Interactions among additives in a lubricant is important. One additive can completely blank-out another additive. Some additive-metal interactions can also be critical. Proper consideration of these factors in choosing the additive/lubricant system during the design phase is crucial in preventing costly premature failures.

Key words: Antiwear additives; properties; interaction; lubricant; metals; zinc dithiophosphate; tricresyl phosphate.

The trend of high performance, long service life in the design of modern machinery has placed a greater emphasis on the quality of the lubricant used. A variety of additives is used in different base fluids to meet ever increasing performance criterion. Modern lubricants usually contain antiwear agent, corrosion and rust inhibitors, oxidation inhibitors. Sometimes detergents, dispersants, viscosity index improvers and other types of additives are also used in certain applications.

Antiwear additives are mainly used to reduce metal to metal contact under highly loaded conditions. Under these conditions, a very thin oil film exists between the surfaces, but intermittent asperity contact occurs. It is generally accepted that the antiwear additive reacts with the metal surface under localized high temperatures to form a surface film to reduce metallic contact and wear.

Most antiwear additives are organic compounds containing one or more elements or functions such as sulfur, chlorine and phosphorus. Carboxyl and carboxylate salts are sometimes used. Recently, the reduction of wear by the use of compounds capable of forming a polymeric film directly on rubbing surfaces have been reported (1-2). One example of this is the use of one percent equimolar mixture of C_{36} dimer acid and C_{16} glycol to form a polyester film. This additive has been shown to be effective in reducing valve lifter wear in bench automotive engines. Wide-spread use of this new additive type is not foreseen in the near future and will not be discussed here.

Common Antiwear Additives

Some of the common antiwear additives used in liquid lubricants are listed below. Excellent review and listing of these additives can be found in (3).

TABLE I

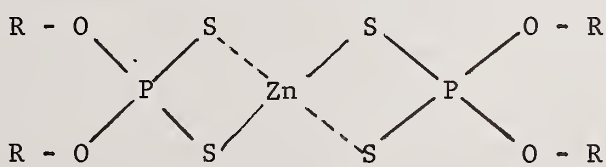
Common Antiwear Additives

Phosphorus	Sulfur
dialkyl phosphites	sulfurized oils
dialkyl phosphates	dibenzyl disulfide
diaryl phosphates	alkyl polysulfides
tricresyl phosphate	zinc dialkyl-dithiocarbamates
tributyl phosphate	

Chlorine	Mixed
chlorinated wax	chlorbenzyl disulfides
chlorinated polyphenyls	zinc dialkyl dithio-phosphates
chlorinated oils	zinc diaryl dithio phosphates

Generally, additives containing sulfur and chlorine are more reactive. They are usually used in highly loaded, high temperature applications to avoid seizure and welding. Sometimes these additives are referred to as extreme pressure (EP) additives. Phosphorus-containing additives are used in moderately loaded situations. The most important commercial antiwear additives are probably tricresyl phosphate (TCP) and zinc dialkyl (aryl) dithiophosphate (ZDDP). These two additives will be examined below in detail.

Zinc Dialkyl (aryl) Dithiophosphate (ZDDP)



Typical properties are listed in Table II.

TABLE II

Physical & Chemical Properties of ZDDP

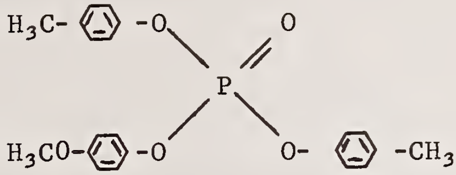
	Alkyl	Aryl
Specific gravity	1.1370	0.9847
Flash point, °C	193	171
Viscosity, cs@210°F	12.7	15
Zn, Wt%	8-10	2-3
P, Wt%	7-9	2-3
S, Wt%	14-17	5-6

ZDDP is extensively used in automotive crankcase oils for gasoline and diesel engines. Its antiwear properties are demonstrated in Figures 1 and 2. These results are obtained in a Shell 4-Ball wear tester with the following conditions: 600 rpm, 75°C bulk oil temperature, 52100 steel on steel ball bearings. Fig. 1 shows that the minimum concentration of ZDDP in a highly refined base oil is about 0.01% by weight. The normal dosage in most applications is from 0.5 to 3.0%(w). Fig. 2 shows the effective load range for ZDDP is from 1 to 20 kg (17,000 kg/cm2). Some extreme pressure protection is shown by ZDDP probably due to the presence of sulfur in the molecules. Forbes (4) had demonstrated by changing the metal of the dithiophosphates, reverse order of performance was observed in antiwear region and EP region. ZDDP showed the best performance in antiwear, moderately loaded region while exhibited the worst performance in the EP region when compared to Bismuth, Tin, Antimony, Lead, Silver, Iron, Nickel and Cadmium dithiophosphates. ZDDP is a very effective oxidation inhibitor. In automotive crankcase applications, it is used to control the oil thickening tendency. In the bench oxidation type of test, a 0.5%(w) ZDDP can reduce the viscosity increase by 2 to 3 times and cut the carbonyl formation by half.

It is generally accepted that ZDDP works by thermal decomposition (5,6). The decomposed products then react with the metal to form a film that is rich in phosphorus but low in sulfur. The oxidation inhibition of ZDDP is probably by peroxide radical decomposition. The thermal decomposition of ZDDP normally is controlled by the alkyl group. There are indications that the lower the thermal stability, the better the antiwear, load carrying capacity. Acids have been shown to catalyze the ZDDP decomposition. Alkaline agents on the other hand retards the decomposition. Its effect on frictional properties is insignificant. Residual stress on the metal surface tends to reduce its antiwear effectiveness (7).

Summing up, the nature of the alkyl group has a marked effect on the antiwear activity of ZDDP. Considerable evidence also shows that the more thermally unstable the ZDDP is, the better its antiwear properties. No conclusive picture has emerged to describe the antiwear mechanism of ZDDP.

Tricresyl Phosphate (TCP)



TCP is a colorless liquid with a viscosity about 30-47 centistokes. Its physical properties are as follows: S.G. = 1.167; Pour point = -25°C; Flash point = 250°C; Autoignition temperature = 590°C; Boiling point at 10mm Hg = 260-275°C. It usually contains about 7.5 to 8.5% phosphorus by weight and the average molecular weight is about 400.

Shell 4-ball wear data are shown in Figures 3 and 4. Fig. 3 shows the minimum concentration required for wear inhibition under 10 kg load (1,000 to 14,500 kg/cm^2 contact pressure), 600 rpm and 75°C. It can be seen that the minimum effective concentration is about 0.5%(w), considerably higher than ZDDP. The normal application range is about 1% to 5%(w). Fig. 4 shows the effective load range of TCP as 1 kg to 40 kg (up to 23,400 kg/cm^2 contact pressure). The higher load data are obtained with a Shell 4-ball E.P. tester. It can be seen that the TCP has no EP wear inhibition under high load, high temperatures. No oxidation inhibition can be observed with TCP. It is also an excellent plasticizer and the compatibility with some vinyl based plastics and elastomers should be checked if the oil will be in contact with them. It has no appreciable effect on frictional properties of the oil. Residual tensile stress on the metal has been shown to increase the antiwear effectiveness of TCP (7).

There have been a lot of studies on the mechanism of TCP (4, 9, 10). Although confusion still exists on the exact mode of action, a general picture has emerged. The polar components in TCP (acid phosphates) are preferentially adsorbed and reacted with the iron surface to form both organo/inorgano phosphate films which carry the load. Subsequent source of the polar acid phosphate material may come from the hydrolysis/thermal decomposition/oxidation of TCP.

General Mechanism for Antiwear Additives

Antiwear additives can be viewed as a functional chemical group attached to a hydrocarbon backbone which provides solubility and compatibility with the base fluids. This also applies to other additives such as detergents (metal sulfonates, phenates, succinimides, etc.). The solution state of these additives in base fluids, due to their polar functional group, is probably in aggregate conglomerate. Watson (11) recently suggested that all polar additives existed in micellar solutions. The critical micelle concentration (CMC) defines the solution state of the additive. In most additive solutions, the bulk of the additive exists in micellar aggregates above the CMC, only 10^{-4} to 10^{-7} moles of the additive exist as individually solvated molecules below the CMC in equilibrium. These molecules then compete for the active adsorption sites of the metal surface. There are also indications that multilayer adsorption takes place. These adsorbed molecules then react with the metal under high temperatures (260°-370°C) and pressures to form organometallic compounds. Sometimes the organic portion of the molecules may be further decomposed/oxidized leaving only the inorgano-metal compounds which constitute the protective film.

Base Oil

The base oil used in a specific application can affect the additive performance. Naturally occurring polar materials such as the oxygenated, nitrogen-, sulfur-containing compounds can compete with the antiwear additives for the metal surface. Aromatics generally have better antiwear properties as a base fluid. In some cases, different base fluid-additive interactions can occur depending on the chemical structures of the base fluid. Fig. 5 shows that in a super-refined paraffinic mineral oil where all the aromatics have been removed, further purification of the base oil gives better antiwear response to TCP. The data are obtained from a Shell 4-ball wear tester at 75°C, 40 kg, 600 rpm and 100 minutes duration. The purification step consists of percolating the white oil through an activated alumina column to remove the polar impurities.

Additive Interactions

Interactions of various additives in fully formulated oils is a serious problem. Recent trend in lubricant formulation centers around longer service life and superior performance, sometimes even at the risk of lower cost/effectiveness. This has prompted more additives at higher concentrations to be used. Additive interaction can be both physical and chemical. The physical interaction consists of competing for the adsorption sites from the standpoint of polarity, micellar aggregation of different additives, solubilization or insolubilization of other additives. Chemical interactions are usually chemical reactions among the additives. Very little is known on this phenomenon. In automotive crankcase oil, the interactions between ZDDP and amine, dispersant and detergent, dispersant and viscosity index improver are known. Additive concentration can also affect the performance. Too high a concentration of a single component can cause multiple interactions that is not observed at lower concentrations. Some antiwear additives are sensitive to water, acidity e.g., ZDDP. When the base oil is changed to a fluid where the major decomposition products are acidic, the effectiveness of ZDDP in that fluid may be seriously impaired.

The composition for active adsorption sites on the metal surface can have marked effect on the additive performance (12). This is shown in Table III.

TABLE III

Effect of Preferential Additive Adsorption

Shell 4-ball wear test with 40 kg load, 600 rpm, 75°C

Base fluid: super-refined paraffinic mineral oil

Additive	Wear Scar Diameter,mm	Iron in Wear Debris,μg		
		Organic Iron	Iron	Iron Oxides
(A) 2.25% TCP	0.60	14	26	78
(B) 2.25% C_{18} Bromide	0.90	106	120	120
(A) + (B)	0.55	12	27	27

Both the degree of wear and the wear product analysis reveal no effect of the octadecyl Bromide when TCP is present.

Other Factors

Variation in atmosphere and moisture content can have strong effects on antiwear additive performance (9). Since this relates to a particular chemical interaction between the additive and oxygen, water, no generalized statement can be made. The presence of oxygen in a lot of cases seems to help.

The metallurgy of the system has profound effect on the efficiency of the load-carrying additives. Rounds (8) has demonstrated quite clearly that different steels can have entirely different responses to the same antiwear additive. Although all of the steels contain 80% or higher iron, the best additive for 52100 steel may not be the best for 440C steel.

The temperature at the wear contact junction is another important parameter. All antiwear agents have their own effective temperature range. The junction temperatures depend on the metallurgy, contact geometry, load, speed and bulk temperatures. If this temperature is too high by improper design, conventional lubricants containing TCP, ZDDP may not work and premature failure occurs. Modified Blok-Archard flash temperature calculations can be used (13, 14) to estimate the junction temperature.

Recommendation

Antiwear additive performance is affected by a wide range of parameters such as load, speed, atmosphere, base oil type, metallurgy, other additives. This article illustrates the complexity and the "art" state of formulating a lubricant. Since very little information is available, the recommended action for a designer is prototype testing or bench simulations where the important parameters mentioned before are carefully matched with the expected service conditions. If special lubricant/additives are required, specification of such should be made. A design in which the lubricant is an integral part of the system can prevent costly premature failures.

REFERENCES

(1) Furey, M. J., "The Formation of Polymeric Films Directly on Rubbing Surfaces to Reduce Wear," Wear, 26, 369-392 (1973).

(2) Zaslavsky, Y. S. et al, "Laboratory Investigations and Industrial Use of Friction Polymer Forming Additive Treated Industrial Oils," JSLE/ASLE International Lubrication Conference, Tokyo, Japan, June 1975.

(3) Smalheer, C. V., Smith, R. K., "Lubricant Additives," The Lezius-Hiles Co., Cleveland, Ohio, 1967.

(4) Forbes, E. S., "Antiwear and Extreme Pressure Additives for Lubricants," Tribology, August (1970).

(5) Rounds, F. G., "Some Factors Affecting the Decomposition of Three Commercial Zinc Organodithiophosphates," ASLE Trans, 18, 79-89 (1975).

(6) Coy, R. C., Quinn, T. F. J., "The use of Physical Methods of Analysis to Identify Surface Layers Formed by Organosulfur Compounds in Wear Tests," ASLE Trans, 18, 163-74 (1975).

(7) Sakurai, T. et al, "Effect of Internal Stress on the Wear Behavior of Steel During Boundary Lubrication," ASLE Trans, 17, 213-23, (1974).

(8) Rounds, F. G., "Influence of Steel Composition on Additive Performance," ASLE Trans, 15, 54-66 (1972).

(9) Goldblatt, I. L. and Appeldoorn, J. K., "The Antiwear Behavior of TCP in Different Atmospheres and Different Base Stocks," Presented at the ASLE/ASME Conference, Houston, Texas, Oct. 14-16, 1969.

(10) Klaus, E. E. and Bieber, H. E., "Effect of P^{32} Impurities on the Behavior of Tricresyl Phosphate-32 as an Antiwear Additive," ASLE Trans, 8, 12-20 (1965).

(11) Watson, R. W., "Petroleum Additives-The Role of Akyl Groups," JSLE/ASLE International Lubrication Conference, Tokyo, Japan, June, 1975.

(12) Hsu, S. M., "The Effect of Chemical Reactions in Boundary Lubrication," Ph. D. Thesis, Chemical Engineering Dept., Pa. State University, University Park, Pa. 16802, 1976.

(13) Bos, A., "The Temperature at the Wear Scars of the Four-Ball Apparatus," Wear, 31, 17-27 (1975).

(14) Francis, H. A., "Interfacial Temperature Distribution Within A Sliding Hertzian Contact," ASLE Trans, 14, 41-54 (1970).

Fig 1
Minimum effective concentration for ZDDP

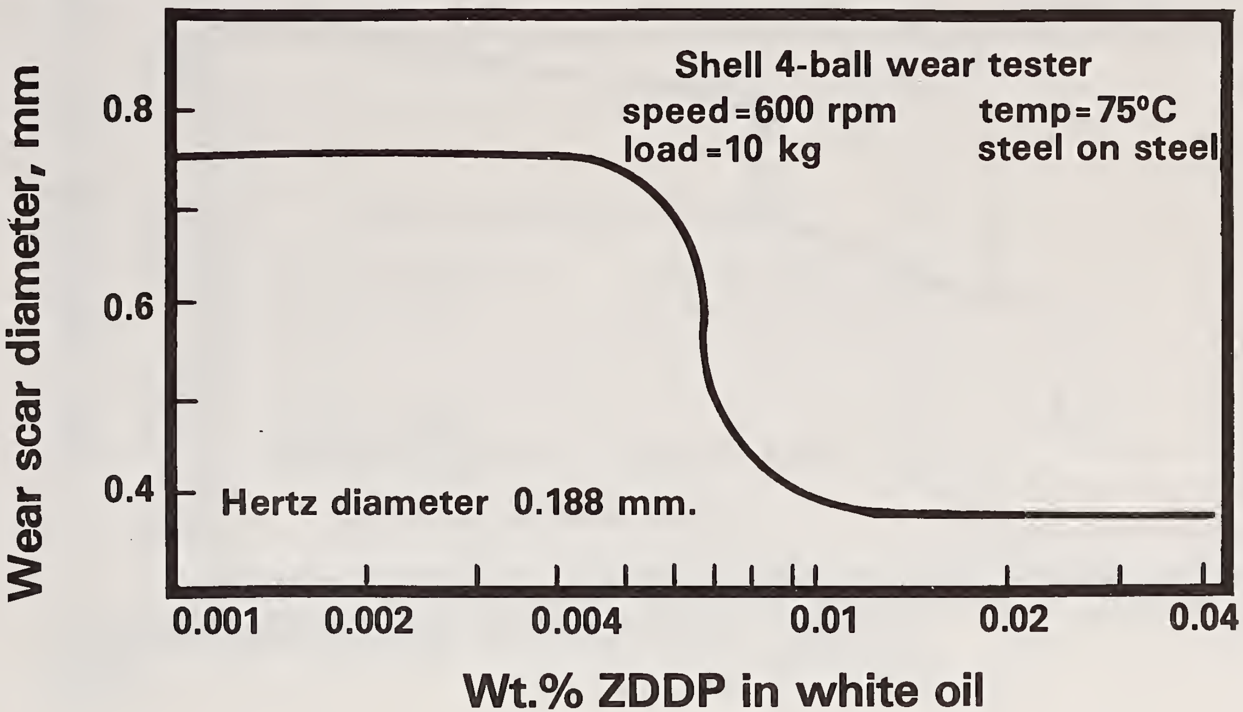


Fig 2

Effective load range for ZDDP

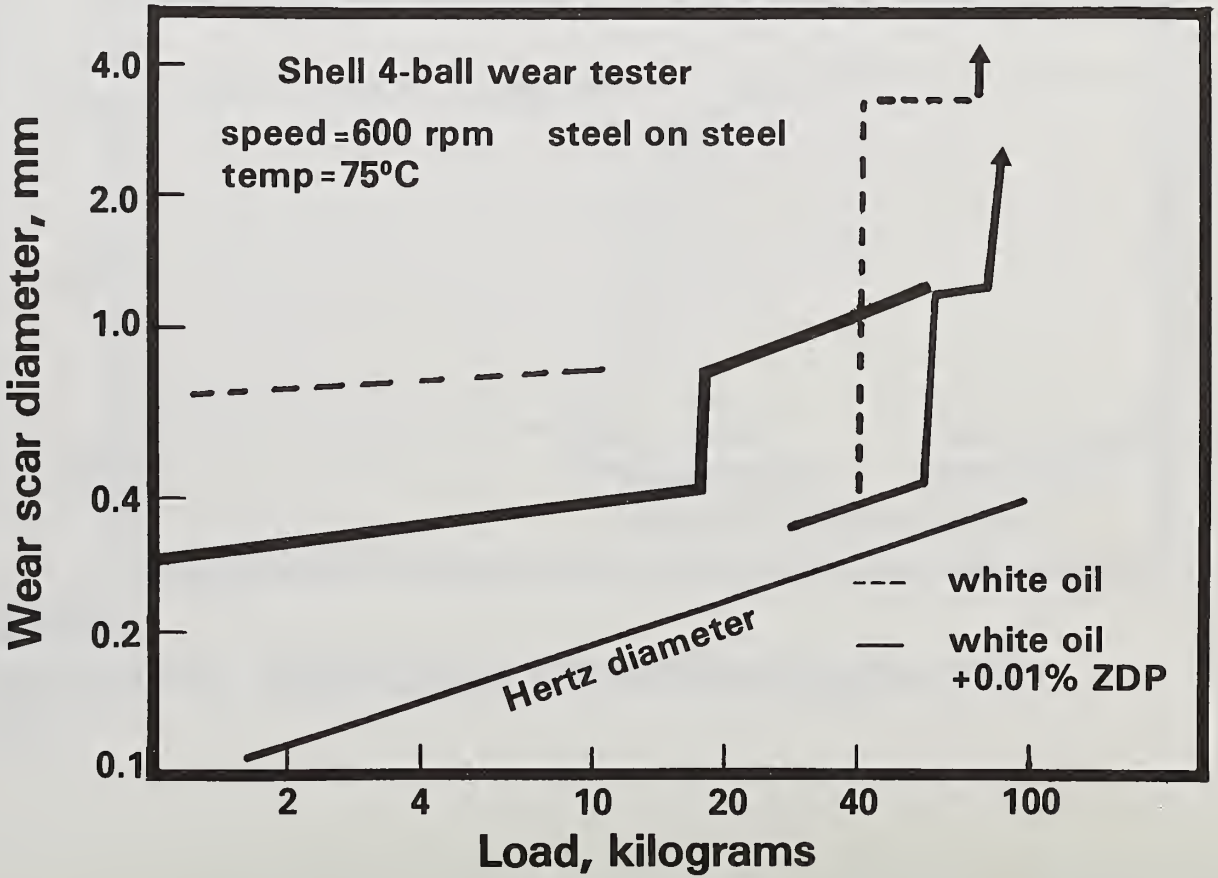


Fig 3

Minimum concentration range for TCP

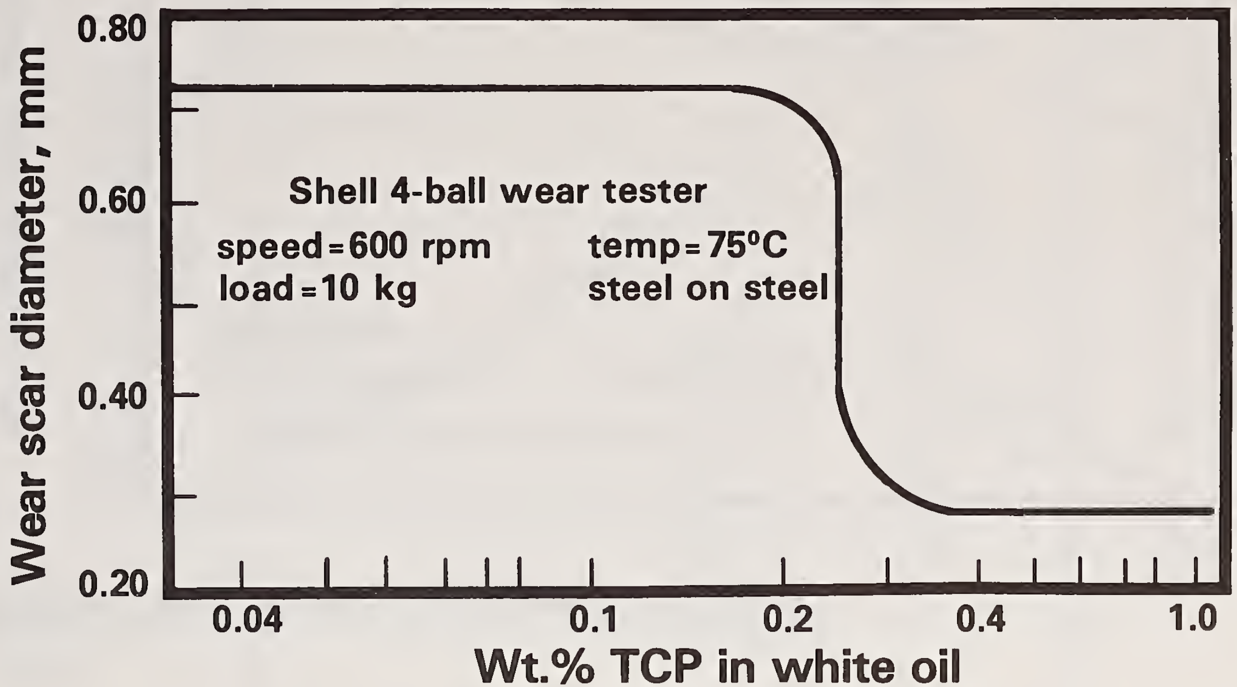


Fig 4
Effective load range for TCP

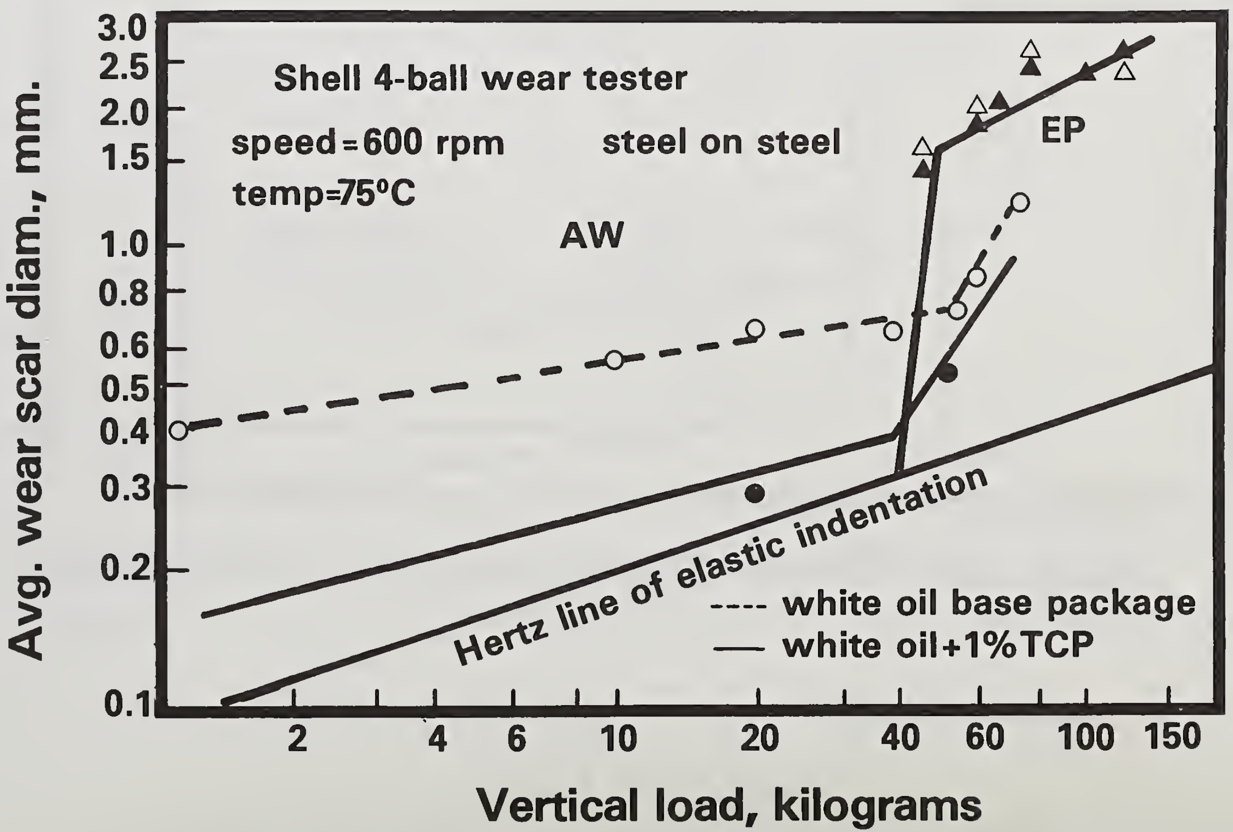
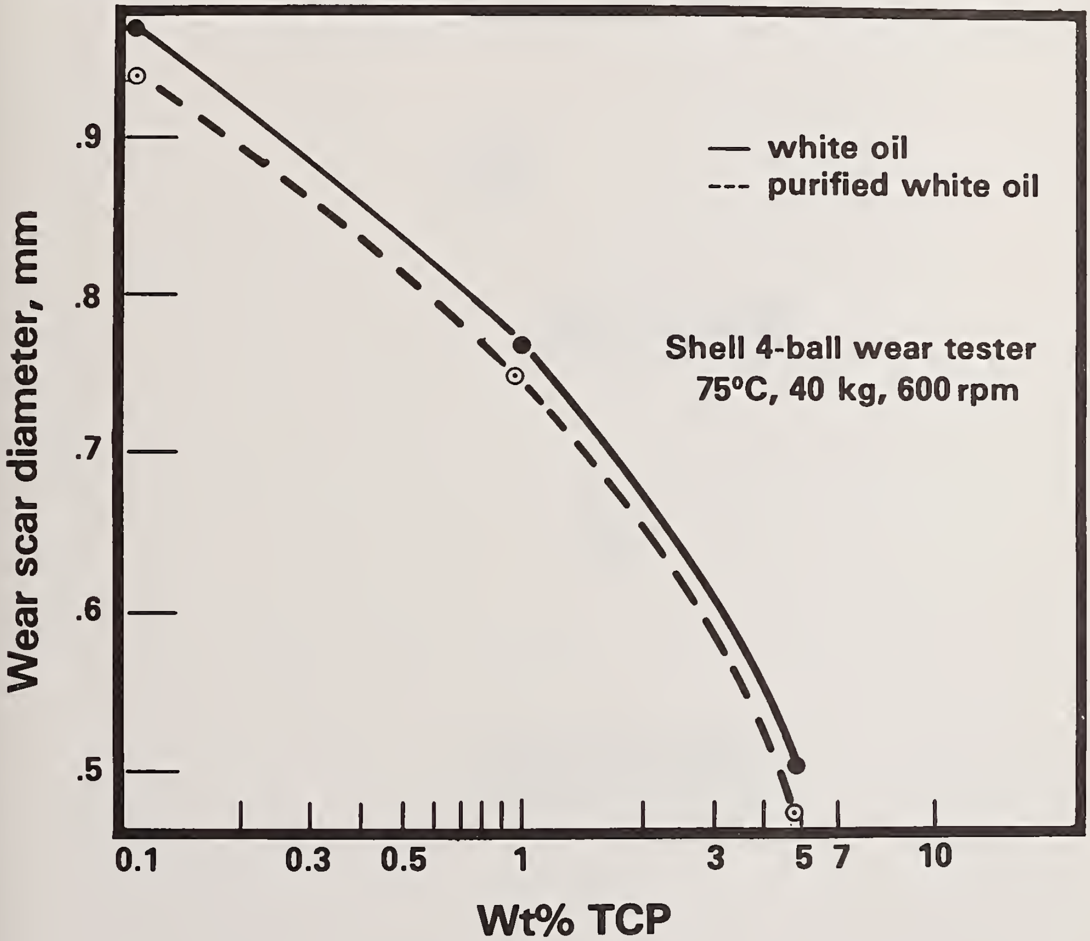


Fig 5

Effect of base oil purity on the wear inhibtion of TCP



SESSION V

VALIDATION

OF

DESIGN

(I)

-TESTING-

Chairman: W. W. Gunkel
General Motors Corporation

PIEZOELECTRIC POLYMERS AND THEIR APPLICATIONS

Seymour Edelman
National Bureau of Standards
Washington, D.C. 20234

Abstract: Polymers with significant piezoelectric activity have been developed. This activity has been used to make measuring instruments of thin polymer sheets. Sensors have been made to measure dynamic stress, strain, acceleration, sound, vibration, and the effects of mechanical impact or shock. Polymer sensors differ from conventional instruments in being light, flexible, thin, highly damped, easily attached to curved, twisted, or compliant surfaces, uniform in response over a wide range of frequencies, and cheap. Neither the instrument or its leads is subject to mechanical fatigue. Polymer instruments have obvious applications in failure prevention. For example, the condition of bearings can be monitored by cementing polymer gages to any available portion of a bearing, the signal analyzed to provide the noise spectrum of the bearing, and the spectrum checked at periodic intervals so that deterioration in the condition of the bearing is noticed in time to prevent catastrophic failure. Other components of machinery can be monitored in similar ways. Typical noise spectra will be shown. The process of poling the material will be outlined. The construction, characteristics, and uses of typical sensors will be described.

Key words: Acoustic emission; damage detection; failure prevention; noise spectrum; piezoelectric polymer; vibration sensor.

The most surprising thing about piezoelectric polymers is that they exist. Polymers differ from conventional piezoelectric materials in so many of their properties that it seems strange that they should be comparable in their piezoelectric response.

Conventional piezoelectric materials are typically crystalline or polycrystalline, hard, stiff, brittle, and dense. It is difficult to machine them into thin layers or to obtain them in sheets of large area and they are expensive. Polymers are compliant, flexible, tough, and light. They are available commercially in layers as thin as six micrometers (one-fourth of one thousandth of an inch); they are available in 1000 meter rolls over one meter wide; and they are relatively cheap.

Measuring instruments using conventional piezoelectric materials usually have metal bases and housings and require threaded holes or specially ground flat surfaces of appreciable area for mounting. Polymer instruments [1,2] typically consist only of the active material and a lead of the same polymer. The active material can be cut to any shape that is suitable for a particular use and can be attached to a surface whose only preparation is cleaning. Rubber cement, cyanoacrylate, epoxy, or other cements can be used. A polymer gage can be attached to curved, twisted, or compliant surfaces.

Conventional instruments tend to have many high-Q resonances both because of their internal spring-mass systems and because of the effect of the mass of the instrument on the elastic compliance of the mounting surface. Polymer instruments are flat with frequency over wide ranges, typically into the megahertz range, their resonances are low-Q, and their mass is so low that they do not resonate with the mounting surface. The piezoelectric modulus, g, governs the usefulness of a material as a sensor. Its value for the usual piezoelectric polymer, poly(vinylidene fluoride) is about six times as large as for a typical lead zirconate titanate ceramic. On the other hand, the value of the d modulus, which governs the usefulness of the material as a generator of motion or as a sound source, is about 21 times larger for the ceramic.

There are some cases of considerable importance where polymer instruments are uniquely suited. Usually the advantages of polymer instruments show up in dynamic measurements. A typical case is measuring the level of vibration at a point on a thin metal sheet. Any point on the metal sheet to which an accelerometer is attached will have significantly greater surface density than the rest of the sheet. The vibration pattern of the sheet will rearrange itself so that the motion at that point is minimized and the measurement is unrepresentative. On the other hand, cementing a small piece of polymer film to the metal sheet need not change the surface density significantly and a meaningful measurement can be made. The amplitude of vibration and the variation of vibration level with frequency measured by the polymer are more nearly representative of what the level at the point would be with no instrument attached. Conversely, both the vibration amplitude and the spectrum of resonances measured by an accelerometer would be affected by the presence of the instrument. Thus, it is practicable to distribute a number of polymer vibration gages over the surface of a panel and to infer from their reading the mode of vibration of the panel under various conditions. This kind of study cannot be performed effectively with conventional vibration measuring instruments.

Our Program Chairman, Jesse Stern, suggested another application for which polymer gages are well suited. This is monitoring the noise spectra of bearings, gear trains, and other transmission systems. Such monitoring has been done with various kinds of vibration measuring instruments. The use of conventional instruments involves a number of difficulties. In some cases, the size of the instrument has prevented it from being mounted directly on the source of the sound and when it was mounted on a nearby structure, it responded to sound from a number of sources. Sometimes its mass affected the motion being measured. Sometimes its own resonances obscured the spectrum of the noise it was measuring. Occasionally, the relatively high cost of the gage made it uneconomic to mount a gage permanently at each point where a spectrum was needed so that a single gage had to be moved from point to point and variations in mounting conditions caused changes in the spectra taken at different times.

Polymer gages eliminate or minimize these difficulties. A polymer gage can be cut to fit whatever surface is available for a given measurement. For example, a crescent-shaped gage can be mounted to the stationary race or the dust shield of a bearing. There is little doubt that it responds to the noise generated directly beneath it. Neither its mass nor its internal resonances affect the signal. The low cost of the gage makes it practicable to mount one permanently at each point of interest so that, if periodic readings of the signal indicate a change in the spectrum, the change can be ascribed to changes in the bearing and not to differences in mounting conditions.

A typical spectrum of the noise of a ball bearing is shown in Figure 1: Two ways of mounting the gage to the bearing are also shown.

Polymer gages are handy for acoustic emission studies. They can be bonded to irregularly shaped surfaces and provide good coupling. Their high internal damping minimizes the effects of ringing. They have a wide useful frequency range extending well into the megahertz region. Their sensitivity is adequate for most work and can be increased by stacking and by the use of bias, if necessary. It is cheap and convenient to use gages in sufficient numbers so that noise sources can be located by triangulation.

The detailed mechanism of piezoelectric activity in a polymer is complicated and controversial. The composition of the polymer, its division into amorphous and crystalline parts, the crystal structure, the presence of surface and space charges and of ionic impurities have all played parts in different explanations. We get useful guidance in preparing material for use in measuring instruments if we consider the polymer to consist of long chains of identical units called monomers. In the case of poly(vinylidene fluoride) (PVF_2), the commonly used piezoelectric polymer, the chains consist of linked carbon atoms and each monomer consists of two carbons, one joined to two hydrogen atoms and the other joined to two fluorine atoms. Each monomer has a strong dipole moment. As the material is received the dipoles are oriented randomly as shown in Fig. 1a. We heat the material, apply a strong electric field, and return the material to room temperature with the field applied as shown in Fig. 1b. We can consider that this process results in the alignment of a significant number of the dipoles normal to the plane of the sheet. The dipoles can be considered to be stiffer than the bonds between dipoles in adjacent chains so that any stimulus that changes the thickness of the sheet will change the surface density of charge. A compensatory flow of charges through the circuit connecting the electrodes forms a signal as shown in Fig. 1c. More detailed analyses have been published [3].

A typical polymer sensor is made from two sheets of the polymer. The active area is a circle, one centimeter in diameter. The active area has evaporated aluminum electrodes on each side of each sheet. The sheets are cemented or fused together so that charges of the same polarity appear on the inner faces and charges of the opposite polarity appear on the outer faces. Each inner face has a thin evaporated aluminum lead running from the active area for about ten centimeters to the center conductor of a coaxial connector. Each outer face is completely covered with evaporated aluminum which is connected to the shield of the coaxial connector. This kind of coaxial construction allows the signal to be completely shielded and all exposed surfaces are at ground potential. The evaporated metal lead on the same sheet of polymer as the active material allows the comparitively massive connector to be far enough from the point at which the measurement is taken to avoid any interference with the measurement. The thin polymer leads can be bent, twisted, or stretched without damage and they are not subject to fatigue. They can be cemented to a supporting surface so that they are not damaged by a blast of air or a reasonable amount of friction and they do not generate noise by their motion.

To avoid leaving you with the impression that polymer gages are perfect in every way, I will describe the effect of temperature [4]. Returning to the picture of a poled polymer as containing a large number of electrical dipoles aligned normal to the plane of the sheet, we can see that lowering the temperature will make the sheet thinner, just as pressure will. In either case, the effect is to change the surface density of charge on the surfaces, giving a signal. Since a signal due to pressure comes from the same phenomenon as a signal due to temperature, there is no way to separate the two effects if they occur at the same frequency. Fortunately, many of our measurements are made at constant temperature except possibly for the effect of adiabatic heating which can be taken into account in the calibration process. In many other cases, the temperature changes slowly compared to the rate of pressure change so that the effect of temperature can be filtered out. Nevertheless, it must be realized that the polymer gages have a very large pyroelectric effect and also that the sensitivity of the pressure gage varies with temperature.

In those cases where the temperature changes at the same frequency as the pressure or where the ambient temperature changes drastically, we have evaporated a bismuth-antimony thermocouple right onto the polymer surface. This polymer thermocouple is an interesting instrument in itself since the metal is only a few hundred angstroms thick giving a very fast response. It is not subject to fatigue and has many of the other advantages of polymer instrumentation. In this application, the thermocouple signal is proportional to temperature but not to pressure. The pressure gage signal is a function of both temperature and pressure. A little arithmetic gives the effect of pressure alone.

The last figure shows a kind of non-invasive, non-destructive measurement that might have some application to personal failure prevention.

References

1. S. Edelman, S.C. Roth, and J.F. Mayo-Wells; U.S. Patent 3,970,862, July 20, 1976.

2. S. Edelman, L.R. Grisham, S.C. Roth, and J. Cohen, J. Acoust, Soc. Am. 48, Part 1, 1040 (1970).

3. F. I. Mopsik, and M. G. Broadhurst, J. Appl. Phys., 46, 4204 (1975).

4. J. Cohen, S. Edelman, and C. F. Vezzetti, Nature Physical Science, 233, No.36, 12, (1971).

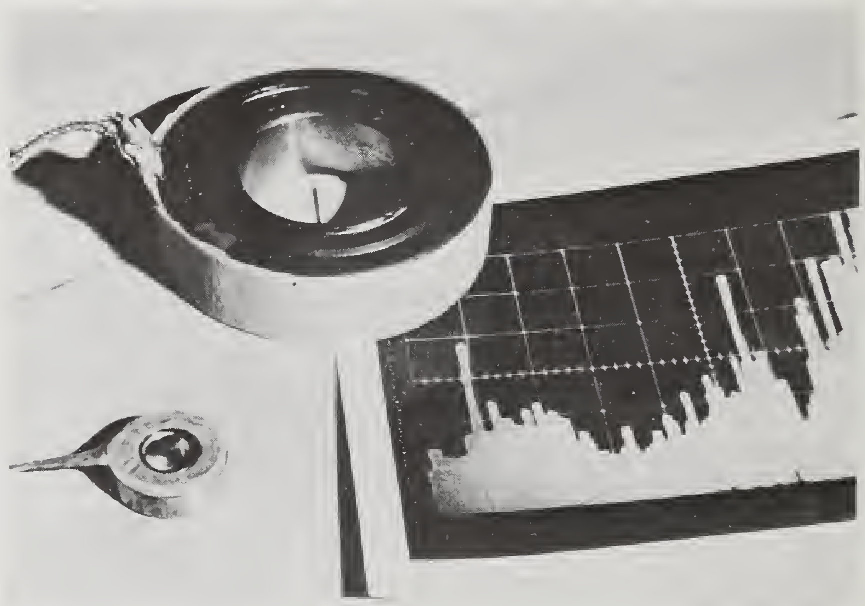


Figure 1 Noise spectrum of ball bearing.

Two methods of attaching polymer gages to bearings.

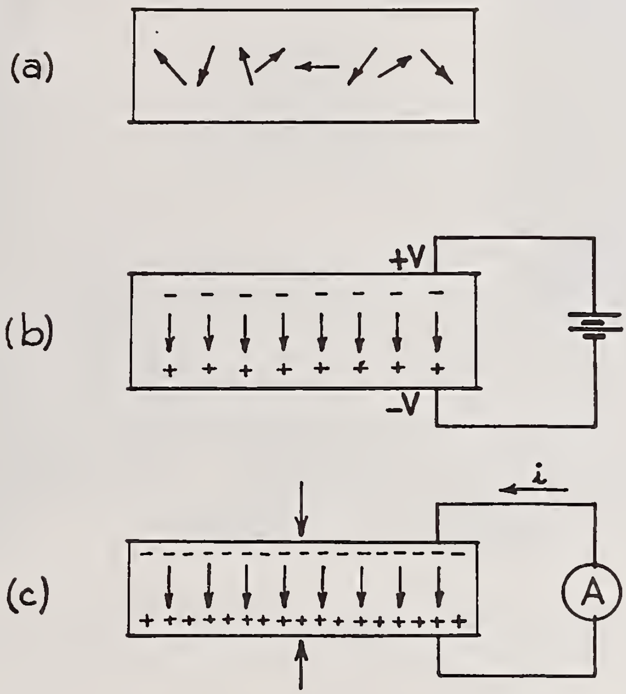


Figure 2a   Unaligned dipoles.

Figure 2b   Dipoles aligned by field.

Figure 2c   Signal due to pressure.

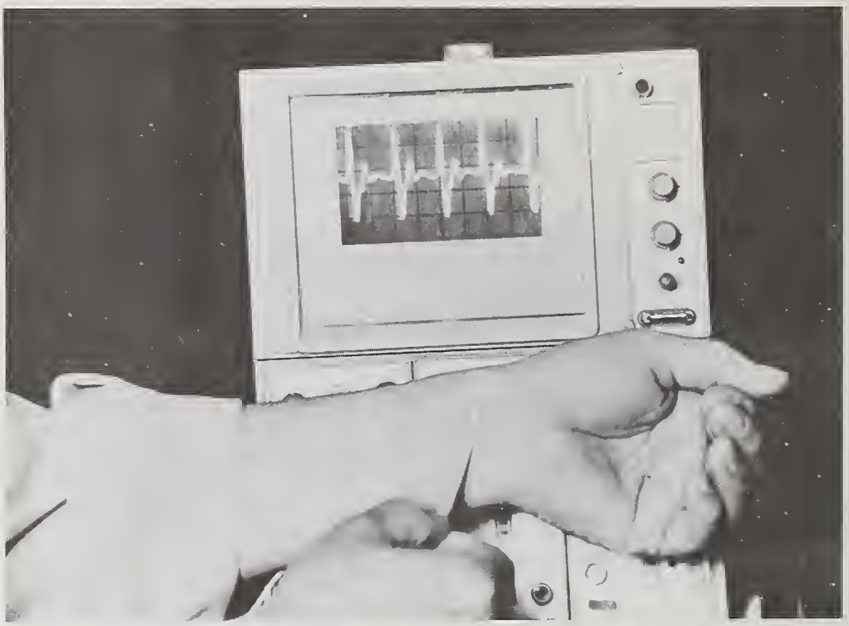


Figure 3 Pulse detected by polymer gage.

EXOELECTRON EMISSION - PAST EXPERIENCE AND FUTURE EXPECTATIONS

W. J. Baxter
Research Laboratories
General Motors Technical Center
Warren, Michigan 48090

Abstract: During fatigue testing of metallic components there is a gradual accumulation of "damage", leading ultimately to total failure. An early manifestation of this damage is the development of surface deformation which eventually provides sites for fatigue cracks to nucleate. However prior to the development of a crack in the metal itself, the surface deformation produces microscopic cracks in the natural layer of surface oxide. Under ultraviolet illumination photoelectrons from the underlying metal can escape more easily through these microcracks, such emission being referred to as exoelectron emission.

Thus exoelectron emission provides a measure of the accumulation of fatigue deformation. By scanning the surface with a spot of ultraviolet radiation the distribution of emission (i.e. deformation) can be displayed, and the severity of the damage assessed from the localized intensity of the emission. After only 2% of life the ultimate fatigue life may be reliably predicted. This technique should be applicable to all metallic materials, but the range of variables such as stress level, yield strength, oxide thickness, etc., has yet to be completely explored.

Key words: Assessment of fatigue damage; exoelectron emission; failure prediction; laser scanner; oxide rupture; surface deformation.

INTRODUCTION

Design engineers ultimately resort to some type of test procedure to validate a structural or engineering component and ensure its reliability in service. This is an essential part of the design process, and can be very costly in terms of the time expended. This is particularly true in applications where the ultimate failure is due to metal fatigue. Such failures occur with little or no warning after apparently satisfactory operation during millions of cyclic loadings at low stress levels. Thus the determination of the fatigue performance requires at least days, but more often weeks or months. At the end of that time the designer obtains a vital (but solitary) piece of information, namely the number of cyclic loadings to cause failure. In some applications it may be necessary to design for an infinite fatigue life, i.e. ensure that the service stresses are below the so-called fatigue limit for a given material. A test in this case is usually eventually terminated after ~10^8 load cycles, the engineer justifiably concluding that the design is certainly adequate. On the other hand the structure may be overdesigned and require further refinement and testing to be cost effective. This situation represents the extreme in terms of the testing time required.

Thus the desirability of an abbreviated fatigue testing technique has long been recognized, and a wide variety of approaches have been explored but with very limited success.[1] The problem is far from simple because of the complexity of the fatigue phenomena. The latter may be conveniently regarded as consisting of two processes i) the initiation of a fatigue crack, and ii) propagation of the crack to ultimate failure. Neither of these processes are completely understood, but they are both known to involve plastic deformation within very localized regions of the metal. Due to the lack of any quantitative measure, or even a useful definition, this deformation is generally referred to as fatigue damage. The objective then is to detect and measure this damage early in the fatigue life, and thereby be able to terminate the test and predict the number of cycles to failure (N_f).

Most techniques which have been tried have been aimed at detecting the initial fatigue crack, which usually develops after about 10% N_f. The exoelectron method on the other hand detects and measures the accumulation of surface deformation after ~1% N_f, i.e. during the crack initiation stage.[2] The origin of this so-called exoelectron emission is now understood, and a scanning technique has been developed which provides quantitative measurements of the all important localized fatigue deformation. This is illustrated by results for metals with quite different yield strengths. Experience with the technique is now sufficient to permit the main requirements to be specified, but there are still some aspects requiring further investigation.

ORIGIN OF EXOELECTRON EMISSION

An early manifestation of fatigue damage is the development of surface deformation which eventually provides sites for fatigue cracks to nucleate. However prior to the development of a crack in the metal itself, the surface deformation produces microscopic cracks in the natural layer of surface oxide, revealing microscopic regions of bare metal. Under ultraviolet illumination, photoelectrons from the underlying metal are able to escape more easily from the freshly revealed metal surface in the microcracks than from the surrounding oxide coated surface. In this way the accumulation of surface deformation increases the photoelectron current, this increase being known as exoelectron emission.

The validity of this model was irrefutably established by direct observation of the exoelectron emission from a specimen deformed in the vacuum of a photoelectron microscope.[3,4] This is illustrated by the photoelectron micrographs in Fig. 1. The exoelectron emission produced by unidirectional tensile deformation of SAE 1018 steel appears as white lines in Fig. 1a. These sources of emission are slip steps, where the very thin (~4 nm) natural oxide has ruptured to reveal the fresh metal surface of the step itself. If the step is allowed to reoxidize by simple exposure to ambient air for a few hours, the exoelectron emission is completely suppressed (Fig. 1b).

The surface deformation produced by fatigue cycling is more complex than that for the unidirectional deformation illustrated in Fig. 1. The deformation is much less homogeneous, and very localized regions of severe deformation develop. The course of events for 1018 steel is summarized by the scanning electron micrographs in Fig. 2, which show the small regions of most severe deformation at various stages of fatigue life. At ~0.5% of the fatigue life, the only visible deformation is the appearance of isolated arrays of slip traces as in Fig. 2a. These traces are in fact very small extrusions of metal, which again will rupture the surface oxide. The surface area of the extruded metal is small at this stage, but as fatigue cycling continues the slip traces develop more structure and broaden (Fig. 2b), and gradually merge together to form a very mutilated surface (Fig. 2c). Eventually i.e. ~6% of life, a localized "honeycomb" surface structure develops with a very large surface area as shown in Fig. 2d. (It is only at this stage that the first fatigue crack develops in the metal itself.) It is this large increase of surface area of extruded metal in these localized regions which results in the large increase of exoelectron emission during fatigue.[2]

ASSESSMENT OF FATIGUE DAMAGE

The details of the localized surface deformation outlined above are of course effectively lost in the initial topography of a real surface. Nevertheless their presence can still be detected by means of the associated exoelectron emission. The procedure consists of first fatigue cycling the sample inside a vacuum chamber, then scanning the surface with a small spot of ultraviolet radiation, and measuring the intensity of the photoelectron emission. As the spot passes over a fatigued region the electron emission is more intense (i.e. exoelectrons), and it is this localized exoelectron emission which provides a quantitative measure of the accumulation of damage.[2]

This technique was initially developed[5] using a simple optical system with a scanning ultraviolet spot of ~70 μm diameter and a scan field of only ~0.3 cm x 0.2 cm. The test samples were correspondingly small and were fatigued in the reverse bending mode. As the spot scanned along the gauge length of the sample the intensity of the electron emission was displayed as a function of position on an X-Y recorder. An example is shown in Fig. 3 of the distribution of electron emission from a sample of 1100 aluminum. The onset and growth of exoelectron peaks identify regions of fatigue damage. The initial background emission in the undeformed areas remains unchanged, and corresponds to the grey areas of intact oxide in Fig. 1a.

From the microscopy studies (e.g. Figs. 1 and 2) it is reasonable to expect that the largest localized emission peak will identify the location of the most severe accumulation of fatigue damage. Furthermore, the intensity of a peak should be some measure of the extent of the local damage. Strong support is provided by the results for SAE 1018 steel shown in Fig. 4. In this graph only the largest peak is plotted for each of twelve samples fatigued to failure. The intensity (I) is normalized with respect to the initial background emission (I_0). Since the samples were fatigued at different stress levels to cover the indicated range of fatigue lives, the abscissa is also normalized and expressed as fraction of life. All the data points fall within a scatter band which is quite reasonable for fatigue experiments.

