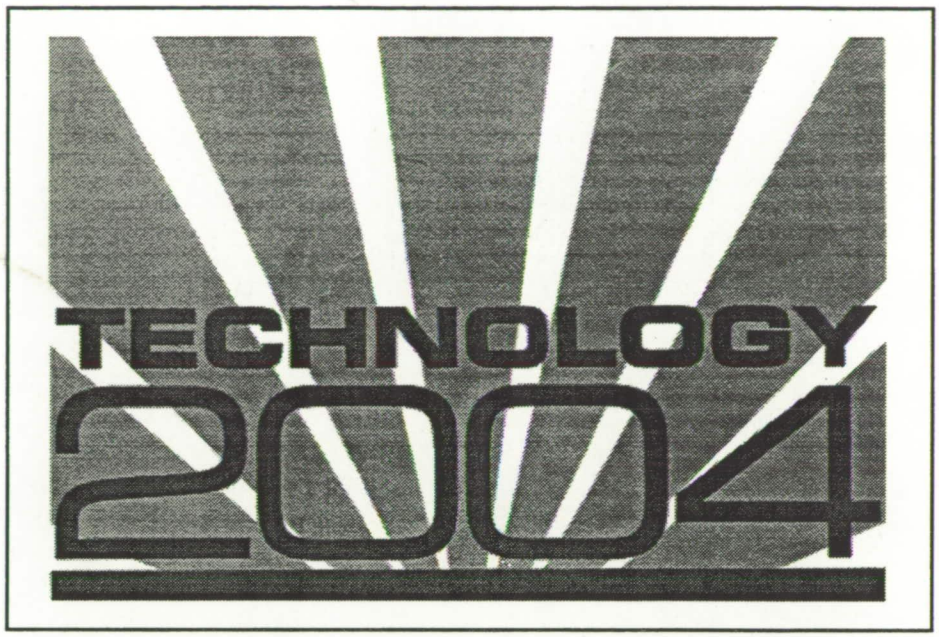
NASA Conference Publication 3313
Volume One



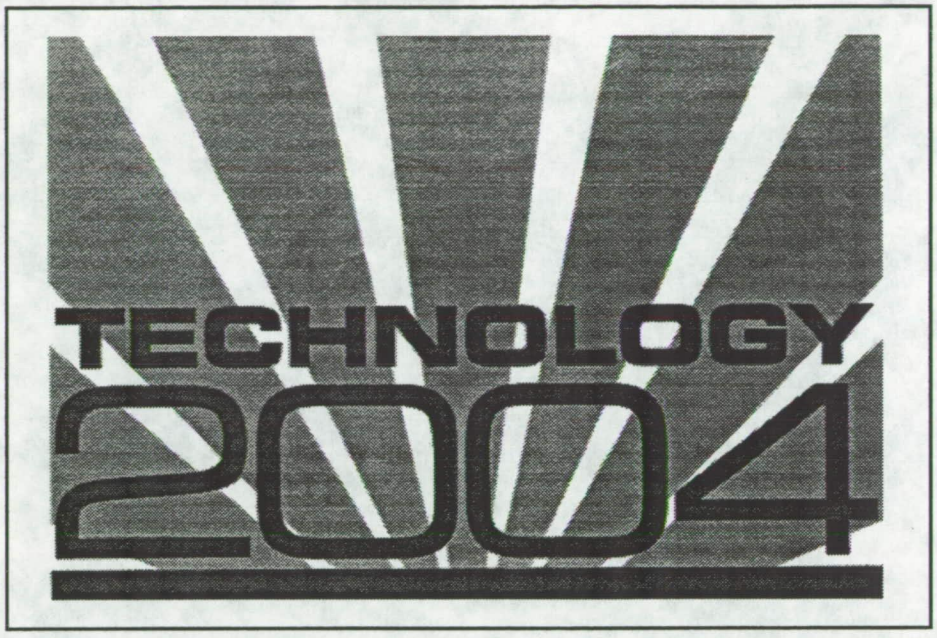
Conference Proceedings

The Fifth National Technology Transfer Conference & Exposition
November 8-10, 1994 • Washington, DC Convention Center