Results of this type provide a calibration curve for the assessment of the extent of accumulation of fatigue damage, from which the remaining number of cycles to failure can be predicted. For example, in this case, if an emission peak develops with a normalized intensity value of 10, then the sample has been damaged to such an extent that between 0.8 and 3% of its ultimate fatigue life has been consumed. Such a measurement constitutes a very abbreviated fatigue test, and predicts the fatigue life as accurately as normally experienced by the time consuming procedure of cycling to failure.

LASER SCANNING SYSTEM

The application of the above technique to the evaluation of the fatigue performance of structural or engineering components requires a small spot of ultraviolet radiation to be scanned over a large area. A frequency doubled argon-ion laser (wavelength=257.3 nm) is well suited to this task, since it has sufficient energy (4.8 eV) for the emission of electrons from metals, and it may be readily focused to a small spot and scanned over a large area. In these initial experiments a spot diameter of 14 μm and a scan path of 9 cm were selected, as appropriate for measurements on cylindrical samples of martensitic steel subjected to torsional fatigue.

The laser scanning system and the procedure for recording the electron emission is shown schematically in Fig. 5, and has been described in detail elsewhere.[6] The electrons photoemitted from the specimen by the incident laser beam are collected by an electron multiplier. The entrance cone of the multiplier protrudes through a grid; the latter improves the collection efficiency of the multiplier particularly for electrons emitted from the ends of the specimen. The intensity of the electron emission can be displayed on a pen recorder, or used to control the beam intensity of an oscilloscope. Since the beam of the oscilloscope is synchronized with the scanning of the laser beam, a photoelectron emission image of the specimen is formed on the oscilloscope screen.

Immediately after passing through the focusing lens (not shown in Fig.5), the now gradually converging laser beam is deflected by the mirrors of an X-Y scanner, and enters the vacuum chamber through a quartz window. There are two modes of scanning. In mode 1 the laser beam is scanned on an X-Y raster while the specimen is stationary, so that a side view is displayed on the oscilloscope. This mode of scanning is suitable for flat surfaces; the example in Fig. 6 shows the emission map of a small specimen of 1018 steel at 8% of its fatigue life. The emission is much stronger from the central portion of the gauge section where the fatigue deformation has accumulated. The intensity of this emission was in good agreement with the earlier results shown in Fig. 4.

In mode 1 it is difficult to collect electrons emitted from the lower portion of the cylindrical specimen indicated in Fig. 5. However in mode 2 the electron emission can be measured and displayed from the entire gauge surface. By observing the beam position on the oscilloscope, the Y position is adjusted and held fixed such that the X scan passes along the centerline of the rod. While scanning the laser beam continuously back and forth along this centerline, the rod is slowly rotated about its axis. The angular position of the rod is monitored by an angular transducer which now controls the Y axis of the oscilloscope. In this way an electron emission map of the entire surface is unfolded onto the oscilloscope screen. The example in Fig. 7 shows the distribution of electron emission from the central portion of a martensitic steel rod after 100,000 cycles of torsional loading. This sample was at 12.8% of its fatigue life and the highly localized nature of the fatigue damage is revealed by the isolated spots of bright emission. This example also illustrates the purpose of the oscilloscope display: it identifies the major regions of intense emission (exoelectrons) where the fatigue damage is most severe. Subsequent quantitative readout on the pen-recorder can now be restricted to these regions of primary interest.

The results from four rods of martensitic steel, fatigued in torsion, are summarized in Fig. 8. For each specimen, only the largest emission peak is plotted regardless of its location. As before (i.e. Fig. 4), the intensities of the peaks are normalized with respect to the initial background emission and plotted as a function of the fraction of fatigue life. The scatter band is quite small, particularly during the first 10% of fatigue life. For comparison the dashed lines represent the scatter band for the earlier results on 1018 steel shown in Fig. 4. The agreement between the two sets of experiments is quite good despite a threefold increase of material strength and a different mode of fatigue loading. In each case the development of exoelectron emission not only reveals the presence of fatigue deformation in less than 1% of the fatigue life, but more importantly provides a quantitative measure of the severity of the damage which can then be expressed in terms of the fraction of life expended, or the useful remaining life.

SUMMARY AND PROSPECTS

The above results illustrate that the laser scanning system can be successfully applied to the early detection and assessment of fatigue damage in small engineering components as well as very small laboratory samples. In this initial demonstration the limitation on specimen size was imposed by the robustness of the readily available vacuum equipment, but this is not a major obstacle to testing of larger structures. The laser beam may be scanned over much larger distances e.g. 50 cm, by scanning with a larger spot of 40 μm diameter.[6] The other important factor is uniform electron collection efficiency, and larger components may require several electron detectors. However the one essential requirement is that the component under test must be mounted in a vacuum chamber and then the cyclic load applied. Otherwise the fresh metal surfaces extruded by the cyclic deformation will reoxidize and no longer be sources of exoelectrons.

The localized intensity of exoelectron emission is clearly quite a good measure of fatigue damage. This result emphasizes the importance of local damage in determining the fatigue life. Furthermore, despite the complexity of the fatigue process, the accumulation of surface deformation is very systematic when viewed on the appropriate microscale. The fundamental material requirements are that the metal under test must i) have a brittle surface oxide and ii) develop regions of microplasticity prior to crack formation in the metal itself. Thus it is not known whether the technique is appropriate for a very brittle metal, but all metals appropriate to fatigue applications should satisfy these requirements. Certainly the fact that the growth of the normalized emission from the martensitic steel torsion specimens is in close agreement with that from bending tests on 1018 steel, is a good indication of the universality of the technique.

Experience with the exoelectron method has so far been confined primarily to the low cycle regime of fatigue (i.e. $N_f < 10^6$ cycles). The applicability to high cycle fatigue should be investigated, because it is in this regime that the greatest benefit is to be derived from an abbreviated test. Of particular interest would be studies close to the fatigue endurance limit. Eventually the quantitative capability of the technique could also be profitably applied to studies of spectrum loading. In all cases the predicted life is of course that in the strictly inert environment of a vacuum. In practice, where the ubiquitous corrosion accelerates crack propagation, the consequent reduction of fatigue life would have to be taken into account.

One aspect requiring immediate investigation is the effect of oxide thickness. The very thin (~4 nm) natural oxide has been shown to rupture precisely along slip traces and thereby permit the details of the surface deformation to be monitored by the exoelectron emission. However it is not known to what extent thicker oxides will rupture in this manner and thereby serve the same function. Clearly a thick oxide scale with weak adhesion to the underlying metal is unlikely to be satisfactory, since it will undoubtedly hide the intricate details of the surface deformation. Also very thick oxides may fracture independently of the detailed distribution of deformation in the metal, in the manner of a stresscoat lacquer.[7] It is not known how this mode of oxide fracture will influence the results.

Finally it is worth noting that the above unique capabilities of this scanning exoelectron technique can be attributed primarily to two fundamental factors:

1) Fatigue damage is usually initiated at or very close to the surface, and electron emission is very sensitive to surface changes.
2) The scanning system can examine all regions of interest on a component or structure and pinpoint the very small region containing the all-important localized severe damage at the site of eventual crack formation.

In this regard, since the most intense source of exoelectron emission corresponds to the most severe accumulation of damage and provides an overall assessment of the "state of fatigue" of the specimen, it is entirely reasonable to regard the distribution of less intense emission as a display of the complex distribution of fatigue damage elsewhere in the sample or component under test. Thus the technique can not only assess the severity of the primary fatigue site, but can also locate and assess regions of secondary weakness. This capability would be particularly useful in evaluations of structural or engineering components, where there are often several locations of potential failure.

ACKNOWLEDGMENTS

The author is very grateful to his colleague, S. R. Rouze, for the photoelectron micrographs shown in Fig. 1.

REFERENCES

1. Barton, J.R. and Kusenberger, F. N., ASTM STP 495 (1971 p. 123.

2. Baxter, W. J., Met. Trans. 6A (1975) p. 749.

3. Baxter, W. J. and Rouze, S. R., J. Appl. Phys. 44 (1973) p. 4400.

4. Baxter, W. J. and Rouze, S. R., J. Appl. Phys. 46 (1975) p. 2429.

5. Baxter, W. J., J. Appl. Phys. 45 (1974) p. 4692.

6. Baxter, W. J., Submitted to J. Testing and Evaluation.

7. Grosskreutz, J. C. and McNeil, M. B., J. Appl. Phys. 40 (1969) p.355.

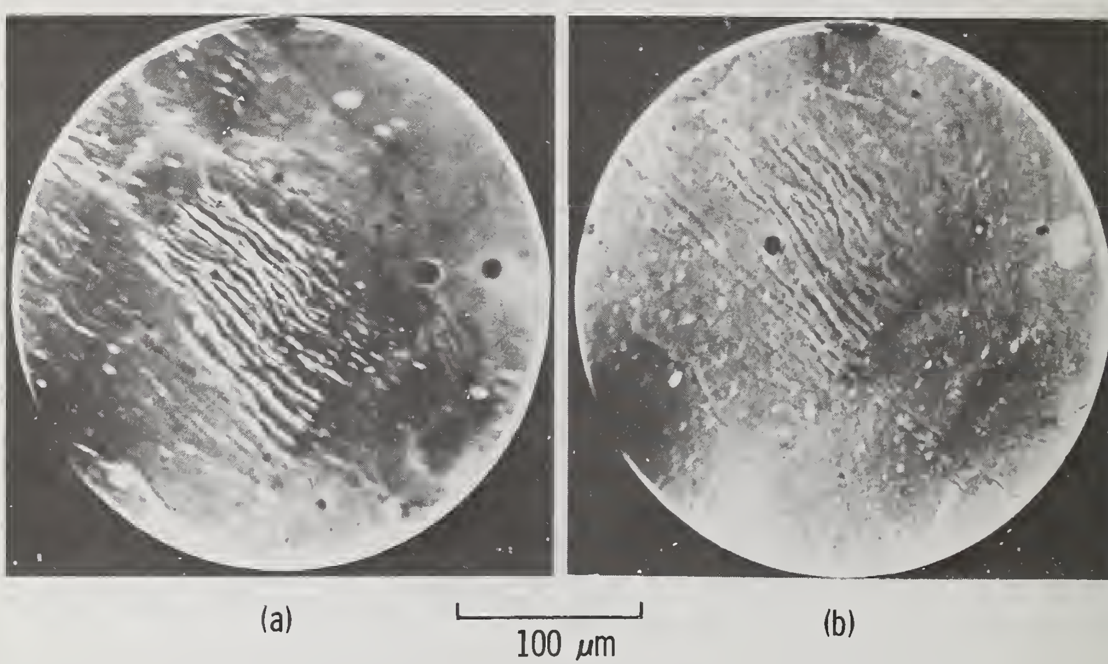


FIG. 1. Photoelectron micrographs of sample of 1018 steel
a) after 4% tensile strain in vacuum,
b) after exposure to room air for 60 hours.

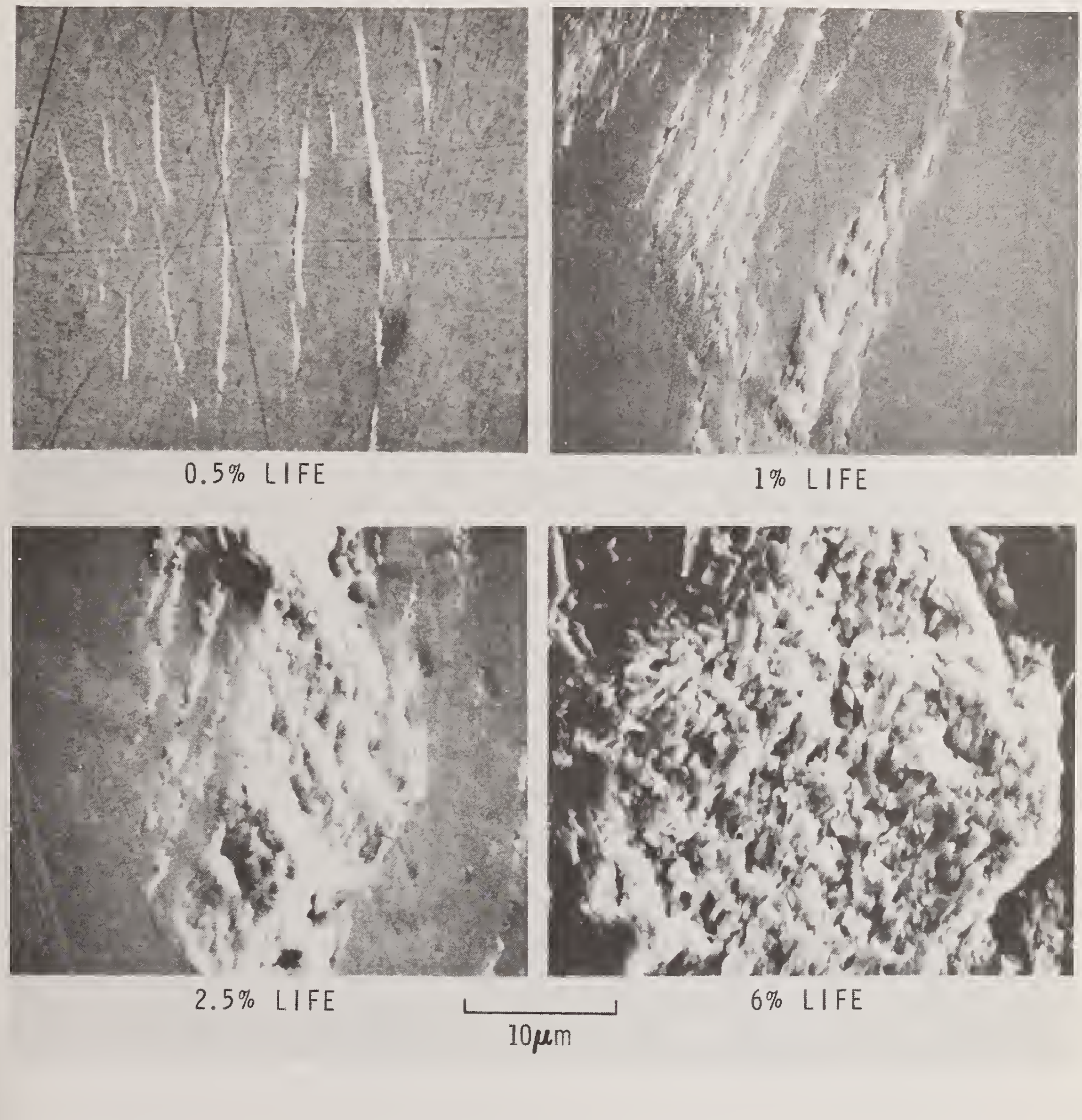


FIG. 2. Scanning electron micrographs of localized surface fatigue deformation during crack initiation in 1018 steel.

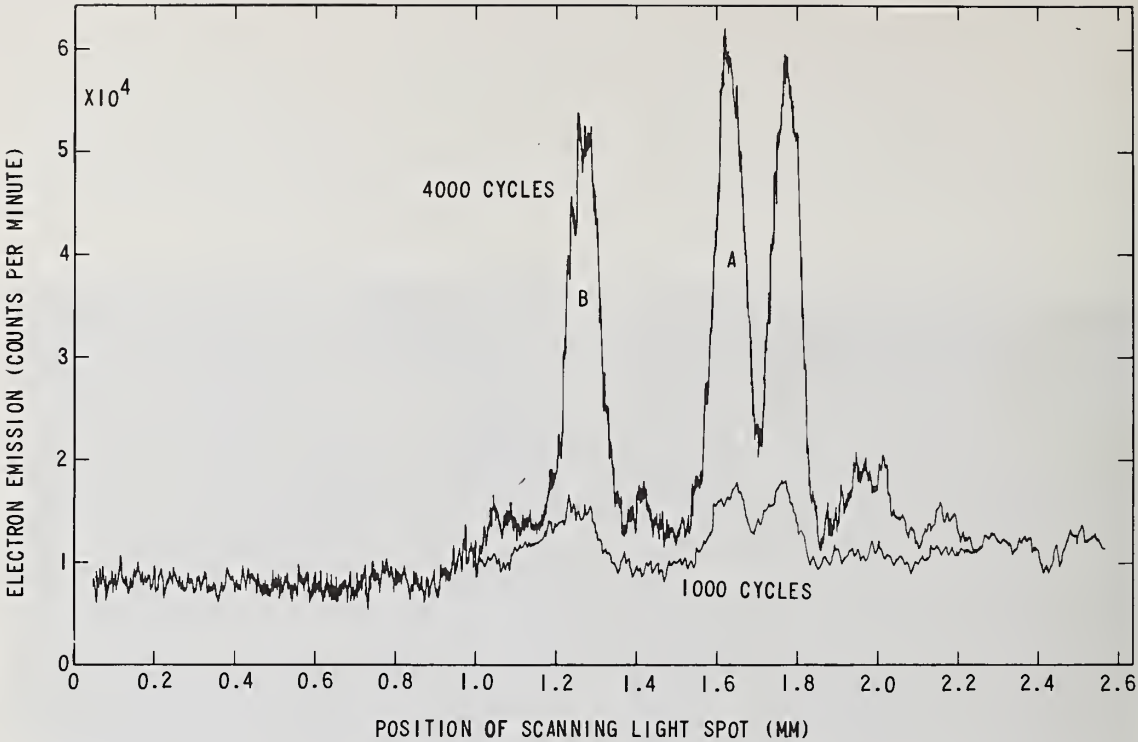


FIG. 3. Exoelectron emission from 1100 aluminum after 1000 and 4000 fatigue cycles. Fatigue life = 140,000 cycles.

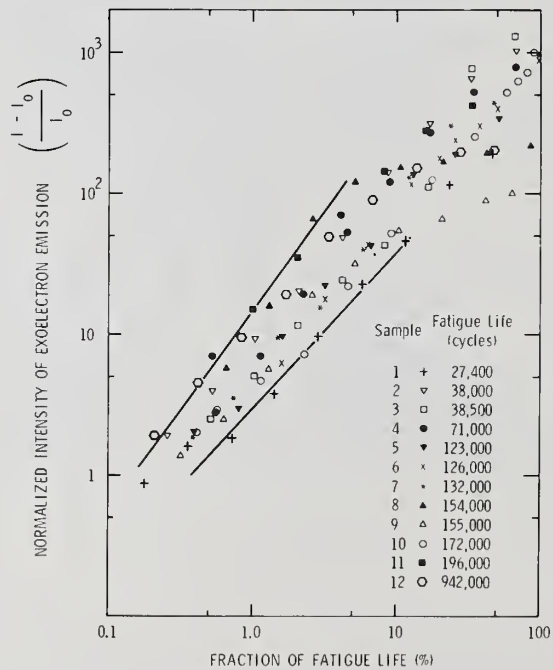


FIG. 4. Growth of the largest exoelectron peaks from twelve samples of SAE 1018 steel with a range of fatigue lives.

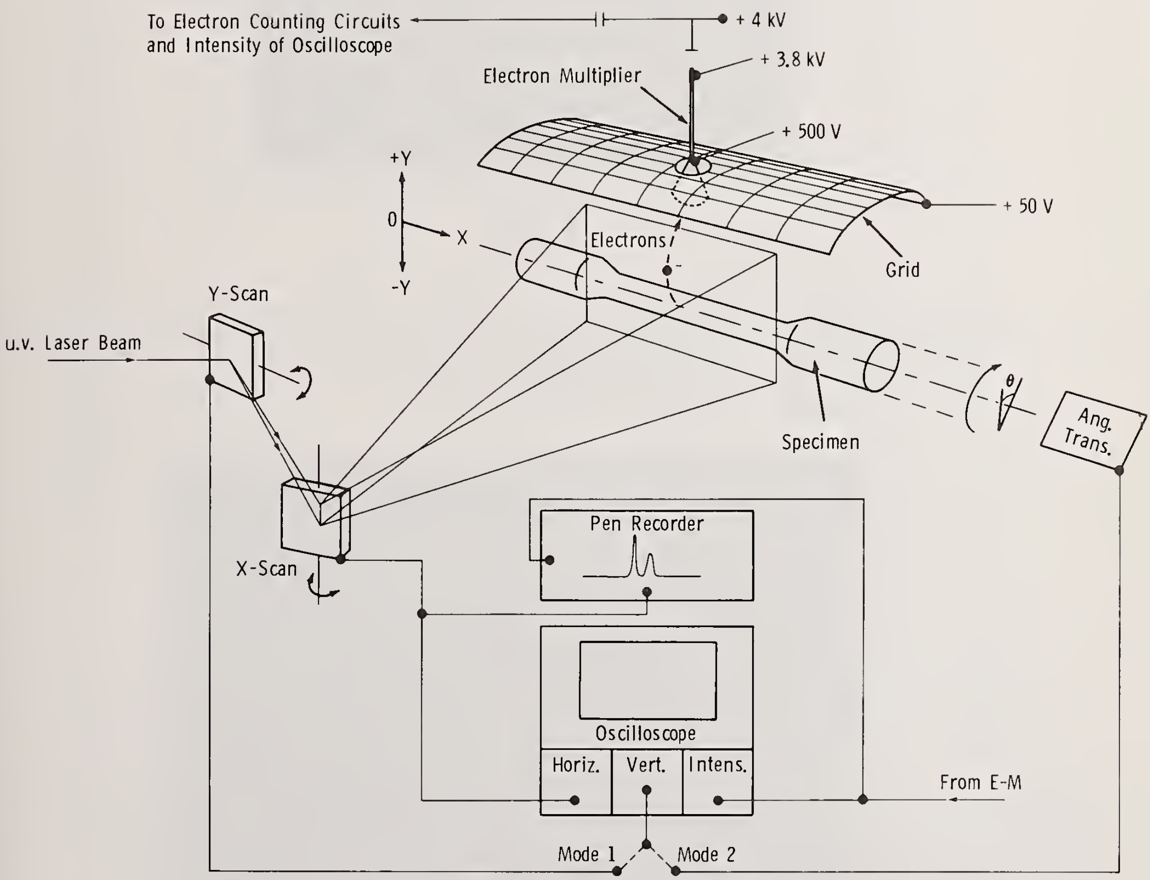


FIG. 5. Schematic diagram of the laser scanning and electron emission recording system.

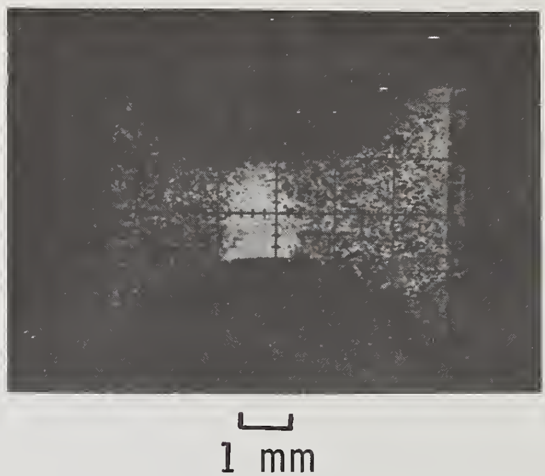


FIG. 6. Photoelectron emission image of gauge section of fatigue sample of 1018 steel at 8% of fatigue life.

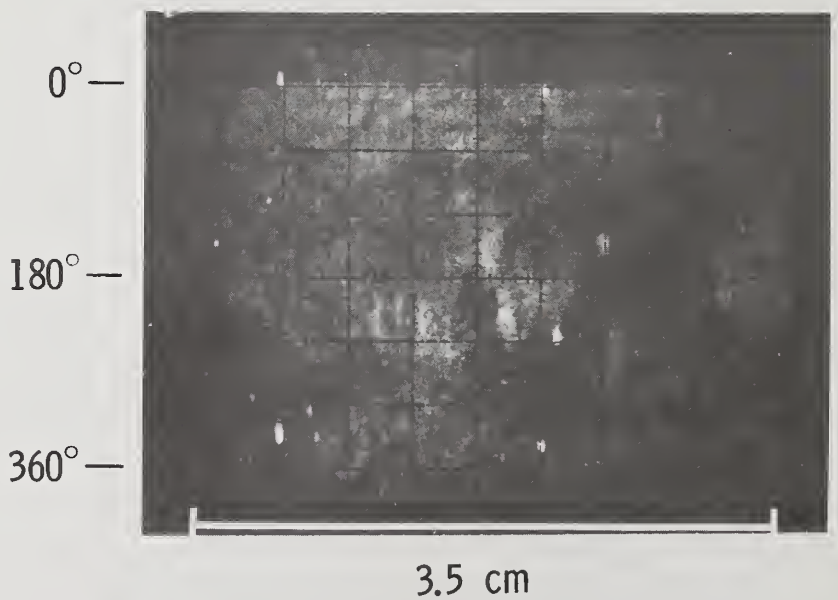


FIG. 7. Unfolded photoelectron emission map of the surface of the central portion of a rod of martensitic steel after 100,000 fatigue cycles (12.8% fatigue life).

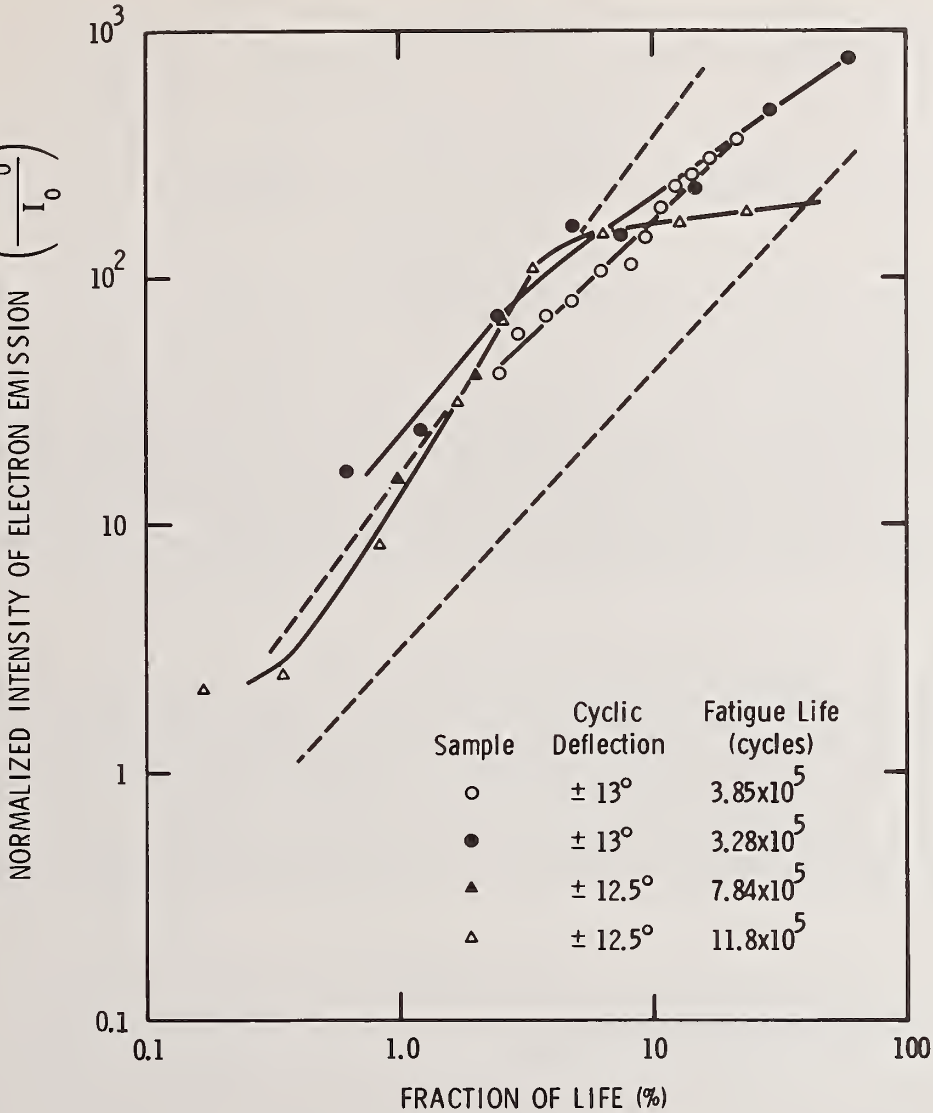


FIG. 8. Normalized intensity of largest exoelectron emission peak from four rods of martensitic steel as a function of the fraction of fatigue life. Dashed lines represent data for 1018 steel shown in Fig. 4.

DETECTING STRUCTURAL DEGRADATION BY ACOUSTIC EMISSION

Michael P. Kelly
Dunegan/Endevco
San Juan Capistrano, California 92675

Robert J. Schlamp
Dunegan/Endevco
San Juan Capistrano, California 92675

Abstract: The application of acoustic emission techniques to the detection and location of structural defects in a variety of materials and geometries is presented. Results obtained during the preservice inspection of the Baltimore Gas & Electric Company Calvert Cliffs Unit 2 Nuclear Power Plant will be discussed. Additional results obtained during proof testing of two petrochemical hydrotreaters will also be presented.

A multichannel computerized source location system was utilized to monitor petroleum drill pipe during proof loading. Several sections of new and used, corroded and sandblasted drill pipe were monitored. Initial results indicate an ability to detect, locate and identify fatigue cracks, coating cracks and corrosion during proof loading. Test results from high reliability petrochemical vessels demonstrate that incremental crack growth during pressurization can be accurately located and assessed.

Key words: Acoustic emission, crack growth, drill pipe, failure, Kaiser Effect, pressure vessels, sources.

INTRODUCTION

Acoustic emission (AE) is the term applied to the spontaneously generated elastic waves produced within a material under stress. Plastic deformation and the nucleation and growth of cracks are the primary sources of acoustic emission in metals. AE is unlike most other nondestructive test (NDT) methods in that the energy comes from the material itself. As a result, AE is more sensitive to growing defects and potentially less reliant on operator interpretation than conventional NDT techniques. Acoustic emission offers a further advantage in the inspection of complex structures which include inaccessible areas. Normally just a few stationary transducers can detect, locate and qualify defects over a very large area, including regions inaccessible by alternate methods.

The most common type of structural degradation in pressure vessels is in the form of subcritical crack growth caused by fatigue, stress corrosion or hydrogen embrittlement. The detection of subcritical flaw growth by acoustic emission can be accomplished in two ways. The first and most obvious is the continuous on-line surveillance of a structure throughout its operating life. This approach is expensive and the transducers and instrumentation are often subjected to hostile environments and/or excessive background noise.

Structural degradation due to subcritical crack growth can also be detected using acoustic emission in conjunction with periodic proof testing. This approach takes advantage of the irreversibility of acoustic emission, which is known as the Kaiser Effect (1). If a pressure vessel, for example, is pressurized to a proof pressure and then operated for a period of time at a lower working pressure, no acoustic emission will occur during subsequent pressurization to the same proof pressure unless structural degradation (crack extension) has occurred in the meantime. If crack propagation occurs in a vessel during service, acoustic emission will be observed above the working pressure during the next overpressurization.

Since proof tests are usually performed during periods of low background noise, such as when a vessel is temporarily out of service, and in a fairly benign environment, the difficulties encountered in continuous on-line surveillance are normally nonexistent.

This report presents the acoustic emission results obtained from a variety of structures, materials, and geometries monitored during proof testing.

INSTRUMENTATION

The portable, computerized multichannel source location system utilized in this work is shown in Figure 1. A block diagram of the system is presented in Figure 2. The system accepts inputs from up to 32 transducers arranged in 4-transducer arrays for planar location and 2-transducer arrays for linear location. The system display oscilloscope (CRT) provides a real-time display of the location of each source in relation to the position of the transducers (array) involved. In addition, a histogram or bargraph showing the overall activity of each active array is constantly updated and displayed on the CRT along with the currently displayed array. Figure 3 represents a hardcopy reproduction of a typical CRT display including the array transducers, active sources, and the histogram of events per array. In addition to the real-time display the important information associated with each event located by the system is printed out by a high speed teleprinter. A typical printout is shown in Figure 4.

EXPERIMENTAL PROCEDURE AND RESULTS

Drill Pipe

The detection of drill pipe defects is of considerable importance to the oil production industry since downhole failures often cause substantial losses. In an effort to improve drill pipe inspection techniques Continental Oil Company sponsored a program to assess the applicability of AE techniques to the detection and location of anomalies in this important class of structures. The method employed in carrying out the program is generally described in Continental Oil Company's U.S. Patent No. 3,911,734. Thirty sections of 114 mm (4.5 in) O.D. and 140 mm (5.5 in) O.D. drill pipe were monitored for acoustic emission during proof loading in two separate studies. In each case, the drill pipe was capped at each end and loaded in tension to the proportional limit of the pipe. Several grades of coated and uncoated pipe were evaluated to obtain a representative sample of various pipe conditions. The tool joint-test machine interface generated considerable extraneous frictional noise during the loading, so special lockout transducers were mounted in each tool joint header to spatially filter out these emissions. Each pipe was instrumented with a total of six 450 kHz resonant frequency transducers; one in each tool joint header to guard against extraneous emission and four more in two 2-channel arrays for linear location along the length of the pipe. Each pipe monitored for acoustic emission was later inspected by an alternate NDT technique for correlation purposes.

Figure 5 shows the transducer locations for a typical drill pipe. The acoustic emission activity expressed in cumulative events as a function of location was plotted for each drill pipe tested. Figure 6 represents actual test data received from sections of four pipes monitored during the initial drill pipe study.

Pipe #11 (Figure 6a) contained a fatigue crack at 175 cm (70 in) in addition to moderate pitting throughout the pipe. The acoustic emission record shows a peak at 179 cm (70.5 in) and moderate activity in several areas demonstrating good correlation between the two inspection techniques. The lack of emission in Figure 6b is typical of a quality pipe without defects and agrees with the post-test inspection which reported no anomalies. Pipe #23 was found to contain several cracks including eight in the 220 cm (87 in) to 241 cm (95 in) region. The AE results in Figure 6c show good correlation in the 200 cm (79 in) to 250 cm (98 in) region and another peak at 102 cm (40). The peak at 102 cm can probably be attributed to another crack or cluster of cracks. Pipe #21 was reported to contain internal pitting and coating cracks at 287 cm (113 in) and 330 cm (130 in). The acoustic emission results in Figure 6d confirm the inspection report.

The results of the first study indicated that while AE was a viable technique capable of detecting drill pipe defects it was difficult to differentiate between internal coating cracks and fatigue cracks. Coating cracks were often interpreted as the more important fatigue cracks and vice versa. In an effort to determine differences in the AE "signature" obtained from these anomalies, a laboratory investigation of CONOCO supplied specimens of coated and fatigue pre-cracked uncoated drill pipe material was initiated. This study revealed that coating crack propagation actually generated higher amplitude signals than fatigue crack extension. A second group of drill pipes was monitored with acoustic emission during proof loading to substantiate the results obtained during the laboratory investigation. Figure 7 represents the acoustic emission activity expressed in events and counts (event size) as a function of location for a 114 mm (4.5 in) O.D. drill pipe monitored during the second study. The post-test inspection reported a fatigue crack at 419 cm (165 in) and coating scratches or cracks in the 355 cm (140 in) to 380 cm (150 in) region. While the coating cracks generated fewer events than the fatigue crack (Figure 7b) there were more counts associated with the coating anomalies (Figure 7a) This confirms the results of the laboratory investigation and indicates that the fatigue and coating crack phenomena can be identified on the basis of magnitude (counts).

High Reliability Petrochemical Pressure Vessels

Acoustic emission source location techniques were applied to the detection of defects in high reliability petrochemical pressure vessels. The purpose of the program was twofold:

1) to prevent the unintentional failure of units during proof testing
2) to enhance the confidence in the structural integrity of units surviving the proof test.

Under American Petroleum Institute (API) regulations, each of these vessels is subjected to at least one proof test to twice the working pressure. Several vessels were instrumented with two arrays of four 140 kHz resonant frequency transducers and monitored during proof testing. Figure 8 is a hardcopy CRT reproduction of the actual sources received from a defect-free vessel.

An artificial defect in the form of a sharp notch was instrumented in the high stress region of one structure to assess the locational accuracy of the acoustic emission instrumentation. Figure 9 presents the data obtained in the hydrostatic proof test of that vessel. The nearly horizontal pattern above the center transducer of the array represents the emission obtained from the artificial defect. The last 50 events received during the proof test are shown in Figure 10. Nearly all of these events originated in the region of the artificial defect.

A third vessel monitored during proof test contained a natural 7.6 cm (3 in) long crack located midway through the 10 cm (4 in) thickness.

Twenty-seven events located in an area 4 cm (1.6 in) X 5.3 cm (2 in) were received from the crack during the test. Figures 11 and 12 present the source activity recorded during the proof test of this vessel. It is interesting to note that although this structure survived a proof cycle to twice the working pressure, it contained a growing subcritical crack. Crack extension actually started well below the vessel working pressure.

Petroleum Hydrotreaters

Two Gulf Oil of Canada petroleum hydrotreaters, manufactured by Japan Steel Works, were monitored with acoustic emission during initial hydrostatic proof testing. The vessels were fabricated from ASTM A-336-F22 steel, normalized and tempered; and included a 6.35 mm (.25 in) stainless steel weld overlay on the I.D. surface. The two cylindrical hydrotreaters, designated R-201 and R-401, measured 37.9 m (124 ft) and 29.4 m (96.5 ft) in length, respectively and were 22 cm (8.7 in) thick. The cylindrical sections were 33.5 m (110 ft) and 24.9 m (82 ft) in length. Each hydrotreater was instrumented with eight arrays of four transducers each, a total of 32 channels. The 140 kHz transducers were attached to the structures with magnetic hold-downs. A thin layer of silicone grease was placed on the surface of each transducer to help couple the stress waves into the transducers.

In accordance with the American Society of Mechanical Engineers (ASME), "Proposed Standard for Acoustic Emission Examination During Application of Pressure" (2), the source location system was calibrated as a complete system prior to and immediately after the inspection of each hydrotreater. Total system operation was verified through the use of pulsing transducers (1 per array). In addition, all acoustically active sources recorded during the tests were graded, using analytical techniques such as event rate, count rate, and source activity as a function of location and/or pressure. The classification of sources by grade requires the following actions:

- Grade A - If it occurs during pressure build-up, pressurization should be halted. Confirm results as relevant by other nondestructive examination methods.
- Grade B - Recordable and reportable for future comparison.
- Grade C - Further evaluation or correlation not required.

Figures 13 and 14 represent graphic displays of all active sources recorded during the AE examination of R-201 and R-401. A tabulation of these sources including grade classification is presented in Table 1.

During pressurization of R-201, several events were received from a relatively small area within the monitoring area of Array 4 (see Figure 15). The events all occurred in the region of a steel measuring tape which was draped circumferentially around the structure. The tape was removed midway through the test and emission from the suspect region ceased. This indicated that the emission was frictional in nature and of no concern. A post-test ultrasonic inspection revealed no defects in the active area, confirming the tape measure suspicion.

Gulf Oil of Canada will periodically requalify these hydrotreaters during their useful life through similar proof tests. The AE results obtained from this initial preservice inspection represent a valuable baseline for future monitoring of these structures. Any crack initiation or propagation which might occur during the operating life of these vessels, will be very evident during subsequent proof tests.

Calvert Cliffs Unit 2 Nuclear Power Plant

Sixty-four channels of source location instrumentation were utilized during an AE examination of the Baltimore Gas & Electric Company Calvert Cliffs Unit 2 nuclear reactor and primary coolant system (3). The primary coolant system consists of four reactor coolant pumps, two steam generators, one pressurizer and associated piping. The AE examination was performed in conjunction with the preservice cold hydrotest of the system, and in accordance with the ASME "Proposed Standard for Acoustic Emission Examination During Application of Pressure".

The 140 kHz transducers were attached to the carbon steel components using magnetic hold-downs and a nuclear approved couplant. An approved adhesive was used for attachment to the stainless steel members. Quad arrays (see Figure 2) were utilized in monitoring the reactor top head, the reactor vessel below the nozzles, the two steam generators, and the system pressurizer. The four reactor coolant pumps, the two steam generator hot legs, the four cold legs, and the pressurizer surge line were covered by multisensor linear arrays. The region of the reactor vessel between the nozzles and the upper flange was also monitored by four transducers in a single contiguous linear pattern.

The reactor vessel and primary coolant system were monitored during the entire pressure excursion to 220 kg/cm^2 (3125 PSI) except for hold periods. The movement of customer personnel visually inspecting for leaks prevented the acquisition of data during hold periods.

The AE examination of Calvert Cliffs Unit 2 revealed 35 insignificant sources. No significant sources were found to be present under the stress conditions imposed by the hydrotest. Table 2 presents the location instrumentation. Figures 16 and 17 represent graphic displays of all active sources received from the reactor vessel and top head.

A post test examination of the Calvert Cliffs Unit 2 primary system is complete in the areas which experienced Grade B and Grade C to B indications during the acoustic emission inspection. These examinations revealed no indications which exceeded ASME code standards. However, minor acceptable indications were found in the regions defined by acoustic emission, which attests to the sensitivity and locational accuracy of the acoustic emission examination method.

CONCLUSIONS

The four tests presented were for widely varying conditions and materials, yet the results showed that standard acoustic emission source location techniques provide a powerful tool for the verification of integrity of pressure vessels and other structures. The results reported here are typical of the results being obtained from bridges, nuclear and petrochemical pressure vessels, pipelines, rocket motor cases, composite materials and aircraft (4-14).

It is our belief that many additional applications will arise in the future when the technique becomes better recognized by the technical community.

REFERENCES

1. J. Kaiser, Untersuchungen uber das Auftreten Gerauschen Beim Zuguersuch; Ph.D. Thesis, Technishe Hochschule, Munich, 1950.

2. "Proposed Standard for Acoustic Emission Examination during Application of Pressure", available from the American Society of Mechanical Engineers, United Engineering Center, 345 East 47th Street, New York N. Y. 10017.

3. M. P. Kelly and R. J. Schlamp, "Acoustic Emission Preservice Inspection of Calvert Cliffs Unit 2 Nuclear Power Plant", Dunegan/Endevco, San Juan Capistrano, California.

4. D. O. Harris and H. L. Dunegan, "Verification of Structural Integrity of Pressure Vessels by Acoustic Emission and Periodic Proof Testing", STP-515, Testing for Prediction of Material Performance in Structures and Components, American Society for Testing and Materials, 1972, pp. 158-170.

5. A. T. Green and H. L. Dunegan, "Acoustic Emission Analysis of Crack Propagation and Fracture in Pressure Vessels and Pressure Vessel Materials", presented at the First International Conference on Structural Mechanics in Reactor Technology, Berlin, Germany, September 1971.

6. N. O. Cross, L. L. Loushin, and J. L. Thompson, "Acoustic Emission Testing of Pressure Vessels for Petroleum Refineries and Chemical Plants", Acoustic Emission, STP-505, American Society for Testing and Materials, 1972, pp. 270-296.

7. A. A. Pollock and B. Smith, "Stress-wave Emission Monitoring of a Military Bridge", Non-Destructive Testing, December 1972, pp. 348-353.

8. M. P. Kelly and R. L. Bell, "Detection and Location of Flaw Growth in the EBOR Nuclear Reactor Vessel", Dunegan/Endevco, San Juan Capistrano, California, DE-73-4.

9. R. Gopal, "Westinghouse Acoustic Emission Studies at EBOR", Final EEI-TVA Rp 79 Meeting, San Antonio, Texas, December 11, 1972.

10. D. L. Parry, "Acoustic Emission Analysis in Oil and Chemical Industry", Exxon Nuclear Company, Richland, Washington, April 1972.

11. M. P. Kelly, D. O. Harris, and A. A. Pollock, "Detection and Location of Flaw Growth in Metallic and Composite Structures", STP-571, Monitoring Structural Integrity by Acoustic Emission, American Society for Testing and Materials, 1975, pp. 221-240.

12. T. F. Drouillard, R. G. Liptai, and C. A. Tatro, "Industrial Use of Acoustic Emission for Nondestructive Testing", STP-571, Monitoring Structural Integrity by Acoustic Emission, American Society for Testing and Materials, 1975, pp. 122-149.

13. R. R. Corle, "The Use of Acoustic Emission to Improve Rocket Motor Case Structural Reliability", presented at AIAA/SAE 9th Propulsion Conference, Las Vegas, Nevada, November 1973, AIAA Paper 73/1258.

14. C. D. Bailey and W. M. Pless, "Acoustic Emission Used to Nondestructively Determine Crack Locations in Aircraft Structure Fatigue Specimens", 9th Symposium on Nondestructive Testing, San Antonio, Texas, April 1973.

TABLE 1

Hydrotreater R-201 Sources

Acoustic Source	Event Density	Counts	Discussion	Grade
S1	Low	High	Minor Nozzle Weld Discontinuity	C
S2	Medium	Medium	Minor Nozzle Weld Discontinuity	C
S3	Low	Medium	Minor Discontinuity	C
S4	High	Medium	Extraneous-Frictional	B
S5	Medium	Low	Minor Discontinuity	C
S6	Low	Medium	Minor Discontinuity	C
S7	Low	Medium	Minor Discontinuity	C
S8	Medium	Low	Minor Discontinuity	C
S9	High	Medium	Support Drag	C to B
S10	Medium	Low	Support Drag	C
S11	Medium	Low	Minor Discontinuity	C
S12	Medium	Medium	Minor Weld Discontinuity	C

Hydrotreater R-401 Sources

Acoustic Source	Event Density	Counts	Discussion	Grade
S1	Low	Medium	Minor Nozzle Weld Discontinuity	C
S2	Medium	Low	Minor Nozzle Weld Discontinuity	C
S3	Low	Medium	Minor Nozzle Weld Discontinuity	C
S4	Low	High	Minor Discontinuity or Support Drag	C
S5	Low	Medium	Support Drag	C
S6	Medium	Low	Minor Discontinuity or Support Drag	C
S7	Medium	Low	Minor Discontinuity	C
S8	Medium	Low	Minor Discontinuity	C
S9	Medium	Low	Minor Discontinuity	C
S10	Medium	Medium	Support Drag	C

TABLE 2

Calvert Cliffs Unit 2 Acoustically Active Sources

Acoustic Source	Event Density	Counts	Location	Discussion	Grade
S1	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S2	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S3	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S4	Low	Medium	Reactor	Minor Nozzle Discontinuity	C
S5	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S6	Low	Medium	Reactor	Minor Nozzle Discontinuity	C to B
S7	Low	Medium	Reactor	Minor Nozzle Discontinuity	C to B
S8	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S9	Low	High	Reactor	Nozzle Discontinuity	B
S10	Low	Low	Reactor	Minor Nozzle Discontinuity	C
S11	Low	Low	Reactor	Insulation Drag or Minor Weld Discontinuity	C
S12	Low	Medium	Reactor	Minor Weld Discontinuity	C to B
S13	Low	Low	Reactor	Minor Weld Discontinuity	C
S14	Low	Medium	Reactor Top Head	Insulation Drag or Minor Discontinuity	C to B
S15	Low	Medium	Reactor Top Head	Insulation Drag or Minor Discontinuity	C to B
S16	Low	Low	Pressurizer	Minor Discontinuity	C
S17	Low	Low	Pressurizer	Minor Discontinuity	C
S18	Low	Low	Pressurizer	Minor Discontinuity	C
S19	Low	Low	Pressurizer	Minor Discontinuity	C
S20	Low	Low	Pressurizer	Minor Discontinuity	C
S21	Low	Low	Piping	Minor Weld Discontinuity	C
S22	Low	Low	Piping	Minor Weld Discontinuity	C
S23	Low	Low	Pump 21A	Minor Discontinuity	C
S24	Low	Low	Piping	Minor Weld Discontinuity	C
S25	Medium	Medium	Piping	Minor Weld Discontinuity	C to B
S26	Low	Low	Piping	Minor Weld Discontinuity	C
S27	Low	Low	Steam Generator 21	Minor Discontinuity in Vicinity of Weld	C
S28	Medium	Low	Steam Generator 21	Minor Discontinuity in Vicinity of Weld	C
S29	High	Medium	Steam Generator 21	Discontinuity in Vicinity of Weld	B
S30	High	Medium	Steam Generator 21	Discontinuity in Vicinity of Weld	B
S31	Low	Low	Steam Generator 22	Minor Discontinuity in Vicinity of Weld	C
S32	Medium	Medium	Steam Generator 22	Minor Discontinuity in Vicinity of Weld	C to B
S33	Medium	Low	Steam Generator 22	Minor Discontinuity in Vicinity of Weld	C
S34	Medium	Medium	Steam Generator 22	Discontinuity in Vicinity of Weld	B
S35	High	Medium	Steam Generator 22	Discontinuity in Vicinity of Weld	B

FIGURE 1. Portable Source Location System

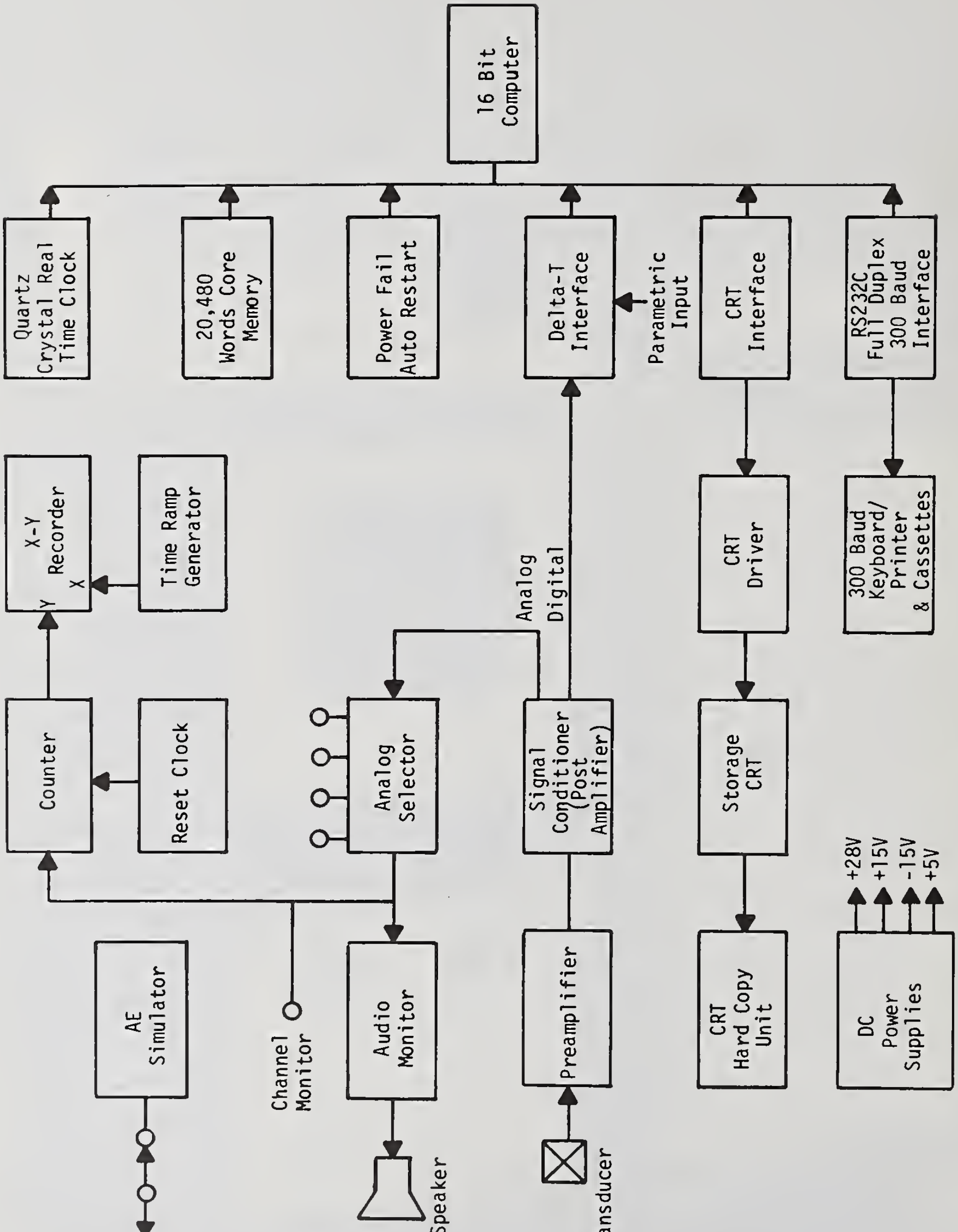


FIGURE 2. Block Diagram of Computerized Source Location System

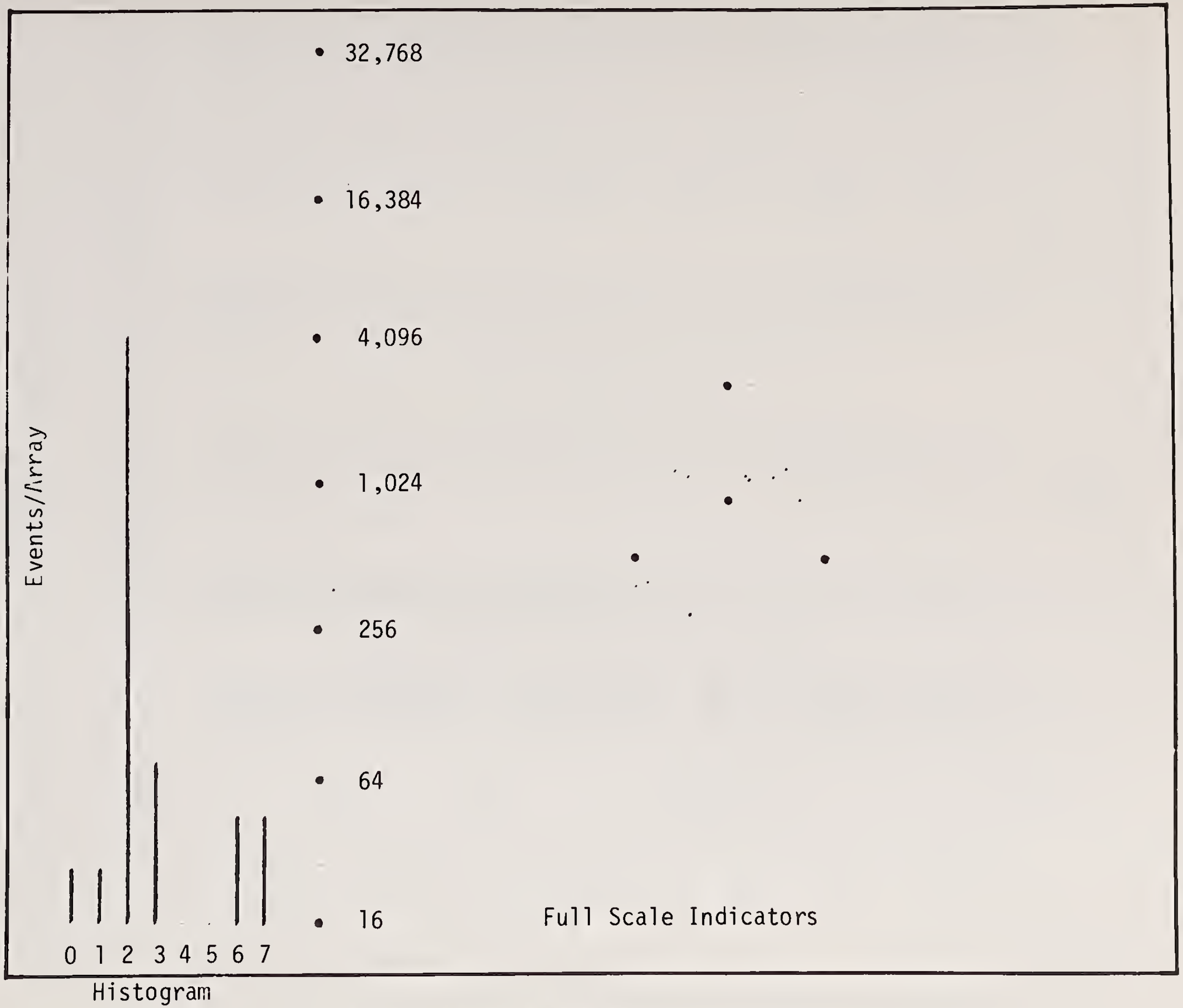
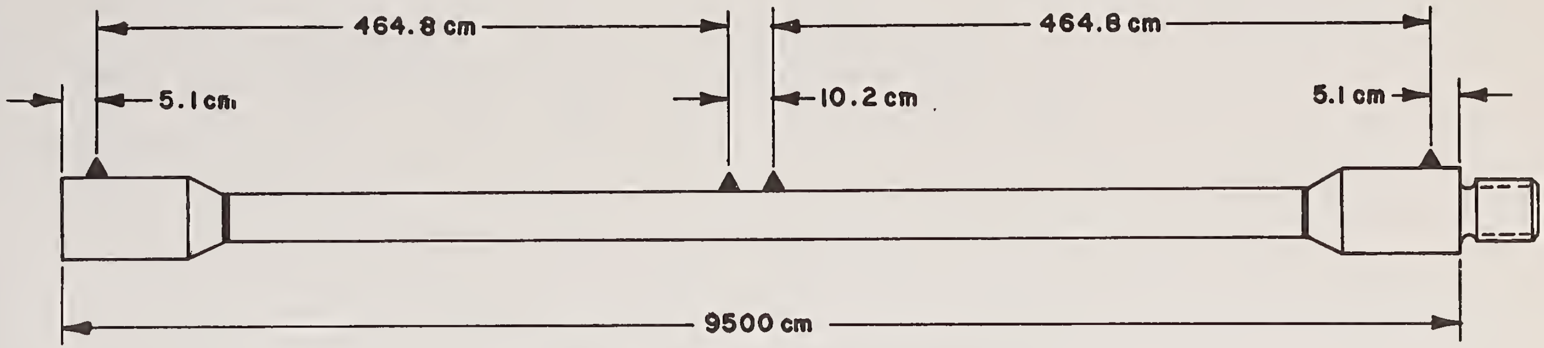


FIGURE 3. Typical CRT Array Presentation

DATA RUN 1

EVENT	ARY	DT0	DT1	DT2	DT3	X	Y	COUNTS	TIME	PARAM
31	0	296	495	1602	0	-11.9	10.6	16	304	1186
35	1	1012	0	1816	2172	-9.1	-9.6	24	321	1214
39	1	490	0	1629	527	-24.2	6.7	111	321	1216
44	1	1535	0	1561	2497	19.1	17.3	128	324	1221
58	0	0	0	3	0	-13.0	7.5	1	380	1256
60	1	1522	0	1563	2263	17.8	14.8	149	414	1279
61	1	561	0	1177	671	-80.0	-5.5	953	414	1279
66	1	729	0	1756	81	38.0	-19.5	12	419	1285
68	1	95	1160	0	1144	7.0	-3.9	41	424	1291
70	0	1300	2165	2227	0	-0.9	20.7	66	424	1291
73	1	118	1433	0	1707	5.5	-5.7	31	425	1293
74	0	0	163	870	1303	-3.9	-6.5	21	426	1294
78	1	545	2579	0	968	11.4	6.3	40	430	1298
79	0	0	584	1728	1116	-7.2	1.3	71	434	1296
84	0	279	0	1046	1490	-5.9	-8.3	6	446	1294
85	0	682	0	1860	1728	-8.9	-3.7	56	451	1297
88	1	490	0	1986	2030	-5.6	-3.6	25	455	1303
89	1	349	0	1222	1558	-6.6	-7.4	18	455	1303
91	0	58	0	1019	1055	-6.7	-4.2	111	455	1303
93	1	1269	2322	0	1906	26.2	-7.3	1	456	1305
94	0	0	872	56	1220	4.7	-6.2	30	457	1306
99	1	48	0	1505	1733	-5.6	-5.4	18	462	1313
102	0	208	0	1454	1377	-7.4	-3.5	48	463	1314
103	1	715	0	1447	1856	-8.8	-10.2	17	463	1315
120	0	0	313	376	982	-0.5	-6.9	387	466	1319
133	1	0	1287	553	1049	4.8	-0.6	100	468	1321
134	0	371	0	1924	90	-19.2	10.5	18	468	1322
141	0	847	0	2459	1396	-11.8	3.5	131	469	1323
144	1	1468	0	2325	2393	-21.4	-13.6	11	470	1324
146	1	200	0	1412	1551	-6.3	-5.0	29	471	1325
152	1	0	866	117	1299	4.1	-6.6	80	473	1328

FIGURE 4. Typical Source Location System Teleprinter Printout



NOTE: Lockout Transducers Not Shown

FIGURE 5. Transducer Locations for a Typical Drill Pipe

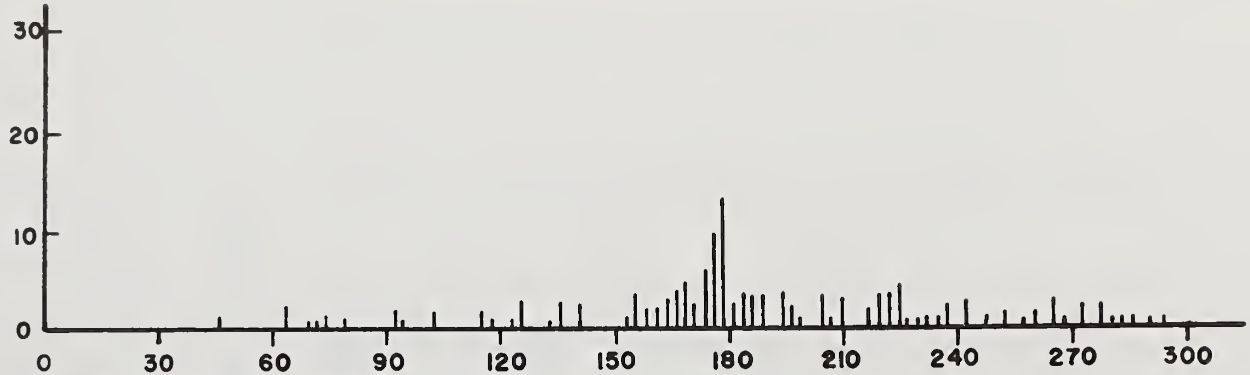


FIGURE 6a. Drill Pipe #11

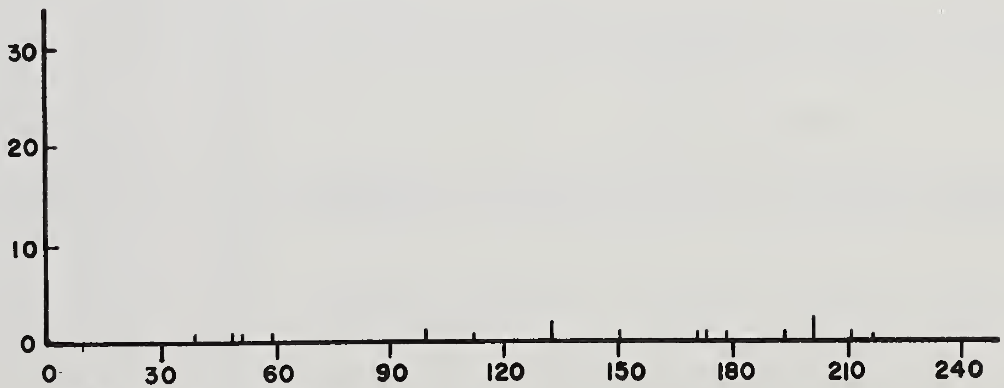


FIGURE 6b. Drill Pipe #13

FIGURE 6. (Continued on next page)

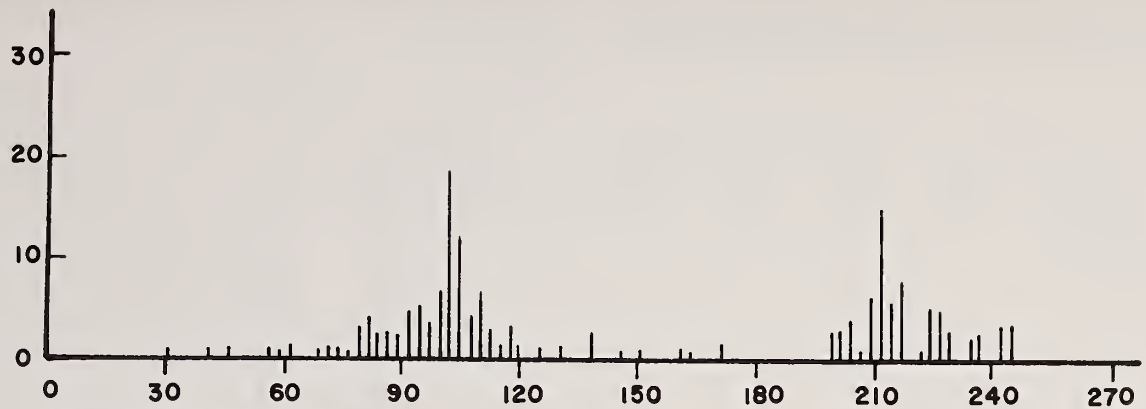


FIGURE 6c. Drill Pipe #23

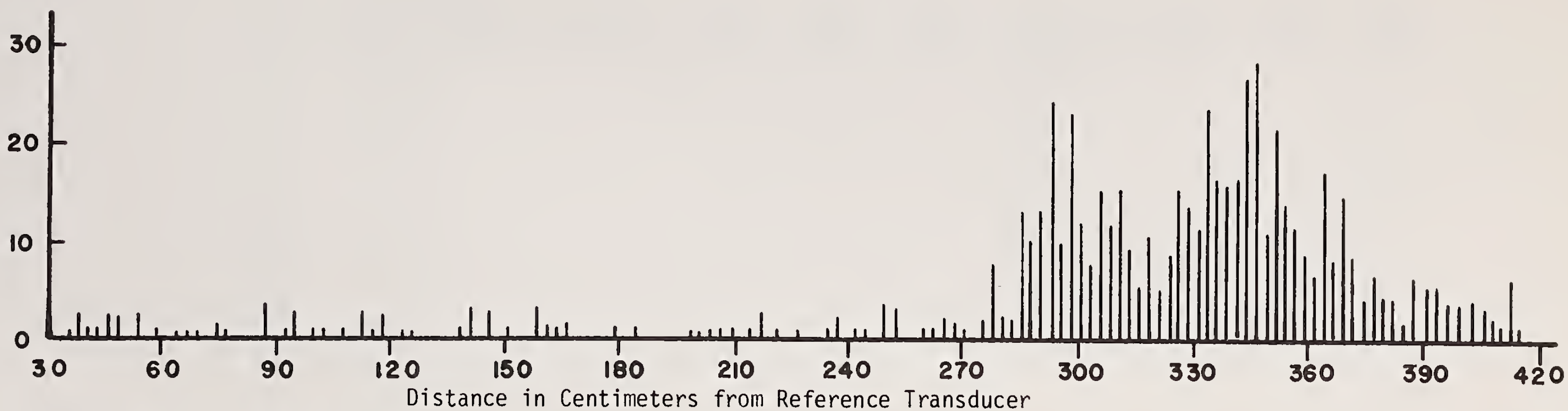


FIGURE 6d. Drill Pipe #21

FIGURE 6. Events as a Function of Location for Four Sections of 140 mm O.D. Drill Pipe

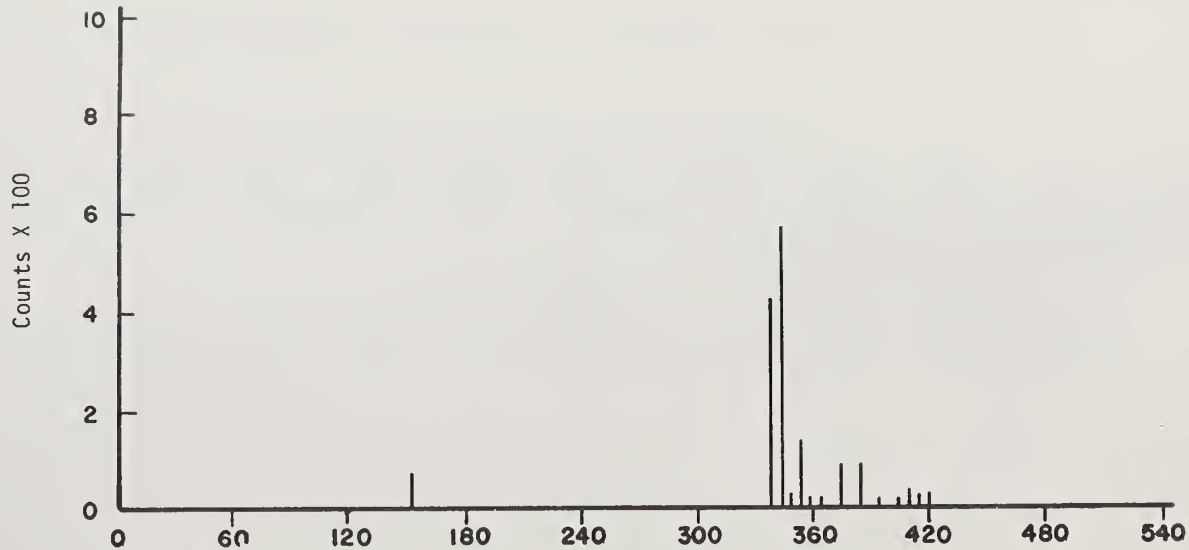


FIGURE 7a

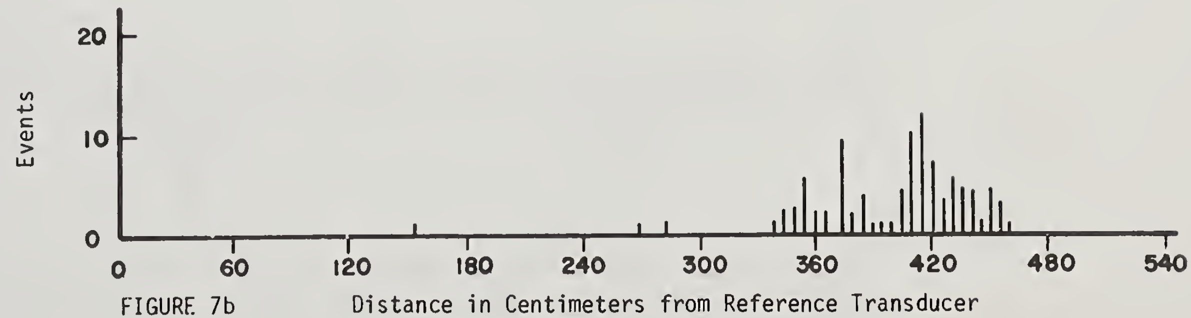


FIGURE 7b

FIGURE 7. Acoustic Emission Activity (Counts and Events) as a Function of Location for a 114 cm O.D. Drill Pipe

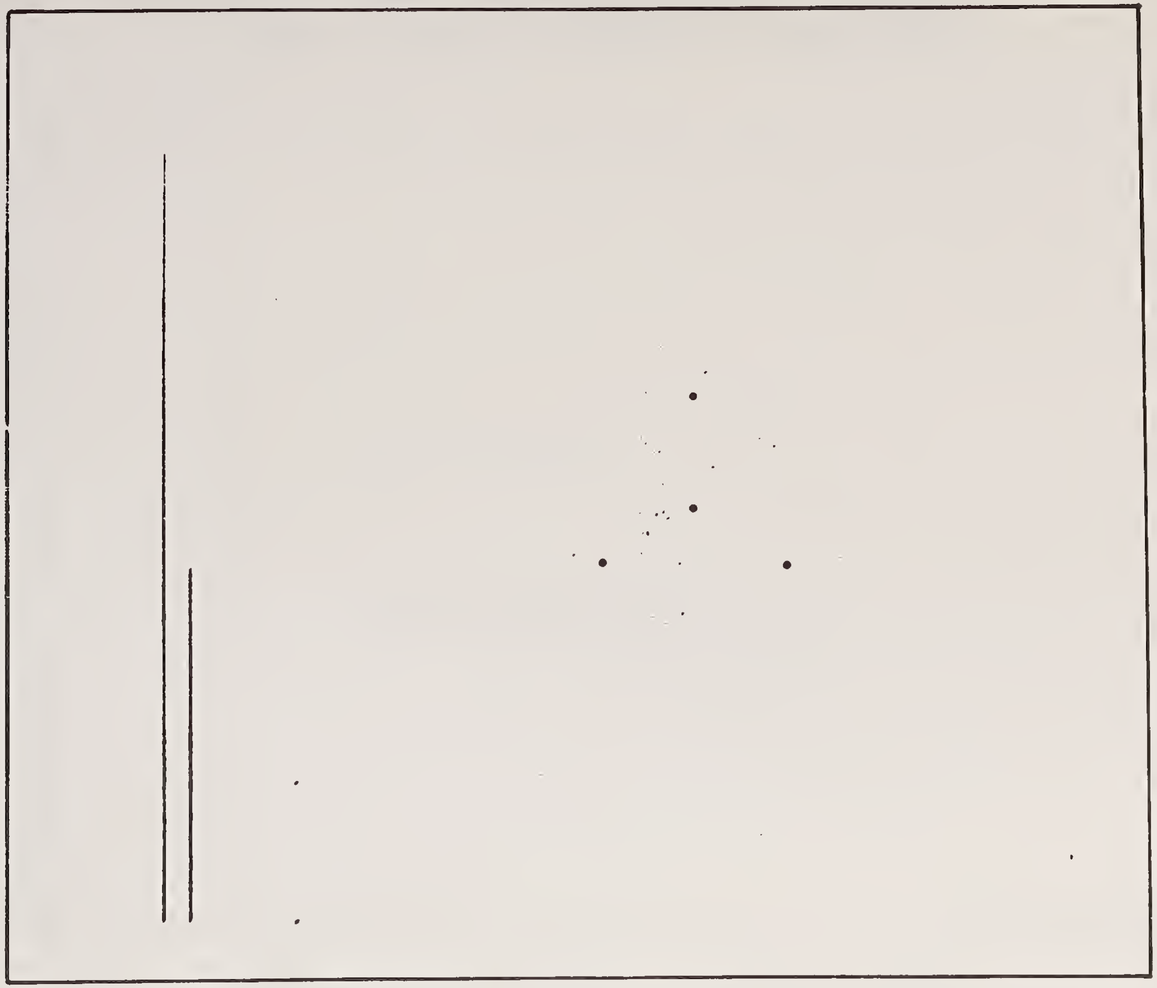


FIGURE 8. AE Sources Received from a Defect Free Petrochemical Vessel

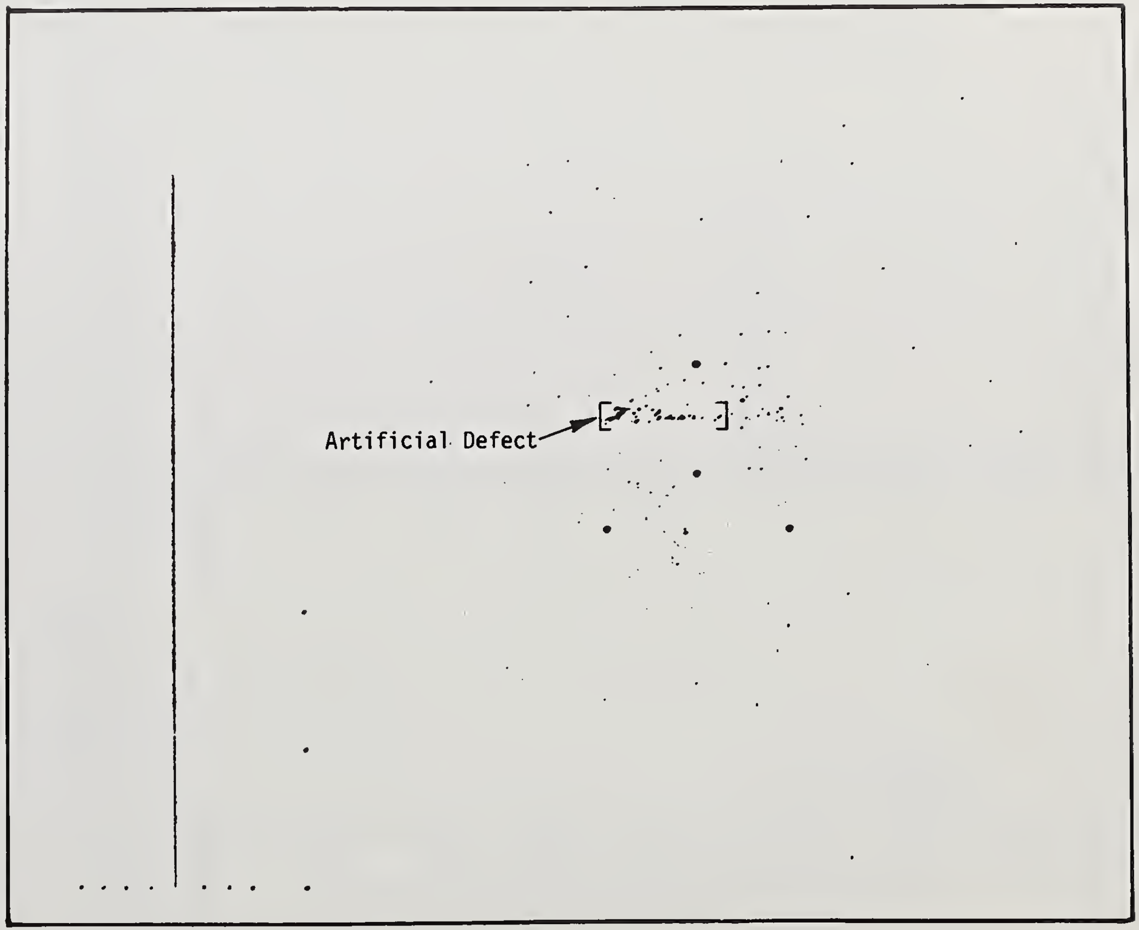


FIGURE 9. AE Sources Received from a Petrochemical Vessel Containing an Artificial Defect - Entire Test

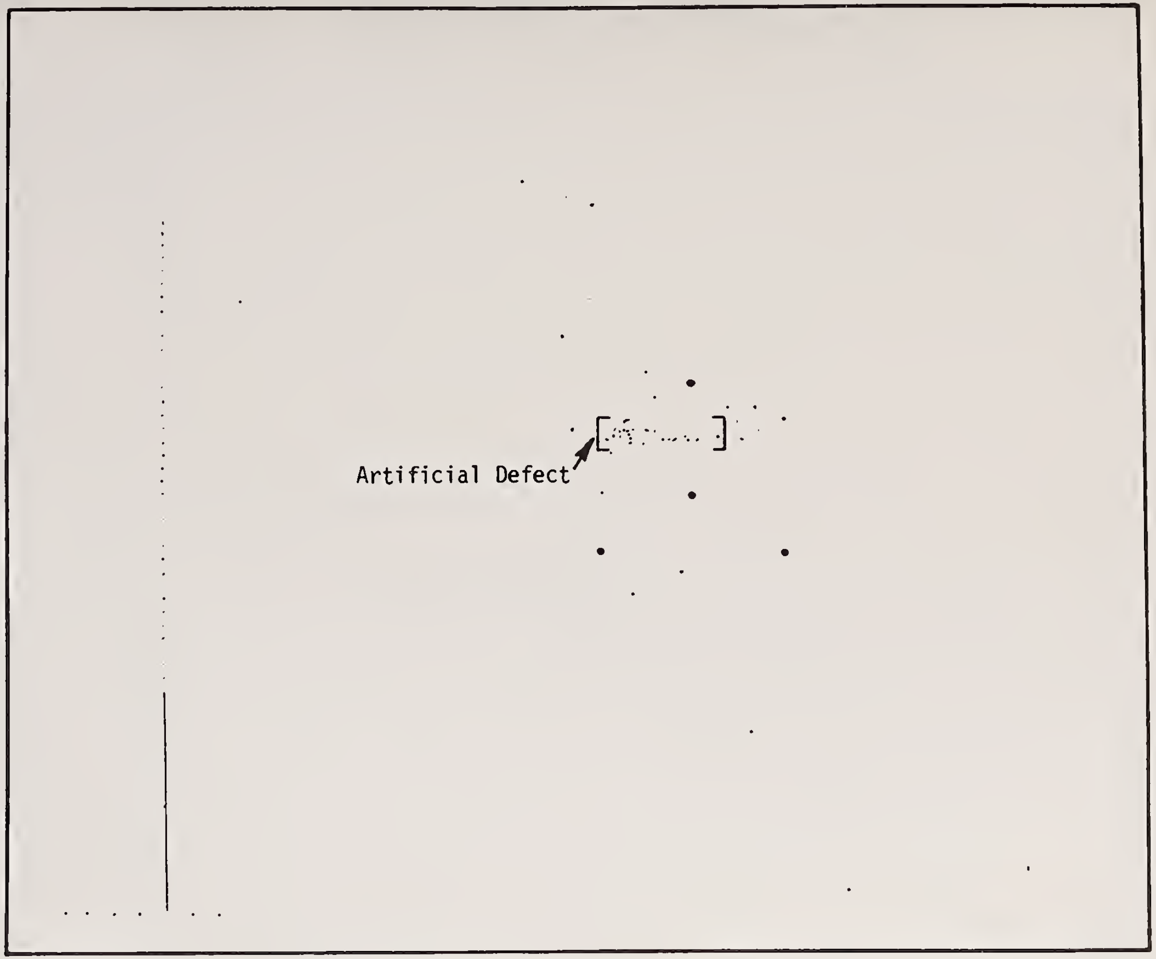


FIGURE 10. AE Sources Received from a Petrochemical Vessel Containing an Artificial Defect - Last 50 Events Received

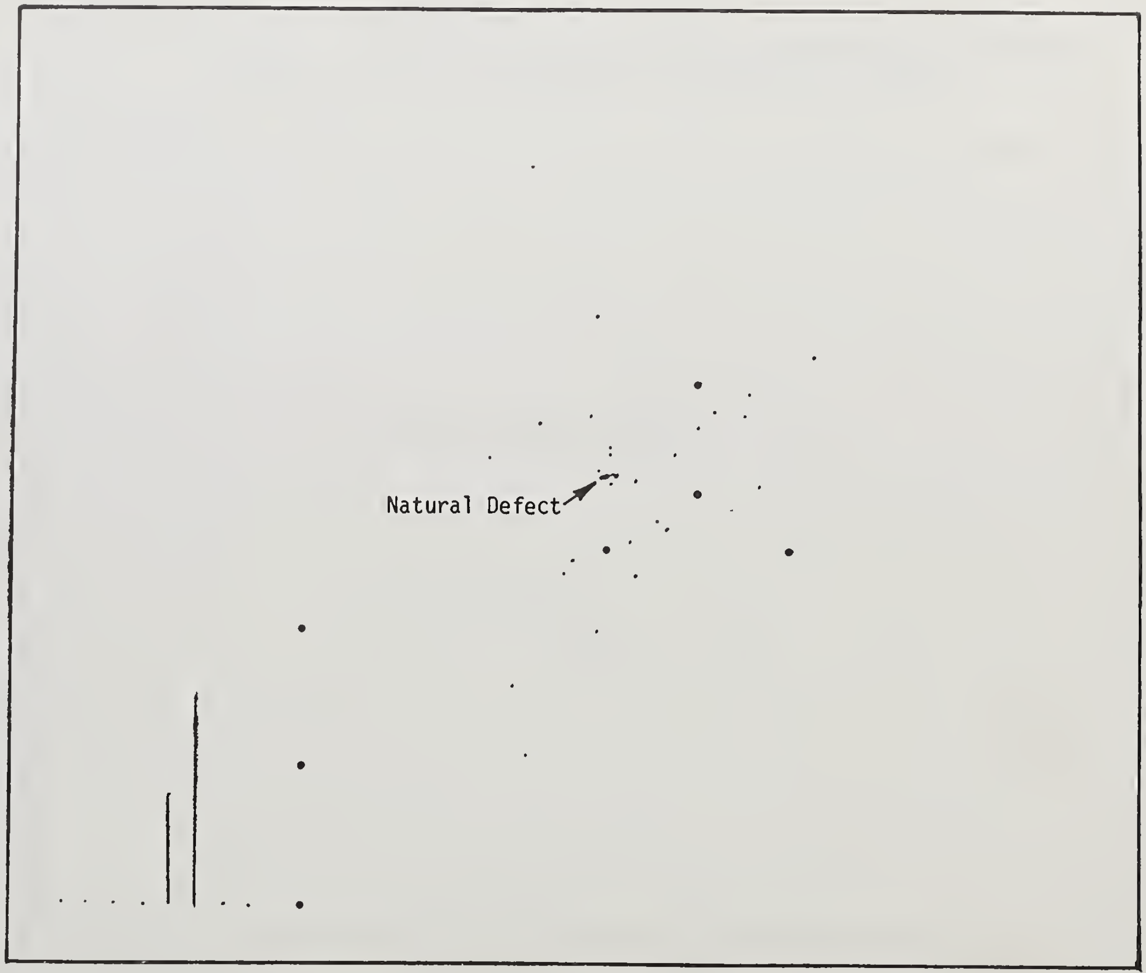


FIGURE 11. AE Sources Received from a Petrochemical Vessel Containing a 7.6 cm Crack - Early in Test

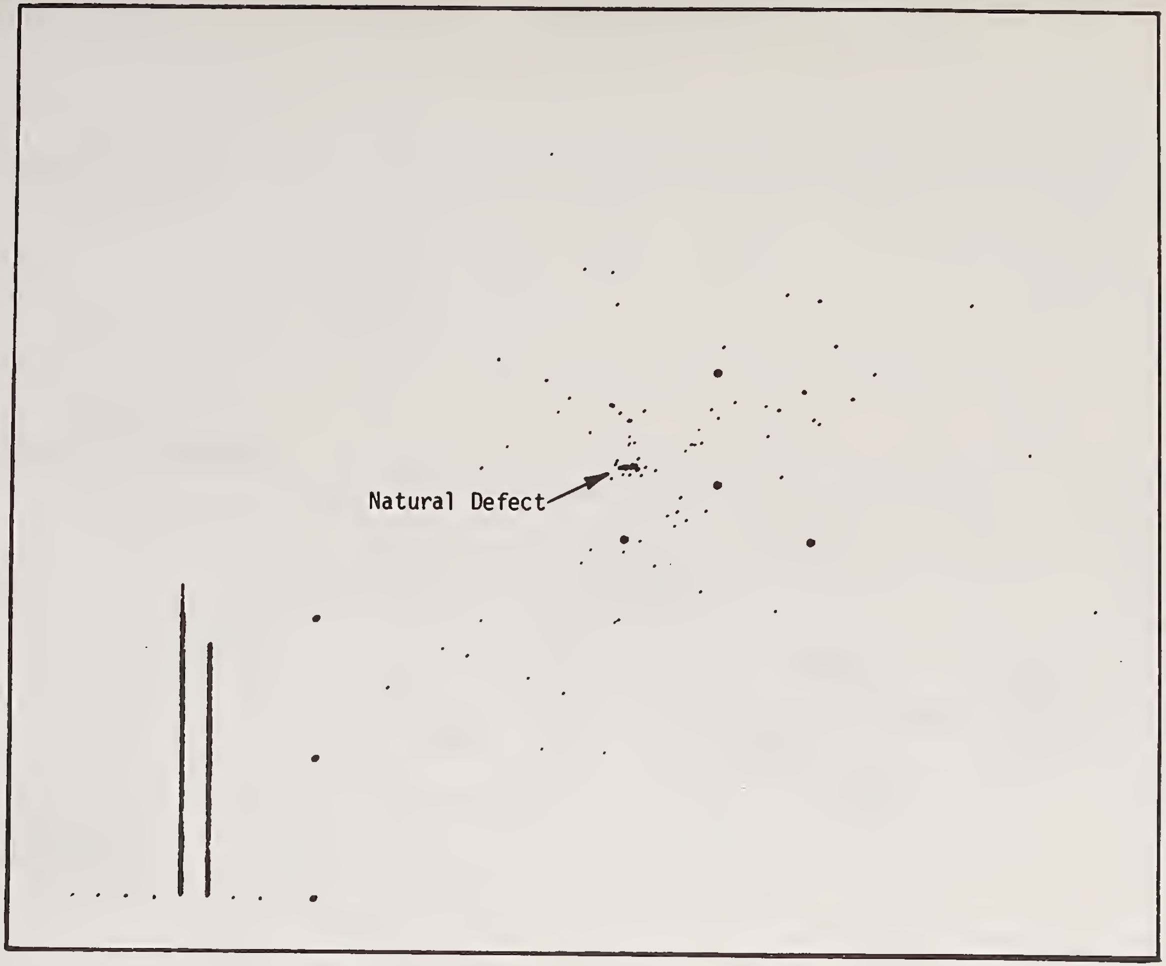


FIGURE 12. AE Sources Received from a Petrochemical Vessel Containing a 7.6 cm Crack - Entire Test

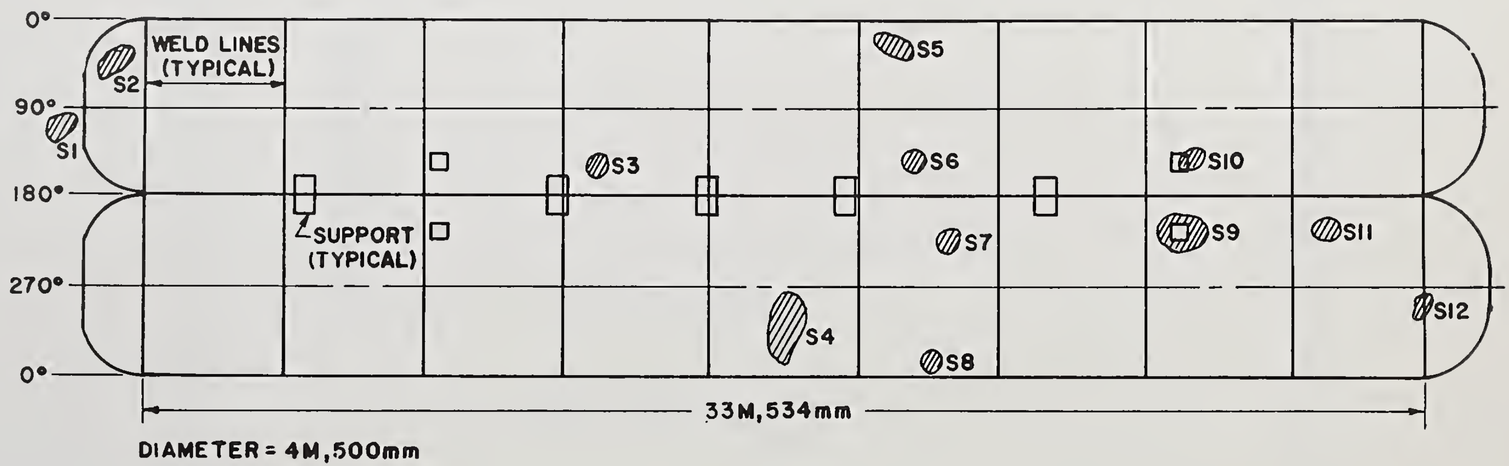


FIGURE 13. Graphic Display of all Acoustically Active Sources for Hydrotreater R-201

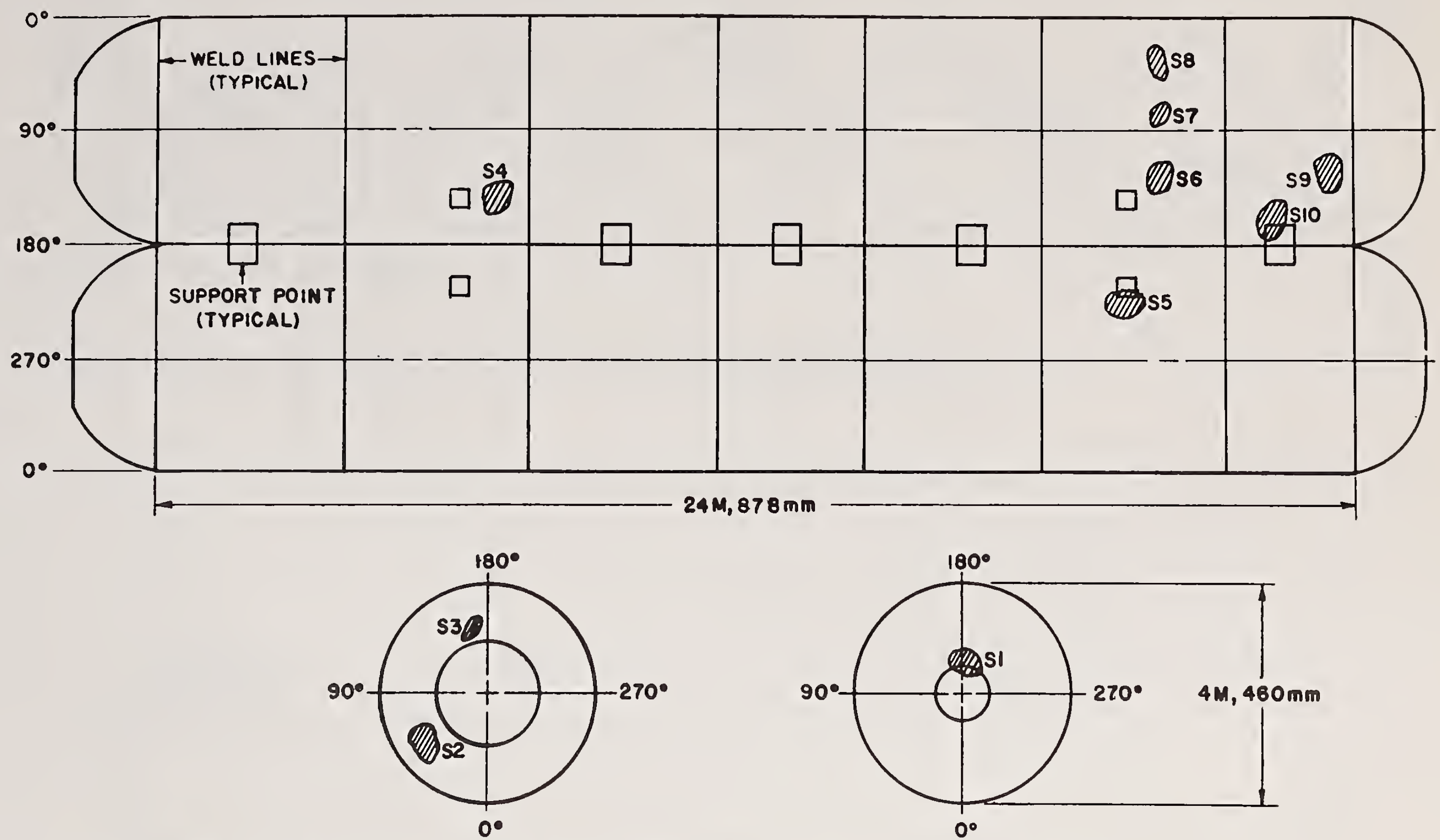


FIGURE 14. Graphic Display of all Acoustically Active Sources for Hydrotreater R-401

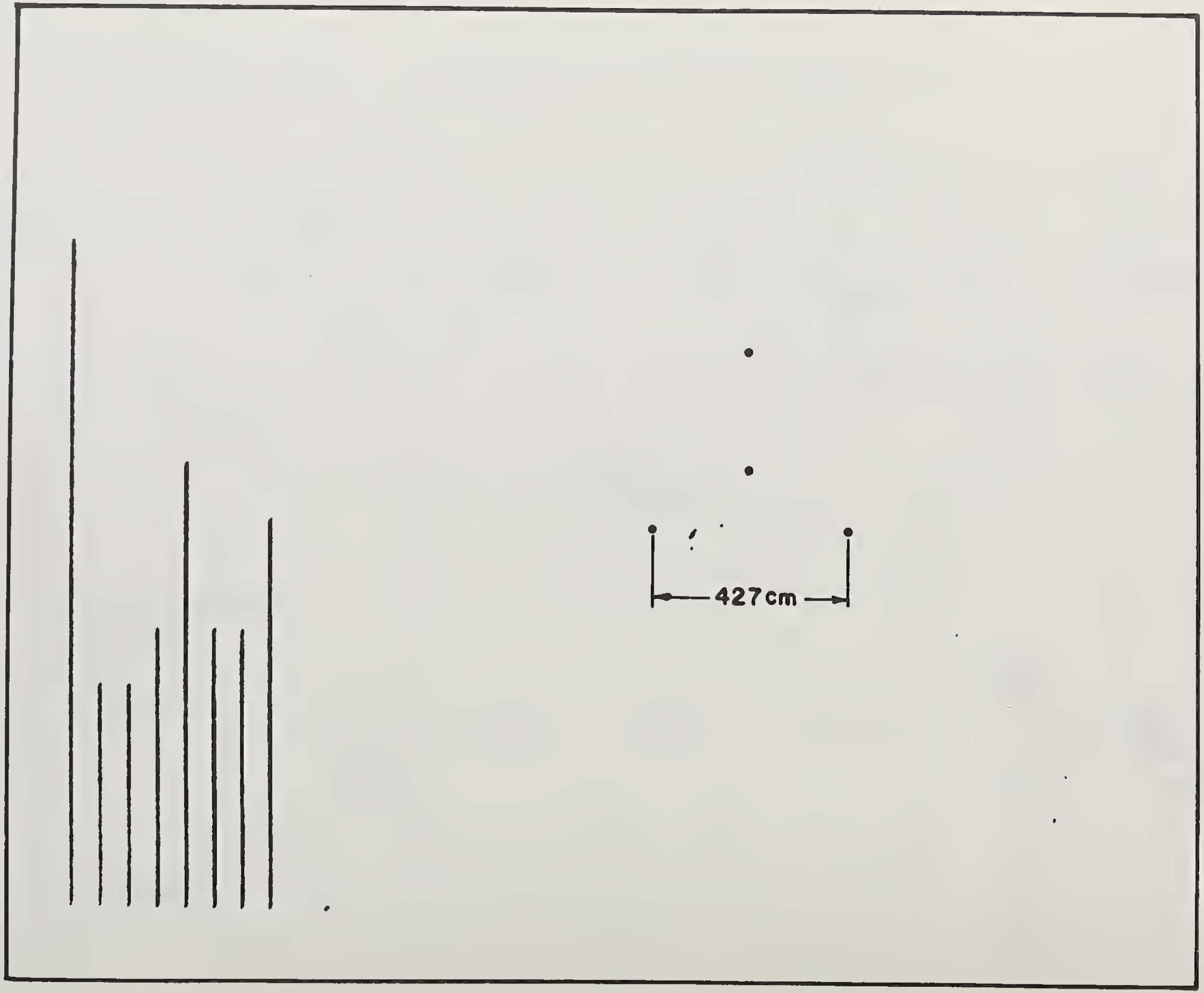


FIGURE 15. R-201 Array 4 Sources

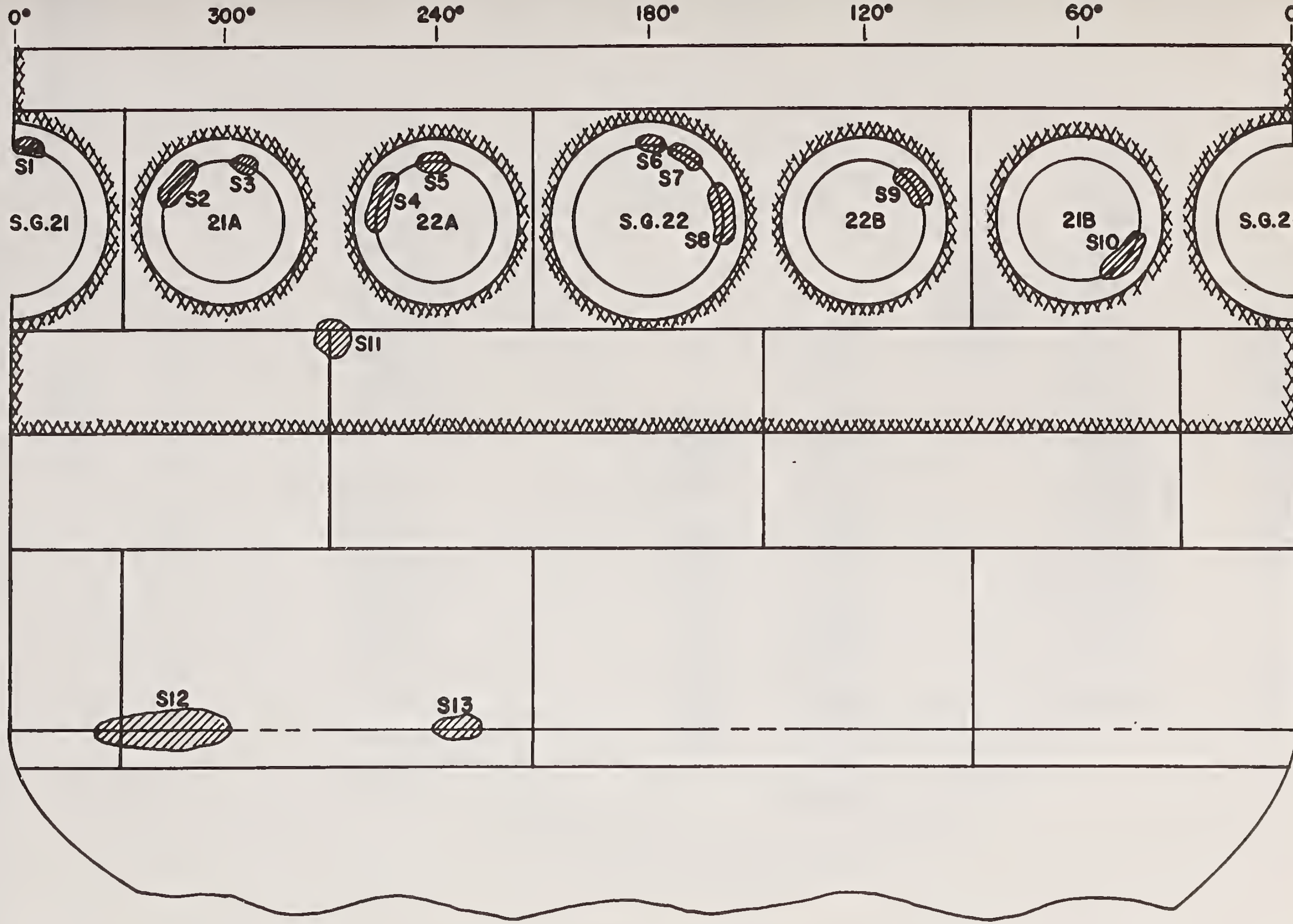


FIGURE 16. Reactor Vessel Sources

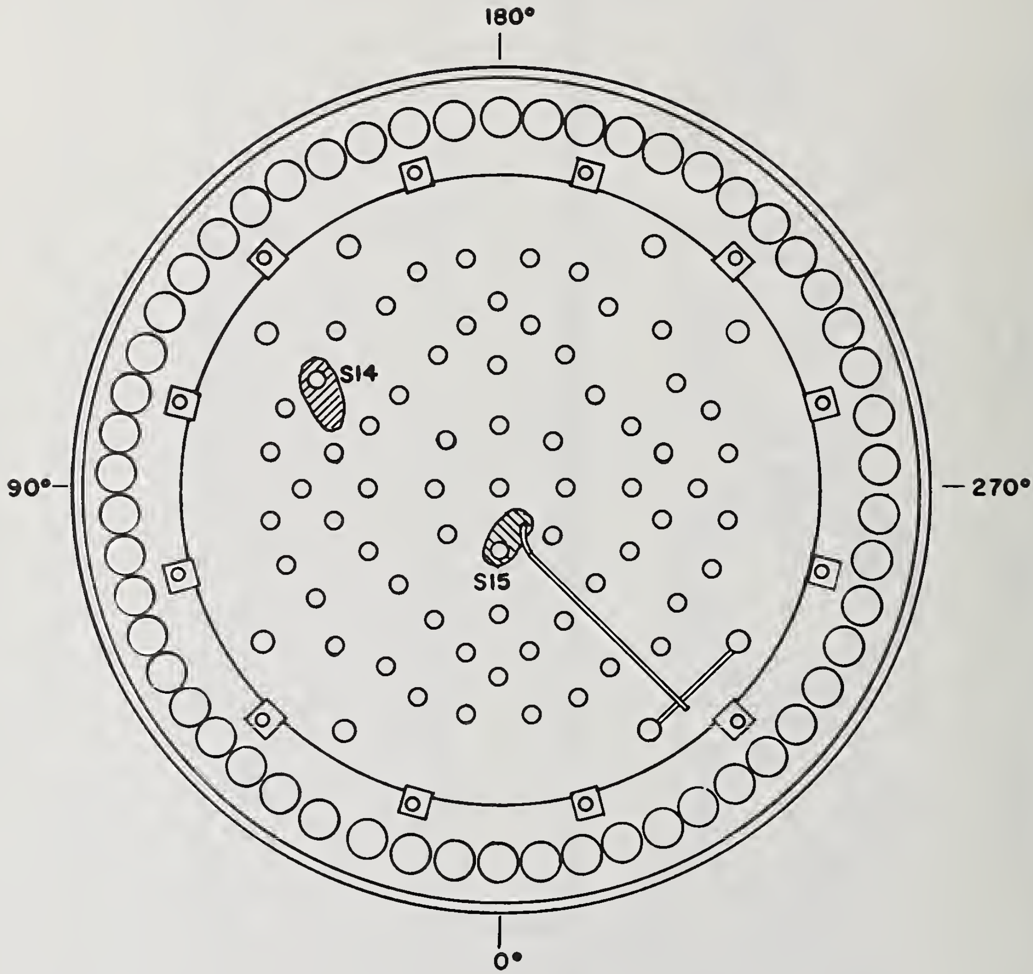


FIGURE 17. Reactor Top Head Sources

SESSION VI

GOALS, PURPOSES AND REQUIREMENTS OF DESIGN

-PRODUCT LIABILITY-

Chairman: J. E. Stern
NASA, Goddard Space Flight Center

NEGLIGENCE CHARGES IN PRODUCTS LIABILITY LAW CAN BE USED IN ENGINEERING AS SAFETY DESIGN CRITERIA

Dr. Raymond R. Hagglund
Mechanical Engineering Dept.
WORCESTER POLYTECHNIC INSTITUTE
Worcester, Massachusetts 01609

Abstract: Society expects manufacturing companies to make products that are safe but the meaning of the word safe is not very clear. In products liability law, the legal requirement is that a product be "reasonably safe". The charges of negligence, as used in products liability suits, can be used as safety design criteria in design and manufacturing to insure that reasonable care will be exercised in producing products that are reasonably safe.