National Aeronautics and
Space Administration

**Scientific and Technical
Information program**

NASA Conference Publication 3313
Volume One



Conference Proceedings

The Fifth National Technology Transfer Conference & Exposition
November 8-10, 1994 • Washington, DC Convention Center

National Aeronautics and
Space Administration

**Scientific and Technical
Information program**

Table of Contents

Volume One

Advanced Manufacturing

Biotechnology

Computer Aided Design and Engineering

Materials Science

Low Cost Fabrication of Silicon Carbide Based Ceramics and Fiber Reinforced Composites

High Temperature Phthalonitrile Resins/Composites

Inorganic-Organic Hybrid Polymers for High Temperature Applications

Advanced Fiber Placement Program for Aircraft Structures

The Commercial Payoffs of the Ballistic Missile Defense Organization's Diamond Technology Initiative: Present and Future

Acoustic Attenuation and Vibration Damping Materials

Development of Environmentally Compatible Solid Film Lubricants

Unicoat—Durable, Environmentally Compliant Polyurethane Paint (Not available at this time)

Magnetic Levitation Transport System for Use in Underground Mining and Materials Handling

Nonlinear Optical Polymers for Photonic Applications

Advanced Materials for Strengthening and Reinforcing Concrete

Higher Durability, Higher Modulus Elastomers Through a Double Network Architecture

Power and Energy

Commercial Technologies from the SP-100 Program

Lithium Rechargeable Batteries for Space Applications (Not available at this time)

Piezoelectric Motor Development at AlliedSignal Inc., Kansas City Division

IMHEX Molten Carbonate Fuel Cells

Magnetic Bearings at Draper Laboratory Air-Lift Hydrohoist: An Innovative Underground-To-Surface Coal Handling System

Hybrid Chemical/Mechanical Heat Pump

Real-Time Utility Metering and DSM via Telephone: DOE-Sponsored Field Trial Results

Test and Measurement

Optical Binary Mixture Meter

Advanced Material Characterization Using Laser Induced Eddy Current Imaging

Microwave Antenna Holography System

Self-Contained Magnetic Field Measuring System

Advances in Commercial High Power Pulsed & CW Diode Lasers

2-Phase Quality/Flow Meter

Spindle Thermal Error Characterization and Compensation System for Machine Tool Accuracy Enhancement (Not available at this time)

A Family of Smart Acceleration Monitoring Systems

Advanced Manufacturing

Computer Aided Optics Manufacturing

Harvey Pollicove, Don Golini and Jeff Ruckman
Center for Optics Manufacturing
University of Rochester

Abstract

Flexibility and time-based performance are just two of the measures that differentiate manufacturers in the aggressive environment of global competition. In most industries, computer numerically controlled machining centers are considered a basic necessity in the battle to maximize flexibility, increase speed and achieve the consistent perfection required by today's quality-conscious buyer. But, while computer-assisted machining has been readily available in most industries, it was not an option for optics manufacturers until recently. In 1993, the computer-controlled *Opticam*® equipment developed at the Center for Optics Manufacturing became commercially available. Today *Opticam*® machining centers are eliminating the specialized tooling, long cycle times and special craft skills required with conventional equipment.

Computer-aided Optics Manufacturing

Harvey Pollicove, Don Golini and Jeff Ruckman
Center for Optics Manufacturing
University of Rochester

Flexibility and time-based performance are just two of the measures that differentiate manufacturers in the aggressive environment of global competition. In most industries, computer numerically controlled machining centers are considered a basic necessity in the battle to maximize flexibility, increase speed, and achieve the consistent perfection required by today's quality-conscious buyer. But, while computer-assisted machining has been readily available in most industries, it was not an option for optics manufacturers until recently. In 1993, the computer-controlled *Opticam®* equipment developed at the Center for Optics Manufacturing (COM) became commercially available. Today, *Opticam®* machining centers are eliminating the specialized tooling, long cycle times and special craft skills required with conventional manufacturing equipment.