Key words: Product Liability; Safety

Designing a "Reasonably Safe" Product

An interdisciplinary engineering team composed of design, material, manufacturing and quality control engineers have the responsibility for creating products which are useful, reliable, reasonably durable, profitable and "reasonably safe". These last words are in quotes and underlined because they are the focus of this paper.

It is fairly easy to establish tests for measuring product usefulness, reliability, durability, and profitability, but it is not so easy to define a test for measuring product safety. Based on my experience as an engineer in products liability litigation, I have found that many engineers have not seriously considered safety. I have also found that some management do not place much emphasis on product safety because safety generally increases the cost of the product. I predict that the increasing number of product liability suits will force both engineers and management to seriously study the meaning of safe and especially the terms "reasonably safe". In analyzing these words, there are two problems. The first is the difference in meaning between safe and"reasonably safe"; the other is the need for developing a set of engineering design criteria, based on these words, which can be introduced into the design and manufacturing process.

Some Background on the Products Liability Problem

It is necessary to have some understanding of the product liability problem in order to fully understand why this paper is placing such emphasis on the terms "reasonably safe".

Today, if someone is injured or killed while using a product, the injured person or his estate, will almost automatically enter a suit against the manufacturer(s), the distributor, and even the retailer, claiming that a defect in the product directly caused the accident. During the last six years there has been a rapid rise in the number of such suits and as the general public becomes more aware, because of newspaper articles, there will be even more suits entered and it is predictable that the number of suits will rise even more rapidly.

When an accident occurs, a suit is filed against a manufacturer but the insurance company carrying the companies products liability insurance actually handles the suit. At the very beginning of a suit, the insurance company investigates the accident and studies the charges of negligence. If the case is black and white and in favor of the injured person, the insurance company will settle; otherwise they will fight. Some cases go to court where a jury makes a decision as to whether the product was defective or not and if the defect, assuming one existed, caused the accident.

I have found that insurance companies are settling many claims because the engineers, who designed and manufactured the product, did not exercise "reasonable care" and, as a result, these engineers would have a hard time convincing a jury that the product was "reasonably safe".

The actual costs of unsafe products is quite clear today. The social costs are the loss of life or the social cost of a person living with an injury that may affect him for life. The financial costs are associated with settlements and jury awards which are causing insurance premiums for manufacturing companies to rapidly escalate. Some companies cannot even get insurance.

In some ways, the products liability problem is getting out of hand. For example, a company which designed and manufactured a machine thirty years ago can be sued in some states even if the machine, which caused an accident, had been resold several times and installed improperly. Insurance companies still have to spend money to investigate and defend these suits. In some states laws have been passed to control this type of suit.

To get back to the main point of this paper, engineers need to understand the meaning of reasonably safe. To do this engineers need to know a little about products liability litigation. Specifically they need to understand the role of the technical expert, the legal strategies used by the attorneys, and the charges of negligence that are used. With this understanding, I feel that an engineer can develop legal-technical criteria which can be used in engineering design.

Understanding the Role of the Technical Expert in a Suit

If a person is injured or killed, an attorney, called the plaintiff's attorney, is generally retained by the injured person or his estate. This attorney calls in a technical expert, an engineer in many cases, to conduct an investigation.

The technical expert attempts to reconstruct the accident and develops technical opinions which answer specific technical-legal questions.

Examples of these general questions are:

1. Was the product defective?
2. Did the defect cause or contribute to the accident?
3. Was the product being properly used?
4. Was the product reasonably safe for its intended use?

If the technical expert develops opinions which show that the product was defective, then a suit is entered against the manufacturer(s). The sued company then contacts its insurance company and the insurance company will perform its own accident investigation and develop some opinions on the charges of negligence which were entered. The insurance company then makes a decision to settle or to fight the suit.

If a decision is made to fight the suit, the insurance company generally brings in a defense attorney skilled at trial law. The defense attorney retains his own technical expert who performs an accident reconstruction analysis and who develops his own opinions about the accident and the product.

While the suit is being developed, the engineering expert may help to educate their respective attorneys in the technical details of the product, draft technical questions (interrogatories), review depositions and give depositions. The experts may even help in developing ideas for presenting the technical-legal case to a jury.

In presenting the case to the jury, each lawyer will use a certain legal strategy. While using their respective strategies, the jury will hear testimony from witnesses, technical experts and medical experts. When all of the testimony is heard, the jury will make a decision in the case.

The following sections give a brief description of the legal strategies used by lawyers at least as understood by an engineer.

General Legal Strategies

An understanding of the strategies that the attorneys will use in the courtroom will develop, in the engineer, more awareness to product safety in general. Also, the knowledge could be very useful if the engineer needs to appear as an expert in a products litigation.

The plaintiff's attorney will attempt to show that the product was defective and that the company was negligent because they did not exercise reasonable care in one or more of the areas of design, materials, manufacturing, testing, quality control, safety devices, user and maintenance instructions, and even in anticipating alternative uses for the product. In addition, the plaintiff's attorney will probably claim breach of warranty stating that the product cannot do what it is supposed to do.

He will also try to convince the jury that the product left the manufacturer in a defective condition, having previously defined one or more specific defects, and that the plaintiff's injuries occurred because the product contained these defects.

On the defense side, the attorney will attempt to show that the company did exercise reasonable care in the areas of design, materials, manufacturing, product testing, quality control, safety features, user and maintenance instructions, and that the company anticipated all reasonable alternative uses of the product. He will also attempt to show that the product can do what it was designed to do.

As a key part of the defense strategy, the attorney will attempt to prove that the person, who was injured or killed, was misusing the product and/or that the user assumed the risk when he used it in an unusual way which he knew to be dangerous.

In the trial, both attorneys will use prepared questions to develop their case so that it will appear to be logical to the jury. They will use information already gathered in interrogatories, depositions and experts' reports to help in framing these questions.

The plaintiff attorney's questions will probe deeply into the areas of negligence which have been charged with the goal of proving that reasonable care had not been exercised, thus resulting with an unsafe product.

The defense attorney will, of course, use questions in an attempt to show that the charges of negligence are not true.

The engineer needs to know what the charges of negligence are in order to develop technical-legal design criteria. These are given next.

Some Charges of Negligence

Some of the specific claims of negligence that might be entered are listed below; you can probably think of others.

Ten possible charges of negligence are:

1. The product was defectively designed.
2. The materials used in the product were defective.
3. The product was defectively manufactured.
4. The product was not tested.
5. The product lacked sufficient quality control testing.
6. The designer was negligent because he did not foresee all possible uses of the product.
7. The product lacked safety devices.
8. The product lacked proper warning signs.
9. Instructions for the proper use and maintenance of the product were not given.
10. The manufacturer did not notify all owners of their products that new safety devices were available, or that a design modification to improve the safety of the product had been made and was now available.

In the trial, the plaintiff's attorney will ask questions and sub-questions in any one or more of the areas where he feels the company was negligent. For example, in the first charge listed, defective design, the plaintiff's attorney will attempt to show that the manufacturer did not use state of the art information; did not use codes such as those developed by OSHA, ANSI, ASTM, etc.; did not prepare proper stress calculations or use special computer programs which were available. These are only a few examples of how the attorney might prove negligence in design.

The plaintiff's attorney will also probe into the record keeping system and he will attempt to show that records, such as calculations, tests, etc. do not exist or that they are not correct. Of course, this type of evidence, if it is found, will suggest to a jury that reasonable care was not exercised by the manufacturer.

Obviously the defense attorney will need correct facts and evidence to show that reasonable care was exercised in all of the areas listed.

Using the Charges of Negligence as Design Tools in Engineering

The charges of negligence, previously listed, can be used in design and manufacturing as a method for defining a "reasonably safe" product. But to be effective, engineering will have to develop defense answers for each of the charges listed and they will have to have records to back up these answers.

Most likely an interdisciplinary team of engineers will have to develop detailed defense type answers for charges of negligence listed and for others that they can think of. Such a team would ensure that the state of the art from other disciplines has not been neglected. If all of these charges can be defended on paper, in a thorough way, then the product can be defined as being "reasonably safe".

These charges of negligence can also be used as a management tool. It is proposed that management request engineering to develop a Safety Design Report for each consumer or industrial good manufactured. An interdisciplinary team of engineers should develop the report and they should devote a chapter to each of the charges of negligence listed.

In each chapter, the engineering team should summarize the state of the art, what was done, what was feasible but not done. For the latter, they should explain why it was not done if it was feasible. The report should be reviewed by company management, and by technical and legal experts outside the company.

If this action is taken, products will be manufactured that are safe - at least "reasonably safe".

Summary

The law of products liability, especially the charges of negligence listed in this paper, can be used in engineering and manufacturing as criteria for determining if a product is "reasonably safe". But in order for these legal-technical criteria for safety to be effective, company management will have to formally request that engineering develop a Product Safety Design Report where this report contains a chapter for each of the charges of negligence listed in this paper.

References

1. Legal: E. N. Swartz, "Hazardous Products Litigation", 1973, Chapter II, Theories of Liability, published by Lawyers Cooperative.

2. Codes and Standards: "Accident Prevention Manual for Industrial Operations", 1974, published by National Safety Council.

PRESCRIPTION TO REDUCE PRODUCTS LIABILITY LOSSES

A. THE DEFENDANT AND PRODUCTS LIABILITY

John P. Arness
Hogan & Hartson
815 Connecticut Avenue, N.W.
Washington, DC 20006

We have taken a brief look at the program for this symposium and find that it is literally filled with highly technical and useful information. I am quite sure that you will not get that kind of information from our talks because Milton Heller and I are just a couple of lawyers who are here to add a little bit of confusion to an area that is already as confused as it can possibly be. I am going to try to approach the situation from the standpoint of a defense counsel representing a target defendant in either a manufacturing or an engineering setting and give you a game plan with which to approach the very perplexing problem of products liability and the ever increasing costs that are incident to it. After I am finished Milton Heller who has been doing a good job for a long, long time scaring the "heller" out of me, is going to do the same to you.

As Mr. Stern told you one of the sidelines that I have is to be a participant in the Defense Research Institute. This is a nationwide organization of some 6,000 individual members and 400 corporate members. Among them are representatives of manufacturing companies -- maybe many of them represented at this conference -- and most of the major insurance companies. It is a most unique organization because it is the only one that I know of that is a real "think tank" organization. It started in 1960 and it is devoted almost exclusively to the production of data and proposals to help target defendants defend themselves against claims and lawsuits that might be filed against them. DRI is particularly interested in the unfounded claim of the type that Professor Hagglund has mentioned to you this afternoon.

Now, you must understand when you think in terms of products liability, and when you think of anything in terms of law, that you are dealing with fluid and imprecise concepts. I know, being scientists such as you are, it is going to be a little bit difficult. But law is a dynamic thing, ever changing. It has no rigid precepts. It has nothing that you can count on from day to day. Because the law tries to accommodate to the ever changing needs of society and in doing that focuses upon every case that comes along. If a particular case has certain factual connotations and cries out for a remedy, the judge is there to try to set things right. If you will analyze the situation I think you would conclude that you would not have it any other way even though the imprecision may offend your scientific mind.

The products liability situation is relatively new. Even as short a time as 25 years ago when I began my practice of law a products liability case was a rare thing. But fads come along and right now we are in that period where it seems fashionable to condemn a manufacturer or an engineer or designer for something that has caused injury. An after the fact Monday morning quarterback finds it easy to place blame particularly upon a deep pocket -- a company or an individual. Notwithstanding what I said about not having rules upon which you can depend in the law, a safe product can be defended. The law, although it does try to accommodate changes in circumstances, is a very reasonable thing. People come into courtrooms before a judge and a jury and once in a while you will have an aberration, but, by and large, that jury and that court is there to see that the fair thing is done. So we are going to try to talk for a few minutes this afternoon about what kind of program you can have to put yourself in the posture when you come before a court to appear reasonable with respect to your product and your conduct.

There are basically three ways in which liability can be affixed upon a manufacturer or design engineer. One is on the theory of negligence, and as Professor Hagglund told you, negligence is a very simple concept. It is defined by saying that one is not supposed to do that which a reasonable person would not do. Or the corollary of that is that you are supposed to do what a reasonable person would do under the same or similar circumstances. The second theory upon which liability can be imposed is breach of warranty. That was originally a contract theory. It merely means that when you sell something, whether to an ultimate consumer or to anyone else, you warrant that that product is reasonably fit for the purposes for which it was intended and if it isn't you have breached your contract. You haven't given the purchaser what you said you would give him. The third theory, and that is the one that is giving people the most trouble these days, is a concept that has become known as "strict liability". Milton Heller is going to tell you a lot more about that a little later but I will just touch on it. Strict liability, as it originated, was not supposed to be what it has come to be. There has been an awful lot of hysterical talk about it. When you say "strict liability" you kindle emotions. Does it mean that I am liable no matter what I do? No -- that is not what strict liability is. Strict liability merely means that when you put a product into the stream of commerce and it is defective in that it contains some feature that is unreasonably dangerous, you will not be heard to say that you were careful in its manufacture and therefore you shouldn't be liable for injuries caused by it. You put it in commerce -- it is unreasonably dangerous -- and you are going to pay for it. That is what strict liability means -- it does not mean you are liable <u>unless</u> it is unreasonably dangerous.

The first thing that a design engineer has to think about in dealing with liability situations is the design of the product as it relates to man; as it relates to the machine; and as it relates to the environmental interface or the circumstances under which that machine or that product is going to be used. I would appreciate it if you would dwell on that concept for a minute, because in my experience in defending products liability cases and in a few cases prosecuting them as well, liability has come about in every instance that I can recall because somebody did not think about how that product was going to interface with society. So that is the first thing you have to do. You have to think about the product as an entire concept, not just the product in its technical sense, but how is that product going to fit in with man and with the function that it is going to perform.

There are many vulnerable areas. Professor Hagglund has touched upon some of them, as a matter of fact he has hit the ten most important ones. But let me review them for you. There are design errors; material selection errors; component selection errors; assembly; testing; instructional materials and techniques; sales literature; labels; advertising; warnings of latent danger; warnings of hazards which cannot be eliminated; inadequate safety measures or features and failure to foresee misuse or abuse of the product. There are some kinds of abuse that you know from your experience people are going to subject the product to. You are required to foresee that which a reasonable person would expect. You must anticipate some alteration to the product or some aborting of safety features. I have seen products, and you have too, where safety features are designed and you have to do this or that to make it work and you know a lazy workman is going to take a shortcut and he is not going to operate that safety feature. You have to make it easy for them to put it on, you can't make it difficult. Liability itself is premised sometimes on the failure to have that safety feature in a proper position or susceptible to easy use. Failure to take protective action after discovery of a defect or failure to properly consider the life of the product and make accurate representations with respect to it are additional vulnerable areas. You have a constant fight between the design engineer and the manufacturer on one hand, and the sales force on the other. You make a product and you know it is going to be good, you know it is going to perform for a certain period of time, but when your salesman attempts to sell it, he is not going to say after 4 or 5 years you should replace this item -- because if he does he is going to meet sales resistance. So after 7 or 8 years and the user is using the product far beyond the life for which it was designed, you have an accident. There is liability because there has been an inaccurate representation with respect to the life of the product and its effective and safe use. Packaging is another area of concern. The product must be packaged properly because sometimes accidents happen when it is not.

Other areas in which liability can crop up involve warranties and the way you go about making repairs and replacement parts available when they are necessary. I am not going to talk about each one of these areas of vulnerability because in this limited time we have this afternoon I would rather talk about the approach to doing away with the liability that can come about from these vulnerabilities. First, I would strongly recommend that you establish in your companies, by whatever name you wish to call it, a design review committee -- design review committee which has interdisciplinary members, not just engineers as Professor Hagglund mentioned. Your design review committee should have people from the sales force on it, personnel from management on it; and it should have lawyers on it. Now, what should the design review committee do? First, it should be given the power to make decisions and to enforce those decisions. A discussion over coffee and doughnuts, once a month, is no good to anyone. The design review committee should be charged with the duty of evaluating all new products and with the duty of re-evaluating, from time to time, old products to see that they conform to current safety standards. The design review committee should be familiar with all current applicable state and federal regulations, as well as industry standards whether they be mandatory or merely recommended. The committee should be charged with the responsibility of keeping abreast of trends; it should be aware of what other manufacturers in the same industry do. It should be particularly alert to safety features to be incorporated into the products. The committee should document all of its deliberations during the design phase and it should explain why a particular modification or safety device, having been considered, was rejected. The same stringent controls should be exercised by the review committee during all phases of production and distribution, even though the design phase may be the most important.

In addition, your companies should have a testing program. The program should be effective so as to thoroughly test the design of any product before the design is finalized. The testing program should also provide effective quality control during the production phases. It should be charged with the duty not only of checking the product that is being manufactured by your company, but should check all components supplied by other manufacturers. In addition, all materials should be checked for suitability.

Next, we recommend that your company have an effective and detailed record retention program. One of the most difficult tasks for a trial lawyer trying to defend a client after an accident has happened is to go back, many times, two, three or four years, to try to find out what the facts are. Any many times he is frustrated to find out that there were no records to show what the facts were. Records are the most important single tool of defense and they are the most important thing you can do to save your company's money in products liability defense situations. What kind of records? Records showing how the design came about. Records identifying the materials and showing how the materials were selected. Assembly drawings, including all the guards and safety features. Records of the suppliers of the component parts. Records showing testing, quality control and inspection. Check lists in shipping to show that warning labels were included, to show that proper instruction material was sent along with the product, and to show that operating manuals were included. Records to show what the state of the art was at the period of time when you are going to be charged with having breached it and reliance upon recollection is hazardous when you are dealing with such a fluid situation as development of the art. I think that type of program will be very, very beneficial.

There are a few additional points I would like to make. By all means, gentlemen, don't let your companies keep the fact that there is a defect in the product a secret. If you discover a defect and then try to cover it up while taking remedial action to see that it does not occur again you are being naive and reckless. It will be discovered. And when it is, gentlemen, a showing of attempted cover-up will lead to an award of punitive damages. You know it is one thing to have somebody injured and to be soaked by a jury for fifty to seventy-five thousand dollars. That is hard enough to bear, but when that jury is told that you found the defect and that you knew about it and you hid it, that is when the millions of dollars get tossed on the top. That is the pill that is really throwing product liability litigation all out of economic reality. It is the millions of dollars in punitive damages that you cannot adequately predict that is driving manufacturers up the wall. When you find a defect, publish it, remedy it, get that product back, send out the warning, take care of it and stop that accident that is going to happen tomorrow. Never keep a defect a secret.

My next subject is personnel. You have to find some way to prevent untrained workers and unqualified personnel from participating in these very important matters. Qualified personnel have to make design decisions. They have to have the technical ability; they have to have the background and the scientific knowledge to make the decisions that are going to be tested later on in court. You just cannot expect unqualified people to make that kind of decision. Yet, history has shown that many manufacturers try to set up design review committees complemented with borderline qualified personnel in order to save money.

Next, pay as much attention as you can to warning situations. I heard a few questions after Professor Hagglund's remarks. We all know situations where liability has been imposed under circumstances which would lead one to ask "How in the name of justice could someone be held to pay damages for such an occurrence? An occurrence like that is almost necessarily going to happen." Another way of thinking about this subject is that some products are impossible to manufacture absolutely safe. Some machinery and some equipment that is designed to do particular jobs are just inevitably going to expose human beings to injury and damage. You cannot make everything safe. You cannot put a guard over a butcher knife blade. So, what can you do? You can adequately warn. You can make certain that there are no latent dangers. You can make certain that you have the right kind of labels on your product so if somebody is going to throw a switch he knows what is going to happen. If you manufacture a product which has a hole with a piston inside of it, you can put a big red sign over the hole which says "Don't poke your head in this hole." Adequate warning is the easiest, safest technique that you can employ to improve the safety of a device or a product that is going to inevitably cause danger if it is not used properly. Put it on your product and save yourself from liability. I cannot believe that a court would hold somebody responsible for an injury that came about from something where there was a great big red sign that said don't do it and yet the person went ahead and did it anyway.

Now, let me briefly mention the need for proper packaging. Some products are dangerous unless they are handled properly, unless they are packaged properly, unless the right kind of warning is put on the package. Be very careful about such details and pay just as much attention to the package as you do to the design.

The last thing I would caution you to heed is the need for prompt investigation. Liability frequently flows from an energetic plaintiffs' attorney who gets to the product before you do. He gets all of his homework done and then files a claim and you sit around and try to pick up the pieces afterwards. Get in there and make the investigation promptly. Save the evidence, identify the witnesses, take their statements and then you will be in a position to set the record straight once a claim is made. In conclusion, I would like to wrap up this part of our presentation by suggesting that, in order to carry out these things that we have recommended, reduce your intercompany procedures to writing so that policies will not be misunderstood, then enforce them rigorously. Establish effective communication to facilitate your loss prevention program. Keep those records and make sure that you have established a mechanism for conducting a prompt investigation.

PRESCRIPTION TO REDUCE PRODUCTS LIABILITY LOSSES

B. THE PLAINTIFF AND PRODUCTS LIABILITY

Milton Heller
Investment Building
1511 K Street, N.W.
Washington, DC 20005

Now you know why John is so successful. He is a gentleman, knowledgeable, and articulate. If any of you become expert witnesses I think those are the three qualities that you should possess. The jury trial often functions as a popularity contest coupled with some good hard facts and theories and a cause of action. Although everything that John says is true, most of the engineers in this country don't work for manufacturers who are willing to adopt the programs John suggested. A lot of you might work for people who cannot afford to adopt all of the safeguards. I doubt very much if even the largest corporations do all of the things that John says they should do. A lawsuit is therefore inevitable when a plaintiff's attorney is retained in a case where a defective product causes injury. If the defect existed at the time the product left the manufacturer's or distributor's hands, the attorney will inevitably find that one or more of the suggestions John made were not adopted by the defendant. If the failure to comply with one or more of the points was a competent contributing cause of the manufacture and distribution of this defective product, and that defective product in turn, without misuse or abuse, caused injury, the plaintiff is supposed to collect.

Regardless of what you read or hear, pursuing a products liability case is a very hard job. It is especially hard to get good experts if you represent the plaintiff, because if you are suing a particular member of an industry like the auto industry, most of the good experts are aligned with the auto industry in one form or another. Those who aren't, and set up their own consulting firms, I have found, as a general rule, are not as knowledgeable with the state of the art and industry. They don't have the facilities and funds to run tests and keep abreast of advances that are so essential to an expert being able to express an opinion on the efficacy and safety of a particular product.

Before I go any further, many terms have been used, like "reasonably safe" and the converse, "unreasonably dangerous." The term "unreasonably dangerous" is the one used in the Restatement of Torts, Section 402A, which is the basis for strict liability. Strict liability is not absolute liability, as John told you, and it requires proof of the defect and injury. An official comment to 402A states that an "unreasonably dangerous" product is a defective product that injures because of a condition of the product not contemplated by the ultimate consumer. Some of the typical examples of defects not contemplated are a steering wheel that comes off while you are driving or wheels that roll off. Those are obvious examples; ones that are not as obvious appear when you get into the failure to warn, or failure to instruct area. Defining terms is most important.

I served on a committee with a group of doctors who was trying to define when a person is legally dead. Is it when the heart stops or brain waves are flat? We kicked it around and finally decided that a man is truly dead when he cannot arise and litigate. The other day I was listening to the Today Show and they were talking about someone who ran for election on a fun ticket in Washington or Oregon, and promised that if he is elected, heads would roll on Capitol Hill. When they asked him what he meant by that, he said he would take portable johns, put wheels on them, and roll them down the hill. Defining terms is so important.

What is a defect? A product can be defective if it doesn't meet design specifications, if it doesn't meet federal regs., other regs., if it doesn't meet industry standards evidenced by association standards. If the product falls below those, and the plaintiff buys the product, and it causes injury, he has a case. Cases like these are few and far between. They are usually settled. Cases you read about in the newspapers are the ones where the product is destroyed, or there is no ostensible defect in the product after some accident or collision that half destroys it. Here we are in a gray area where expert witnesses come in and quite logically and precisely opine that a defect caused the accident. They can reconstruct, based upon the way it happened, and come up with a very good explanation. Those cases are rarely settled because for every one of those experts, the industry can find ten to refute him and call him a prostitute or whatever.

The tenth factor that Professor Hugglund mentioned, a continuing duty on the part of the manufacturer to modify or warn, is the hot area in the field of products liability at this time. But you must limit that concept to a product that has a long life expectancy. It can be a punch press, an airplane, truck, or industrial equipment. When it is manufactured, distributed and sold it must not be designed or produced in a negligent way. By that I mean the manufacturer has exercised reasonable care. Putting it another way, the manufacturer has done everything that a reasonably prudent manufacturer would have done. In addition, it was not defective according to the state of the industry or the state of the art. But then, you have a Catch 22 situation. If the manufacturer remains in the business, to be competitive, he has to continue to improve his product, research, and keep abreast of improvements in the field by his competitors. When he does that he eventually becomes aware that the product he sold ten years ago can be vastly improved from a safety standpoint by a modification that costs very little. Now, the other situation is when the manufacturer, after ten years or so, is informed that his product is causing a number of injuries and deaths, but knows of no reasonable method of improving the product. In the latter situation the manufacturer has a duty to inform his purchasers that there is a danger which he does not think can be obviated, so that the user or purchaser can act appropriately. In the initial instance, where the manufacturer learns that the product can be improved, he has a duty to offer a modification that can improve the safety of that product, or in the alternative, at least inform the purchaser that he can purchase a new product with the new safety features. There aren't many appellate cases around the country on it, but there are a few, one in Illinois, one in Massachusetts that I am aware of, one in Maryland. I don't know of any recent opinion that holds that a manufacturer, once he sells a product and title has passed, has no obligation or duty to that purchaser to keep him informed when the manufacturer knows that the product can be improved from a safety standpoint. Now, some examples. Take a manufacturer of an anesthesia machine used in an operating room that hooks up to an outlet in the wall. Another manufacturer has sold and installed a main central system that feeds oxygen and other gases through the walls to let's say nine operating rooms in that hospital. The mechanisms hook up in the operating room. They work by pressure, although there are electrical ones too. Now, for years an anesthesiologist was a computer. He had to watch the vital signs periodically, and at the same time, periodically glance over at the flow meter to see that the nitrous oxide and the oxygen were flowing. Anesthesiologists are well skilled. However, not all operating rooms have anesthesiologists on duty at all times. Some hospitals have nurse anesthetists or dental residents or other quasi-medical personnel who are inducing patients and watching them. They are not as skilled. Occasionally a hose kinks, something goes wrong, the connection is poor, and the nitrous oxide flows but the oxygen doesn't. The person administering the oxygen doesn't glance over at the flow meter, the patient has brain damage and eventually dies.

The manufacturer knows the tube kinks, and knows that many of the hospitals around the country are teaching hospitals with trainees who are not as qualified as the M.D.'s. The manufacturer knows that since the gas flows by pressure, a little whistle that costs about $2-$3 could be put on the machine so that when the gas pressure falls below a certain level, the whistle will go off and warn them that the gas is not flowing.

Now, to me it is negligent for the manufacturer not to inform the hospital that they have this little device on their newer models, because the burden of telling the hospital personnel that they have the whistle is very slight. The manufacturer regularly does business with hospitals, they have detail men going into every hospital at least once a month. The cost of the whistle is slight, that's the burden, compared to the benefit of saving a life.

Another illustration appears in the motorcycle industry. As you know, a driver has available to him a hand throttle, manually operated by his right hand. If he is propelled backward out of his seat, the motorcycle racer is prevented from being propelled too far back by a "sissy" back. In older models, a return spring was not installed which would have automatically reverted the hand throttle to neutral. In newer models, the return spring is installed. For some reason certain manufacturers failed to include the return spring, and in others the return spring is defectively designed. Should manufacturers inform purchasers of older models that return springs are readily available at a reasonable cost for installation on older models?

In the failure to warn field the most "far out" case of its kind was handed down by the Court of Appeals of Maryland a year or so ago. A well known company makes perfumes and colognes that are flammable. They also incorporate their fragrances into candles. The candles are a popular item for the pre-adolescent kids who enjoy the fragrance at birthday parties. Some 12 or 13 year old kids were having such a party, but the candles were unscented. The children were accustomed to scented candles. They took mommy's cologne or perfume and poured it on the lit candle. A child was horribly burned. The Court of Appeals held that the manufacturer had a duty to place a warning on the perfume or cologne cautioning that it be kept away from fire. Although the case might sound extreme, it contained abundent evidence that such a warning was on the cologne aerosol bottles, and since the kids were conditioned to the fragrant candles, it was foreseeable that children would engage in the devious, mischievous and ingenuous play that caused the injury.

DISCUSSION

Question: Many catalogs of manufactured equipment or instruction manuals contain a warranty statement something to the effect that whatever is warranted against applies to defects, workmanship and material for a period of such a length of time or, in the case of an automobile, so many miles. All other warranties express or implied are null and void. Another statement that often appears is something to the effect that in keeping with its policy of constant product improvement the manufacturer reserves the right to change specifications from those noted in the catalogs and manuals. Are either of these statements legally meaningful in today's legal climate?

Mr. Heller: With regard to the first statement, in most jurisdictions, that is usually an attempt on the part of the manufacturer to limit their liability for breach of warranty to replacement of the parts. The typical auto warranty says we guarantee this to be free from defects in material and workmanship and we will replace parts. That is a commercial warranty, and although they go further usually and say "We are not liable for anything else", the Courts have held this disclaimer unenforceable and most statutes among the States specifically exclude cases for personal injury or consequential damages. That attempt to limit their liability has no effect at all in personal injury cases. It does in commercial cases.

Mr. Arness: Yes, I would agree with that except for one thing. I spoke in my remarks earlier about the life of the product. This is a concept that hasn't been brought to the fore. However, I think it is a very viable concept once you get your sales force under control. Many products just shouldn't be used beyond a certain period. I think a warranty that is weasel worded is just an attempt to limit legitimate liability and might not be enforced. But if a manufacturer clearly said this product should be used for four years and that anyone who uses it beyond four years is taking his life in his own hands, I think that type of limitation would be heeded and would provide a measure of protection in a court of law. On your other question pertaining to the other type of disclaimer, you just cannot rely on any kind of a disclaimer automatically. You are still going to be held to a standard of reasonableness. You can put disclaimers on all you want to, but if you have violated one of these reasonable standards you are going to be held liable for the injury consequences of that violation.

Question: The question I had about the disclaimer is more directed toward the follow-up, toward getting information back to the ultimate user. The manufacturer is trying to tell the user that he can expect that the manufacturer is probably going to change the product next year, next week, five years from now, because the manufacturer may find in the future that what he is selling today is not going to be up to tomorrow's standards.

Mr. Arness: You are trying to put the burden on the user to follow up whereas Milton Heller told you the burden is on the manufacturer. You have to get to the user and you cannot put something on there that says ". . . Look we may change this next month, or next year, and you get it when we do." The manufacturer is going to have the burden, the reasonable burden. I do not subscribe to those who say that a manufacturer who puts out millions of products is going to have to find where everyone of those millions of products are and get word to the person who has it every time there is a modification. You are going to have sales and resales and things of that nature and tracing the ownership of these products is an impossible burden. But, in the case of capital goods, normally that transaction is easy to trace. In the case of large industrial equipment, that is going to be relatively easy. In the case of vehicles, that is going to be easy. In the case of smaller products, the manufacturer, in order to comply with the burden, is probably going to have to resort to mass media advertising or something of that nature. But they are going to have to do something.

Question: I notice something was carefully avoided in this presentation about the changes to the product to cover safety features which were missing from the original product. A manufacturer found that a handy, dandy whistle, costing $5, would make his product a lot safer. Is he responsible for paying for the $5 or can he go to the users and say "Look we have now found a safer way to do it and we can sell you a handy, dandy whistle for $50?" Does that cover him automatically?

Mr. Heller: Oh yes, I didn't mean to suggest that you had to furnish the whistle free. In fact it is good business, when you modify and improve an obsolete product or an outmoded product, to sell that improvement to the old users. I think the wise thing to do is to keep a record of the fact that you called on that person and warned him of it. You can't force him to purchase it.

Mr. Arness: I think that is absolutely right. But I want to tell you something. I represent one manufacturer that really follows these things to a fetish. He has developed a very interesting technique. He sends out with his product those postcards which have the return card attached. Then he has a follow-up system, if they do not return the card, he follows up with them. Then if there are subsequent improvements he sends out another card or letter. There are all kinds of ways to do it.

Question: I believe it was Attorney Arness who in passing referred to products liability as the latest legal fad. Does this imply that this is something which you think is going to run its cycle and is going to be de-emphasized in the future?

Mr. Arness: I certainly do and I will tell you why. There is such a hue and cry already abroad in this land that I am certain something is going to be done to keep product liability litigation within proper bounds. As Professor Hagglund told you, the insurance premiums are going out of sight. I am told that certain manufacturers' premiums have tripled in this last year alone. Whenever that kind of a syndrome occurs and things get out of joint, this nation has the imagination to deal with it and there will be changes made that will stop it. I do not know what the changes are going to be, but legislators will make it unprofitable for people to bring products liability cases. But we have had trends in the law ever since the law has been used as a vehicle for solving human disputes. Plaintiff's lawyers 40 years ago, had one kind of toy, now they have products liability and five years from now they will have something else to play with. In my role as defense attorney I probably will be standing before some group 10 years from now and talking about something else because there will be a rash of those different cases to defend.

Mr. Heller: Let me comment on that if I may. What John says might be true, but I don't think so, because the law of torts, especially in the products liability field, has been a force for good. It has been found that, but for some of these big verdicts and liberal laws, the manufacturer would not voluntarily make a safe product. Mind you, I am not a great fan of Ralph Nader. But if he had not pushed that National Highway Safety Act or whatever, we wouldn't have these recalls that we see and what I am saying is reflected in the laws that are coming out from the legislatures around the country. Every act in this field that passes through Congress, specifically states that this does not mean that a person does not have a common law right to sue for products liability recognized prior to the enactment of this law. They specifically reserve to the public their right to sue when injured by a defective product.

Mr. Arness: I think that is absolutely accurate, but that is not contrary to what I said. What I said is, manufacturers are now being exposed to a much greater number of product liability litigations at a rapidly increasing rate - much more so than they were five years ago. The answer certainly is that rigid controls are going to be imposed and manufacturers are not going to put out unsafe products with impunity. Once the science develops better, safer designs and manufacturing techniques, and improved quality controls, then products liability volume will be cut down. It is going to be too hard to make out a products liability case. Then something else will be easier and come into vogue so as to replace the current products liability fad.

Question: What effect would a change in model number have on product liability? In other words, would a yearly change in model limit the liability for discontinued models?

Mr. Heller: Generally on your duty to design from a negligence standpoint, you have a duty to design the products that employ all of the knowledge that the industry knows at the time. Now, the model T goes out, but then if you find that the model T is unreasonably dangerous and you know how to improve it, I would say there is a duty to warn the purchasers. If you find that it cannot be modified and yet you still know it's unreasonably dangerous you have a duty to warn. Something that is obsolete and was manufactured in accordance with a standard of care different from present standards, is, in itself, no indication of negligence. This is not the kind of question that can be answered yes or no. You have to talk about a specific part of the model T. If you are talking about the brakes or something else then I could be more specific.

Question: I am just curious, I will put this to both of you. In the future is this a possibility - say someone is driving a car and they get hurt. They feel that their injury was due to a defect in the car, hire an attorney and file suit. The jury finds in favor of the defense. Can the manufacturer countersue the plaintiff's lawyer and the person injured or his estate for defamation of character or the equivalent, in any way. Is that a method that will be there maybe five years from now to limit the suits?

Mr. Arness: No, and the reason for that is very simple. I think that one of the greatest - and now we are waxing mellow, I grant you - but I think one of the greatest privileges that we have in this country is open resort to the Courts for solving disputes. I do not believe that our society will ever tolerate a penalty for that access. Now if you ever had that kind of penalty, people with resources could keep people without resources from an attempt to vindicate their rights in a civilized manner. It will never come about.

Mr. Heller: There has always been a civil abuse of process lawsuit available to anyone who is sued without any basis. Proving such a case is very difficult. Now a doctor in Chicago proved it and got an $8,000 verdict on a counterclaim. I do not know what has happened to it after that.

Mr. Arness: That verdict has not been tested yet, it is still in Court. It could never come about when a person was truly injured and tried to get a recovery. The only way it could come about would be if somebody wasn't really injured, but came into Court and fraudulently represented that they were injured and fraudulently manufactured a claim then maybe you would have civil abuse of process, but it would have to be almost willful. A recovery for civil abuse of process would never be imposed on someone making a bona fide attempt to get a recovery even though their theory was way out.

Question: What is going to happen to cut down this spiral of lawsuits and insurance premiums? Will the government be the insuror of last resort?

Mr. Arness: There is a presidential commission meeting this week. As a matter of fact, they met on Monday and Tuesday. This commission, composed of representatives from all types of manufacturing concerns, as well as lawyers from both sides of the street, was meeting in an attempt to grapple with this product liability situation. I am certain they will come out with some kind of recommendation. What it will be would take a crystal ball to predict.

Mr. Heller: Let me say one thing, one of the arrangements that has been criticized a great deal is the contingent fee that lawyers take these cases on. Of course, you read about $3,000,000 recovery and the lawyer gets $1,000,000. First of all, the contingent fee, as far as I am concerned, is probably the most equitable fee arrangement that an attorney can have with his clients. If he loses the case, the attorney gets nothing, but more important, if you have a good injury case, it matters not to the attorney whether you are the president of a bank or that president's janitor. Sometimes you would prefer to represent the latter. All that matters is the merit of your case, and that person can walk into any law office and hire the attorney he wants without putting up a nickel. To me that is very fair.

Question: Would either of you like to comment on the proposals to change the process for trying products liability cases. That is to impose a technical evaluation, a technical pretrial before the litigation goes to the jury.

Mr. Arness: That is one of literally dozens if not scores of proposals for handling the current products liability crisis. As a matter of fact that is one of the more viable alternatives to the situation that we have now. Particularly with respect to strict liability. Strict liability is supposed to be a doctrine that makes you pay when the product is found to be defective. A defective product is one which, when you put it out in the stream of commerce and when used as it is supposed to be used, has the ability to hurt someone. You cannot really test that premise in an arena where you have someone lying in a pool of blood after they have had an accident. It is not really fair to judge on a Monday morning after someone got hurt the night before. If you are going to test the product to find out if it is in fact defective and if you are going to ascertain whether the product, in the environment in which it was supposed to be used, is dangerous, you do it objectively. You do it scientifically. The fact that somebody happened to get hurt might be a factor, but it certainly would not be the compelling factor that it is now in the lawsuit trial. The people who recognize and appreciate that fact of life say, "Let's have a bifurcated trial - let's try the product first in front of a jury - let the plaintiffs' lawyer come in and say, "This product is a dangerous product - it's got this wrong with it - it was designed in such a way that someone is liable to get hurt"; and let the jury look at the case and evaluate the product."

Their focus will be on the product and not upon the unfortunate plight of the person who happened to get hurt. That approach, objectively speaking, is a fair way to approach the situation. However, the American Trial Lawyers Association and plaintiffs' groups all over the land are going to fight such proposals like wounded eagles and it probably won't get through.

Mr. Heller: I say that we should have a bifurcated trial. Bear in mind it really doesn't affect anything. If you represent someone with both legs off he will be in the courtroom because he has a right to be present and the jury will see him anyhow. Trying the case outside of the Court, and I think that is what was being suggested, is not necessarily a satisfactory practice since you can't deprive a person of his right to come to the Court. Besides, I am not so sure it is a good idea, because, like in the medical malpractice field, it is very difficult to get objective experts, or specialists to sit as a jury. The argument that juries are sympathetic to people who are injured, is not true. Let me cite an example. I tried a case, for three and a half months against an automobile manufacturer on the design of one of their cars. I represented a quadraplegic girl, 17 years old, couldn't be prettier or sweeter and had everything going for her; couldn't move anything except her head. We had a lot of good testimony. We went to the jury. We lost. We didn't get a nickel. Evidently the automobile manufacturer had better expert testimony. What I am saying is that there are some jurisdictions that have emotional jurors, but there are many that do not.

Question: Referring to the cologne case, I was wondering if there were any other perfume manufacturers who indicated on the container that the product was flammable. If they did not, wouldn't the sued manufacturer have some recourse?

Mr. Heller: That is a good point, and some of the evidence showed that the sued manufacturer itself put a warning on its spray cologne - "Don't spray in the presence of flame". But there was no evidence that any of the manufacturers put "Flammable" on any of their perfume or cologne. Again, just because no one in the industry does it doesn't mean that the industry is safe. If it is unreasonably dangerous and the whole industry adheres to that standard, then any one of them is liable. In other words, what you are saying is that they were not negligent, true, but in addition to negligence there are other theories. I think Professor Hagglund might not have made it as clear as John did, that there are three theories of recovery; there is negligence, if you adhere to the standard of care of your competitors you are not negligent; but in addition there is breach of warranty, regardless of negligence; and there is strict liability, regardless. So strict liability means that if it is unreasonably dangerous even though everybody puts it out that way, you are liable.

SESSION VII

VALIDATION OF DESIGN (II)

-RELIABILITY-

Chairman: G. Hurt
Detroit Diesel, Allison Division

SAFETY ANALYSIS: QUALITATIVE, QUANTITATIVE AND COST EFFECTIVE

H. D. Wolf
Grumman Aerospace Corporation
Bethpage, New York 11714

Abstract: No system involving man can be 100% safe since he is not error-free. The objective of any safety analysis effort is to identify what to "worry" about in the system and how to take steps to eliminate that "worry". With the proper attention to (1) past history of similar systems; (2) development of safety controls in all program phases; (3) reviews of program efforts; (4) aggressive accident/incident analysis, corrective action and closeout; and (5) an educational program for all project personnel from corporate managers on down, a system can approach "failure-free" conditions. The various approaches to safety analysis are described in this paper, including preliminary, system and operational hazard analyses. The advantages and limitations of the several methods are discussed. Examples are used to emphasize specific points and to enhance the reader's understanding. Some ideas on safety feature decision-making are included.

Key Words: Accident prevention; hazard analysis; risks; safety.

Introduction: Prevention of accidents and incidents is the primary function of safety analysis. This means preventing personnel injuries and damage to equipment from the point where the idea is germinated until the ultimate user relegates it to the scrap heap. The techniques are simple, rigorous, time consuming; they can be applied to a railroad wheel, to a space vehicle or to a nuclear reactor. They require a different point of view from other system analyses. To provide the essence of safety analysis and describe what it can accomplish, this paper has been organized into two major sections: Safety Analysis Techniques and Risks/Benefits.

Safety Analysis: The primary tools of an early or preliminary safety analysis are the worksheet, a hazard log book, a checklist on hazards, a system description and a "challenging" nature. Using the work done by W. Hammer (Reference A) in categorizing potential hazards and historical information on how similar systems survived, each element of the system to be studies is challenged by each hazard category to determine which hazard category applies to which system element. The tentative results are listed on the worksheet, Figure 1, indicating where potential safety problems may be. Simplifying the first cut is done by constructing Essential Hazardous Situations (EHS); these are the fundamental safety issues to address. A logbook is established with files on each of these EHS; tracking of the hazard to closeout becomes essential once the design is frozen.

Identification of a hazard is only the first step. In a Preliminary Hazard Analysis, Figure 2, an iterative process is initiated where design approaches, safety devices, warning devices or crew procedures are used to eliminate, control or resolve the hazard situation. The method(s) for handling the hazard are compared with the user's requirements and resolved to the user's satisfaction. This can be done formally or informally. If a specification is used, the spec is changed to reflect the new design. At a logical point in the system evolution, a design review is performed. The purpose of this review, from safety's viewpoint, is to assess how and to what extent the system designers have satisfied the safety requirements. It is not meant to be a contest between safety and engineering as to how a design requirement was satisfied, but rather to answer the question: Does the selected design satisfy the "no personnel injury/no equipment damage" criteria? New hazards may be identified, which may require additional design effort.

At this point the safety program has accomplished several things:

1. identified general hazards to worry about; i.e. design boundary conditions

2. established additional design/operational requirements to include in the design.

3. identified where additional fertile areas for "improvement" can be found.

4. provided detailed procurement specification data/sections.

5. created a data bank that can subvert (and potentially support) any legal action taken in association with this design.

The logbook on hazards data is expanded and embellished as the design is refined and as more detailed information is developed. At specific points, the analyst can say, "I've carried the the hazard closeout as far as possible for this portion of the design." That means, for example, the specification includes the proper requirements or the detailed drawings include the proper material specifications. Ultimate closeout may be when the finished product successfully completes an acceptance test or the user successfully opens his beer can without cutting a finger.

For complicated systems, each segment of the design (system, subsystem, assembly, component, part) can be broken down into its basic elements and be subjected to a comparable hazard analysis. Figure 3 presents the next hazard analysis level, a System Hazard Analysis. Normally, this is conducted during the period prior to freezing a final design and starting final manufacturing and assembly operations. The technique used is similar to a Preliminary Hazard Analysis except the level of detail is greater because the analyst is digging more deeply into the basic design. To maintain traceability, hazards identified in one category are maintained in that category throughout the program. Resolution of hazards now may require more than just simple discussions with the cognizant engineer. Program management may be brought in as the cost of implementing a safety feature or redesigning may require a trade study to determine the cost/benefits. Just prior to the design freeze, another design assessment is conducted to set the stage for the final push to completion. The purpose at this point is to identify all known safety problems and establish a schedule, budget and approach to their resolution. One approach may be to "accept the risk." The ramifications of such a decision are financial, contractual, legal, and at times, political; stranding two American astronauts on the moon in full view of the world is not a tenable position for a company. If your customer can agree that such a risk is acceptable, you have at least shared the responsibility.

Data that have been collected on system safety problems serve the program in ways other than just design. The manufacturing and test personnel use the data bank as a source of information on things to be careful about during their efforts. Failed open valves that startle a user by their "noise" can also cause the test operator to jump if a system is pressurized and a valve fails. If the user can flick switches in an incorrect sequence applying power to the wrong circuit, so can the test man. It also works the other way. Problems that occur during test can be analyzed by asking the questions: What if this happened during normal use? How would the user react? What is the impact? Also, accident and incident analysis can identify generic problems whose resolution results in design improvements. This type of information is fed into the design change loop and also into the next type of hazard analysis: Operating Hazard Analysis.

People make mistakes; even the best trained, most highly motivated personnel can err. Operating Hazard Analysis (Figure 4) serves the purpose of identifying those situations where personnel "Failure" is likely. Such situations can stem from a design compromise that dictated "its too late to redesign now" to an operator limitation: man cannot exert x lbs. of force on an escape system handle under a centrifugal force of 2 g. Though the technique is similar to the other aforementioned analyses, the results are quite different. Design changes are highly unlikely at this point in the development, so crew procedures are used. The hazards identified are compensated by the crew. To do this they must know how to identify the situation (cues to hazard onset), the backout steps to take to prevent the situation if timely action is taken or, at least, to minimize the impact of the situation if it cannot be prevented. The output, then, of this analysis are these types of data:

- procedural steps
- procedure sequence constraints
- caution notes
- warning notes
- reference data bank: cues, clues and data
- support equipment, if required

Fed into operator manuals, training courses, crew procedures documents, test program plans and procedures and technical management planning documentation, human failure situations can at least be controlled even if they cannot be completely eliminated.

At this point in an idealized program, the design has satisfied safety requirements where feasible, the product has been manufactured and tested verifying the quality and the safety margins of the design, and the unit is ready to "go on the air." Complex systems are subjected to one more review: a readiness review. A blue ribbon panel is formed of customer and corporate "experts" to review the design, the test results, the closeout of open issues and the safety risks that have been identified and to render a judgement. Is the product really ready to go?

Special studies may be directed before go-ahead is given. A Warning Time Analysis may be one of them. Though normally done in parallel with the Systems and Operating Hazard Analyses, it can be done at any time. Time critical situations identified in prior analyses are compared with system and human response times. The worksheet is used to assure no hazard categories are forgotten and to help understand the causes and effects of these time critical situations. Judgments are made as to whether the design and procedures are adequate or not. In some cases the result is traumatic. For example, human reactions are too slow to initiate escape mechanisms if the fireball of explosion travels at speeds of detonating hydrogen. Automating such a system and assuring "No False Initiations" is a career unto itself. The results of the Warning Time Analysis are elimination of time critical sequences, automation of inadequate responses and identification of potential residual risks (see Figure 5).

Residual risks require constant re-evaluation. Prior to each opportunity for personnel injury or equipment damage, it is essential that attention is focused on the risks. User and program management must re-decide based on current data if the risk is still worth taking. It is obvious that these risk discussions can become thorns in management's side. Diplomacy is the only way to avoid rapid ejection from a program under such circumstances.

The various types of hazard analyses have been presented above. Related analyses, FMECA and Fault Tree, will be covered in the following papers. Several basic questions always asked about these hazard analyses are: What do we gain and why do it? Figure 6 summarizes the pros and cons of performing such analyses.

<u>Risks/Benefits:</u> The second part of this paper deals with some thoughts on safety risks and benefits. If we define Expected Loss Index (ELI) as the product of the probability of event occurrence and the dollar amount at risk, we have a potential decision-making tool for accepting/rejecting safety features. Let's assume for the moment that an acceptable safety risk is one that has a probability of occurrence of one in a million (10^{-6}). Figure 7 is a log plot of dollars at risk versus the probability of the event occurring. If we decide that an ELI of 1.0 is allowable, then Figure 7 is divided into Area A - event probability less than 10^{-6}; Area B - ELI less than 1.0; and Area C - ELI greater than 1.0. If, for a given design, the ELI is 1000, your goal is ELI=1 and the event probability is 10^{-4}; then your approach is either to reduce the probability of the event by eliminating or limiting the exposure to the hazard and/or reducing the dollars at risk, for example, by moving a high cost element into a low risk area.

There are several problems with this approach:

1. Price tags on men's lives are emotional issues and highly variable.

2. Dollars at risk include elements difficult to pin down: user costs, training costs, test costs, etc.

3. Event probabilities are only estimates which range from "substantiated" by accident data to "best guess."

To resolve this question, when is it beneficial to add a safety feature or make a change for safety reasons, the emotional element of the cost of men's lives must be eliminated and a realistic accident data base must be constructed to improve event probability guesses.

Conclusions: Safety hazard analyses are another tool set in the kit of the assurance disciplines. They are capable of identifying boundary design conditions, controlling the safety risks, alerting management and supporting any corporate actions to resolve legal suits. Methods and techniques for deciding when it is beneficial to add a safety feature are still in the early development stages.

References:

(A) Hammer, Willie, "Progressive Qualitative Hazard Analyses" AIAA Fourth Annual Meeting and Technical Display No. 67-935, October 1967.

WORKSHEET

(SYSTEM ELEMENTS IDENTIFIED)			
HAZARD	COMPONENTS POS. AFFECTED	CRITICAL OPER. SEGMENTS	REMARKS
1. ACCELERATION			
2. CONTAMIN-ATION			
3. CORROSION			
4. DISLOCATION, CHEMICAL			
5. ELECTRICAL			
6. EXPLOSION			
7. FIRE			
8. HEAT & TEMP.			

(21 CATEGORIES OF HAZARDS)

FIGURE 1

PRELIMINARY HAZARD ANALYSIS

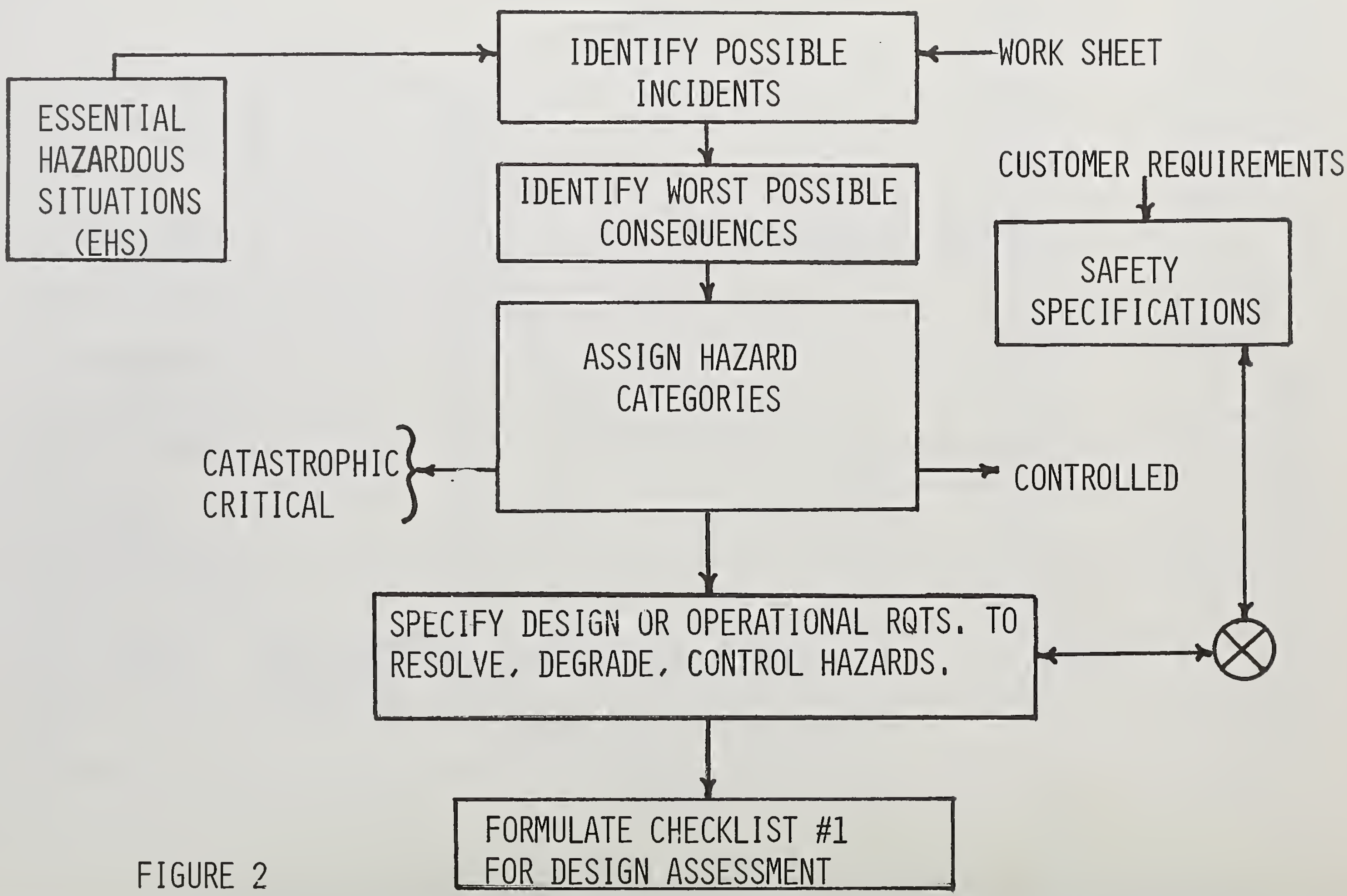


FIGURE 2

SYSTEM HAZARD ANALYSIS

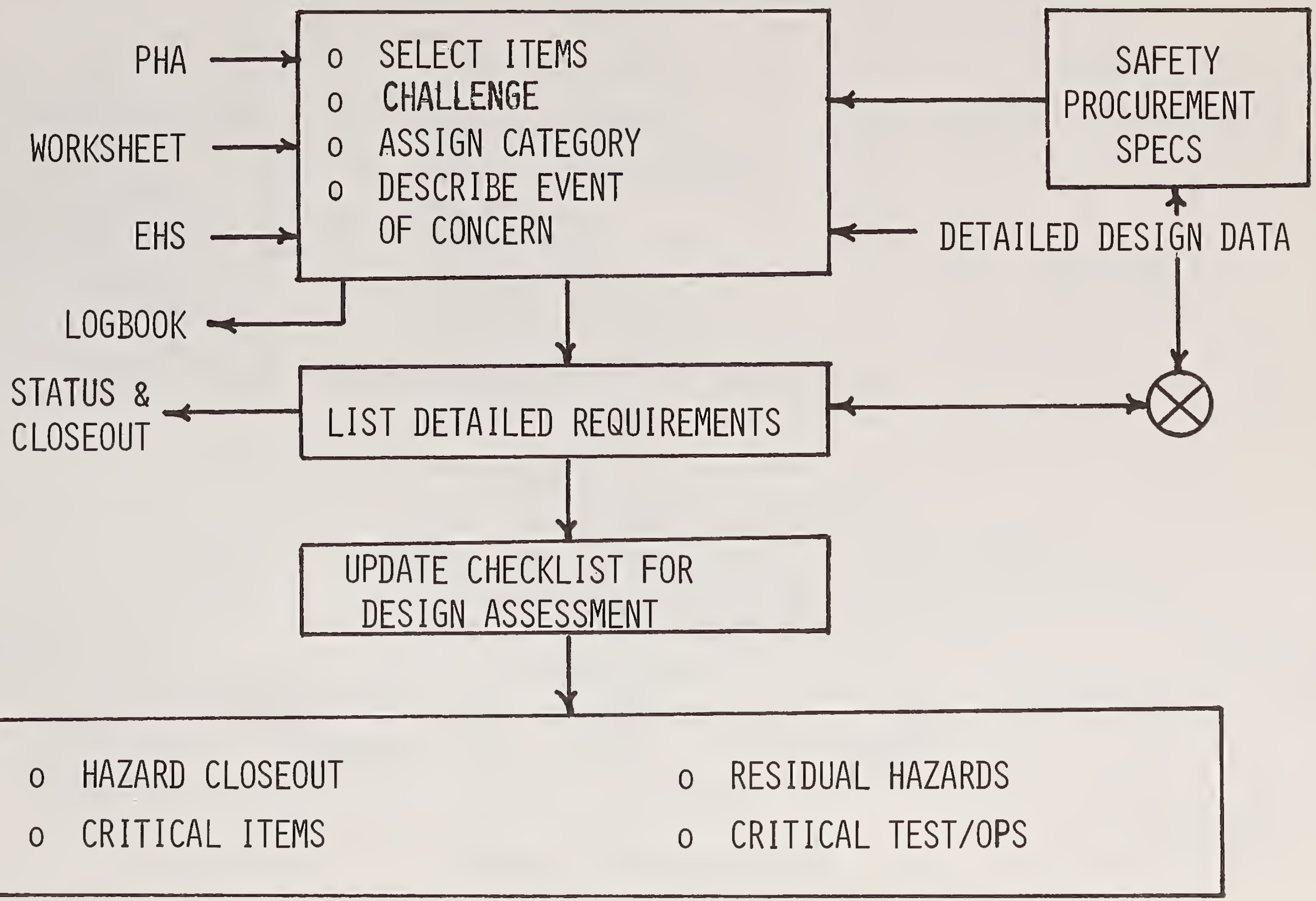


FIGURE 3

OPERATING HAZARD ANALYSIS

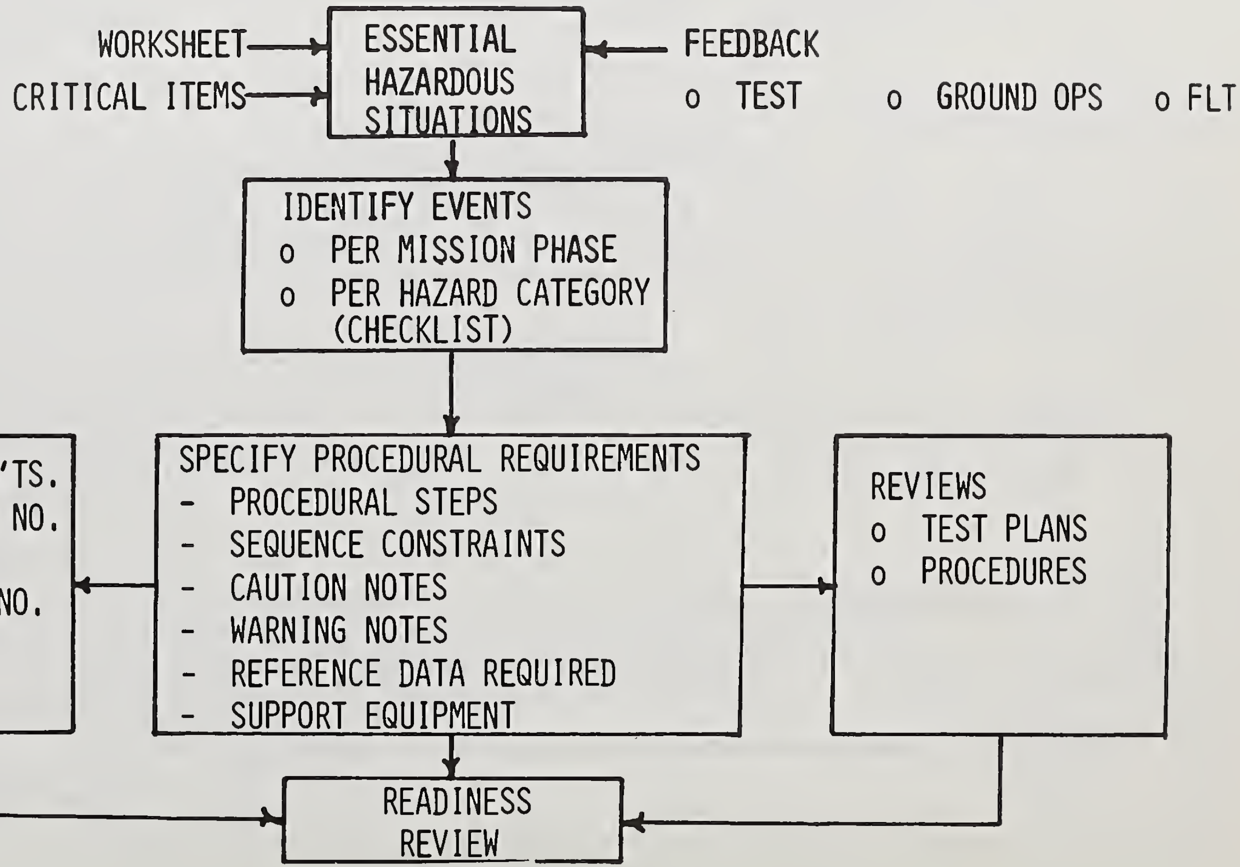


FIGURE 4

WARNING TIME ANALYSIS

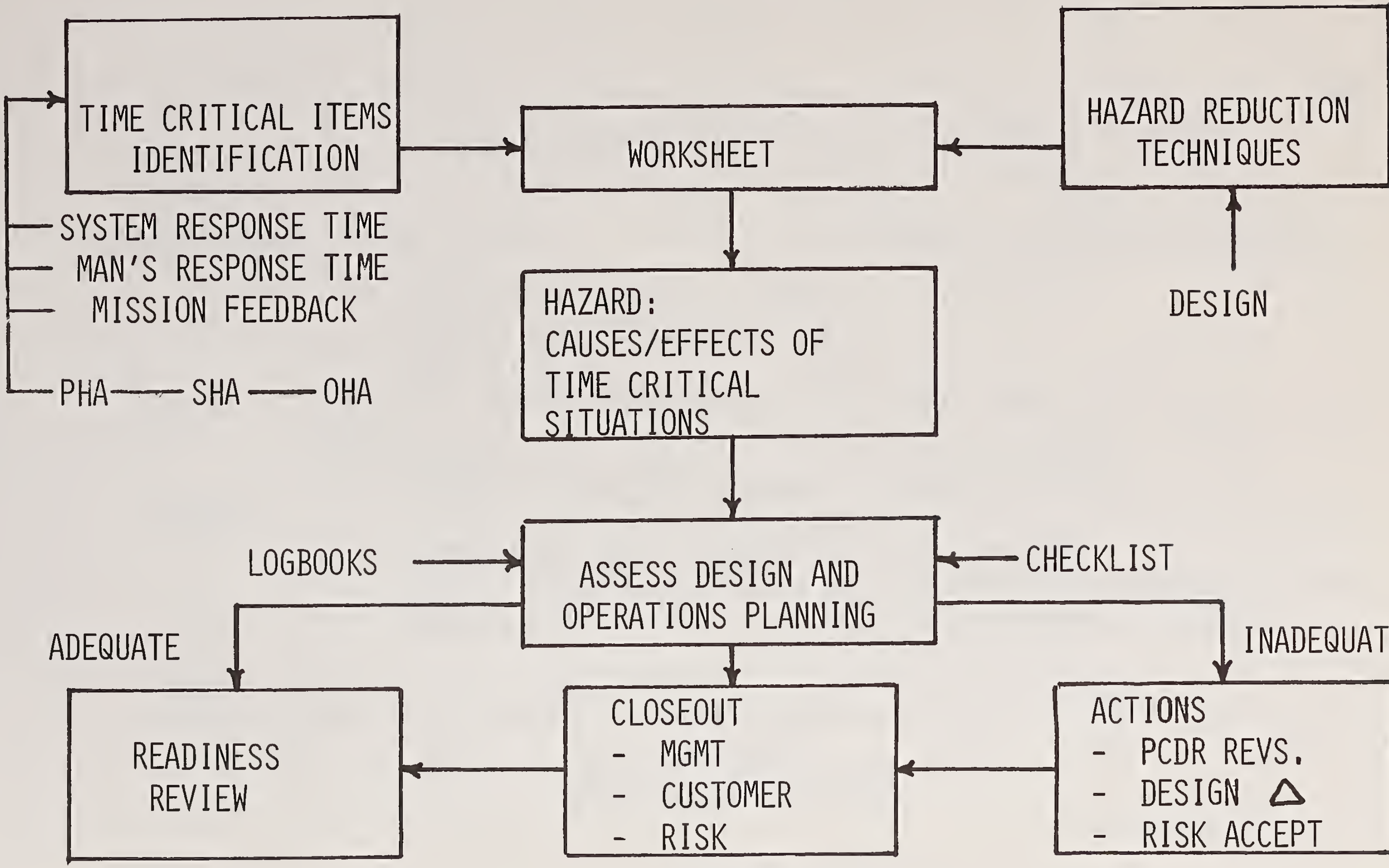
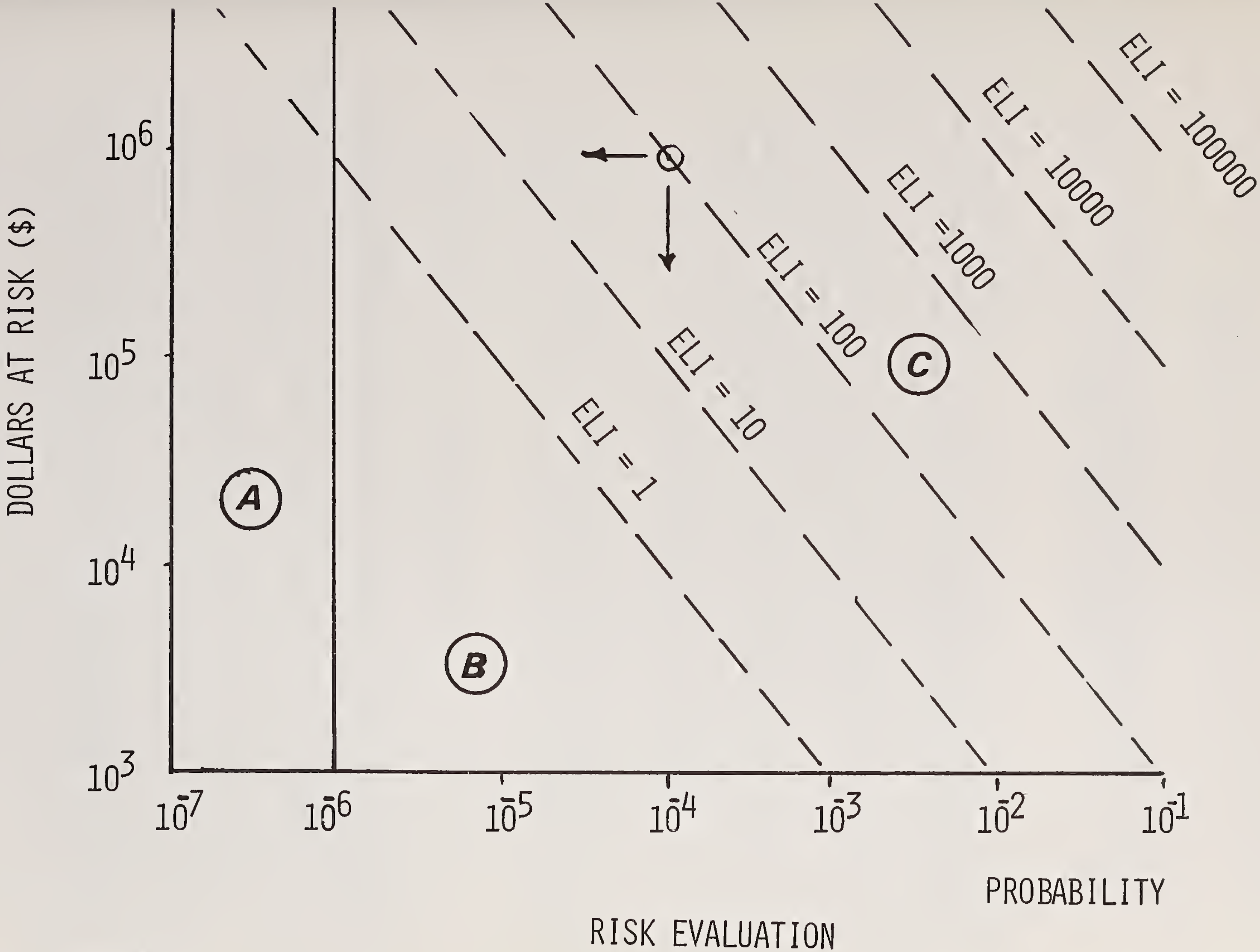


FIGURE 5

HAZARD ANALYSIS ASSESSMENT

FIGURE 6

FACTOR	PRO	CON
RIGOR	o QUALITATIVE o CHECKLIST LIMITS FORGETFULNESS o STANDARD PROCEDURES; ITERATIVE	o DEPENDS ON ANALYST CAPABILITY o HAZARD CATEGORIZATION NOT OPTIMUM o TIME CONSUMING
VALIDITY	o ACCIDENT DATA SUPPORTS FINDINGS o ELIMINATE ONE ACCIDENT	o ANALYST BIASES RESULT o PAY FOR INSURANCE
OUTPUT	o DESIGN BOUNDARY CONDITIONS o SAFETY REPORT CARDS o ASSURANCE HAZARDS CONTROLLED o OPERATING CONSTRAINTS	o PAPER, PAPER, PAPER
SAFETY PROGRAM SCHEDULE	o DO IT EARLY	o SAFETY CAN BECOME RUBBER STAMP
LABOR COST	o ¼ - ½% OF ENGINEERING HOURS	o 1% OF ENGINEERING HOURS



DOLLARS AT RISK ($)
10^6
10^5
10^4
10^3
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
PROBABILITY
RISK EVALUATION
ELI = 100000
ELI = 10000
ELI =1000
ELI = 100
ELI = 10
ELI = 1
A
B
C

FIGURE 7

A FAILURE MODE AND EFFECT ANALYSIS PROGRAM TO REDUCE MECHANICAL FAILURES

Richard R. Landers
Reliability and Maintainability Department
Gould Ocean Systems Division
Cleveland, Ohio 44117

Abstract: Described is an interdisciplinary technique to reduce the number of mechanical failures in an ongoing torpedo program through the use of failure mode and effect analysis techniques. The paper describes the procedure and forms for completing the analysis and provides representative examples. FMEA's are performed at the component level early in the design phase to find possible ways that equipment can fail and to determine the effect of such failures on the system. The FMEA is used to assure that: all component failure modes and their effect have been considered and either eliminated or controlled; information for design reviews, maintainability analysis and quantitative reliability analysis is generated; data for maintenance and operational manuals are provided; and that inputs to safety and hazard analyses are available in a Critical Item List (CIL).

Key Words: Failure mode and effect analysis; reliability; failure reduction; critical item list; system safety.

INTRODUCTION

The old maxim, "an ounce of prevention is worth a pound of cure" is quite applicable to today's products. Mechanical failures can be prevented through the application of fairly recently developed techniques. One such technique is the failure mode and effects analysis, which is usually abbreviated into the acronym FMEA. This analysis method provides for early evaluation of a design and, as applicable, generation of recommendations for design improvement.

This paper is organized in the following manner. The following major section is written as a standard practice as if it were to be put in a organizational manual. The second major section provides an example of an FMEA to illustrate the method. The paper concludes with a summary of observations regarding the practice and application of this analytic technique.

STANDARD PRACTICES FOR FAILURE MODE AND EFFECT ANALYSIS

PURPOSE

The failure mode and effect analysis (FMEA) is a design technique to provide a degree of assurance that an equipment and system are safe and reliable. The design office specifies which equipments/systems are to be subjected to an FMEA.

DEFINITION OF FMEA

A failure mode and effects analysis is a reliability design evaluation procedure initiated early in the design phase and periodically updated which:

a. Documents possible failure modes in a system design which fall within specified ground rules.

b. Determines by analysis the effect of each identified failure mode on system operation.

c. Identifies items critical to mission success and safety.

d. Assists in the evaluation of design features, such as redundancy, failure detection, override, and fail-safe characteristics.

e. Generates recommendations for design improvements or contingency options.

f. Provides an input to reliability analysis and design review data packages.

g. Allows priorities to be met for testing, procurement, inspection, etc.

h. Provides information for fault isolation and maintainability analysis.

i. Classifies each potential failure effect with a criticality category and assigns a probability of occurrences for the potential failure.

j. Provides a data base for subsequent comparison in a closed-loop failure reporting and corrective action program.

k. Provides information for system safety analyses.

l. Provides information for operation and maintenance manuals, handbooks and training programs.

Thus, the FMEA is an engineering technique which allows evaluation of critical failure sources by predicting the observable reaction of an item to a failure within it and the effect which the failure has on performance and operational requirements. A failure is defined as the inability of an item to perform its required functions within specified limits. (Defects and discrepancies that do not adversely affect function are not considered as failures.) A failure mode is the physical description of the manner in which a failure occurs (e.g., open, leak, structural failure). Figure 1 shows a typical completed FMEA form.

RESPONSIBILITIES

The design office is responsible for the preparation of FMEA lists. These lists identify by work breakdown structure the items requiring FMEA. Design engineering is responsible for reviewing and approving the FMEA's. Reliability Engineering reviews the FMEA's and prepares a critical item list (CIL) for subsequent hazard analyses. As the system becomes operational, field results are compared with analyses and adjustments made as required. Figure 2 shows an FMEA procedures flow chart.

PROCEDURES

1. Establish Ground Rules

   FMEA's generally are prepared by a number of persons. Therefore, it is important that a comprehensive set of rules be provided to assure consistency and completeness. Among these rules are treatments for:

   a. Operational and environmental requirements.
   b. Lowest level of hardware tier to which analysis is to be performed (generally down to the piece part level)
   c. Work breakdown structure and other coding systems for FMEA item identification.

2. System Description

   A functional statement is prepared for each system, a subsystem and assembly being analyzed giving a narrative description of the operation of each item for each operational mode. Schematics are included. A system block diagram is prepared showing series and redundant relationships along with inputs and outputs. Separate block diagrams are prepared for each operational phase or mode as applicable.

3. Potential Failure Mode Analysis

   Potential failure modes for each hardware item are analyzed and their effects on the system determined. The FMEA considers multiple failures only where functional redundancy and/or component redundancy exists. Interfaces within redundancy and isolation techniques for redundant elements are analyzed to ensure that the desired redundancy is not negated due to a failure of any of these interfaces or isolation techniques. In addition to hardware failure modes analyses, the FMEA includes an analysis of the effects due to loss of inputs. The analysis provides for each failure mode:

a. The effect of the failure on the operation of the subsystem and complete system.

b. The indication(s) of failure presented to the operator(s) or maintenance personnel during use or checkout.

c. The failure isolation and repair procedure required in use and checkout and corrective action available to operator(s) and maintenance personnel.

d. The controls or features in effect to eliminate or reduce failure mode frequency and/or criticality of occurrence.

Analyses for various products would likely require varying degrees and kinds of the above items. For example, in FMEA's prepared for aircraft it is desirable to include information detailing: means of indication of failure to crews; means to isolate faults; other failures that produce similar indications; variation due to flight phase; etc. It may also be desirable to have a preliminary analyses early in the design stage and then modify the analysis as more detailed information becomes available.

4. Criticality Categorization

Each failure effect is categorized into one of the following:

a. Critical - Loss of life or a disabling accident or extensive damage to equipment resulting in the inability to perform the mission.

b. Major - Not critical, but loss of reserve capability or early (but safe) termination of mission.

c. Minor - Not major, but causes a non-serious delay in start or completion of mission or an unscheduled adjustment or other action.

Specific characteristics of a program probably require modification of the above categories to better describe and tailor criticalities to the product and its use.

5. Loss Frequency

A frequency (or probability) of loss is assigned for each failure mode. In most cases this frequency or probability is simply the hourly failure rate of the component part in its operational mode and environment. In some cases it may be desirable (or sufficient) to merely categorize as "high," "medium" and "low." However, this simplified categorization should have defining rules for determining into which category a failure falls, such as greater than 1 chance in 10, 1 chance falling between 10 and 1000, less than 1 chance in 1000.

6. Critical Item List

Once the FMEA is complete, a critical item list (CIL) is prepared. The CIL identifies those failures whose effects exceed a predetermined threshold of seriousness. Generally, this threshold is the product of the criticality and loss frequency. For example, if there were three criticality and three frequency classifications, there would be nine categories of seriousness. The CIL threshold could then, say, be set at the "major criticality and medium frequency" level and include all failure events at and above this level (i.e., all "critical" plus "medium and high major") as shown by the shaded portion of Figure 3.

The CIL becomes an input to the safety analysis effort. That is, each item on the CIL is examined in detail to determine the nature and extent of the safety hazards associated with the item. This analysis aims at the elimination or reduction and control of hazards.

EXAMPLE OF A FMEA

A somewhat abbreviated example of a FMEA is provided below. The example follows the six steps listed in the procedure section.

1. Ground Rules

   a. The purpose of the torpedo is to seek out and destroy assigned targets. The operational environment is 10 to 1000 feet depth, -55°C to 105°C, and a launch shock of 10 g. The torpedo is designed to meet its mission requirements without redundant elements and is not to pose a hazard to its operators or to other equipment.

   b. This FMEA is performed at the subassembly level and, as such, considers the effect of each piece part failure mode on neighboring items, the mission, personnel, and associated equipment. Possible modes of failure are considered in the activation, operational, handling and storage phases. The failure frequency and criticality criteria are presented in IS 2345.

c. A block diagram of the torpedo as it relates to this portion of the FMEA is shown in Figure 4.

2. Torpedo Steering Subsystem Functional Description

Four three-position servovalves are mounted on the manifold forward face. The servovalve controls the hydraulic fluid flow to and from twin hydraulic control piston cylinders. The servovalve receives a fin drive signal from the gyro control unit (GCU) which positions the valve as commanded. A combination of hydraulic fluid flow is mechanically converted into control surface movement, thus altering the course of the torpedo. Each piston pair is connected to a rocker arm which pivots on a rocker arm shaft which extends through a hole in an aft tailcone stabilizing fin. A control surface (fin) is spline connected to the rocker arm shaft. As rocker arm position varies, the linear transformer provides a fin feedback position voltage to the GCU. As the control surface is moved either port or starboard, the gear movement varies the output voltage of the linear transformer which is fed back to the GCU fin control servo loop.

CONCLUSION

Methods and details for preparing and applying FMEA's may vary from company and, indeed, from program to program within a company. The important consideration here is that the FMEA procedure be tailored to specific needs. To be avoided is the preparation of FMEA's as a ritualistic exercise where form is considered more important than content. Also to be avoided is a cookbook approach whereby previously prepared FMEA's are lifted in total with little consideration for currentness or applicability to altered requirements.

Experience indicates that FMEA's are best prepared as a cooperative effort between the design and reliability engineers. The design engineer generally is best qualified to identify failure modes and establish the effects. The reliability engineer, as an objective participant, usually can best evaluate the results and make beneficial recommendations. Also, the reliability engineer usually is in the best position to perform the expected frequency and criticality of occurrence portion of the analysis whereby the seriousness of the failure can be established and categorized.

When properly prepared, the FMEA provides an orderly, early documentation of product elements important to performance, reliability and safety. This provides a strong data base for program planning and a basis for design improvements or contingency options to eliminate or control the effect of failure.

FAILURE MODE AND EFFECT ANALYSIS

SYSTEM Torpedo P/N 2793 PREPARED BY R. Landers DATE 11/5/76

ASSEMBLY Steering P/N 9372 CHECKED BY J. Smith DATE 11/5/76

SUBASSEMBLY Hyd. Actuator P/N 2397 PAGE 5 OF 16

FAILURE MODE	FAILURE CAUSE	FAILURE EFFECT	CORRECTIVE ACTION TO MINIMIZE OR ELIMINATE FAILURE / EFFECT	CRITICALITY	FREQUENCY
1. Leakage	1. a. "O" ring damaged during assembly. b. "O" ring not compatible with skydrol fluid. c. Inadequate design for 3000 psi. d. Poor finish of sliding surfaces. e. Corrosion of surfaces. f. Electrolytic action between parts. g. Damage to "O" ring by excess friction from exposed rod.	1. a. Hydraulic fluid will leak into control transformer causing possible damage to electrical wiring. b. Severe leakage will deplete hydraulic supply.	1. a. To assemble, "O" ring must pass by sharp corner of orifice opening. Design change will be incorporated to chamfer sharp edges. b. EPR "O" ring compatible with skydrol. c. Proven gland standards utilized in design. Qualification program will further prove adequacy. d. 16 finish adequate. e. Exposed rod chrome plated. Skydrol will protect interior surfaces. f. Identical metals used throughout. g. Rod retracted high percentage of operating time. Little opportunity for excessive dryness.	Critical	High
2. Structural Failure	2. a. Fatigue. b. Inadequate safety margin.	2. a. Steering assembly will be unable to function within specified limits. Torpedo could go off course.	2. a. Fatigue analysis performed. All corners of adequate radius. b. Stress analysis proves adequate margin of safety. Test programs will prove structural adequacy.	Critical	Low

4-1285-10

F-1299 7-76

Figure 1. Typical Completed FMEA Form

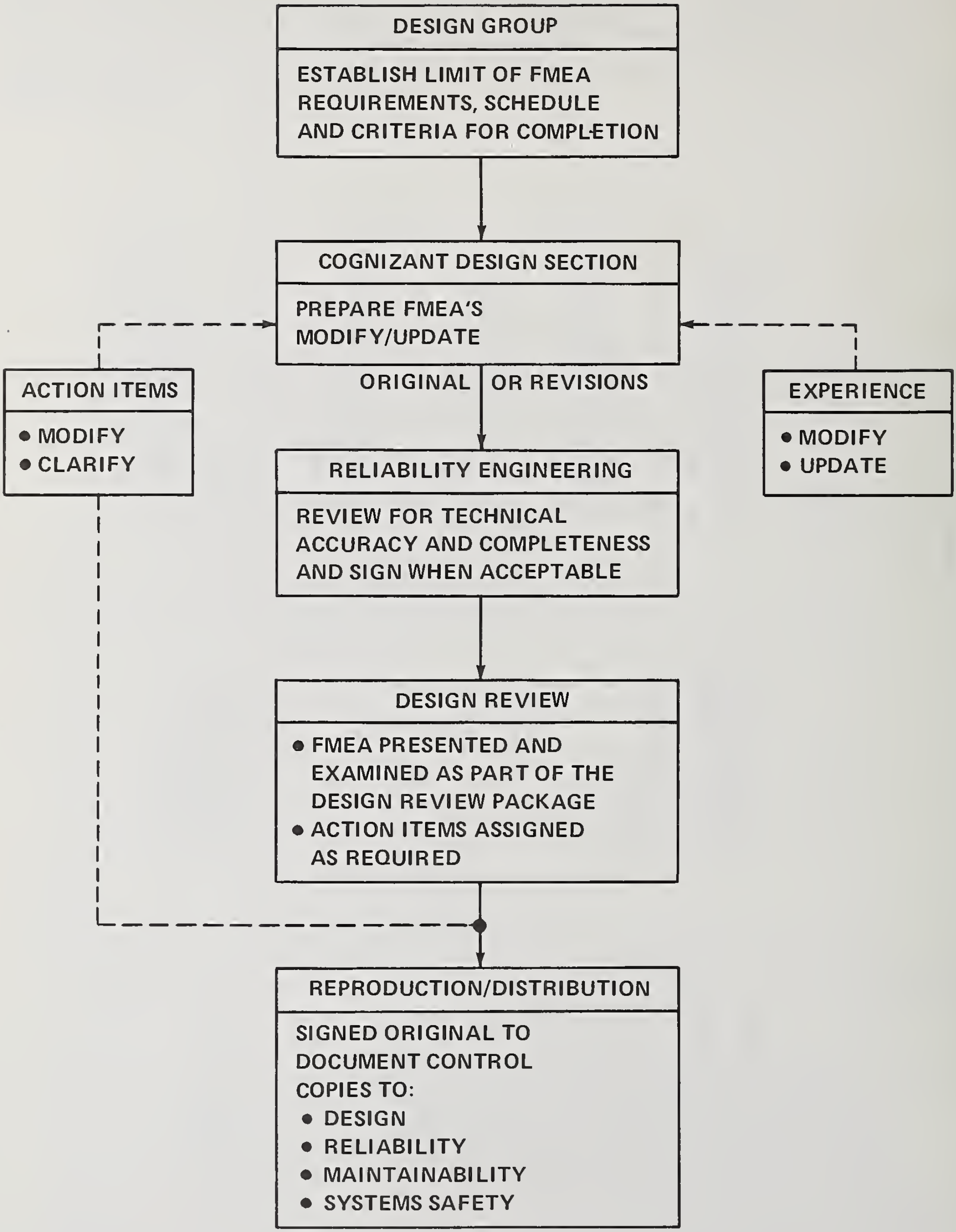


FIGURE 2. FMEA PROCEDURES FLOW CHART

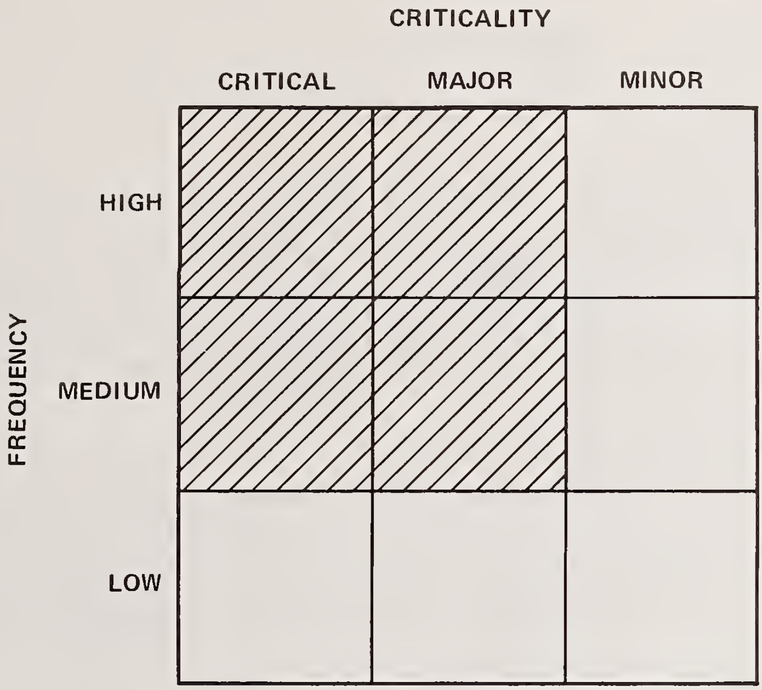


FIGURE 3. CRITICAL ITEM LIST CRITERIA

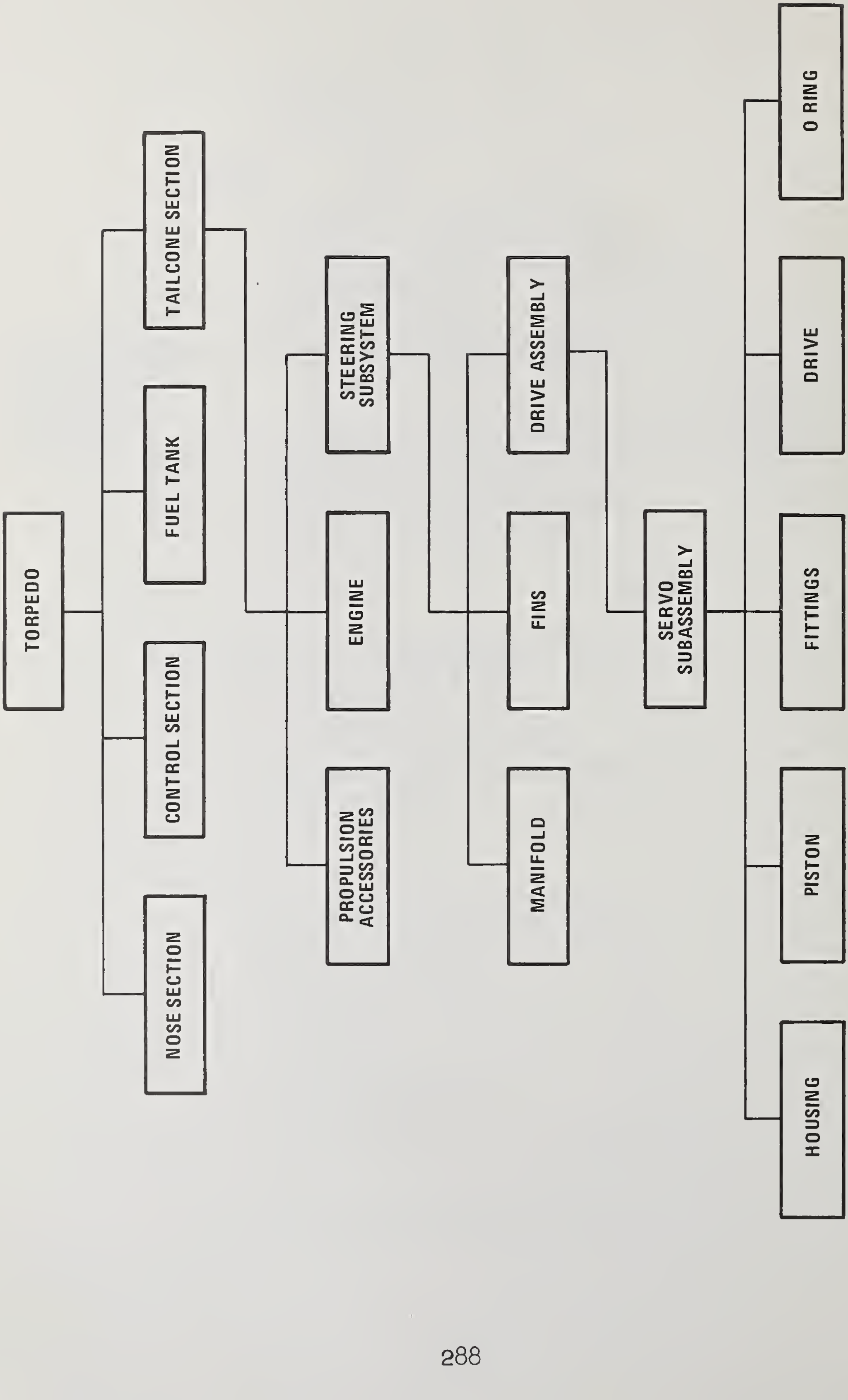


FIGURE 4. TORPEDO BLOCK DIAGRAM

FAULT TREE ANALYSIS AS A PART OF MECHANICAL SYSTEMS DESIGN

J. B. Fussell and D. P. Wagner
The University of Tennessee
Knoxville, Tennessee 37916

Abstract: Fault tree analysis is a method of system reliability and safety analysis that is rapidly gaining favor as a design tool. Fault tree analysis has the potential of offering an objective basis for analyzing system design, performing trade-off studies, analyzing common cause failures, demonstrating compliance with safety requirements, and justifying system changes and additions.

In this paper emphasis is placed on the use of fault tree analysis in mechanical system design with particular concern for:
(1) Introductory concepts
(2) Schemes for integration into the design process
(3) Limitations
(4) Related application

In addition, a review of the latest production computer programs is given.

Key words: Availability; fault tree analysis; probabilistic analysis; reliability; systems analysis.

1.0 INTRODUCTION

Fault tree analysis is a technique of reliability and probabilistic safety analysis and is generally applicable to complex dynamic systems. Fault tree analysis provides a versatile, mathematical tool for analyzing complex systems. Its application can include a complete plant as well as any of the systems and subsystems. Fault tree analysis provides a basis for analyzing system design, performing trade-off studies, analyzing common cause failures, demonstrating compliance with safety requirements, and justifying system changes or additions.

The logic of the approach makes it a visibility tool for both engineering and management. The fault tree method is concerned with assuring that all critical aspects of a system are identified and controlled. The fault tree itself is a graphical representation of Boolean logic associated with the development of a particular system failure, called the TOP event, to basic failures, called basic events. For example, the TOP event could be the failure of an automotive braking system to stop the vehicle during an emergency with the basic events being failures of the individual braking system components.

In 1961 the concept of fault tree analysis was originated by Bell Telephone Laboratories as a technique with which to perform a safety evaluation of the Minuteman Launch Control System [1]. At the 1965 Safety Symposium, sponsored by the University of Washington and the Boeing Company, several papers were presented that expounded the virtues of fault tree analysis [2]. The presentation of these papers marked the beginning of a widespread interest in the possibility of using fault tree analysis as a reliability tool. In the early 1970's progress was made in the solution of fault trees to obtain reliability information about relatively complex systems [3-7]. The collection and evaluation of failure data is still of the utmost importance [8-11].

Fault tree analysis is of major value in:

(1) Providing the analyst with insight into system behavior
(2) Pointing out the aspects of the system important in respect to the failure of interest
(3) Providing options for qualitative or quantitative system reliability analysis
(4) Providing a graphical aid giving visibility to those in system management who are removed from the system design changes
(5) Allowing the analyst to concentrate on one particular system failure at a time.

The cost of development in first-time application to a system is substantial. As in the development of engineering drawings, the cost is somewhat offset by future application of the models in accident prevention, maintenance scheduling, and system modifications. The expense is justified by detail of the qualitative or quantitative analysis resulting from fault tree analysis. Another aspect of fault tree analysis that limits its application at this time is the relatively few persons skilled in its techniques. Even skilled personnel might develop a fault tree for a given system in different ways.

Although certain single failures that can result in several component failures simultaneously, called common cause failures, can be pointed out by a detailed fault tree analysis, the analyst must properly account for common cause failures, specifically at some point in a detailed analysis. The analyst should be aware that fault tree analysis does not inherently ferret out all common cause failures. Additional limitations are discussed in Section 4.0.

2.0 CONCEPTS OF FAULT TREE ANALYSIS

2.1 Fault Tree Terminology

Fault tree construction is the logical development of the TOP event. As the construction proceeds, each fault event is also developed until basic events are reached. A fault event is a failure situation resulting from the logical interaction of basic events. The development of any fault event results in a branch of the fault tree. The event being developed is called the base event of the branch. The branch is complete only when all events in the branch are developed to the level of basic events.

Fault events are classified as either primary faults, secondary faults, primary failures, or secondary failures. A primary failure is defined as the change of a component from the working to the non-working state for which the component is held accountable, and repair action on the component is required to return the component to the working state. A primary fault is defined as the change of a component from the working to the non-working state for which the component is held accountable, and repair action on the component is not required to return the component to the working state, i.e., the component is "self-healing". Secondary faults and failures are the same as primary faults and failures except the component is not held accountable for its change of state.

The graphical symbols used in a fault tree fall, basically, into two categories: logic symbols and event symbols [1,12].

The logic symbols, or logic gates, are used to interconnect the events that contribute to the specified main event, or TOP event. The logic gates that are most frequently used in fault trees are the basic AND and OR Boolean expressions. The AND gate provides an output event only if all input events occur simultaneously. The OR gate, specifically the inclusive OR gate, provides an output event if one or more of the input events are present. More than one variation of AND and OR gates exist. A priority AND gate requires that the input events must occur in a specified order before an output is passed. An exclusive OR gate passes an output when any single input occurs, however, if more than one input event occurs, the exclusive OR gate does not pass an output. The third type of OR gate, the mutually exclusive OR gate, also passes an output when any single input occurs, however, the occurrence of more than one input must be physically impossible if the mutually exclusive OR gate is to be applicable.

The more frequently used event symbols are the rectangle, circle, and diamond. The rectangle represents a fault event resulting from the combination of more basic faults acting through logic gates. The circle designates a basic system component failure or fault input that is for the purposes of quantitative analysis usually assumed to be independent of all other events designated by circles and diamonds. The diamond symbol describes fault inputs that are considered basic in a given fault tree. The fault tree is not developed further, either because the event is of insufficient consequence or the necessary information is unavailable. In order to obtain a quantitative solution for a fault tree, both circles and diamonds must represent events for which reliability information is input. Events that appear as circles or diamonds are treated as basic events.

The triangles shown in Figure 1 are not strictly event symbols although they have traditionally been classified as such. The triangle indicates a transfer from one part of the fault tree to another. A line from the side of the triangle (transfer out triangle) denotes an event transfer out from the associated logic gate. A line from the apex of the triangle denotes an event transfer into the associated logic gate from the transfer out triangle with the same identification number.

The TOP event of a fault tree occurs when the system passes from the unfailed to the failed state. A given basic event can appear in many places in a fault tree. A fault tree representation is shown in Figure 2. An explanation of the symbols used in fault trees is given in Figure 1.

2.2 Problem Definition

System definition is often the most difficult task associated with fault tree analysis. Of primary importance is a functional layout diagram of the system of interest showing all functional interconnections and identifying each system component [a]. An example might be a detailed

---

[a] For some systems that are not hardware oriented such a diagram may not exist and, indeed, the fault tree itself can be the only feasible diagrammatic system representation. For the discussion presented the emphasis is directed toward hardware-oriented systems. However, the implications and terminology extend to all fault tree analyses.

electrical schematic. Physical system bounds are then established focusing the attention of the analyst on the precise area of interest. A common error is failure to establish realistic system bounds and thereby a diverging analysis is initiated.

Sufficient information must be available for each of the system components to allow the analyst to determine the necessary modes of failure of the components. This information can be available from experience of the analyst, from the technical specifications of the components, or from a library specifically for this purpose.

A further step in the system description is to establish the system boundary conditions. These analytical boundary conditions should not be confused with the physical bounds of the system. System boundary conditions define the situation for which the fault tree is to be drawn. A most important system boundary condition is the TOP event. For any given system, a multitude of possibilities for TOP events exists. The selection of the "correct" TOP event is sometimes a difficult task. The system initial configuration is described by additional system boundary conditions. This configuration must represent the system in the unfailed state. Consequently these system boundary conditions depend on the TOP event. Initial conditions are then system boundary conditions that define the component configurations for which the TOP event is applicable. All components that have more than one operating state generate an initial condition. System boundary conditions also include any fault event declared to exist or to be not-allowed for the duration of the fault tree construction. An existing system boundary condition is treated as certain to occur and a not-allowed system boundary condition is treated as an event with no possibility of occurring. Neither existing nor not-allowed system boundary conditions appear as events in the final system fault tree.

2.3 Fault Tree Construction

An example is given here to demonstrate some of the fundamental aspects of fault tree construction. The example system is a dual hydraulic braking system. The system physical bounds include the master cylinder assembly, front and rear brake lines, and the wheel cylinder and brake shoe assemblies. The system boundary conditions are:

TOP Event ≡ Loss of all braking capability
Initial Conditions ≡ Brakes in released position
Not-Allowed Event ≡ Failures due to effects external to the system
Existing Events ≡ Parking brake inoperable
Tree Top ≡ Shown in Figure 3.

Figure 3 reflects the inductive reasoning that all braking capability is lost if the brakes on all the individual wheels fail to operate.

From a knowledge of the components, the fault tree shown in Figure 4 is constructed. The event "right front brake fails to operate" occurs if the right front brake shoes fails to operate or the right front brake shoes are not applied. The event "right front brake shoes are not applied" occurs if the right front wheel cylinder fails to operate or the wheel cylinder is not activated. This logical development continues until the fault tree is completed to the level of basic events, as shown in Figure 4.

2.4 Qualitative Analysis

Qualitative analysis includes determining (a) the system modes of failure and (b) the components of the system that share an alliance such that they are candidates for a common cause failure. A system mode of failure, called a minimal cut set, is a group of basic component failures, called basic events, that are collectively sufficient to cause the TOP event to occur. The occurrence of each basic event in the minimal cut set is necessary if the occurrence of the TOP event is a result of the failure of the minimal cut set in question. The minimal cut sets are the minimum amount of information required to construct the complete truth table for the logic model. Computer programs for determining minimal cut sets from existing fault trees are available [3, 13,14,15].

A common cause failure [16] is a secondary cause of failure that is applicable to the development of more than one component malfunction. For the purpose of discussing qualitative analysis of common cause failures, definition of several terms is necessary.

A significant common cause event is a secondary cause that is common to all the basic events in one or more minimal cut sets. If, in addition, all the components represented by the basic events in that minimal cut set share a "common location", that minimal cut set is a common cause candidate. Components share a common location if no barriers are present that are capable of insulating the components from the secondary cause of failure. Components may share a common location irrespective of the physical distance separating them.

On occasion, a significant common cause event may not be specified for a common cause candidate, but rather the common cause candidate is identified based solely on a "significant common cause condition". A significant common cause condition is a condition that closely links all the basic events in the minimal cut set. For example, all components indicated by the basic events in a minimal cut set being produced by the same manufacturer is a significant common cause condition. The common cause candidate is then identified without concern about a common location. Other significant common cause conditions arise from components being tightly linked by a common location. Components in the same electrical circuit, chemical flow loop, or even tightly clustered in a cabinet can give rise to common cause candidates based on significant common cause conditions rather than as the result of specifying the secondary cause susceptibility and location of each component.

2.5 Quantitative Analysis

Since the introduction of fault tree analysis, the area receiving the most research and development effort has been the evaluation of fault trees [5,7,17]. The evaluation of a fault tree is obtaining reliability information about the TOP event and perhaps the minimal cut sets from the data supplied for the failure of the basic components.

Quantitative reliability analysis is of value for analysis of system design, demonstration of compliance with safety requirements, and justification of system changes or additions. From minimal cut sets and the component reliability characteristics, the reliability characteristics for a particular system failure can be calculated. Component reliability characteristics are completely described by their time-dependent failure rate (hazard rate) and their time-dependent repair rate. In practice, both failure rates and repair rates are often assumed to be constant.

From these component failure rates, component repair rates, and minimal cut sets, the time-dependent characteristics for a particular system failure that can be calculated include:

(1) Availability and unavailability
(2) Reliability and unreliability
(3) Rate of failure
(4) Expected number of failures
(5) Failure rate (hazard rate)
(6) Repair rate.

Availability and Unavailability. The availability is the probability that the system failure does not exist at some specified time in the future. The unavailability is the probability that the system failure does exist and numerically is equal to unity minus the availability. The general assumption used in the availability calculation is that no system component failures exist at time t = 0.

Availability is usually associated with repairable systems. In which case, the availability at time t contains no information about system failures before time t. However, the availability is a useful characteristic if the occurrence of the specified system failure is tolerable at least some expected fraction of the time. The average unavailability during a time interval is the expected fraction of the time the system failure exists. Therefore, the availability is a useful characteristic when system downtime has economic implications.

As an example of a system for which unavailability is a factor of merit, a building fire extinguishing system that will extinguish any building fire, given it functions, is considered. The unavailability is the probability that the system will not be available to extinguish a fire because a sufficient number of system components are failed. Although for the extinguishing system to fail to operate during a fire would probably be considered intolerable, the unavailability still has merit because the situation requiring the system to be available is expected to exist only a small fraction of the time. The unavailability at time t can be interpreted here as the probability that the extinguishing system is failed on demand.

In summary, availability and unavailability are:

(1) Generally reported only for repairable systems
(2) Useful characteristics if the occurrence of the system failure is tolerable at least some expected fraction of the time.

Techniques for calculating availability and unavailability are well documented and appear in Reference [13].

<u>Reliability and Unreliability</u>. The reliability is the probability that the system has experienced no failures to a given time t. The unreliability is the probability that the system has experienced one or more failures to time t and numerically is unity minus the reliability.

Reliability is generally a pertinent characteristic of both repairable and nonrepairable systems. For nonrepairable systems, the unreliability is numerically equivalent to the unavailability since if the system fails, it is thereafter unavailable. For nonrepairable systems, reliability is conventionally the reported quantity rather than availability.

Repairable system reliability calculations account for the possibility of a redundant component failing and being repaired without system failure necessarily occurring. Repairable system time-dependent unreliability, called the distribution of time to first failure by mathematicians, cannot be calculated exactly for the general case. Conservative, bounding approximations for repairable systems do exist, however, for the distribution of time to first failure [18,19].

System reliability is in general an interesting system characteristic and is the pertinent factor of merit when the system failure is considered intolerable. For catastrophic system failures, such as explosions, system reliability is reported; system availability has limited engineering meaning for such failures.

Rate of Failure and Expected Number of Failures. The rate of failure as a function of time is the expected number of failures per unit time at time t. The rate of failure should not be confused with the failure rate (a misnomer) to be discussed later. The changes in the rate of failure as a function of time are of interest from a qualitative point of view as well as from a quantitative point of view. Also, the rate of failure is used to calculate the expected number of failures.

The expected number of failures during a specified time interval is given by the integral of the rate of failure over that time interval. If the cost of each system failure is known, the expected number of failures multiplied by the cost per failure gives the expected cost (risk) during the time interval due to the specified system failure.

The expected number of failures is in a sense a more global system characteristic than either availability or reliability because it contains information about the system with respect to an entire time interval. Unlike probabilistic quantities, the expected number of failures for repairable systems can be greater than unity.

Failure Rate and Repair Rate. The failure rate as a function of time t is the probability of failure during t to t + dt, given no failures before t. The term failure rate is a misnomer; hazard rate is more fitting and is sometimes used. The term failure rate is so intrenched in the literature, however, that little chance exists that it will be discarded.

The time-dependent system failure rate completely describes the density of time to first system failure and, consequently, the distribution of time to first system failure (unreliability). If the repairable system failure rate is constant, the mean time between system failures is given by the inverse of the system failure rate.

The repair rate as a function of time is the probability of repair during t to t + dt, given no repair before t. The repair rate is a misnomer also for the same reason as is the failure rate. The time-dependent system repair rate completely describes the probabilistic system repair characteristics. If the repair rate is constant, the mean time to repair is the inverse of the system repair rate.

From the system failure rate and repair rate, the time-dependent system reliability, availability, and expected number of failures can immediately be determined. Also, the failure rate and repair rate supply the quantities needed to treat a system as a component of another system in a subsequent more global analysis.

3.0 INTEGRATION INTO THE DESIGN PROCESS

Qualitative and quantitative fault tree analysis results should be integrated into the design process and can provide valuable assistance in upgrading system design and performance. An organized approach to this integration of methods is shown by the flow chart in Figure 5.

As shown in the figure, the first reliability and safety models are formed after preliminary design and capability evaluation. These models are then analyzed qualitatively and quantitatively. Qualitative analysis provides information about the system modes of failure and potential common cause failure candidates. Single failures and potential common cause failures which cause an undesirable situation can be identified and corrected.

4.0 LIMITATIONS OF THE METHODOLOGY

The limitations of the system reliability methodology fall into two categories: limitations occurring during implementation and limitations in the theory.

In spite of the large effort that has been directed toward the theoretical aspects of system reliability methodology, many individuals feel that, in general, implementation of system reliability methodology is "still in the academic stages" [20]. Computer programs used to determine the minimal cut sets cannot always handle the large fault trees encountered in practice because of astronomical computer storage requirements or excessive computer execution time. Similarly, even if only several hundred minimal cut sets are considered to adequately represent the TOP event in question, exact quantitative solutions are not often possible (nor desired) because of computer restrictions.

Theoretical system reliability methodology is limited because of the binary assumption - either an event is failed or it is working. Degraded performances, other than totally failed, cannot be considered. Priority (sequential) AND gates and cold standby cannot be treated quantitatively by analytical solutions in the general case; but rather, Monte Carlo solutions must be used.

A methodology for quantitative treatment of secondary failures has not been published. A secondary failure is failure of a system component that is due to that component being forced to operate outside its design envelope, that is, failure due to excessive environmental or operational stresses. Fault trees that contain secondary failures are generally not coherent [21] because repair of all the basic events that have failed does not guarantee repair of the TOP event. In short, fault trees that display development of secondary failures are not coherent because removing the secondary cause of failure does not repair the component that has failed.

No exact solution for repairable system reliability exists for the general case. For TOP events that represent hazardous conditions, the reliability is often the quantity of interest rather than the availability. Common practice has been to consider all basic events nonrepairable and to use the resulting TOP event unavailability as an upper bound on unreliability. The unreliability is thereby largely overpredicted in most cases because the possibility of redundant basic event failure and then being repaired without the TOP event occurring is not taken into account.

Although considerable effort has been made to catalogue data that reflect the probabilistic failure characteristics of components [22-24] inapplicable or insufficient data continue to be a problem for the system reliability analyst. Fortunately, a great deal of information is available from a system reliability analysis without the benefit of good data. Qualitative analysis does not rely directly on data and is often the source of most of the useful information obtained from an analysis. Ranking components in order of importance is possible even if data are only relatively correct. However, good data are required to calculate the expected number of system failures, availability, and reliability. System reliability methodology suffered a major setback in the nuclear industry in the late 1960s because of poor or inapplicable data being applied to complex systems and an effort made to convince system management the quantitative analysis results were valid [25].

5.0 RELATED APPLICATIONS

System reliability methodology is applicable to a wide range of problems. Formal analysis methods are relatively new. The fault tree technique originated at Bell Laboratories in 1961 and was later improved by the aerospace industry with individuals from the Boeing Company making many published contributions [4,17,26]. Until about 1970, publicized applications of the technique remained almost entirely hardware oriented. Then in 1970, Crosetti and Bruce [27] announced application to administrative and management systems as well. Examples given included evaluations of marketing alternatives and optimization of freeway traffic control systems.

With system reliability methodology being applied to intangible systems, industry began to realize the full potential of the technique. By organizing pitfalls in a logical manner in a graphical representation, system management is better able to evaluate risk and communicate managerial reasoning to peers, supervisors, and subordinates. Decision making is consequently carried out more objectively, and a powerful aid to selling ideas results.

In 1973, a tome entitled MORT - The Management Oversight and Risk Tree [28] was made available. The volume is concerned with application of the fault tree technique to tangible and intangible systems as an aid to evaluating risk and to making management decisions. While the MORT methodology will undoubtedly be extended, the volume is a worthwhile addition to any management library. Recently, a large scale study was undertaken to evaluate the risk associated with nuclear power plants. This study, the Rasmussen study [29], made extensive use of system reliability methodology to evaluate the probabilities necessary to determine risk. The overall approach used in the Rasmussen study is subject of cause-consequence methods [30-32].

6.0 CONCLUSIONS

Fault tree analysis is a versatile reliability tool that has rapidly won favor with those involved in reliability and safety calculations.

Fault tree analysis is a relatively new subject, continuously developing and expanding. Concepts and techniques of fault tree analysis have been developed over the past decade and now results from this type analysis are important considerations in the design of many systems such as aircraft, ships and their electronic systems, missiles, and nuclear reactor systems.

Fault tree analysis is applicable to problems concerning tangible and intangible systems. Application of the technique to complex hardware systems encountered in practice is emerging from the academic stages. Considerable research is needed to complete the theory associated with fault tree analysis. Methods for more efficient analysis of fault trees must be formulated before system reliability engineering becomes a fully mature engineering science. In spite of the limitation of fault tree analysis, it remains the best approach available for a large number of the systems analysis problems encountered in practice.

Fault tree analysis can be the most simple or sophisticated analytical reliability tool depending on the needs of the analyst. Although the concepts of fault tree analysis are outstandingly simple, constructing suitable fault trees for complex systems requires thorough understanding of the principle of fault tree construction and evaluation as well as an understanding of the system being analyzed.

7.0 REFERENCES

1. Haasl, D. F., "Advanced Concepts in Fault Tree Analysis", System Safety Symposium, June 8-9, 1965, Seattle: The Boeing Company [a].

2. "System Safety Symposium", Proceedings of symposium sponsored by the University of Washington and the Boeing Company, Seattle, Washington, June 8-9, 1965 [a].

3. Semanderes, S. N., "ELRAFT A Computer Program for the Efficient Logic Reduction Analysis of Fault Trees", IEEE Transaction on Nuclear Science, Seattle, Washington, November 1970, p. 79.

4. Nagel, P., "Importance Sampling in System Simulation", IEEE Transactions on Nuclear Science, Seattle, Washington, November 1970, p. 101.

5. Vesely, W. E., Analysis of Fault Trees by Kinetic Tree Theory, IN-1330, (October 1969).

6. Vesely, W. E., "A Time-Dependent Methodology for Fault Tree Evaluation", Nuclear Engineering and Design, 13(2), August 1970.

7. Crosetti, P. A., Computer Program for Fault Tree Analysis, DUN-5508, (April 1969).

8. Crosetti, P. A., "Fault Tree Analysis with Probability Evaluation", IEEE Transactions on Nuclear Science, Seattle, Washington, November 1970, p. 132.

9. Greger, G. E., D. A. Snyder, P. D. Gross, "Description and Uses of a Critical Systems Data File for Nuclear Plants", The American Society of Mechanical Engineers, United Engineering Center, 345 East 47th Street, New York, New York 10017, January 9-12, 1972.

10. Yarosh, M. M., "Accident Analysis", Nuclear Safety, 3(4), 1962.

11. Leonardini, L., "The Third Reliability Meeting at Riso", Nuclear Safety, 11(4), July-August 1970.

12. Mearns, A. B., "Fault Tree Analysis: The Study of Unlikely Events in Complex Systems", System Safety Symposium, June 8-9, 1965, Seattle: The Boeing Company [a].

13. Vesely, W. E. and R. E. Narum, PREP and KITT: Computer Codes for The Automatic Evaluation of a Fault Tree, IN-1349, August 1970 [b].

14. Worrell, R. B., Set Equation Transformation System (SETS), SLA-72-0028A July 1973. Available from Sandia Laboratories, Albuquerque, New Mexico 87115, USA.

15. Fussell, J. B., E. B. Henry, and N. A. Marshall, MOCUS - A Computer Program to Obtain Minimal Sets from Fault Trees, ANCR-1156 August 1974, [b].

16. Fussell, J. B., et. al., A Collection of Methods for Reliability and Safety Engineering, ANCR-1273, April 1976 [b].

17. Michels, J. M., "Computer Evaluation of the Safety Fault Tree Model", System Safety Symposium, June 8-9, 1965, Seattle: The Boeing Company [a].

18. Esary, J. D. and F. Proschan, "A Reliability Bound for Systems of Maintained, Interdependent Components", Journal of the American Statistical Association, 65, No. 329 March 1970, pp. 329-338.

19. Ross, S. M., Reliability and Fault Tree Analysis, SIAM 1975, pp. 331-350.

20. Vesely, W. E., "General Modeling Techniques Applied in the Rasmussen Safety Study", Reliability and Fault Tree Analysis, SIAM 1975, pp. 775-804.

21. Esary, J. D. and F. Proschan, "Coherent Structures of Nonidentical Components", Technometrics, Vol. 5 1963, 191-209.

22. "Reliability Stress and Failure Rate Data for Electronic Equipment", Military Standardization Handbook, MIL-HDBK-217B. Available from the Chief Bureau of Naval Weapons, Department of the Navy, Washington, D.C. 20360, USA.

23. Failure Rate Data Handbook (FARADA), U.S. Naval Fleet Missle Systems Analysis and Evaluation Document #SP 63-470, April 1970. Available from Naval Weapons Station, Seal Beach, Corona, California 91720, USA.

24. Ablitt, J. F. , "An Introduction to the 'SYREL' Reliability Data Bank", UKAEA REPORT SRS/GR/14 1973.

25. Griffin, C. W., "The Fault Tree as a Safety Optimization Design Tool", ANS TOPICAL MEETING ON WATER-REACTOR SAFETY, CONF-730304 March 1973 [b].

26. Nagel, P. M., "A Monte Carlo Method to Compute Fault Tree Probabilities", System Safety Symposium, June 8-9, 1965, Seattle: The Boeing Company [a].

27. Crosetti, P. A., and R. A. Bruce, "Commercial Application of Fault Tree Analysis", Proceedings of Reliability and Maintainability Conference, Vol 9, 1970,pp. 230-244.

28. Johnson, W. G. MORT - The Management Oversight and Risk Tree, U.S. Atomic Energy Commission, SAN 821-2, UC-41 (February 12, 1973 [c].

29. "Reactor Safety Study", WASH-1400 (NUREG-75/014), October 1975 [b].

30. Nielson, D. S., "Cause-Consequence Diagrams", NARS PUBLICATION 2, December 1972 [d].

31. Nielsen, D. S., "The Cause/Consequence Diagram as a Basis for Quantitative Accident Analysis", RISO-R-5-71 May 1971 [d].

32. Nielsen, D. S., "Use of Cause/Consequence Charts in Practical System Analysis", Reliability and Fault Tree Analysis, SIAM 1975, pp. 849-880.

[a] Available from the University of Washington Library, Seattle, Washington.

[b] Available from the NTIS, 5285 Port Royal Road, Springfield, Virginia.

[c] Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402, USA.

[d] Available from the Library of the Danish Atomic Energy Commission RISØ, Dk-9000, Roskilde, Denmark.

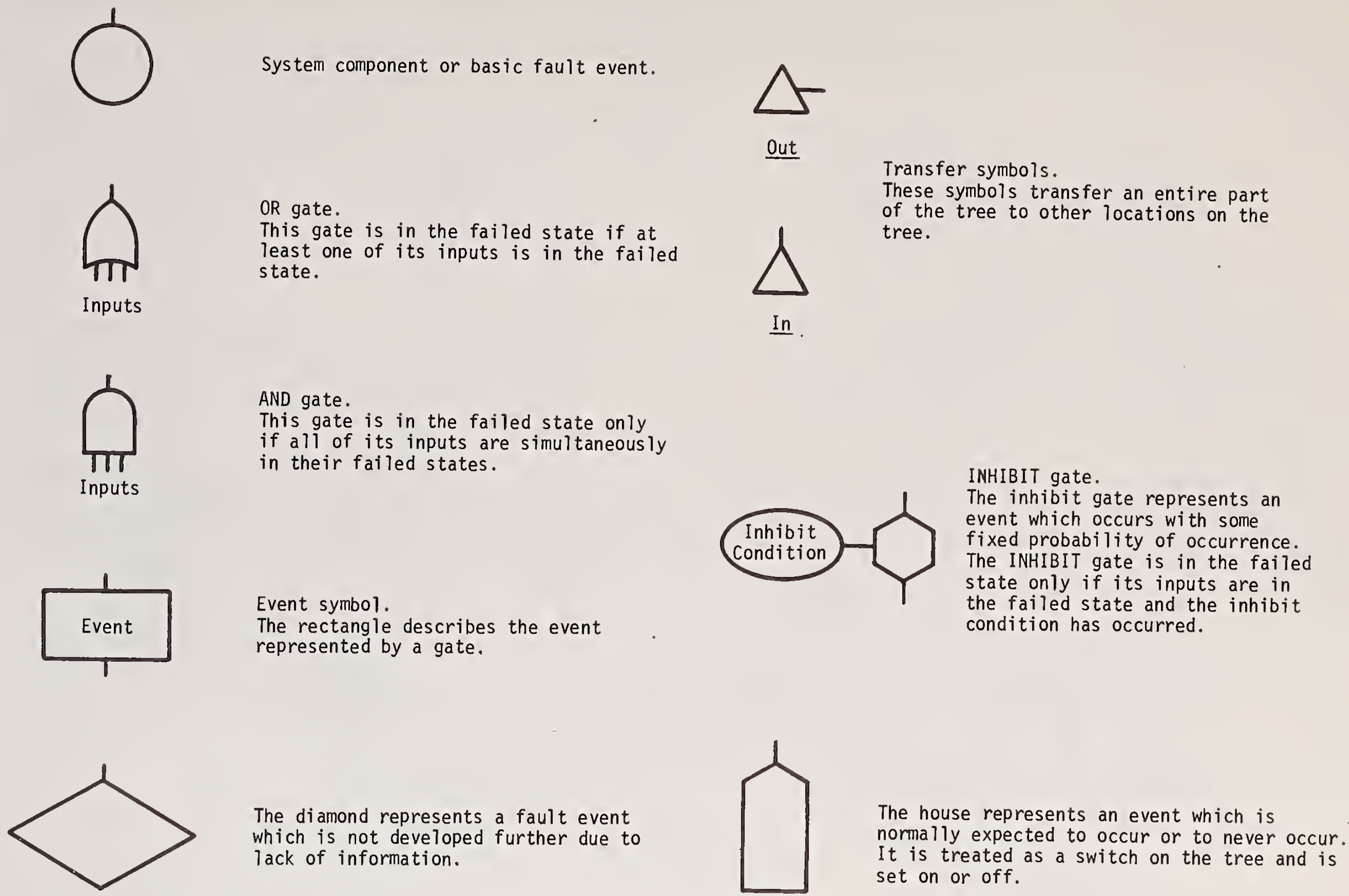


Figure 1: Fault Tree Symbols

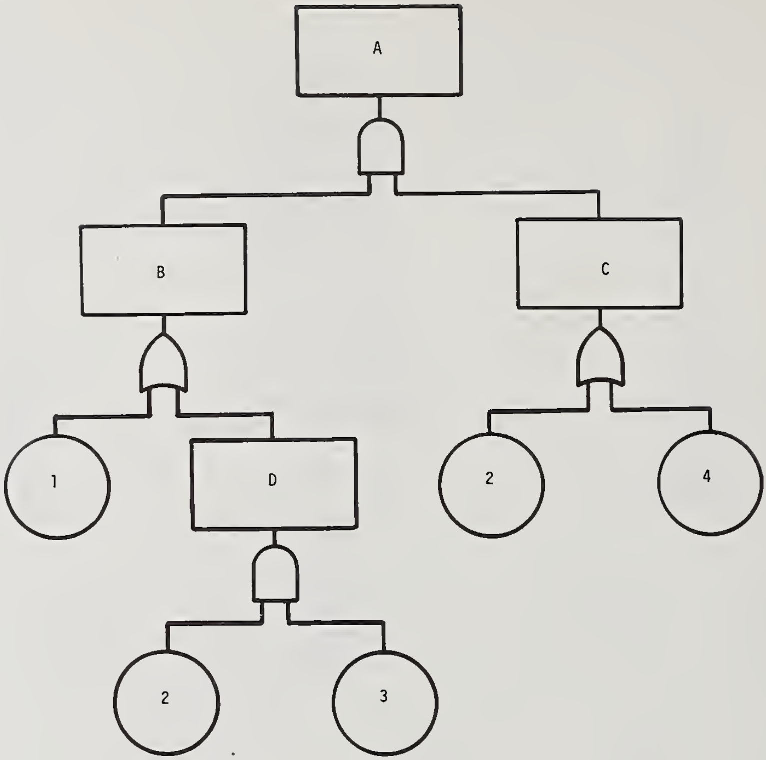


Figure 2: Sample Fault Tree Representation

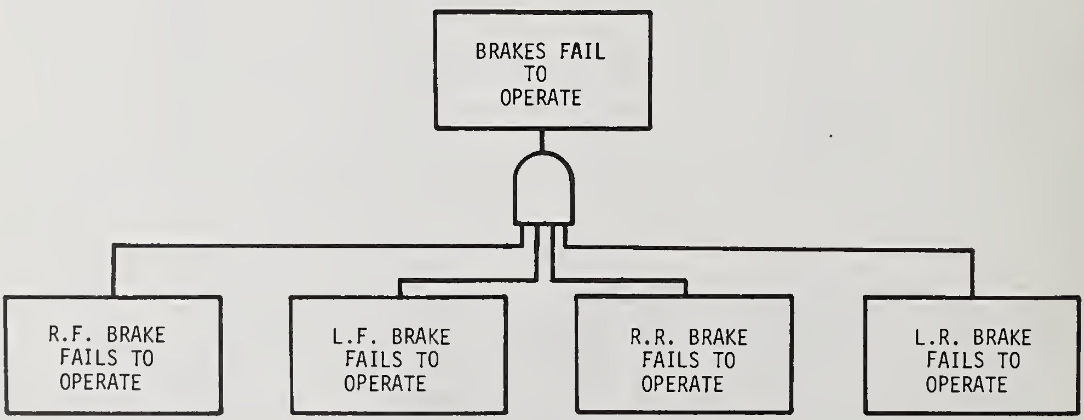


Figure 3: Tree Top Boundary Condition For Braking System Example

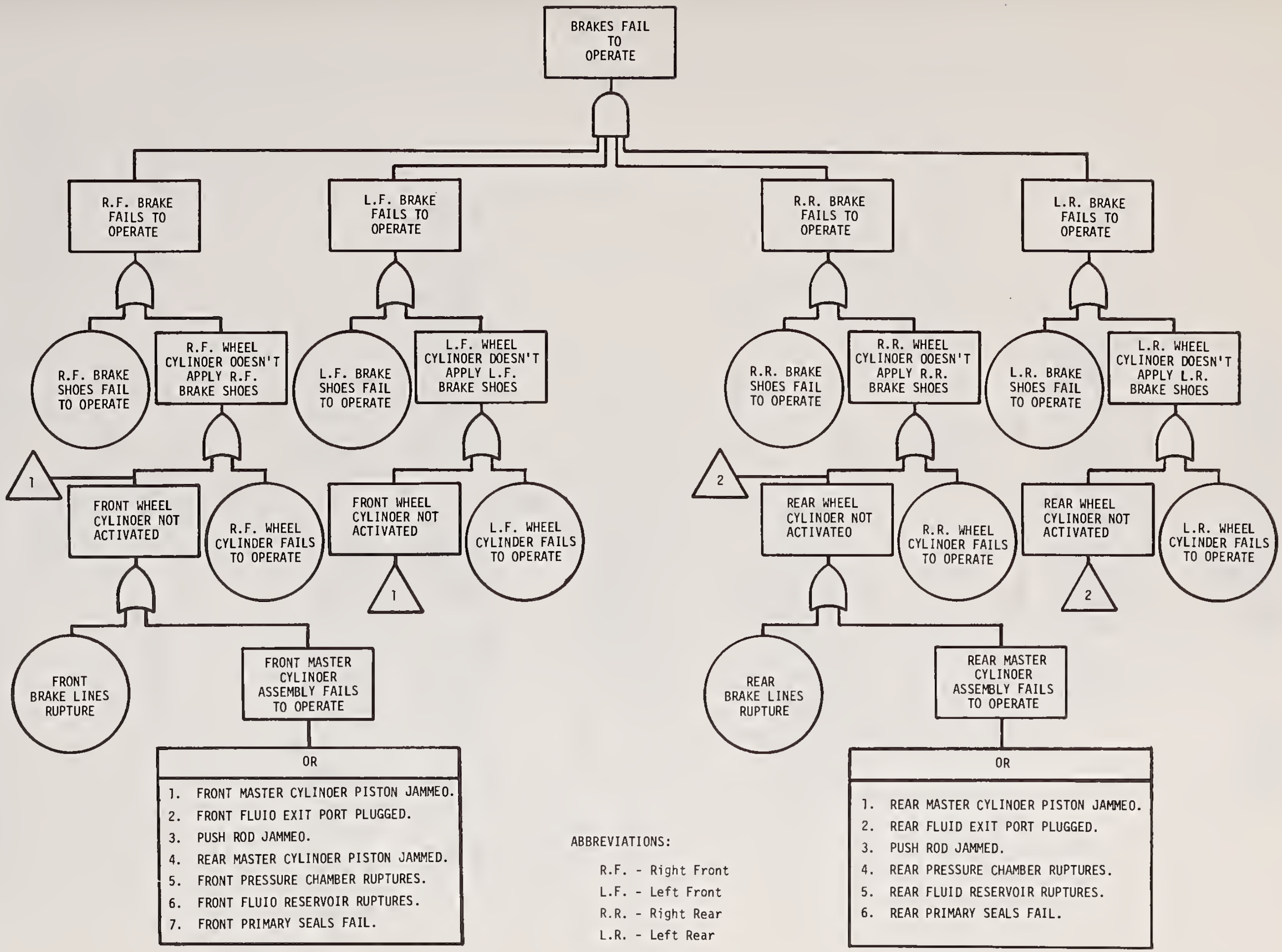


Figure 4: Fault Tree For Braking System Example

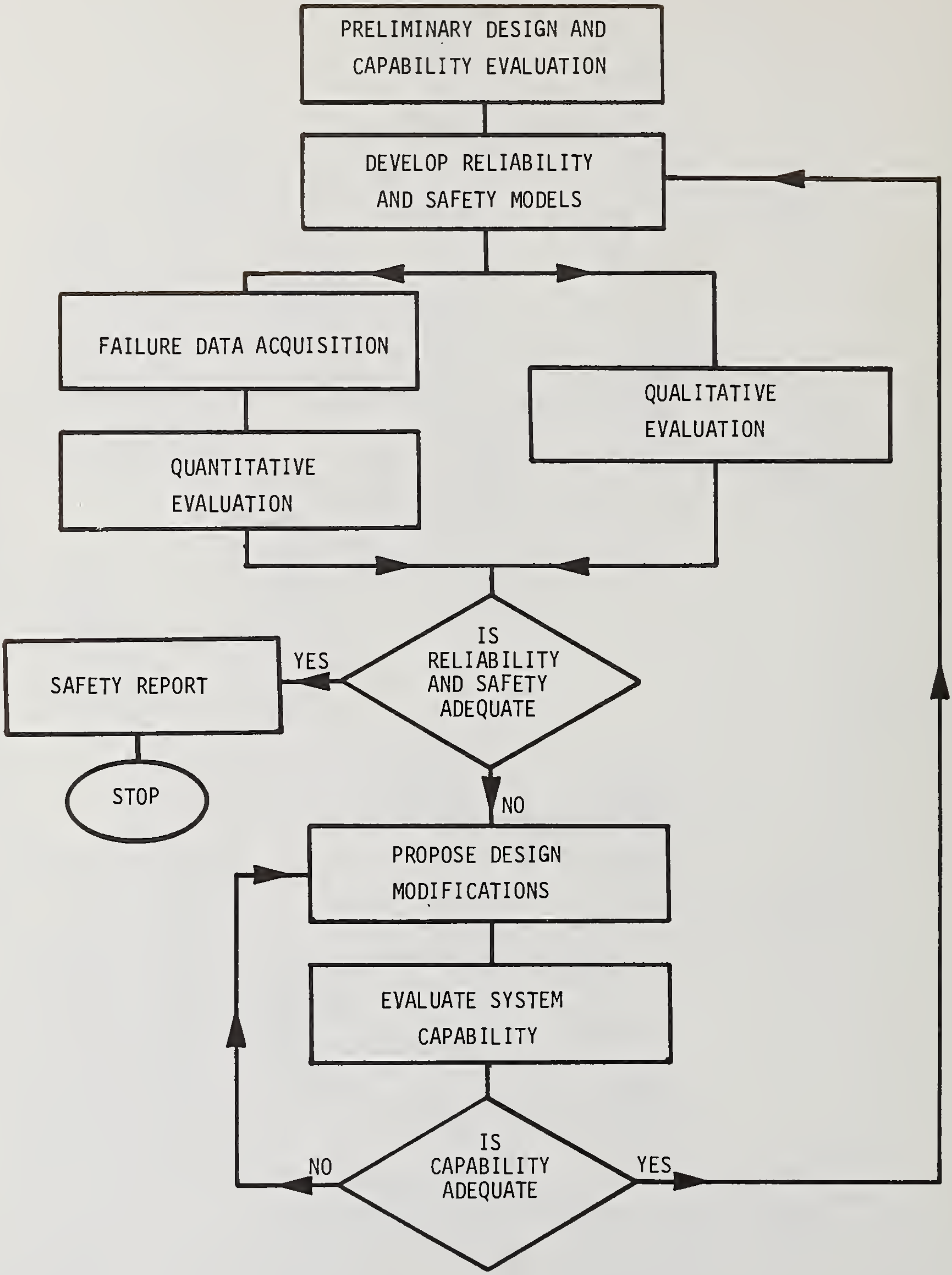


Figure 5: Flow Chart For Integration Of Fault Tree Analysis In The Design Process

SESSION VIII

MATERIALS OF DESIGN (III)

-COATING MATERIALS AND TECHNIQUES-

Chairman: A. K. Wolff

T. R. Mallory and Company

ION PLATING - CONCEPTS AND APPLICATIONS

D. M. Mattox
Sandia Laboratories, Albuquerque, N. M. 87115

Abstract: Ion plating is a generic term applied to deposition processes in which the substrate is subjected to a flux of high-energy ions prior to and possibly during film deposition. This flux of ions cleans the substrate surface to allow intimate substrate-film interactions. Ion bombardment during deposition may be used to control the properties of the deposit and to increase the "throwing power" of the deposition technique. Recent advances in the understanding of ion plating processes and applications will be reviewed. These include sputter cleaning, reactive plasma cleaning, gasless (or vacuum) ion plating, and several new source configurations. Related deposition processes, such as activated reactive evaporation, will be discussed and specific large scale and small scale industrial applications of ion plating will be detailed.

Key words: Adhesion; coatings; films; gas discharges; sputtering.

Introduction: Failure of materials often begins with surface defects. Surface processes such as corrosion, erosion, wear, microcracking, etc., can often be controlled by the proper application of coatings which protect by their resistance to failure or by sacrificial means. Coatings may be applied by a variety of techniques, each of which has its strong and weak points. Some desirable coating properties are good adhesion, high durability, low cost, and the ability to provide the functional properties desired. The coating process itself must not degrade the properties of the base material. Examples of materials degradation caused by the coating process are the hydrogen incorporation in high strength steels during electrodeposition giving hydrogen embrittlement and the microstructural changes in the base material resulting from the high temperature required for some chemical vapor deposition processes.

The ion plating deposition process was developed[1,2,3] to obtain good results in vacuum processing where adhesion was normally a problem because of poor film nucleation, barrier layers (contaminants or natural) or film/substrate material insolubility.[4,5] Ion plating is a generic term applied to atomistic film deposition processes in which the substrate surface is subjected to a flux of high energy ions sufficient to cause appreciable sputtering before and during interface formation. The ion bombardment may, or may not, be continued during film growth. Ion plating is usually performed with the substrates immersed in a gas discharge while a dc or rf potential is applied to the surface to be coated. Sources for the film atoms include[2,3]; (1) resistively-heated filaments, (2) electron beam heated crucibles (e-beam ion plating), (3) rf-heated crucibles, (4) sputtering cathodes (high-bias sputter deposition), (5) metal-bearing gases (chemical ion plating), and (6) sources in which vaporized atoms pass through an electron cloud or plasma to enhance ionization of the metal atoms.

From the standpoint of adhesion, the principal benefits of the ion plating process are: (1) cleaning is an integral part of the deposition process and the surface is maintained "clean"[6] until the interfacial region is formed, (2) a high energy flux to the substrate surface during deposition enhances diffusion and chemical reaction in the surface region without necessitating bulk heating, and (3) alteration of the surface and interfacial region in a manner conducive to good adhesion by creating high defect concentrations[7] developing a compressive stress[8] and creating a "pseudodiffusion"-type of interface.[5] In addition, it has been shown that ion bombardment during film growth can improve surface coverage, alter the morphology of the deposited film, and affect the physical properties of the coatings.[9]

Figure 1 shows a typical ion plating system using a thermal vaporization source and a dc diode gas discharge to provide ion bombardment of the substrate. In a typical operation, the precleaned substrate is placed into the vacuum system, the system evacuated to an ultimate of $< 10^{-6}$ Torr, the evaporant premelted and cooled with the shutter closed. The system is then backfilled with argon to a pressure of 0.1 - 1 ("bleed-back pressure") m Torr. The vacuum pump is then throttled to give a discharge pressure of 1 - 20 m Torr, and a dc gas discharge is established at 2 - 5 kV dc and $\sim$ 0.5 m A/cm^2 substrate current density. The surface is then sputtered for a period of time which depends on the material surface condition, and level of background contamination.[10,11] At the end of the sputter cleaning portion of the process, the evaporator filament temperature is raised until the film material begins to evaporate. A shutter between the source and substrate is then opened, allowing the material to deposit on the substrate. The important step is the formation of the surface-coating interface without interruption of the ion bombardment. Of course, for a film to form, the deposition rate must exceed the sputtering rate.

Figure 2 schematically depicts the processes that occur in a dc gas discharge used in ion plating an electrically conducting substrate. The primary source for ionization in the discharge gas is electron-atom collision $e^- + G^0 \rightarrow G^+ + 2e^-$. The gas ions that are produced may then become part of the plasma and ultimately lost to the chamber walls, or they may be accelerated across the cathode dark space to the cathode. Very few of these ions reach the cathode with the full cathode fall potential. In transit they undergo charge exchange collisions with gas atoms $G_1^+ + G_2^0 \rightarrow G_2^+ + G_1^0$, producing a spectrum of high energy neutrals and high energy ions, which bombard the cathode surface and surrounding fixturing.

TABLE 1

Film-Substrate Couples

Electrical Contact	Corrosion Coatings
Ag-Si	$A\ell$-U
Ag-Ge	$A\ell$-Steel
Cu-Si	C-Steel
Cu-Ge	Au-Steel
$A\ell$-Si	
V-Si/Ge	
In-CdS	
Cu-Steel	

Lubrication Coatings	Joining
Au-Steel	Cu-Mo
Pb-Steel	Cu-Ta
MoS_2-Carbon	Cu-Nb
Ag-Steel	$A\ell$-Be
	Ag-Fe
	Cu-Ti
	Cr-Mo
	Ag-Be

TABLE 2

Reactive Ion Plating	
CdS	HfN
Si_3N_4	HfC
TiC	$A\ell_2O_3$
In_2O_3	ZnTe
TiN	

The high energy atoms/ions that impinge on the cathode may (1) produce secondary electrons, (2) sputter cathode atoms (S), (3) sputter surface contaminants (sputter cleaning), (4) become neutralized and reflect from the cathode surface as high energy neutrals or as metastable atoms (G^*), and/or become incorporated into the cathode surface modifying the surface properties. In ion plating, the impinging atoms may sputter film atoms (M).

The secondary electrons that are produced are accelerated back across the cathode dark space, causing ionization which sustains the gas discharge.

The sputtered surface atoms (film, substrate, or contaminant) can be (1) scattered back to the cathode, (2) ionized by electron-atom collision or metastable-atom collision $S^0 + G^* \rightarrow S^+ + G^0 + e^-$ (Penning ionization and accelerated back to the cathode or lost from the cathode region, and (3) transported as high energy neutrals to some surface (sputter deposited). Penning ionization is probably the most important ionization process for sputtered cathode atoms. The energy distribution of sputtered atoms is higher (1 - 100 eV) than that of thermally vaporized atoms (0.2 - 1 eV). The return of sputtered atoms to the cathode surface is a function of many discharge parameters[10] as well as the system throughput. This return is important in sputter cleaning,generation of surface features (such as cones and whiskers), surface coverage and also in the sputter deposition of alloy material. It has been estimated that at 0.1 Torr gas pressure, 90% of the sputtered atoms are scattered back to the cathode. It should be noted that only neutral or negatively-charged particles can escape from the cathode surface.

The gas atoms that are incorporated into the cathode surface may affect the surface properties, such as sputtering yield (number of cathode atoms sputtered per incident ion). The gases trapped in the surface during sputter cleaning may be released after film formation and cause poor film adhesion. Gases trapped in the surface or deposited film may be released by heating.[12]

If the sputtering gas (or impurities in it) can react with the cathode surface, the ion bombardment can be used to form a compound or alloy surface. This makes sputter cleaning the surface of a reactive metal difficult if a reactive gas is present. Contaminant gases in the discharge are often more reactive than they are in the simple gaseous state since a portion of them are disassociated and/or ionized. If reactive gases are present during film deposition, compound films may be formed (reactive ion plating).

In ion plating, film atoms are injected into the discharge from a source such as a resistively-heated filament electron beam heated crucible, sputtering target, or rf-heated crucible. Such a source is shown in Fig. 2. Here the film atoms M may be (1) scattered back to the filament, (2) deposited as neutrals on the substrate (cathode), (3) ionized by electron-atom or metastable-atom collision and accelerated to the substrate, or (4) nucleated in the vapor phase to form small particles (gas evaporation) which become negatively charged and may be collected on a positively charged surface or lost to the system walls. In any case, the film atoms act as foreign particles in the gas discharge and can cause changes in the discharge parameters.

It should be noted that in a sputtering discharge only a small percentage of the gas atoms are ionized, so the cathode and other surfaces in the system are being continuously bombarded by thermal neutrals. In addition, the discharge acts as a rather intense source of ultraviolet radiation $h\nu_1$ which can affect organic surfaces and possibly insulator surfaces. The cathode surface will become heated during sputtering, since most of the energy of the bombarding ions is given up as heat and will radiate energy in the infrared, $h\nu_2$. Soft x-rays may be emitted from the cathode surface and affect semiconductor materials. The cathode current is not totally due to ion current, but is a combination of ion current and secondary electron current, It is, however, valuable to know the total cathode current density in order to establish reproducible ion bombardment.

Cleaning: Even though cleaning is an integral part of the ion plating process, it is important that gross contamination be removed prior to insertion into the ion plating system. In our experience, carbon is one of the worst contaminant materials since it reacts to form carbide layers which prevent diffusion.[6] In ion plating, the in situ cleaning can be performed by (1) sputter cleaning[10] by ion bombardment using inert gas or (2) reactive plasma cleaning[11] using a reactive species in the discharge gas which will chemically react with the surface materials to give volatile species.

Sputter cleaning is a well established method for obtaining atomically clean surfaces in a clean environment. Sputter cleaning is usually accompanied by heating in order to desorb sputtering gas incorporated into the substrate surface and to restore the surface crystallography. Sputter cleaning in a less clean environment can pose problems due to recontamination and the reactive nature of the sputter-cleaned surfaces. It has been found using in situ surface analysis by soft x-ray appearance potential spectroscopy (SXAPS) that the best way to sputter clean metal surfaces is to use rf sputtering at a low discharge pressure.[10] Insulator surfaces can also be cleaned using rf sputtering.

Reactive plasma cleaning[11] uses a reactive species in the gas discharge and may or may not use a potential applied to the surface to be cleaned. Gases having halide species have been found to be effective in cleaning metal surfaces. Figure 3 shows the cleaning of a titanium surface by argon in sputter cleaning and reactive plasma cleaning using $HC\ell$ or $CC\ell_4$ vapor as measured by SXAPS carbon peak height. Using low power inputs, the reactive plasma cleaning is more effective than sputter cleaning and $HC\ell$ is better than $CC\ell_4$ apparently due to residual carbon left by the $CC\ell_4$. Reactive plasmas may be formed using gaseous species of chemicals which would normally be used to etch the material and which form volatile species. Such reactive species might be the halides (F, $C\ell$) or hydrogen.

For carbon removal, a reactive plasma containing oxygen is very effective, but may result in excessive oxidation of easily oxidizable materials. The oxygen plasma produces ultraviolet radiation which causes bond scission in carbon compounds and active oxygen species react with the resulting fractions to form volatile gases.[13] Ion bombardment also increases bond scission and reaction.[14] The oxygen plasma may be used to clean insulator and gold surfaces, and may also be used to remove hydrocarbon and carbon contamination from metals which form passive oxides. An oxygen plasma is also useful in cleaning and desorbing gases from the walls of stainless steel vacuum systems.[13,15]

Energy Deposition: High energy particles bombarding a surface will lose their energy by (1) sputtering surface atoms, (2) generating heat, and (3) forming defects in the near-surface region. Since sputtering yields (ratio of sputtered atoms to incident ions) are rather low and the energies of the sputtered atoms are far less than that of the incident particle, only a small percentage of the incident energy is carried away by the sputtered atoms.

Little can really be said about the distribution of lattice vibrations (heat) in the near-surface region. One could speculate that atoms in the first few layers are subjected to vibrations equivalent to extremely high temperatures.

It has been shown that ion bombardment of surface creates a large concentration of point defects.[16] These defects can be of such great number as to destroy the surface crystallography and to trap diffusing gas atoms. Thus the surface defect concentration and entrapped gas is dependent on the incident particle energy and dose. This surface condition probably influences the sputtering yield.

Good adhesion is promoted by[5] (1) strong bonding within the interfacial region, (2) low local stress levels, (3) absence of easy deformation or fracture modes, and (4) the lack of long term degradation modes. These properties depend on the nature of the interfacial region. The most desirable type of interface is one which exhibits a gradual change (grade) of material properties across the interfacial region and which contain no defects which act as stress concentrators or crack initiators. This type of interface can result from interdiffusion or the formation of a "pseudodiffusion" type of interface. A pseudodiffusion interface is one formed by energetic processes such as melting/quenching, ion plating, ion implantation, codeposition, or processes involving nonequilibrium defect concentrations. Even a very thin (10 - 1000 Å) interfacial region may be effective in distributing interfacial stress.

In ion plating, the surface is bombarded by ions prior to film deposition. This bombardment cleans the surface of contaminants and surface layers which might inhibit diffusion or chemical reaction. The bombardment also generates high defect concentrations which may accelerate diffusion (radiation enhanced diffusion) and introduces stresses in the surface region which may promote diffusion. The surface heating from ion bombardment will enhance diffusion and chemical reaction and aid in the solution or disruption of contaminants which remain on the surface. Bombardment by ions during interface formation will give an enhanced nucleation density.[17] Bombardment by high energy ions/neutrals of the film material will give "ion implantation" into near-surface region. Ion bombardment will cause film atoms to be injected into the substrate surface.[18] Heating will also enhance diffusion and chemical reaction. The return of backsputtered substrate and film material to the surface will tend to generate a graded interface by codeposition and implantation.[19]

Film Properties: Deposition in a gas atmosphere and under ion bombardment has been shown to improve surface coverage over that obtained by vacuum deposition. This is due to several effects; namely, (1) gas scattering,[20] (2) backsputtering and redeposition,[19] (3) forward sputtering,[9] and (4) ionization and direction along field lines. Gas scattering alone normally gives poor adhesion and a low density film.

Ion bombardment during deposition may be used to affect a number of film properties. Figure 4 shows the intrinsic film stress and density of sputter-deposited chromium as a function of substrate bias during deposition.[9] The density of both dc and rf sputter-deposited chromium are shown. In general, ion bombardment during deposition will give a more dense film both for metals and ceramics.[9,21] Under ion bombardment, thick deposits do not develop the columnar morphology often associated with thick vacuum or sputter deposits.[9,22] Film porosity is also decreased.

Gas may be incorporated into the growing film and may be deliberately used to achieve high gas concentrations in deposited films.[12] This gas may usually be desorbed at rather low temperatures during deposition. Ion bombardment during deposition may give compositional changes in glass,[21] alloy, or compound materials.

An important parameter in the modification of films by ion bombardment is the ratio of impinging high energy particles to depositing film atoms. If this ratio is small, obviously the effects will be small.

Applications: Identifying applications of the ion plating process is often rather difficult since it is called by other names such as "ion vapor deposition," "glow discharge physical vapor deposition," "ionization electrostatic plating," "bias deposition," and others. The use of electron beam evaporation sources (e-beam ion plating) and rf-heated crucible sources[23] has greatly expanded the field of applications.

Japan has led in the development of the ion plating process for commercial applications.[24] The Japanese have been interested in the ion plating process as one of the "dry plating" processes which may be used to replace electrodeposition and its associated pollution problems. The process is being used for depositing both functional and decorative coatings.

In the United States most applications have been rather specialized where other deposition techniques have not succeeded. Ion plating of aluminum on enriched uranium reactor parts has been successful in preventing corrosion.[25] Aluminum coating of high strength steel bolts has been developed into a large scale commercial application.[26] Metal lubricant has been ion plated on titanium, zirconium, and other such metals to aid in subsequent joining. Ion plated metals have also been used as a "strike" to allow subsequent electroplating.[28] In many applications, ion plating is of interest for its ability to uniformly cover steps and complex geometries. An example is the aluminum metallization of silicon semiconductor devices.[29]

Table 1 summarizes some of the applications of the film-substrate systems listed. Table 2 gives some materials that have been deposited by reactive ion plating processes.

References

1. D. M. Mattox, Electrochem. Technol. 2 295 (1964).
2. D. M. Mattox, J. Vac. Sci. Technol. 10 47 (1973).
3. D. M. Mattox, (Proc. 6th International Vacuum Congress 1974), Jpn. J. Appl. Phys. Suppl. 2, Pt. 1 p. 443 (1974).
4. D. M. Mattox, "Interface Formation and the Adhesion of Deposited Thin Film" Sandia Laboratories Report SC-R-65-82 (Jan. 1965), Available from NTIS*
5. D. M. Mattox, "Thin Film Adhesion and Adhesive Failure" to be presented to ASTM Symposium on Adhesion Measurement of Thin Films, Thick Films, and Bulk Coatings, Philadelphia, Pa.,(11/2-4/1976). To be published in the symposium proceedings.
6. D. M. Mattox, "Surface Cleaning in Thin Film Technology," Sandia Report SAND74-0344 (Jan. 1975). Available from NTIS*
7. R. R. Hart, H. L. Dunlap, and O. J. Marsh, J. Appl. Phys. 46 1947 (1975).
8. G. Dearnaley, Appl. Phys. Lett. 28 244 (1976).
9. R. D. Bland, G. J. Kominiak, and D. M. Mattox, J. Vac. Sci. Technol. 11 671 (1974).
10. G. J. Kominiak and J. E. Uhl, J. Vac. Sci. Technol. 13 1970 (1976).

11. G. J. Kominiak and D. M. Mattox, "Reactive Plasma Cleaning of Metals Metals," to be published in Thin Solid Films (Sept. 1976)

12. D. M. Mattox and G. J. Kominiak, J. Vac. Sci. Technol. 8 194 (1971).

13. R. R. Sowell, R. E. Cuthrell, R. D. Bland, and D. M. Mattox, J. Vac. Sci. Technol. 11 474 (1974).

14. L. Holland and S. M. Ojha, Vacuum 26 53 (1976).

15. R. M. Lambert and C. M. Comrie, J. Vac. Sci. Technol. 11 530 (1974).

16. E. V. Kornelsen, Radiation Effects 13 227 (1972).

17. M. L. Tarng and G. K. Wehner, J. Appl. Phys. 43 2268 (1972)

18. L. E. Collins, J. G. Perkins, and P. T. Stroud, Thin Solid Films 4 41 (1969).

19. G. J. Kominiak, "Metallization Adhesion Layer Formed by Backscattering during dc Sputter Etching," to be published J. Vac. Sci. Technol (Sept. 1976).

20. Kurt D. Kennedy, Glen R. Schevermann, and Hugh R. Smith, Jr., Research/Development Magazine, p. 40 (Nov. 1971).

21. D. M. Mattox and G. J. Kominiak, J. Electrochem. Soc. 120, 1535 (1973).

22. T. A. Thornton, J. Vac. Sci. Technol. 11 666 (1974).

23. T. Spalvins and W. A. Brainard, "Ion Plating with Induction Heated Sources," NASA Report TM-X-3330 (Jan. 1976)*

24. Chikara Hayashi, Jpn. J. Appl. Phys. Suppl. 2 Pt. 1, p. 241 (1974), (Proc. 6th International Vac. Congress 1974).

25. R. D. Bland, Electrochem. Technol. 6 272 (1968).

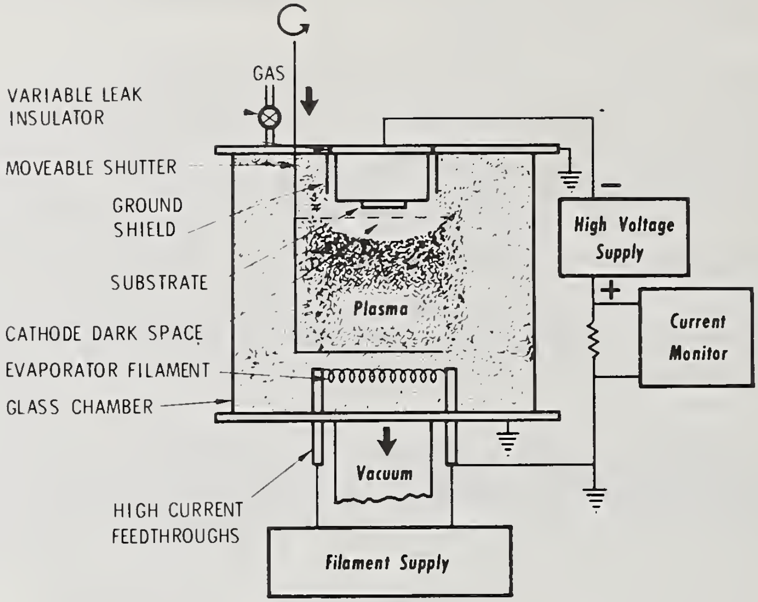
26. K. E. Steube (McDonnell-Douglass), "Glow Discharge-Tumbling Vapor Deposition Apparatus," U.S.Patent 3,926,147, U.S.Pat.Off. Gaz. 941, 1128 (12/16/75).

27. N. Ohmae, T. Nakai, and T. Tsukizoe, Wear 30 299 (1974).

28. Bernard C. Strupp, Plating 61 1090 (1974).

29. J. L. Vossen, "Method of Metallizing Semiconductor Devices," U.S.Patent 3,640,811 (2/8/72) RCA.

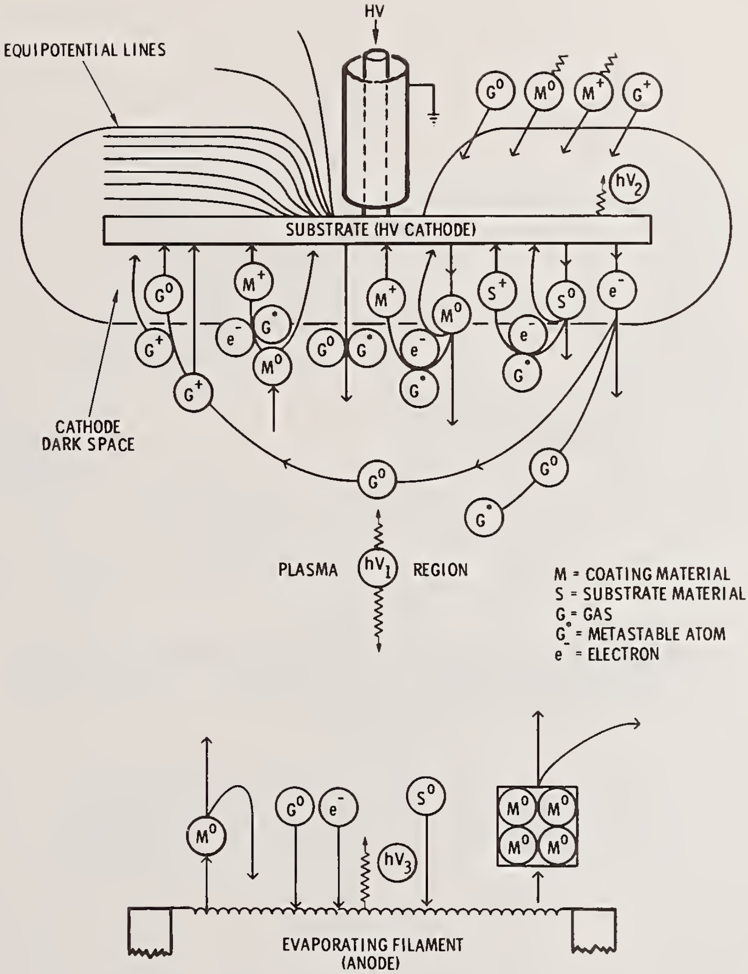
*National Technical Information Service
U. S. Department of Commerce
Springfield, VA 22161



APPARATUS FOR ION PLATING

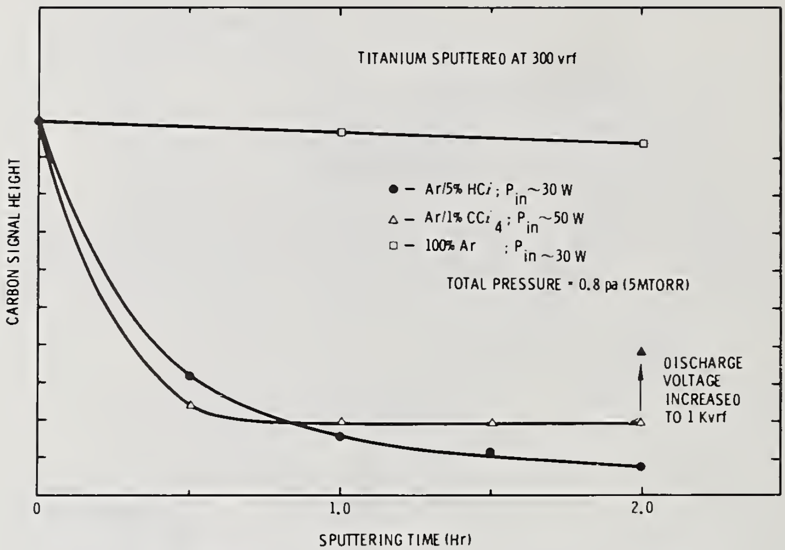
TYPICAL ION PLATING APPARATUS USING A DC POTENTIAL ON THE SUBSTRATE AND A RESISTIVELY-HEATED VAPORIZATION SOURCE

FIGURE 1



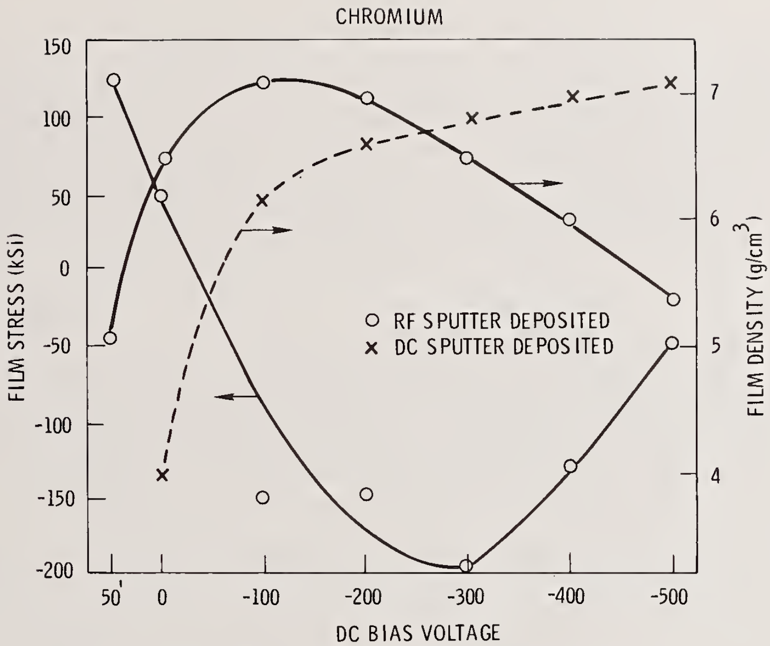
SCHEMATIC REPRESENTATION OF THE PROCESSES WHICH OCCUR IN A DC GAS DISCHARGE WITH A THERMAL VAPORIZATION SOURCE.

FIGURE 2



CLEANING OF TITANIUM BY RF SPUTTER CLEANING AND REACTIVE PLASMA CLEANING (Ref. 11)

FIGURE 3



AN EXAMPLE OF HOW FILM PROPERTIES MAY DEPEND ON DEPOSITION CONDITIONS. DATA TAKEN FROM REF. 9.

FIGURE 4

SPUTTERING

T. Spalvins
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

Abstract: This paper primarily reviews the potential of using the sputtering process as a deposition technique; however, the manufacturing and sputter etching aspects are also discussed. Since sputtering is not regulated by classical thermodynamics, new multicomponent materials can be developed in any possible chemical composition. The basic mechanism for dc and rf sputtering is described. Sputter-deposition is described in terms of the unique advantageous features it offers such as versatility, momentum transfer, stoichiometry, sputter-etching, target geometry (coating complex surfaces), precise controls, flexibility, ecology, and sputtering rates. Sputtered film characteristics, such as strong adherence and coherence and film morphology, are briefly evaluated in terms of varying the sputtering parameters. Also described are some of the specific industrial areas which are turning to sputter-deposition techniques.

Key words: Sputtering; Thin films; Protective coatings

INTRODUCTION

Sputtering is a process where a surface, when immersed in a plasma such as an ionized inert gas, is bombarded with energetic particles that cause ejection of surface atoms. The heart of the sputtering process is essentially the disintegration of the target material under ion bombardment. These ejected sputtered atoms can be collected onto a substrate to form a film.

Using the sputtering process as a film deposition method is receiving an enormous amount of interest and acceptance. New ways of applying the sputtering process are constantly being found for those areas where the more conventional deposition methods can not solve the problems encountered. In addition, as a film-deposition method, the sputtering process is used in sputter fabricating intricate mechanical components which are difficult or almost impossible to manufacture by machining, casting, or powder metallurgy techniques. The sputtering process is also used as a universal, nonchemical etching technique - for example, sputter etching, ion milling, or micromachining.

Especially in the film deposition area, the unknown potential for sputtering originates from the knowledge that it is possible to prepare new multicomponent materials (such as new alloys and compounds in any chemical proportions and relations) and also, if needed, to form a continuous compositional change or concentration gradient. Since the sputtering process is not regulated by classical thermodynamics and Gibbs phase rule relationships, one is not forced to remain within the framework of the rigid regulations. All of these unexplored features show that sputtering is really in its infancy; therefore, at present the potential future impact of sputtering on industry is inestimable.

The purpose of this paper is to acquaint the mechanical engineer with the potentials of the sputtering process, particularly with the physical rf-sputtering technique. The following sections on sputtering modes, unique features of sputtering, characteristics of sputtered films, and some practical applications of sputtered films are written in an introductory nature in order to bring out and emphasize the key features of the method and the resulting films. For more detailed information one is referred to Refs. 1 to 6.

SPUTTERING MODES

There are two basic types of sputtering, depending on whether the glow discharge plasma is generated by direct current (dc) or radio frequency (rf) fields. To induce sputtering, a negative surface charge has to be built up on the target. Besides the two basic dc and rf diode-type techniques, the sputtering modes and configurations are also as varied as the number of sputterers in the field. All of these sputtering modes and configurations arise essentially from (1) the way in which plasma is created (dc, rf, auxiliary electrodes), (2) the target and substrate positioning and their geometrical configurations, (3) the number of sputtering targets in the system, (4) the type of gases used (inert or reactive), and (5) the utilization of magnetic fields. A few modes are dc diode and triode, rf and dc/bias, dc-rf combinations, reactive sputtering, and magnetron sources.

dc Sputtering

The simplest sputtering configuration is the dc diode type, which consists of two electrodes. The coating material is the target or cathode, and the specimen to be coated is the anode. When these electrodes are immersed in an inert gas (argon) atmosphere of 10 to 100 μm and a dc potential of 500 to 5000 V, is applied across the two electrodes, ionization of the gas occurs, and generates a glow discharge plasma. The positive argon ions are accelerated toward the target with a kinetic energy high enough to knock off or sputter atoms from the target surface.

The oldest, simplest, and least expensive method is dc sputtering. The limitations to this method are that only conductors and certain semiconductors can be sputtered. Nonconductors (ceramics, glass, etc.) cannot be sputtered by the dc method. The reason is that a positive charge accumulates on the target's surface and acts as a barrier to further ion bombardment and thus the sputtering process. This problem, however, can be overcome by using an rf potential.

rf Sputtering

In rf sputtering a high frequency potential in the low megacycle range is directly applied to the metal electrode behind the target. The target (an insulator) is bonded to the metal cooled electrode, thus forming a capacitive coupling. Since the rf voltage is applied to the target assembly through rf fields generating plasma, the sputtering current is controlled by the rf voltage. Efficient plasma generation requires frequencies above 10 megacycles, and the resulting rf diode sputtering system operates usually at the allowed band of 13.56 megacycles. The insulator surface in contact with the plasma is alternately bombarded by electrons and positive ions during each rf cycle. When the surface is positive it attracts electrons, when it is negative it attracts ions. Since the electrons in the plasma have a higher mobility than the ions, the electron current to the target surface is initially much greater than the ion current. The cathode acts as a diode and charges the coupling capacitor to the peak value of the rf input voltage and then attains a negative bias.

If a voltage greater than 500 volts is applied, the surface bias will be sufficient to accelerate ions from the plasma to an energy high enough to cause sputtering of the insulator. A simplified schematic of the sputtering process is shown in Fig. 1.

All rf diode systems use the previous mechanism which makes the technique universally applicable. In addition to sputtering insulators, rf-sputtering can operate at lower argon pressures (1 to 5 μm) and attain higher sputtering rates than dc sputtering.

rf Sputtering with dc bias

Of the many possible sputtering modes and arrangements which are used today, rf sputtering with a dc bias will be described very briefly. Fig. 2 shows schematically and photographically a sputtering system which consists of two independently operated sputtering processes: rf sputtering (deposition) and dc sputtering (cleaning). The dc sputtering process is used strictly for cleaning or sputter etching metallic substrates before rf sputter deposition. The specimen is thus capable of sequential substrate cleaning or etching followed by sputter coating or simultaneously etching (biasing) while sputter depositing.

DISCUSSION

Unique Features of Sputtering

The unique, advantageous features of rf sputtering make it the most versatile of all the deposition methods used today. From an industrial point of view, for many applications the following features are of great significance: versatility, momentum transfer, stoichiometry, sputter-etching, target geometries - coating complex surfaces, precise controls, flexibility, ecology, and sputtering rates.

Versatility. - Virtually any solid material, regardless of its chemical complexity, can be sputtered in the same stoichiometry on practically any type of specimen. Alloys, intermetallics, inorganic compounds, glasses, ceramics, selected organics such as teflon (PTFE), polyimides, and cattle bone have been successfully sputter deposited.

Momentum transfer. - The sputtered atoms from the target are transferred by a momentum transfer process. The sputtered atoms are dislodged by impact evaporation as opposed to thermal evaporation. Sputtering will occur only when the actual energy by the bombarding argon ions transferred to the target surface exceeds the usual lattice binding energy of 3 to 10 eV. Sputtering does not depend on the vapor pressures of the constituent elements, and since there is no direct heating involved, it is sometimes referred to as a "cold process."

Stoichiometry. - When the sputtering parameters are carefully controlled, multicomponent solids can be deposited with the same chemical composition. The target must be water cooled at all times to avoid thermal evaporation and bulk and surface diffusion since during sputtering less than 5 percent of the kinetic energy of the bombarding ions goes into the kinetic energy of the sputtered atoms, the other 95 percent goes as heat into the target.

If the target is kept at a sufficiently low temperature to avoid evaporation and diffusion, the composition of the sputtered material will be identical to the composition of the target, even though the chemical components have different relative sputtering yields. Stoichiometry in the sputtered coating is retained because the component having the highest sputtering yield cannot diffuse from the bulk to the surface as fast as it is removed by sputtering. Following this, the lowest yield component will be sputtered along. Very soon a steady-state condition is reached in which the material is transferred from the target to the substrate in the same composition.

An important property in sputtering is the sputtering yield which is defined as the number of atoms ejected from the target per incident ion. Carbon has the lowest sputtering yield, less than 0.1 atom per ion, while silver has the highest, 2.7 atoms per ion. Sputtering yield, which is synonymous with sputtering rate, will be discussed separately in the last section.

Sputter-etching. - Instead of applying the potential to the target, the potential is applied to the substrate to induce sputtering, which is essentially reverse sputtering. Before the substrate is sputter deposited it is sputter etched. The purpose is to clean the surface of contaminants, oxides, and skin effects of cold working that may be produced by mechanical polishing. Most sputtering systems are capable of sequential substrate cleaning or etching or of simultaneous etching while sputter depositing.

Target geometries - coating complex surfaces. - Targets can have any size and shape such as planar, cylindrical, hemispherical, or some other suitable configuration. It can be made from one material or several sections of different material composition. Planar targets (12.7 to 20.32 cm diam.) are the most widely used. When such targets are sputtered, atoms leave the target at all possible angles as shown in Fig. 3. As a result, most of the specimen surface is in direct line of sight with some portion of the target. Also, because sputtering in a plasma occurs at relatively high pressures (5 to 25 μm), the mean free path of the sputtered atoms is relatively short (<1 cm). Thus, the sputtered atoms will be scattered in random directions by collisions with particles in the plasma; such scattering enables sputtered atoms to reach surfaces that are not in a direct line of sight with the target. As a consequence, irregular, nonsymmetrical surfaces can be coated in cavities and around corners without any rotation in just one operation.

Precise controls. - Sputtering offers an extraordinary control in terms of deposition rate, film thickness, uniformity, density, and film morphology. Tolerance requirements can be controlled to a millionth of a centimeter.

Flexibility. - Sputtering offers many options of various parameter combinations. As a result, the chemical composition of the coating can be controlled in any desired ratio from stoichiometric to nonstoichiometric compounds. In addition, graded compositions, laminated structures (composites), dispersion strengthened structures, and insoluble additions can be formed. These various compositional changes can be accomplished in a number of ways; for example, multitargets can be used where controlled mixing of materials is performed by sputtering from different targets or by reactive sputtering where the flow rate of the gases carrying one of the reaction product constitutents is carefully controlled.

Ecology. - Sputtering does not create any disposal problems.

Sputtering rates. - Sputtering has the disadvantage that deposition rates are relatively low. The average rate is 0.005 to 0.3 μm/min. However, equipment improvements are gradually yielding higher deposition rates. In specially designed high rate sputtering systems rates up to 250 μm/hr have been achieved. In order to increase the rate, efficient ways must be found to cool the target sufficiently with increased power input.

The sputtering rate (yield) to a first approximation can be related to the type of the material being sputtered and the sputtering parameters (bombarding ion flux density, angle of incidence, gas pressure, target and substrate temperature, etc.). The general trend as to how the sputtering parameters affect the rate are shown in Fig. 4. The rate increases with rf-power input to a maximum value and then falls off. Above the ion bombarding energy of ∿7 keV, there is no gain in sputtering yield. An increase in spacing between the target and the specimen decreases the deposition rate. Increases in the target and substrate temperatures also decrease the deposition rate. Although an increase in the gas pressure initially increases the sputtering rate, it eventually causes a decrease in the deposition rate because of a decrease in the ion mean free path. Applied magnetic fields increase the plasma density and thus increase the deposition rate.

The low deposition rates also have certain advantages; for example, they do afford a high degree of film control. The slow rates have a tendency to form denser films as compared to higher rates.

CHARACTERISTICS OF SPUTTERED FILMS

Functionability of a coating, regardless of its intended use, depends primarily on the degree of adherence, coherence, and coating morphology. Sputtered coatings grow in a complex plasma environment. The coating adherence, coherence, and morphology are directly affected by (1) sputter etched or biased surface, (2) kinetic energy of the sputtered species, (3) plasma conditions, and (4) substrate temperature and topography.

The strong adherence normally obtained with sputtered coatings can be attributed basically to the surface cleanliness and the relatively high arrival energies of the sputtered material. These energetic submicroscopic sputtered particles have certain activation energies which not only favorably affect the surface adherence but also increase the cohesion between the sputtered particles. The strong particle to particle cohesion is responsible for the formation of high density films. The submicroscopic particle size is important in both compactability and final density of the coating as well as strength. Strength is generally related to the final grain size - the smaller the grain, the stronger the compact. Due to these characteristics, relatively thin films in the 0.2 to 10 μm range can be used where previously thicker films were required. It can be implied, therefore, that the coating attachment to the surface is more important than the volume of coating present.

It is well known that stress induced peeling, which is caused by internal stresses in the film, increases with film thickness. Therefore, in sputtered films, which are usually very thin (<1 μm), the stress induced peeling effect is minimized.

A diverse range of coating morphologies and properties can be reliably produced by controlling the various sputtering parameters and the substrate conditions. The nucleation and growth of the sputtered film can be varied and this would in turn affect the film properties. Morphological changes (preferred texturing: columnar, equiaxed and even epitaxial growth) in the coating can be initiated. The grain size can be decreased substantially, and this reduction will increase the density of the coating by simply bias sputtering. One of the most obvious bias related properties of many coatings is the optical reflectance at ground potential and dull and matte appearance at bias potential. Since a detailed discussion of this subject is outside the scope of this paper, one is referred to Refs. 1 to 9.

SOME PRACTICAL USES OF SPUTTERING

The sputtering process is primarily used in sputter deposition technology. However, high rate sputtering techniques have been developed which can achieve sputtering rates up to 250 μm/hr. Thicknesses up to, but not limited to, 0.63 cm have been achieved. As a result, commercial interest has developed in thick deposits. Fabrication by sputtering of free standing shapes such as sheet, tubing, and inner and outer cylindrical structures with coolant passages for thrust chambers has been developed as illustrated in Fig. 5.

Sputter etching provides an alternate technique to chemical etching. The outstanding features are that it is universally applicable to all materials, it eliminates undercutting, and it is widely used for pattern delineation.

Of all the specific industrial areas the greatest sputtering activity is in microelectronics and microminituarization such as integrated thin film circuit technology. In the last 10 years sputtering has not only spread rapidly in the mechanical area but also in practically all areas which require films that are difficult or impossible to handle by other means. Just a few areas where sputtered coatings have an increasing impact are corrosion and high temperature oxidation protection, reduction of friction and wear, solid film lubrication, decorative purposes, replicating techniques, biomaterials for surgical implants, and solar cell development.

Several specific examples will illustrate how sputtered films function in these areas. When solid film lubricants such as MoS_2 are sputtered on sliding surfaces, only 0.2 μm thick films are required for effective lubrication in vacuum or dry air [10, 11]. An endurance life evaluation is shown in Fig. 6 where sputtered MoS_2 films are compared to two other application techniques. The endurance lives of bearings sputter coated with a duplex coating (0.1 μm thick underlayer of Cr_3Si_2 and subsequently with 0.6 μm of MoS_2) were greatly improved over the lives of those bearings which had MoS_2 films directly applied as shown in Fig. 7. Teflon PTFE (polytetrafluoroethylene) has been successfully sputtered with excellent adherence on metal, glass, paper, and wood surfaces. Fig. 8 shows 0.1 μm thick PTFE films sputtered on hypodermic needles. Special formulations of modified silicon carbide and other refractory carbides have been sputtered on airborne components such as first-stage compressor blades. These components operate in severe environments and have to be protected from corrosion, oxidation, erosion, and abrasion

[12]. An interesting approach is to sputter cattle bone on metallic prosthetic devices used as surgical implants for hip bone replacements. The sputtered bone film promotes bone growth and attachment to living bone [13].

SUMMARY OF RESULTS

The sputtering process is primarily used in thin-film deposition technology. Because higher sputtering rates have been attained, sputtering is now used for fabricating intricate mechanical components (tubing, cylinders with passages, etc.). Finally, the sputtering process is being used as a universal nonchemical etching technique.

In coatings technology sputtering has yet unexplored potentials due to the fact that almost any material can be sputtered in proportions and relations that are not regulated by classic thermodynamics and phase interrelationships. Many sputtering modes and configurations can be used to achieve chemical compositions which range from stoichiometric to nonstoichiometric. Coatings with gradual changes in composition and laminated and dispersion strengthened structures can also be obtained using these configurations.

The unique sputtering features of versatility, momentum transfer, stoichiometry, sputter etching, target geometry, precise controls, flexibility, and sputtering rates contribute to the outcome of the desired coating.

The sputtered films generally exhibit strong adherence and coherence, and the coating morphology can be controlled by the sputtering parameters. As a result, thin sputtered films having 0.2 μm thicknesses are sufficient for many applications. Sputtered MoS_2 lubricant films 0.2 μm thick exhibit superior performance during sliding friction, and sputtered Teflon (PTFE) and sputtered cattle bone have found new useful applications.

REFERENCES

1. Wehner, G. K., and Anderson, G. S., Handbook of Thin Film Technology, Chap. 3, L. I. Maissel and R. Glang, eds., McGraw Hill, 1970, pp. 3-1 to 3-38.

2. Miassel, L., Handbook of Thin Film Technology, Chap. 4, L. I. Maissel and R. Glang, eds., McGraw Hill, 1970, pp. 4-1 to 4-44.

3. Carter, G., and Colligow, N. S., Ion Bombardment of Solids, Chap. 7, Heinemann Educational, London, 1968, pp. 310-353.

4. Mattox, D. M., Sputter Deposition and Ion Plating. American Vacuum Society, 1973.

5. Brown, S. C., Introduction to Electrical Discharges in Gases, John Wiley and Sons Inc., 1966.

6. Vossen, J. L. and O'Neill, J. J., RCA Review, Vol. 29, pp. 149-173, 1968.

7. Chopra, K. L., Thin Film Phenomena, McGraw-Hill, 1969.

8. Thornton, J. A., Journal of Vacuum Science and Technology, Vol. 11, No. 4, pp. 666-670, 1974.

9. Notis, M., High Frequency Heating Review, Vol. II, No. 8, Lepel High Frequency Laboratories, Inc., pp. 4-6.

10. Spalvins, T., "Sputtering - A Vacuumed Deposition Method for Coating Material," Paper No. 72-DE-37, presented at the Design Engineering Conference and Show of the ASME, May 8-11, 1972.

11. Spalvins, T., ASLE Proceedings - International Conference on Solid Lubrication, American Society of Lubrication Engineers, 1971, pp. 360-370.

12. Gentner, K., "Potential of Sputtered Films as Protective Coatings," NADC-75038-30, Naval Air Development Center, 1975 (N76-17251).

13. Shaw, B. J. and Miller, R. P., "Sputtering of Bone on Prosthesis," Patent 3,918,100, United States, Nov. 1975.

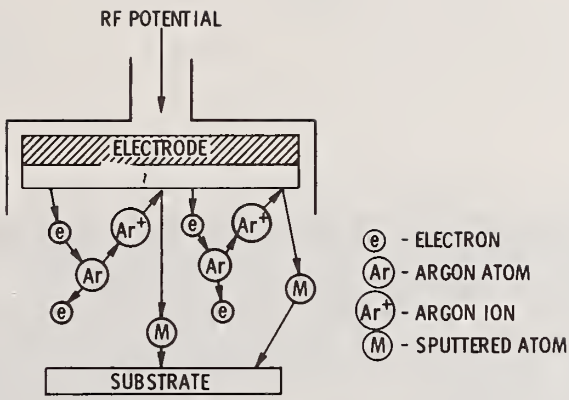
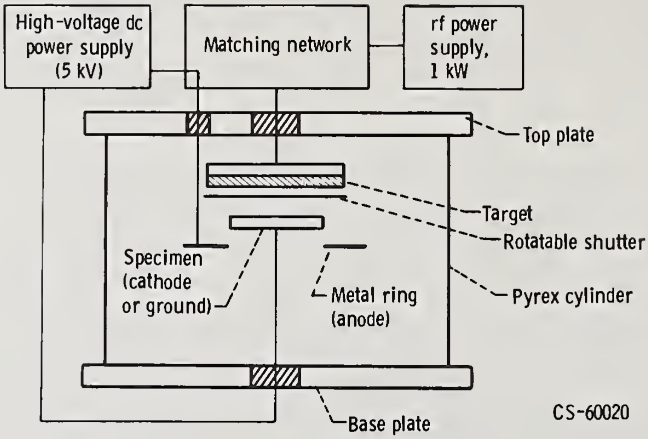
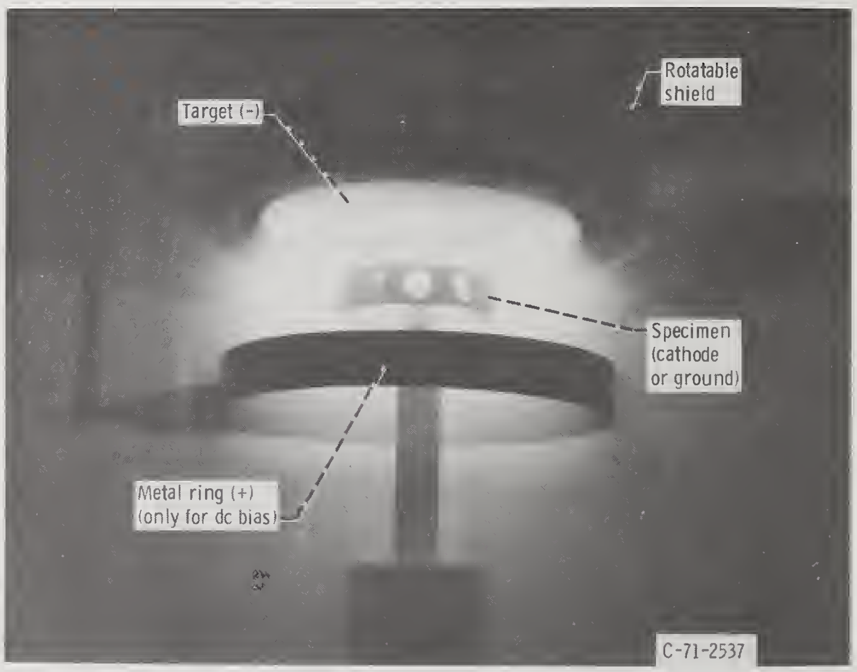


Figure 1. - Schematic of sputtering process.



(a) Schematic diagram.



(b) View of apparatus during sputter coating.

Figure 2. - Radiofrequency diode sputtering apparatus with direct-current bias.

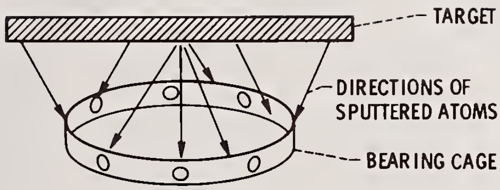


Figure 3. - Schematic of sputter coating complex surfaces.

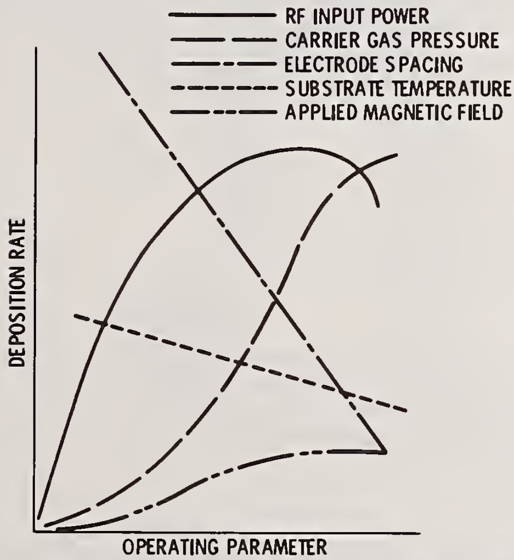


Figure 4. - Deposition rate as a function of various operating parameters.

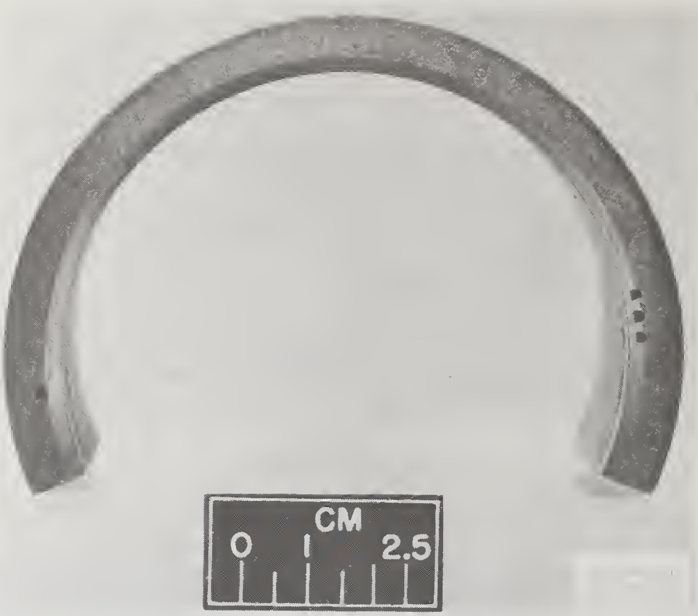


Figure 5. - Fabrication of cylinders by sputtering.

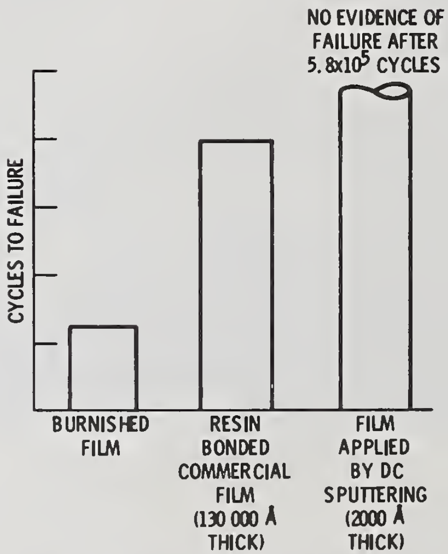


Figure 6. - Endurance lives of MoS_2 films applied by various techniques.

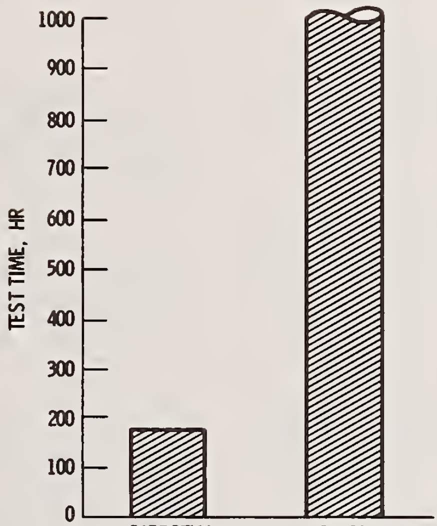


Figure 7. - Endurance lives of 440C stainless-steel ball bearings with sputtered MoS_2 films on races and cage - with and without a Cr_3Si_2 underlayer.

CS-74111

Figure 8. - Hypodermic needles and protective housings with sputtered teflon.

MECHANICAL FAILURES PREVENTION GROUP

25th MEETING

Engineering Design

List of Registrants

A. Anderjaska
Federal Aviation Admin.
800 Independence Avenue, SW
Washington, DC 20591

J. Arness
Hogan and Hartson
Washington, DC 20006

B. P. Bardes
Bimba Manufacturing Co.
101 E. Main Street
Monee, IL 60449

W. J. Baxter
General Motors Research Lab
12 Mile and Mound Roads
Warren, MI 48090

D. K. Benson
Midwest Research Inst.
425 Volker Boulevard
Kansas City, MO 64110

E. P. Blyskal
Eastern Airlines
International Airport
Miami, FL 33148

A. Branum
DECdesign Marketing
200 Forest Street
Marlboro, MA 01752

L. G. Buhr
Hughes Aircraft Company
Bldg. 117, MS 1
P. O. Box 90515
Los Angeles, CA 90009

H. C. Burnett
National Bureau of Standards
Metallurgy Division
Washington, DC 20234

J. B. Cardoso
Aerospace Corporation
2350 E. El Segundo
El Segundo, CA 90245

O. Casarella
Washington Navy Yard
Naval Weapons Engineering Support Act., Bldg. 219.2
Washington, DC 20374

J. B. Christian
Wright-Patterson Air Force Base
Lubricants and Tribology Branch
Dayton, OH 45433

A. A. Conte, Jr.
U.S. Naval Air Development Center
Code 30212
Warminster, PA 18974

R. A. Coulombe
Naval Ship Engineering Center
Code 6146, NC-4, Rm. 268
Washington, DC 20382

R. M. Davison
Naval Weapons Engineering Support Act.
9205 W. Stayman Drive
Ellicott City, MD 21043

P. R. deBruyn
Industrial Liaison
National Bureau of Standards
Washington, DC 20234

O. Decker
Mechanical Technology, Inc.
968 Albany-Shaker Road
Latham, NY 12110

E. J. Deemer
Naval Air Systems Command Rep. Atlantic
NAS Norfolk, Bldg. R-52 (ASCR 332)
Norfolk, VA 23511

M. G. Dhawan
Naval Ship Engineering Center
National Center 4, Code 6145F
Room 378
Washington, DC 20382

S. Edelman
National Bureau of Standards
B-324, Polymers Building
Washington, DC 20234

G. Ellis
Potomac Research, Inc.
7655 Old Springhouse Road
McLean, VA 22101

R. M. Evan-Iwanowski
Syracuse University
141 Link Hall
Syracuse, NY 13210

S. Flanagan
Federal Aviation Admin.
800 Independence Avenue, SW
Washington, DC 20591

P. M. Fleming
National Bureau of Standards
Metallurgy Division
Washington, DC 20234

J. B. Fussell
University of Tennessee
Knoxville, TN 37916

W. W. Gunkel
General Motors Corporation
EAS-PA, GM Tech. Center
Warren, MI 48090

R. R. Hagglund
Worcester Polytechnic Inst.
Worcester, MA 01609

F. Hall
Washington Navy Yard
Naval Weapons Engr. Support Act.
Bldg. 219.2
Washington, DC 20374

P. Hallick
Federal Aviation Admin.
800 Independence Avenue, SW
Washington, DC 20591

H. R. Hegner
GARD, Inc.
7449 North Natchez Avenue
Niles, IL 60648

M. Heller
519 Investment Building
Washington, DC 20005

P. Henry
Peoples Gas Light & Coke Co.
122 S. Michigan Avenue, Rm. 1725
Chicago, IL 60603

J. D. Hoffman
Director, Inst. for Materials Res.
National Bureau of Standards
Washington, DC 20234

S. M. Hsu
Amoco Chemical Corporation
Naperville, IL 60540

G. Hurt
Detroit Diesel
Allison Division
Indianapolis, IN 46201

R. B. Johnson
National Bureau of Standards
Institute for Materials Res.
Washington, DC 20234

G. Kardos
Carleton University
Ottawa, Canada K1S 5B6

W. D. Keeley
Vibration Specialty Corp.
100 Geiger Road
Philadelphia, PA 19115

M. P. Kelly
Dunegan/Endevco
Rancho Viejo Road
San Juan Capistrano, CA 92675

H. K. Keswani
Gulf Research & Development Co.
P. O. Drawer 2038
Pittsburgh, PA 15230

E. E. Klaus
Pennsylvania State University
108 Fenske Laboratory
University Park, PA 16802

R. A. Kuehl
John Deere Plow & Planter Works
501 Third Avenue
Moline, IL 61265

R. R. Landers
Gould Ocean Systems Division
Cleveland, OH 44101

H. P. Lee
NASA
Goddard Space Flight Center
Greenbelt, MD 20771

P. Lewis
Mechanical Technology, Inc.
968 Albany-Shaker Road
Latham, NY 12110

R. E. Maringer
Battelle Columbus Labs
505 King Avenue
Columbus, OH 43201

J. H. A. Mason
British Embassy
Defence Research & Dev. Staff
3100 Massachusetts Avenue, NW
Washington, DC 20008

D. M. Mattox
Sandia Laboratories
Albuquerque, NM 87115

R. S. Miller
Office of Naval Research
Arlington, VA 22217

R. Misialek
U. S. Navy
6762 NAVSEC Phila. Div.
Philadelphia, PA 19112

E. Passaglia
National Bureau of Standards
Institute for Materials Research
Washington, DC 20234

R. Pepper
Fiber Materials, Inc.
Biddeford Industrial Park
Biddeford, ME 04005

J. Perso
Whirlpool Corporation
Monte Road
Benton Harbor, MI 49022

M. Peterson
Wear Sciences
32 Sutherland Drive
Scotia, NY 12302

A. G. K. Pong
Consumer Product Safety Comm.
5401 Westbard Avenue, Rm. 935
Bethesda, MD 20207

C. Portillo
John Deere Plow & Planter Works
501 Third Avenue
Moline, IL 61265

R. Predmore
Goddard Space Flight Center
Code 313
Greenbelt, MD 20771

M. Rayl
RCA Corporation
RCA Labs
Princeton, NJ 08540

H. Schaeffer
University of Maryland
College Park, MD 20740

A. Schmudde
Caterpillar Tractor Co.
Industrial Division-Engr.
Peoria, IL 61629

J. Scialdone
Goddard Space Flight Center
Greenbelt, MD 20771

G. W. Scott
Oak Ridge National Laboratory
P.O. Box X, Bldg. 4500-S, Rm. Y-15
Oak Ridge, TN 37830

T. R. Shives
National Bureau of Standards
Metallurgy Division
Washington, DC 20234

W. H. Skewis
Telcom Systems, Inc.
2300 South 9th Street
Arlington, VA 22204

L. Sibley
SKF Industries
King of Prussia, PA 19406

C. O. Smith
University of Nebraska at Omaha
P. O. Box 688
Omaha, NB 68101

J. R. Sodorquist
Federal Aviation Admin.
800 Independence Avenue, SW
Washington, DC 20591

T. Spalvins
NASA
Lewis Research Center
Cleveland, OH 44101

J. E. Stern
NASA, Goddard Space Flight Center
Code 721
Greenbelt, MD 20771

J. N. Strohlein
National Bureau of Standards
Building 304, Room 136
Washington, DC 20234

J. W. Taylor
David W. Taylor Naval Ship Research
and Development Center
Annapolis, MD 21402

W. A. Troha
U.S.A.F. (AFAPL/TBP)
Wright-Patterson AFB, OH 45433

C. E. Vest
NASA, Goddard Space Flight Center
Code 313
Greenbelt, MD 20771

W. A. Willard
National Bureau of Standards
Metallurgy Division
Washington, DC 20234

W. O. Winer
School of Mechanical Engineering
Georgia Institute for Technology
Atlanta, GA 30332

H. Wolf
Grumman Aerospace Corporation
Bethpage, NY 11714

A. Wolff
T. R. Mallory and Company
N.W. Industrial Park
Burlington, MA 01803

MFPG PUBLICATIONS

Both printed and microfiche copies of the following MFPG publications (1-11) whose catalog numbers start with "AD" or "COM" may be obtained from the NTIS.

National Technical Information Center
5285 Port Royal Road
Springfield, VA 22161

1. Glossary of Terms . . . . AD 721 354

2. Proceedings of Meetings 1-9 (set of 5) . . . . AD 721 359

   Meeting Nos. 1-5 Papers and discussion on failure analysis and control.

   Meeting No. 6 "Detection, Diagnosis and Prognosis", December 6, 1968.

   Meeting No. 7 "Failure Mechanisms as Identified with Helicopter Transmissions", March 27, 1969.

   Meeting No. 8 "Critical Failure Problem Areas in the Aircraft Gas Turbine Engine", June 25-26, 1969.

   Meeting No. 9 "Potential for Reduction of Mechanical Failure Through Design Methodology", November 5-6, 1969.

3. Proceedings of Meeting No. 10 . . . . AD 721 912
   "Vibration Analysis Systems"
   January 21-22, 1970.

4. Proceedings of Meeting No. 11 . . . . AD 724 475
   "Failure Mechanisms: Fatigue"
   April 7-8, 1970.

5. Proceedings of Meeting No. 12 . . . . AD 721 913
   "Identification and Prevention of Mechanical Failures in Internal Combustion Engines"
   July 8-9, 1970.

6. Proceedings of Meeting No. 13 . . . . AD 724 637
"Standards as a Design Tool in Surface Specification for Mechanical Components and Structures"
October 19-20, 1970.

7. Proceedings of Meeting No. 14 . . . . AD 721 355
"Advances in Decision-Making Processes in Detection, Diagnosis and Prognosis"
January 25-26, 1971.

8. Proceedings of Meeting No. 15 . . . . AD 725 200
"Failure Mechanisms: Corrosion"
April 14-15, 1971.

9. Proceedings of Meeting No. 16 . . . . AD 738 855
"Mechanical Failure Prevention Through Lubricating Oil Analysis"
November 2-4, 1971.

10. Proceedings of Meeting No. 17 . . . . AD 750 411
"Effects of Environment Upon Mechanical Failures, Mechanisms and Detection:
April 25-27, 1972.

11. Proceedings of Meeting No. 18 . . . . AD 772 082
"Detection, Diagnosis and Prognosis"
November 8-10, 1972.

12. Proceedings of Meeting No. 19 . . . . .COM-74-50523
"The Role of Cavitation in Mechanical Failures"
October 31 - November 2, 1973.

Printed copies of the following MFPG publications (stock numbers starting with SN) may be obtained from the Government Printing Office.

Superintendent of Documents
U.S. Government Printing Office
Washington, DC 20402

Microfiche copies of these publications (catalog numbers starting with NBS-SP) may be obtained from NTIS.

13. Proceedings of Meeting No. 20 . . . . SN003-003-01451-6
"Mechanical Failure - Definition of the Problem"
May 8-10, 1974.
Cost: Printed copy $6.10

14. Proceedings of Meeting No. 21
"Success By Design: Progress Through Failure Analysis"
November 7-8, 1974.
Cost: Printed copy $3.50 . . . . SN003-003-01639-0
Microfiche copy $2.25 . . . . NBS-SP-433

15. Proceedings of Meeting No. 22
"Detection, Diagnosis and Prognosis"
April 23-25, 1975.
Cost: Printed copy $4.25 . . . . SN003-003-01556-3
Microfiche copy $2.25 . . . . NBS-SP-436

16. Proceedings of Meeting No. 23
"The Role of Coatings in the Prevention of Mechanical Failures"
October 29-31, 1975.
Cost: Printed copy $2.65 . . . . SN003-003-01664-1

17. Proceedings of Meeting No. 24
"Prevention of Failures in Coal Conversion Systems"
April 21-23, 1976.
Cost: Printed copy $3.00 . . . . SN003-003-01760-4

APPENDIX

NATIONAL MATERIALS POLICY IN THE 95th CONGRESS

CURRENT AND POTENTIAL GOVERNMENT IMPACTS ON AVAILABILITY

Franklin P. Huddle
Congressional Research Service
Library of Congress
Washington, D. C. 20003

An interesting and constructive circumstance about national materials policy as well as the broader subject of national science and technology policy, is that it is very largely bipartisan. We are fortunate that in both Houses of Congress and in the relationship between the legislative and executive branches of our national Government, there is a healthy consensus on these matters. Thus, in the elections that took place this week, the question relevant for this meeting was not over issues of materials or science policy, but whether the winners in the various campaigns will be motivated to preserve this same bipartisanship, and will be equipped to deal knowledgeably with the subjects in question.

As far back as 1970, the National Materials Policy Act setting up the National Commission on Materials Policy, was jointly supported by seven Democrats and five Republicans, including the Chairman and the Ranking Minority Member of the Senate Subcommittee that brought the legislation into being. There was no opposition to the bill in either House of Congress.

Similar bipartisanship has characterized other materials and science policy legislation. The National Commission on Supplies and Shortages was created in response to the joint initiative of the Senate Majority and Minority Leaders, Senators Mansfield and Scott. The four legislative members of the Commission are a Democrat and a Republican from each House of Congress.

Also, the National Science and Technology Policy, Organization, and Priorities Act of 1976, approved by the President last May 11, was originally introduced by Chairman Teague and Ranking Minority Member Mosher, of the House Committee on Science and Technology. Once differences in legislative approach had been ironed out between the House and Senate versions, there was no opposition to the measure and the White House accepted it with ceremony.

Shortly thereafter, on last June 17, a bill to establish a national materials policy was offered in the House of Representatives as H.R. 14439, under the sponsorship of James Symington and Charles Mosher, Chairman and Ranking Minority Member, respectively, of the House Subcommittee on Science, Research, and Technology.

For one last illustration of this healthy bipartisan consensus, I offer the example of the Office of Technology Assessment, whose Board consists of three Democrats and three Republicans from each House of Congress, and whose studies are responsive to the interests of congressional Members of both political parties.

I suggest that it is beneficial to the scientific, the technological, and the materials communities of the United States to cherish and preserve this state of affairs. We should also recognize--as did the National Commission on Materials Policy--the need to deal with policy for materials, energy, and environment as one interlocked and interconnected subject. Similarly, we should deal with the global flow of materials and scientific knowledge as a subject of shared interest of all nations, a matter of inescapable interdependence. And thirdly, as the authors of the Constitution of the United States proclaimed in the Preamble, it is necessary to strike a balance between the demands of the present and the needs of the future--to serve both "ourselves and our posterity." These should not be partisan matters, and there is evidence to show that they have not been.

Then, what ought to be the role of the industrial community, the materials community, and the science-and-technology community with respect to national policy for materials, energy, and environment? I suggest two broad areas of activity in which you, as members of one or more of these three communities, can best serve.

First, you can help to guide the Congress and the President in the effective implementation of important legislation passed by the 94th Congress, whose term of office ends at the close of 1976.

Second, you can help to guide the Congress to design future legislation and action programs in the remaining unfinished business of the 94th Congress which is likely to become the priority business of the 95th Congress, beginning next January.

Implementing Policies Promulgated by the 94th Congress

Among the actions of Congress up to the close of the 94th Congress are several of great importance to the materials community.

First, there is the Act that created the National Commission on Supplies and Shortages, whose final report is expected by the close of 1976. The Act calls for advice to the Congress on means to anticipate and deal with future important shortages of industrial materials and agricultural products. What actions this Commission will propose, and how the Congress will respond to the recommendations will not be known until 1977, but should warrant the careful scrutiny of the materials community.

The scope of the Commission's report is expected to encompass the management and analysis of supply and demand information, stockpiling decisions, and materials conservation programs.

Second, in Title III of the National Science and Technology Policy, Organization, and Priorities Act of 1976, approved May 11, a "President's Committee on Science and Technology" is created and instructed to report in two years on 13 policy questions. These include: government organization for science and technology, the management of scientific information, technology assessment, the relations of the Federal Government with the States and private industry, elimination of red tape obstructing innovation, support for basic research and education, manpower training, the strengthening of community uses of science, and long range plans for the management and use of science and technology. It will be of interest to this audience to learn how the President's Committee proposes to resolve these important policy questions, and how the Congress responds to its recommendations.

Three other subjects decided by the 94th Congress will be of interest next year and thereafter. One is the implementation of the Toxic Substances Control Act, and another is the administration of the new Resource Conservation and Recovery Act of 1976, approved October 22. Both of these two pieces of legislation are so complex and broad in scope that all I can do here is commend them to your careful study. How they are carried out will have much to do with the successful response of industry to the policies these Acts prescribe.

Finally, one legislative decision was reached by non-action. That was the response of the Congress to the earlier proposal of the President to liquidate most of the remaining reserves in the National Stockpile of Strategic and Critical Materials. For several years and more, the Congress had deferred action on this matter. Finally, last October 1, the Federal Preparedness Agency, which now has jurisdiction over stockpile planning and management, announced the results of an extended study of stockpile objectives. According to a press release from FPA, the new objectives policy was recommended by the National Security Council and approved by the President. There were three changes:

1. Planning will be based on the first three years of emergency of indefinite duration, compared [with] one year under the previous guidelines.

2. The civilian portion of the economy will be provided for after some reasonable allowances for "belt tightening." Previous policy did not separately consider civilian needs.

3. For each year used in planning, stockpile needs will be estimated separately for defense and civilian requirements.

In addition, Director Bray of the Agency promised that goals and policies would be reviewed at least every four years, or sooner if required.

The issue remaining to be examined in 1977 is whether this policy modification in the Strategic and Critical Materials Stockpile is sufficient to meet future national industrial needs, or whether a national economic stockpile or international buffer stocks should also be established. The first alternative has been reported on by the Office of Technology Assessment, and will also be discussed in the final report of the National Commission on Supplies and Shortages. The second alternative was suggested by the Secretary of State, Dr. Kissinger, in his major address to the United Nations General Assembly in its special session in September 1975.

Unfinished Congressional Business in Materials for 1977

I have already mentioned the prospect that the next Congress--the 95th--will be called on to implement proposals of the National Commission on Supplies and Shortages, and the science proposals of the President's Science Advisory Committee. The Congress will also be concerned with monitoring the execution of other materials and science legislation, including the several important measures adopted in 1976.

What may be the most important item of unfinished business of interest to the materials community is the bill proposed in the House as H.R. 14439 and introduced separately in the Senate by Frank E. Moss of Utah as S.3637. Title I of this bill sets forth a proposed national policy for materials. Title II would create a principal presidential adviser on materials and a policy board. To coordinate implementation of national policy there would also be created a commission representing principal Federal agencies with existing functions in materials management or materials research. Paralleling the commission in scope would be two congressional committees on national materials policy. The purpose in introducing this legislation was explained by Congressman James Symington of Missouri, last June 17. It was, he said, "designed to promote some serious thinking about where we go from here with the remainder of Earth's resources. What are we going to do with the stuff which still remains accessible to us? How can we best use it?" The bill was intended, he went on to say, "primarily for the purpose of stimulating discussion on what is probably the most widespread and serious problem facing modern civilization--lack of any cohesive policy to govern, or at least to guide, the use of materials." Such a policy was central to any modern technological society, a prime source of opportunity for energy conservation, and a basis for achieving environment quality.

At the request of the House Committee on Science and Technology I have been inviting and collecting views of informed persons in the materials community on this proposed legislation, and would welcome further contributions of opinions from this source. I should add that last week I had occasion to discuss the subject with a senior executive of one of the larger companies in the field of plastics, who suggested that perhaps he had been too harsh in his comments, and that for that reason they might not be welcomed. I assured him that the purpose in sending the bill out for criticism was to invite it, and that compliments were of no particular value. What the Committee is looking for is a consensus in which all can share.

Incidentally, I received another letter from an industrial society declining to comment on the bill, and explaining that such comment was considered premature. My reaction to that statement is that the best time to comment on legislative proposals is in the early stages of their formulation, instead of waiting until they are, so to speak, frozen in concrete. Comments are most useful in the earliest stages. If you wait until the end of the line, you may find yourself in the position of "take it or leave it" without any real opportunity to help shape the final product.

A number of materials policy issues have been under study for the past two years or so in the Office of Technology Assessment, and are the subject of reports either issued or about to be. These are (1) Economic Stockpiling, (2) the Management of Materials Information, and (3) Access to Minerals in the Ground. All three subjects are likely to become of interest to the 95th Congress.

In addition, the O.T.A. has been planning or starting up other studies to meet congressional committee requests in such additional subjects as the following:

- Resource Recovery, Recycling, and Re-Use
- Conservation through Reduced Wastage
- Mineral Exploration Technologies
- Alternative Responses to Materials Shortages
- Substitution Alternatives for Critical Imported Materials
- Existing Federal Coal Development
- Conservation of Energy through Materials Management

These, too, are likely to become of interest to the 95th Congress, in the years 1977-78.

The Legislative Situation in 1977

As I suggested at the outset of this statement, there are some difficulties in forecasting what the interest of the Congress is likely to be in 1977. This week, you and I went to the polls to vote for a President and a Vice President, one of the 435 Members of the House of Representatives, and perhaps also one Member of the roughly one-third of the Senate selected by the voters in 1976.

The 95th Congress will be very different from the 94th, particularly in its leadership. Many Members of both Houses who have taken a leading role in developing national materials policy did not run for reelection. These included Senators Mansfield and Scott, who were responsible for the initiative that led to the National Commission on Supplies and Shortages; House Speaker Carl Albert, who joined with Mansfield and Scott in urging the President to work with the Congress on national materials policy; Congressmen James Symington and Charles Mosher, the sponsors of the national materials policy bill; and a number of other very important Members.

A number of other members of the 94th Congress were defeated at the polls, including Senators Tunney and Brock, both members of the National Commission on Supplies and Shortages, and Senator Frank Moss, who sponsored S. 3637, the proposed national materials policy act, in the Senate.

Thus, when the new Congress is convened in January, there will be a number of new faces. The opportunity--indeed, the necessity--will be at hand for the materials community to help build a new national consensus on materials policy.

In the 94th Congress upwards of 500 separate legislative proposals were introduced that involved national materials policy. Surely this was some kind of a record! But will the 95th Congress have the same recognition of the significance of materials for economic health and progress, energy conservation, and environmental quality? I suggest that the time will soon be at hand for the materials community in the United States to help the Congress to assess the importance of national materials policy. What are the elements of this policy that the materials community would wish the legislature to promulgate and implement? How is the national wellbeing, with respect to materials policy, to be determined and acted upon?

It is the constitutional function of the Congress to shape the national policy on important questions like these. But it is the constitutional function of you, the electorate and the informed technical community, to advise the Congress wisely on these matters.

What mechanisms and means are available to you to exercise this constitutional function? I have observed quite a number of these. As a matter of fact, a very useful and proper relationship can be built up between the Congress of the United States and the scientific/technical community. Among the appropriate mechanisms that characterize this relationship are -

- The appointment of scientific or technological interns to serve as voluntary staff members in congressional offices.

- The organization of legislative liaison committees to entertain requests for technical information and advice, and to channel such requests to qualified respondents within the technical society.

- The organization of ad hoc committees to assess and evaluate particular legislative proposals.

- The acquisition and dissemination of congressional literature of particular interest to the members of the technical society.

- Establish liaison with the four policy advisory services to the Congress (the Congressional Research Service, the General Accounting Office, the Office of Technology Assessment, and the Congressional Budget Office) and provide these services with resumes of technical people best qualified to provide advice on special categories of technological subjects, problems, and issues.

It is important to note that the legislative function generates an insatiable demand for scientific and technological information and advice. The advisory offices to the Congress are constantly called on to satisfy this demand, manifestly, we cannot do this without recruiting a great staff of experts in all imaginable fields. So the alternative--which we follow--is to establish ourselves as a bridge between the Congress and the various technical communities. That is why I am here tonight, and it is the essence of my message to you. In sum, the Congress needs your help, and I ask you to supply it.

TESTING FOR SUCCESS

John C. New
National Aeronautics and Space Administration
Goddard Space Flight Center
Greenbelt, Maryland

In 1959 the Goddard Space Flight Center was created as a major new space flight center of NASA. It was franchised to oversee the design, development, testing and flight of unmanned spacecraft. It was my privilege to be a participant in the creation, construction, and operation of a new test facility for providing assurance that these spacecraft would be successful in space. It is the purpose of this discussion to review the fifteen year record of the test and evaluation program developed at the Center.

On the earth, we deal with equipment and materials which cost a few dollars per pound. The costs of space hardware are upwards of $15,000 - $50,000 per pound, delivered into space. The complexity of this space material can be likened to insertion of a scanning electron microscope into space orbit, operating it by remote control and then transmitting its data back to ground. Generating confidence, then, that these types of complex systems will function reliably after the launch and in the space environment, is the purpose of the test facility and test program at Goddard Space Flight Center.

In the early sixties, the objectives of NASA were to explore space; gain scientific data on the space environment, planetary and astral bodies; develop applications and instruments for communications and weather observations; survey and observe earth resources; place man in space and return him safely. This period will be remembered as one when the technical supremacy of the United States was being challenged. The technological emphasis was on achievement and the avoidance of failure. The eyes of man world-wide were focused alternately upon the US and USSR space programs as they raced for a moon landing. This paper will deal only with the unmanned space program; the manned program deals with problems of a different order of magnitude.

Some of the different types of spacecraft over which Goddard Space Flight Center has cognizance cover the range from the initial U.S. spacecraft Vanguard, a relatively simple system, to the highly sophisticated scientific and applications satellites such as the Orbiting Solar Observatories and the LANDSAT (Earth Resources Technology Satellites). Also included are such spacecraft as Echo, OGO, OAO, TIROS, NIMBUS, SYNCOM, and several foreign satellites. The information covered in this paper was derived from 57 missions. Some of these missions investigated energetic particles, cosmic rays, sun corona, solar energy, radiation belts, weather phenomena and communication techniques. Most of the spacecraft carried a large number of scientific or application types of experiments.

Figure 1 is a typical spacecraft. It is actually one of the early Explorer type spacecraft. The figure shows distribution and the subsystems complement of the spacecraft. In the early Vanguard days launching an eight pound spacecraft was considered a problem. Today we are still concerned with the weight of spacecraft, but we are now dealing with thousands of pounds, orders of magnitude greater than the early days. A typical unmanned spacecraft may weigh from a few hundred pounds to a few thousand pounds. The largest, heaviest unmanned spacecraft weighed 4600 pounds. This is the Orbiting Astronomical Observatory, with a one meter telescope. It is still in space collecting data that is unavailable from earth because of the ultraviolet light absorption by the earth's atmosphere.

Figure 2 shows the various steps in the spacecraft development cycle from the spacecraft concept through the dissemination of the scientific information derived from the spacecraft experiments. The length of time in each step varies with different spacecraft but the overall time is typically about two years. Many of the steps are performed concurrently, in parallel, in order to meet demanding and critical schedules. In some cases, the scientific information being sought is tied to a planetary event to which the launch schedule is timed. Missing the schedule means delaying the launch for the next planetary cycle, a delay of many months or even years. Insofar as the test programs are concerned, there is generally ample time to plan and replan, but little time to execute them. This is due to the fact that more time is generally used for the concept, design and development than is allotted for these stages. Consequently the test program period is squeezed between the fixed launch date and the increasing time demands of the development stage.

Product assurance requires an organization which has the test equipment and trained personnel to generate techniques and develop a facility that will execute the evaluation program and do it correctly the first time. The invested cost of a spacecraft, at the time it is being tested, is large --- several millions of dollars. For the Orbiting Astronomical Observatory, it was about 60 million dollars. This did not include the costs of the scientists' time and their years of research up to this point, it was just the labor and hardware costs to produce the spacecraft system. One thinks twice about rehandling and retesting a massive device that has this kind of invested cost. Also it becomes more obvious why the tests should be done correctly the first time when one considers that delays in schedules at this point --- just prior to launch --- are very costly (several thousands of dollars per hour) since project manpower loading costs are at their peak.

The various operating conditions and environments, natural and induced, to which a spacecraft is exposed are shown in Figure 3. The unmanned spacecraft test program is concerned with the environmental conditions of prelaunch, powered launch, and orbital spaceflight. The test philosophy developed by and insisted upon by the Test and Evaluation Division at Goddard Space Flight Center was to test the flight spacecraft, the one-of-a-kind unit, at test levels that were expected to occur during launch and in orbit; and to test the complete system in it's flight assembled condition. This test approach was unique in the formative days of the space program and it generated controversy and debate before it was accepted. This philosophy and practice was pursued at Goddard Space Flight Center from it's inception and has resulted in a success rate of 97% for 57 spacecraft for which GSFC was responsible. For the 33 spacecraft which passed through GSFC's Test and Evaluation Division facilities and personnel over the 15 year review period, no failures were experienced. A 100% mission success rate was achieved. This is not to say that failures did not occur. They did both in ground tests and in space flight. On the ground they were fixed; in space flight, redundancy and software work-arounds permitted mission success to be achieved.

This is a record of which the Test and Evaluation Division is extremely proud. It was accomplished by use of excellent facilities, a positive and highly supportive attitude by Center and NASA management and the dedication of the Test and Evaluation technical staff.

Our system test program has several objectives. These include proving the hardware design and applications, eliminating weak links, locating and identifying subsystem interactions, training launch personnel and developing design guidance.

It has been the experience at GSFC, as it probably is and has been elsewhere in the engineering world, that engineering design is a process of successive approximations, a series of iterations converging toward a practical optimum. It is exceedingly rare to find a complex mechanism which works perfectly the first time. New devices are developed through use and through the correction of known malfunctions. In space projects there is no way to correct mistakes once the spacecraft is launched; there is no way to measure what may malfunction or fail. If one knew where to locate the failure detector, the design could have been corrected so it would not have failed in the first place. Very little direct failure information has been brought back from space. Generally, failures in space have been deduced from the telemetered data indirectly related to the failure itself.

The reliability of spacecraft and assurance of flight success was found in environmental testing, where the spacecraft was exposed to conditions simulating those that were expected during launch and in space. General Environmental Test Specifications were developed and became the "design criteria" for the spacecraft project. The weak links in the system were discovered and eliminated. It was found that there was considerable interaction between various subsystems including electronics, mechanical, control, telemetry and power subsystems. The interactions were identified; their effects were evaluated; and the systems were modified to remove or correct those interactions which interferred with the satisfactory functioning of the spacecraft system. Teams of people were trained with the actual flight hardware, so that they were familiar with the spacecraft functions and operations. These teams could then satisfactorily recognize and respond to launch or flight emergencies through use of redundant systems or software reprogramming. Out of this space flight experience design-guide-concepts were developed and fed back into the later programs most often through a revised General Environmental Test Specification.

Based on the launch environment, which is measurable and predictable, a philosophy for flight units was adopted in which the test severity was set at a level of two sigma above the mean. At that level the probability of missing a failure is about five percent. The overall record is somewhat better than that. For prototype units, the test severity was increased to 150 percent of the flight test levels, which, statistically, is about the three sigma level. This choice of prototype test level was more fortuitous than scientific, a kind of engineering factor of ignorance based on instinct tempered with past experience. Figure 4 graphically describes the test level selections.

While the time spent in the launch environment lasts for only a few minutes, the time in the orbital environment can extend over a period of months or years. Based on spacecraft system failures that can be expected in orbit, a test philosophy was developed whereby infant mortality in space was virtually eliminated. This was achieved by exercising the spacecraft on the ground for a sufficient length of time so that the early failures were discovered on the ground. The failure was analyzed; corrections and modifications were incorporated in the design and the failure mechanisms were eliminated. Figure 5 shows the effects of the test program and space environments on the occurrence of failures in spacecraft systems.

The environmental test program centered on simulating the space environment on the ground, exercising the spacecraft systems long enough under space environment conditions until the infant mortality was virtually eliminated and failure rates were reduced to an acceptable level. Studies are being pursued regarding the length of time a spacecraft system should be operated in a thermal-vacuum chamber (space environment simulator). The present thinking is that thirty days would be an ideal exposure. However, the economics of testing must be considered. A balance must be made between the costs of testing and the costs of launching a spacecraft with high confidence that a minimum failure rate has been achieved.

One of the problem areas in spacecraft systems is the heat transfer in the space environment. The normal gas pressure environment internal to the spacecraft is less than 10^{-4} torr. This precludes convective heat transfer and we must depend upon conductive and radiative processes. Figure 6 shows a typical electronic part failure caused by poor heat transfer.

Although there was radiative heat transfer between the electronic part and its environment, the adequacy of the thermal design depended upon heat conduction to the base plate or heat sink. Poor thermal contact between the part and heat sink caused the temperature of the part to exceed safe limits and a failure occurred. It is difficult to insure good contact between mating surfaces. This is a function of the contact area, the physical and chemical character of the contact surface, the bearing or contact stresses, the nature of the base material, the torque applied to the fastenings. Add to this the people problems in quality control and you find that it is necessary to environmentally test the design on the ground and determine whether a space failure is likely.

Let us now look at a unique type of problem that occurs in spacecraft components. A loose metal particle, generated during part manufacture or during environmental test, migrated and shorted out two adjacent leads. Now this is not a big bother in a television set on the ground. One can use the standard terrestrial solution, kick the set, and the particle will generally move out of the way. The gravity field does our work. However, in a zero gravity field, the particle can float around, under small spacecraft disturbances and maneuvers, creating opportunities to cause frequent malfunctions.

Various types of spacecraft subsystem problems that were uncovered by environmental testing are shown in Figure 7. A problem --- malfunction or failure --- is defined as a performance which does not meet or surpass functional or operational specifications. The P and F stand for the Prototype or Flight test level severity. It is quite evident that the test program had isolated many problems that needed correction and that a marked reduction in problems has occurred when the modified and improved flight units were retested. It also shows that the largest source of problems are due to the exposure of the subsystems to thermal-vacuum environment.

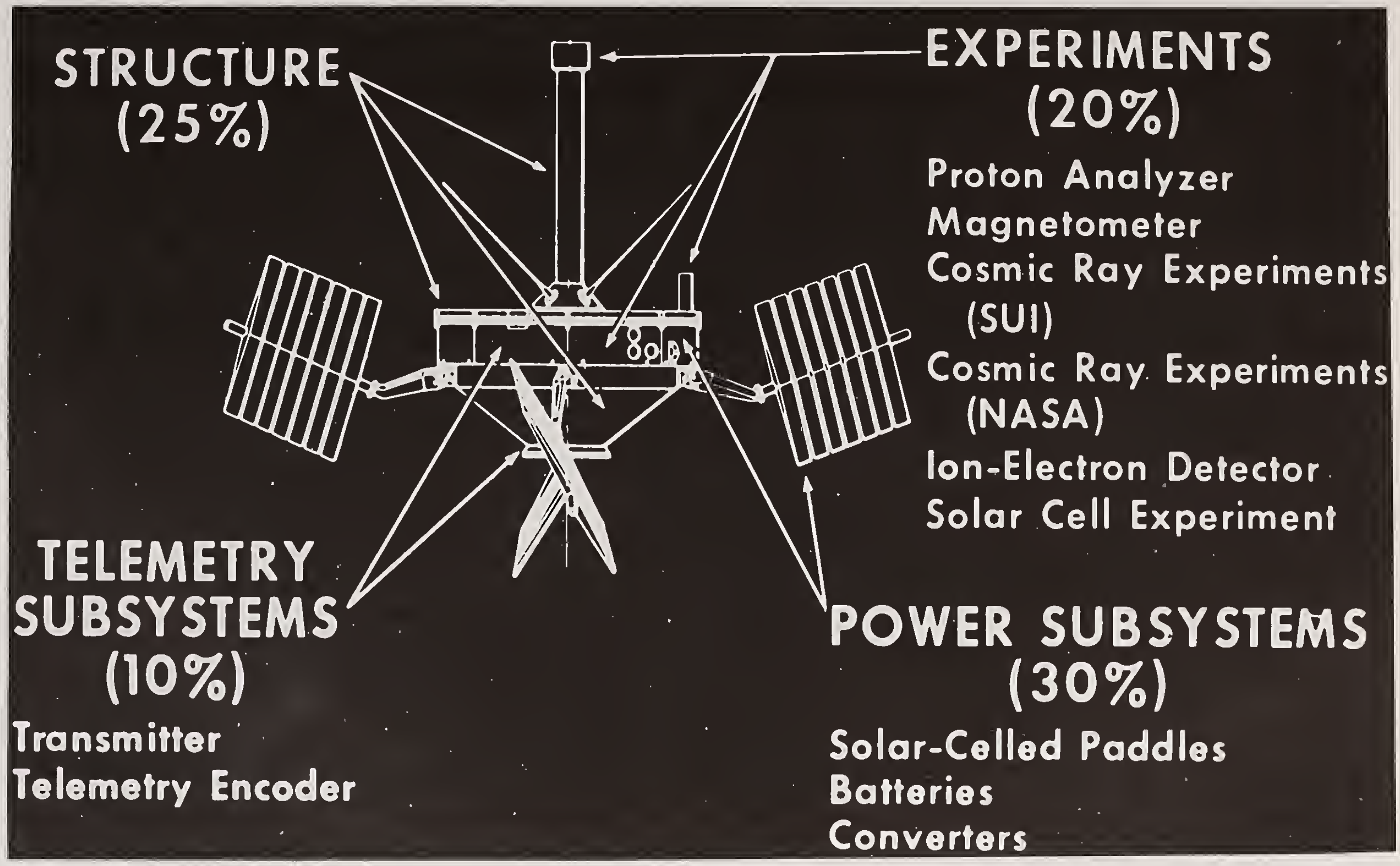
A flight subsystem which has had more than its fair share of problems is the tape recorder. The typical flight tape recorder is a device made up of all kinds of materials --- metals, polymers, organic and inorganic components of wide variety, and is a composite of electrical, electronic and mechanical elements. It is expected to have a useful life, for as long as five years, during which time it must faultlessly record and play back data acquired from flight experiments. A typical recorder records at one speed and plays back at a speed about 100 times the recording speed. With such stringent demands upon the tape recorder, we find that design improvements must be made to overcome the failures observed in space and in testing on the ground.

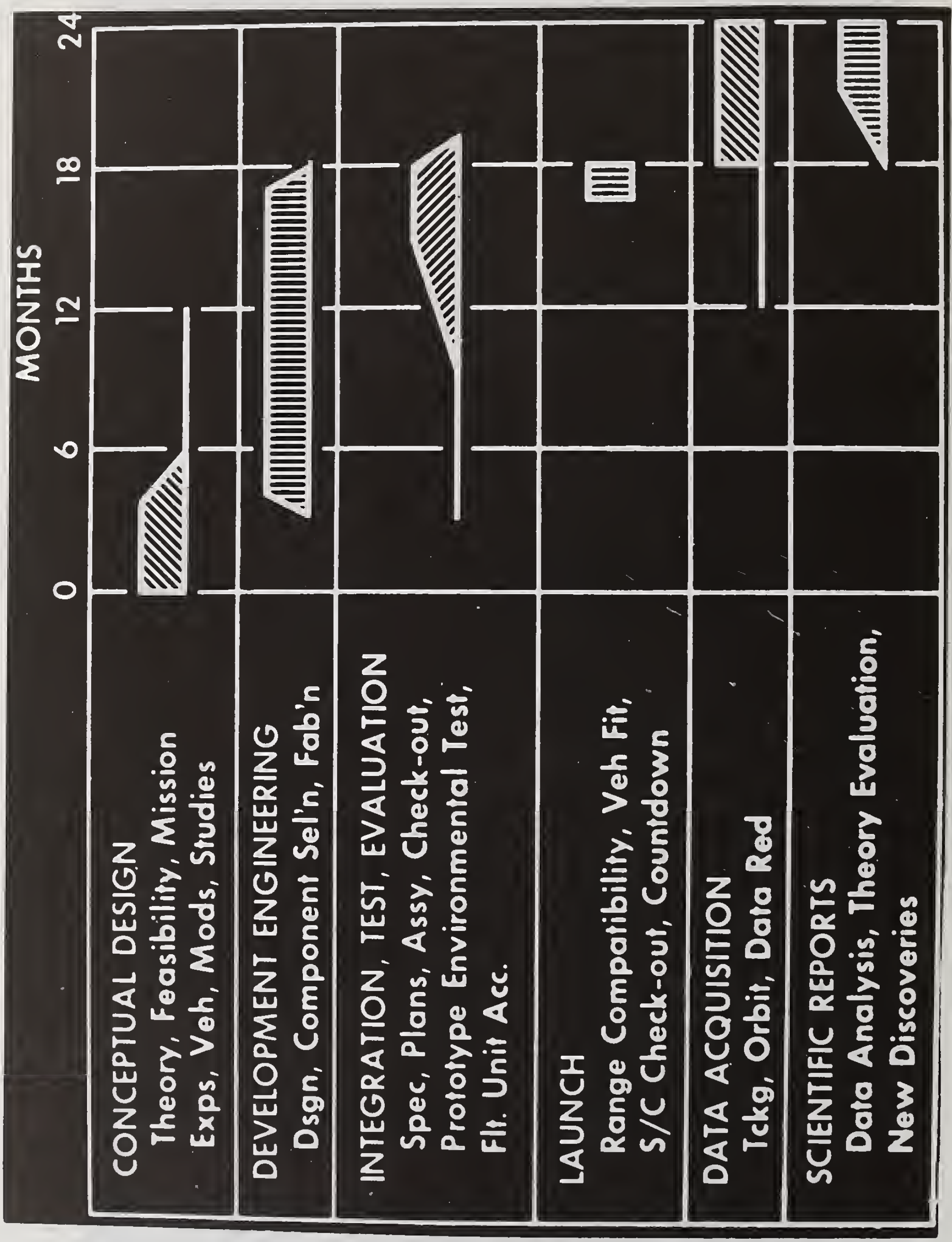
The tape is the source of failure due to wear and load; bearings are a source of trouble due to lubrication problems; lubricants can generate contamination problems; the recording play back head suffers wear degradation. Environmental and life testing has helped solve some of these tape recorder problems. Continued design changes and life testing are the bases for improving this necessary device for spacecraft flight systems.

Figure 8 shows how well we have done in space. We find that about 50% of the first month's problems show up on the first day. It is conjectured that the causes for this high percentage of first day problems is due to the first exposure to a weightless condition, not to the launch environment. The weightless condition is conducive to the movement of loose particles. These particles have the opportunity to upset electrical and electronic circuitry as previously indicated, and to interfere with the functioning of mechanical elements by lodging in bearings, joints, etc. Upon reaching orbit a space observatory consisting of several hundred thousands of parts is first turned on after experiencing weightlessness, energetic particle bombardment, outgassing pressure changes, and radiative heat transfer conditions. Is it surprising that 50% of the malfunctions occur on this first day? The other 50% of the problems are spread over the remaining days of the month.

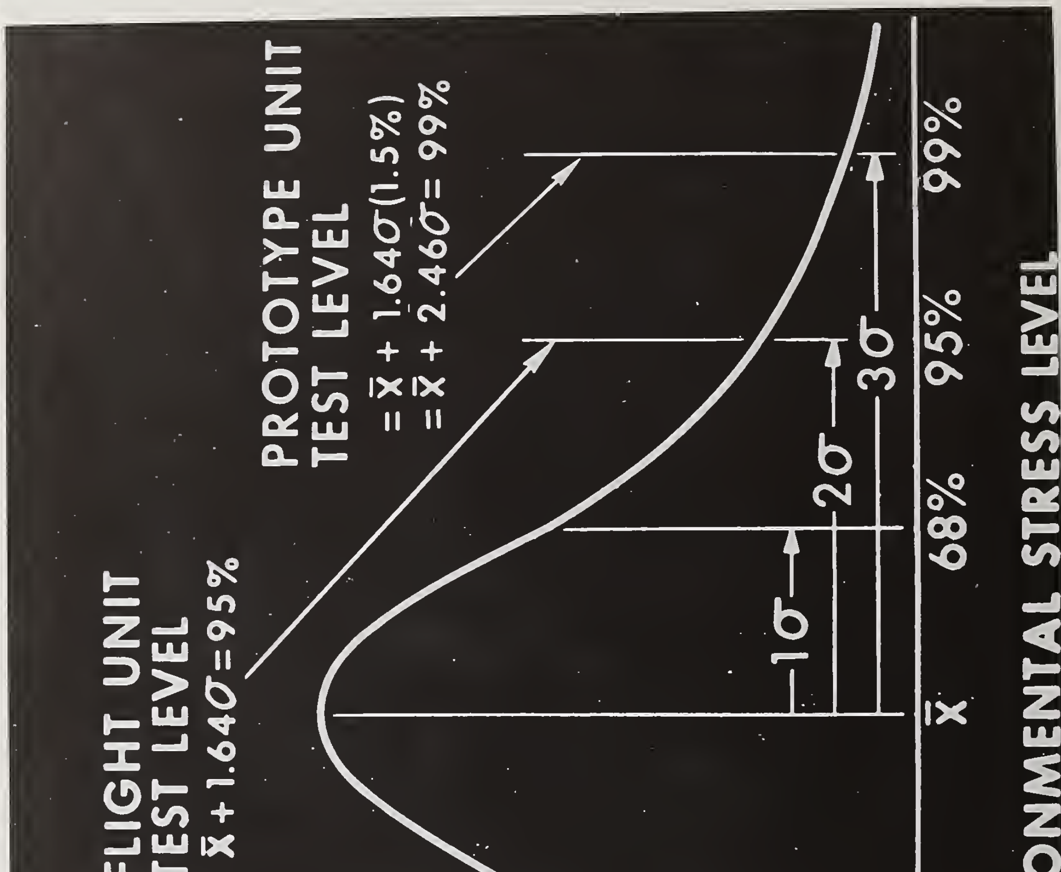
The success of our test program is indicated in Figure 9. The comparison between the malfunctions uncovered during system test and those subsequently occurring in flight indicate the effectiveness of our test efforts. A success rate of 97% has been achieved. We do not get all the malfunctions --- the test screen is not 100% effective. However, it is interesting to note that between the system test malfunctions and the flight malfunctions there is a reduction of 10:1 on all devices, 12:1 on electrical devices; 13:1 on electromechanical devices.

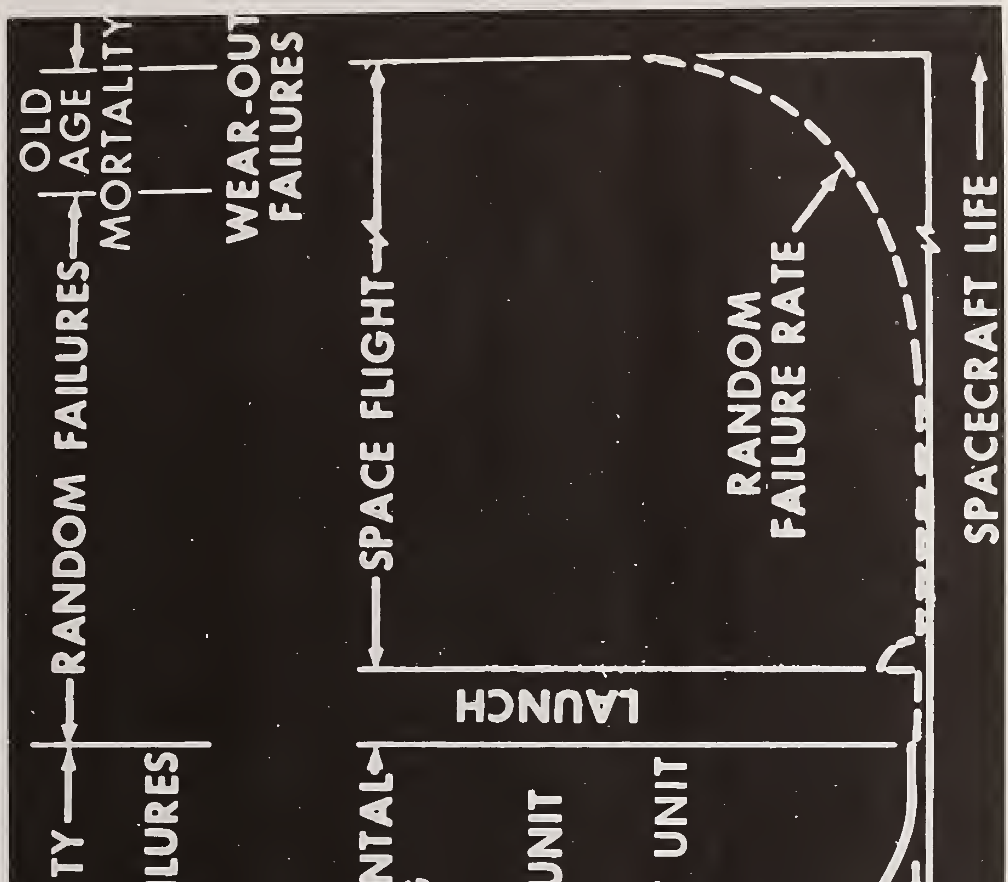
Based on our successful test program experience, it is felt that the concept of full systems test at expected levels has been proven, especially for the space program where the quantities we deal with are not large number production units but single systems, one of a kind type, of a research and development nature.


STRUCTURE
(25%)
EXPERIMENTS
(20%)
Proton Analyzer
Magnetometer
Cosmic Ray Experiments
(SUI)
Cosmic Ray Experiments
(NASA)
Ion-Electron Detector
Solar Cell Experiment
TELEMETRY
SUBSYSTEMS
(10%)
Transmitter
Telemetry Encoder
POWER SUBSYSTEMS
(30%)
Solar-Celled Paddles
Batteries
Converters


MONTHS
0
6
12
18
24
CONCEPTUAL DESIGN
Theory, Feasibility, Mission
Exps, Veh, Mods, Studies
DEVELOPMENT ENGINEERING
Dsgn, Component Sel'n, Fab'n
INTEGRATION, TEST, EVALUATION
Spec, Plans, Assy, Check-out,
Prototype Environmental Test,
Flt. Unit Acc.
LAUNCH
Range Compatibility, Veh Fit,
S/C Check-out, Countdown
DATA ACQUISITION
Tckg, Orbit, Data Red
SCIENTIFIC REPORTS
Data Analysis, Theory Evaluation,
New Discoveries


PLANETARY LANDING & DWELL
LANDING IMPACT
VIBRATION
AERODYNAMIC NOISE
AERODYNAMIC HEATING
PLANETARY ATMOSPHERES
PLANETARY RADIATION BELTS
PLANETARY PARTICLES (DUST)
VARIABLE GRAVITY
VACUUM, RADIATION, AND THERMAL CONDITIONS
TEMPERATURE EXTREMES
RE-ENTRY CONDITIONS
ACCELERATION
VIBRATION
AERODYNAMIC NOISE
AERODYNAMIC HEATING
THERMAL SHOCK
IMPACT OR LANDING SHOCK
WATER IMMERSION (IF APPLICABLE)
EXPOSURE TO NATURAL ELEMENTS PRIOR TO RECOVERY
PRE-LAUNCH CONDITIONS
TEMPERATURE & HUMIDITY
SHOCK & VIBRATION
HANDLING
STERILIZATION
R-F RADIATION
STORAGE DURATION
POWERED LAUNCH
SHOCK & VIBRATION
ACCELERATION-THRUST, GUIDANCE, WIND SHEAR
AERODYNAMIC NOISE
AERODYNAMIC HEATING
PRESSURE DECREASE
CORONA
ORBITAL & SPACE FLIGHT
SPACE VACUUM
SOLAR RADIATION
3 K HEAT SINK
EARTH RADIATION AND ALBEDO
RADIATION BELT & SOLAR FLARES
TEMPERATURE EXTREMES, CYCLIC VARIATION
SEPARATION AND DESPIN
WEIGHTLESSNESS
ATTITUDE CONTROL
ENGINE RESTART, VIBRATION
NO AIR DAMPING
MAGNETIC TORQUES
METEOROIDS
ON-BOARD NUCLEAR SOURCES


EXAMPLE:
SINGLE ENDED,
NORMAL DISTRIBUTION
FLIGHT UNIT
TEST LEVEL
$\bar{X} + 1.64\sigma = 95\%$
PROTOTYPE UNIT
TEST LEVEL
$= \bar{X} + 1.64\sigma(1.5\%)$
$= \bar{X} + 2.46\sigma = 99\%$
1σ
2σ
3σ
$\bar{X}$
68%
95%
99%
ENVIRONMENTAL STRESS LEVEL
FREQUENCY OF OCCURRENCE


INFANT MORTALITY
RANDOM FAILURES
OLD AGE MORTALITY
DESIGN
WORKMANSHIP
MATERIAL
FAILURES
WEAR-OUT FAILURES
INTEGRATION & BENCH TESTING
ENVIRONMENTAL TESTING
LAUNCH
SPACE FLIGHT
PROTOTYPE UNIT
FLIGHT UNIT
RANDOM FAILURE RATE
SPACECRAFT LIFE
NUMBER OF FAILURES

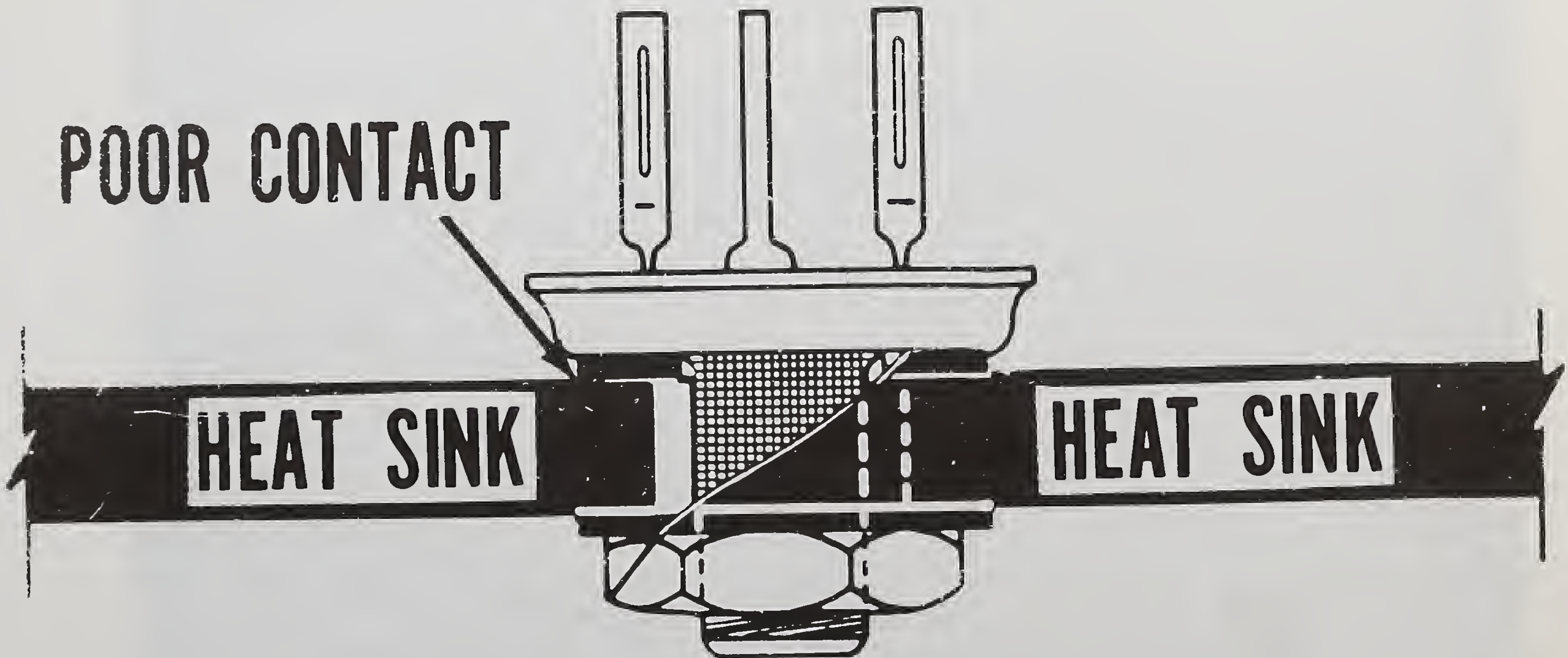


Figure 6. Heat Sink. Electrical Failure (Thermal Runaway) Revealed During Thermal-Vacuum Tests. Inadequate Heat Sink or Poor Thermal Contact are Common Failure Causes.

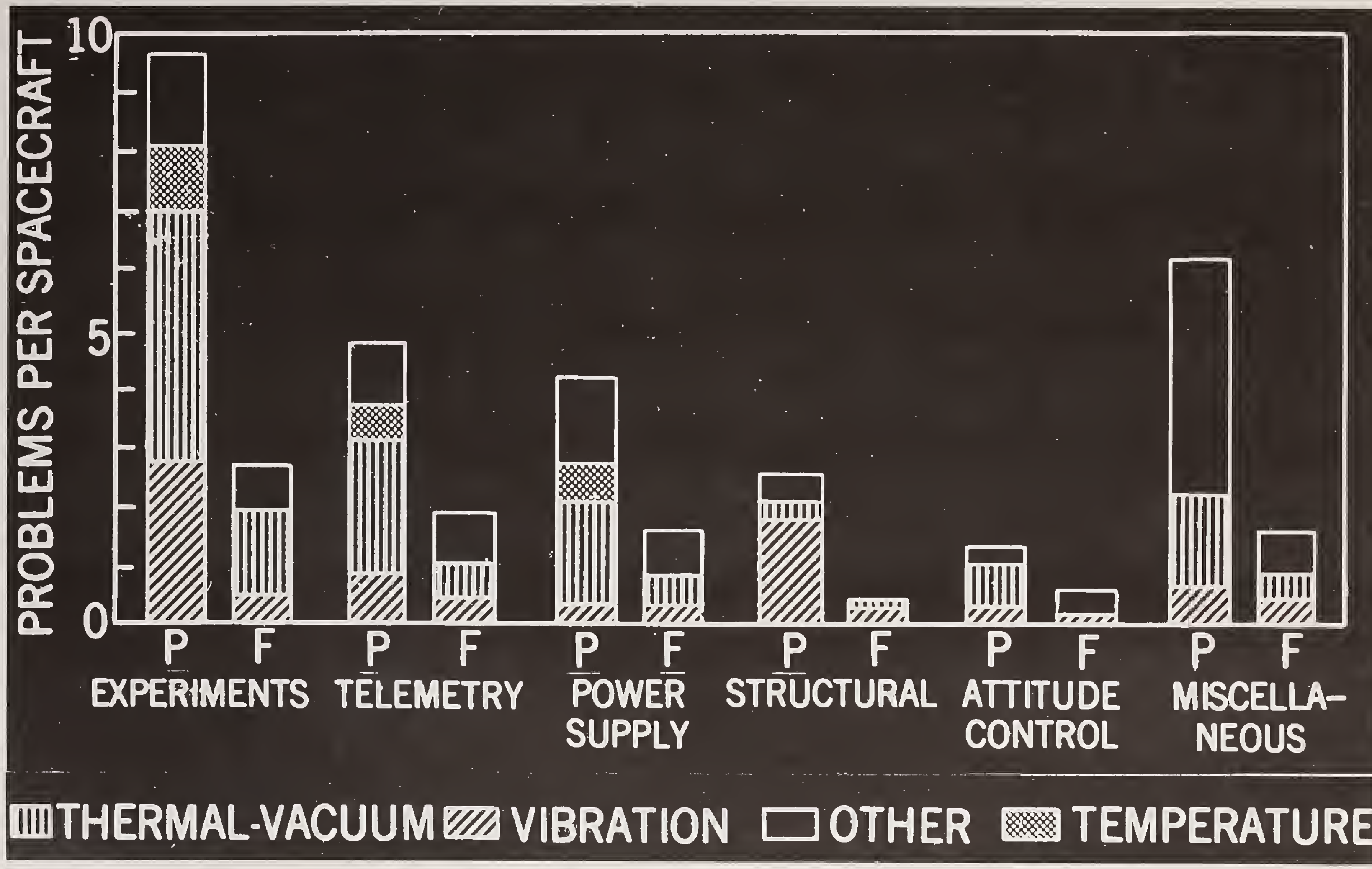
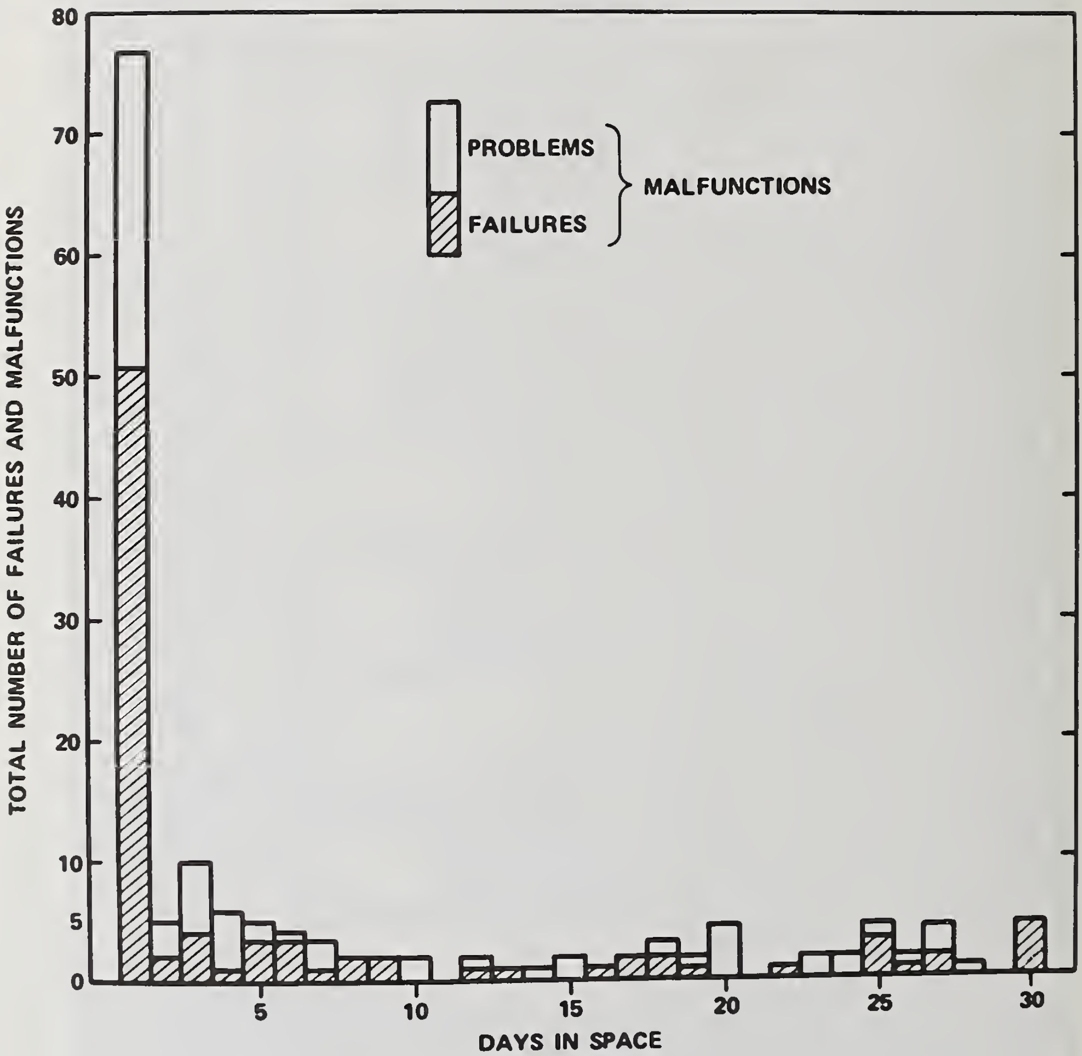
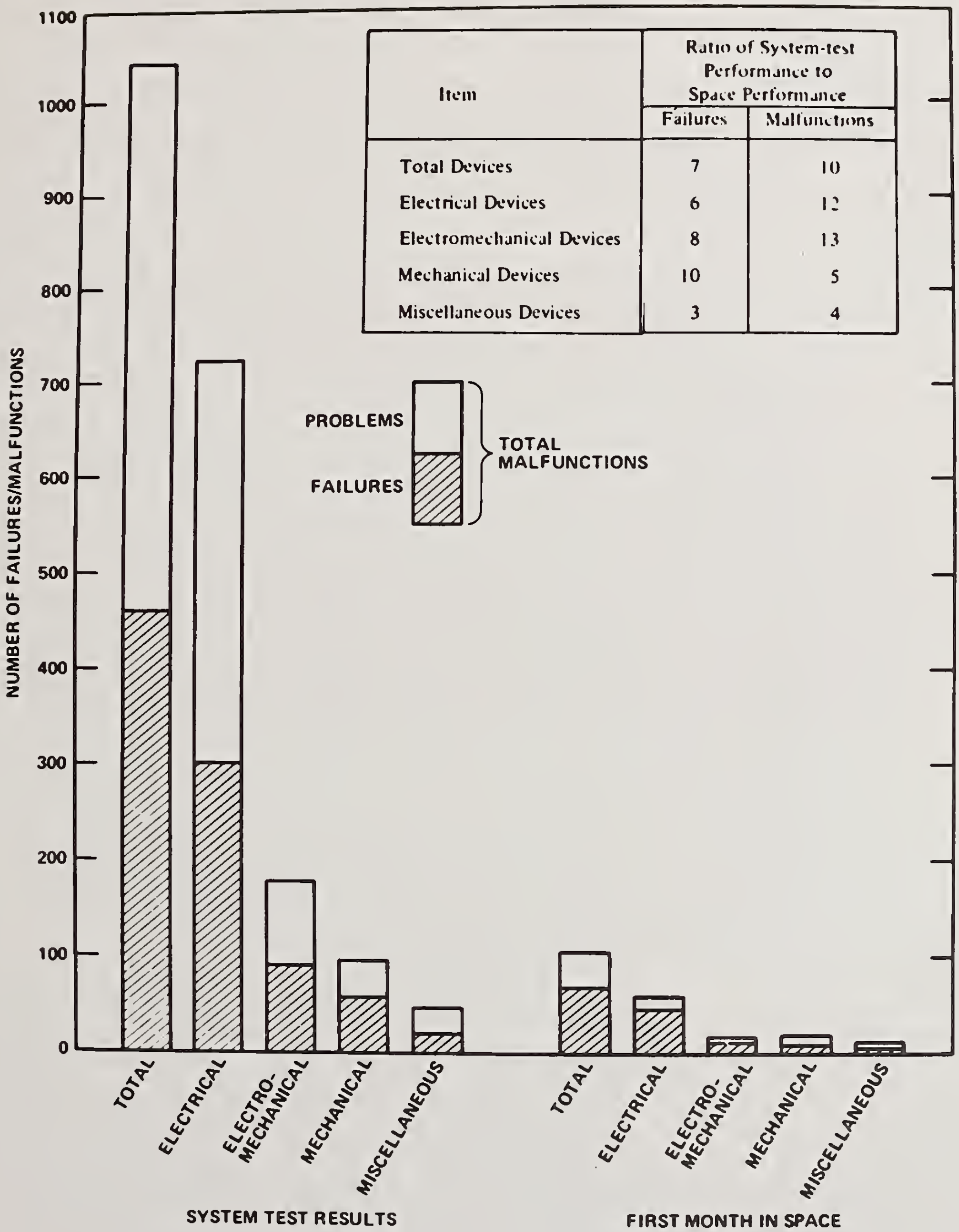


Figure 7. Spacecraft Test Problems (As Related to Subsystems).


80
70
60
50
40
30
20
10
5
0
TOTAL NUMBER OF FAILURES AND MALFUNCTIONS
PROBLEMS
FAILURES
MALFUNCTIONS
5
10
15
20
25
30
DAYS IN SPACE



Item	Ratio of System-test Performance to Space Performance	
	Failures	Malfunctions
Total Devices	7	10
Electrical Devices	6	12
Electromechanical Devices	8	13
Mechanical Devices	10	5
Miscellaneous Devices	3	4

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBS SP 487	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE Engineering Design--Proceedings of the 25th Meeting of the Mechanical Failures Prevention Group	5. Publication Date August 1977 6. Performing Organization Code
7. XXXXXXXX EDITORS T. Robert Shives and William A. Willard	8. Performing Organ. Report No.
9. PERFORMING ORGANIZATION NAME AND ADDRESS NATIONAL BUREAU OF STANDARDS DEPARTMENT OF COMMERCE WASHINGTON, D.C. 20234	10. Project/Task/Work Unit No. 11. Contract/Grant No.
12. Sponsoring Organization Name and Complete Address *(Street, City, State, ZIP)* Inst. for Matls. Res., NBS, Washington, DC 20234; ONR, Arlington, VA 22217, NADC, Warminster, PA 18974; Frankford Arsenal, Philadelphia, PA 19137; FAA, Washington, DC 20591; ERDA, Washington, DC 20545	13. Type of Report & Period Covered 14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number : 77-21810

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

These Proceedings consist of a group of twenty-three submitted papers from the 25th meeting of the Mechanical Failures Prevention Group which was held at the National Bureau of Standards in Gaithersburg, Maryland on November 3-5, 1976. The central theme of the Proceedings is the application of design failure prevention, with emphasis on the techniques of design, materials of design, design validation, and the goals, purposes, and requirements of design as related to product liability.

17. KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

Coating materials; composite materials; design; failure avoidance; failure prevention; lubricants; materials of design; product liability; reliability of design; techniques of design; validation of design

18. AVAILABILITY	19. SECURITY CLASS (THIS REPORT)	21. NO. OF PAGES
[XX] Unlimited [ ] For Official Distribution. Do Not Release to NTIS	UNCLASSIFIED	367
[X] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13.10:487 [ ] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151	20. SECURITY CLASS (THIS PAGE) UNCLASSIFIED	22. Price $5.25

USCOMM-DC 29042-P74

☆ U. S. GOVERNMENT PRINTING OFFICE : 1977--240-848/288

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, and chemistry. It is published in two sections, available separately:

- **Physics and Chemistry (Section A)**

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $17.00; Foreign, $21.25.

- **Mathematical Sciences (Section B)**

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $9.00; Foreign, $11.25.

To begin publication as a single journal, July-August, 1977

DIMENSIONS/NBS (formerly Technical News Bulletin)—This monthly magazine is published to inform scientists, engineers, businessmen, industry, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on the work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing.

Annual subscription: Domestic, $12.50; Foreign, $15.65.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a world-wide program coordinated by NBS. Program under authority of National Standard Data Act (Public Law 90-396).

NOTE: At present the principal publication outlet for these data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St. N.W., Wash. D. C. 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The purpose of the standards is to establish nationally recognized requirements for products, and to provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order **above** *NBS publications from: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.*

Order **following** *NBS publications—NBSIR's and FIPS from the National Technical Information Services, Springfield, Va. 22161.*

Federal Information Processing Standards Publications (FIPS PUBS)—Publications in this series collectively constitute the Federal Information Processing Standards Register. Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Services (Springfield, Va. 22161) in paper copy or microfiche form.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service. A literature survey issued biweekly. Annual subscription: Domestic, $25.00; Foreign, $30.00.

Liquified Natural Gas. A literature survey issued quarterly. Annual subscription: $20.00.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $30.00. Send subscription orders and remittances for the preceding bibliographic services to National Bureau of Standards, Cryogenic Data Center (275.02) Boulder, Colorado 80302.

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
Washington, D.C. 20234

OFFICIAL BUSINESS

Penalty for Private Use, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE
COM-215

U.S.MAIL

SPECIAL FOURTH-CLASS RATE
BOOK