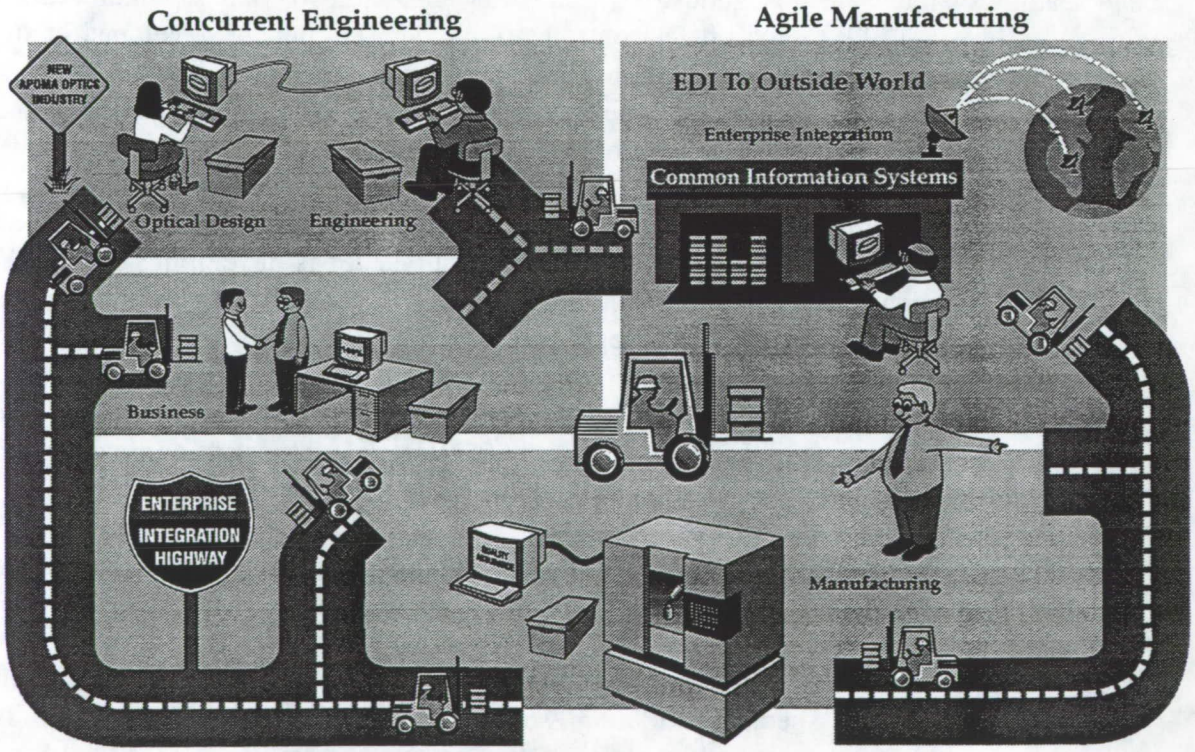
Under the sponsorship of the American Precision Optics Manufacturers Association (APOMA) and the U.S. Army Materiel Command, the Center is developing computer-aided machining centers and deterministic microgrinding techniques to automate optics manufacturing. Using *Opticam®* technology, precision optics can be produced faster, and with consistent precision. These accomplishments have been recognized by the Department of Defense (DoD) *Manufacturing Technology Achievement Award* and the industry's *Photonics Circle of Excellence Award.* COM's most important commendation has been *Opticam®'s* immediate industry acceptance and high implementation rate.

Opticam® computer-aided technology maximizes optics manufacturing flexibility and improves productivity. Trials by APOMA industry members, who use their own production parts to directly compare their in-house manufacturing time to *Opticam®* process time, routinely achieve cycle time reductions of 50% and decrease lead times even more dramatically. Specialized tooling is eliminated, process times are substantially reduced, and yields are dramatically improved. It is not unusual to attain 100% yields on trial lots, which are run by CNC machine operators, not skilled opticians.

Forecasting that future products will have more rigorous specifications, need more difficult shapes, require higher quality, and that customers will demand shorter delivery times is easy. Future business success is more difficult to predict, but a critical factor will be the technology chosen to meet the ever-increasing demands of the ever-more discriminating customers. The choice for the optics manufacturer is clear: continue using high-cost, non-deterministic, labor-intensive conventional processes; or implement cost-effective deterministic processes using computer-aided manufacturing technology.

Computer Integrated Manufacturing

Center development initiatives are redefining the manufacturing and competitive capabilities of the precision optics base. COM's development approach is much more comprehensive than simply performing isolated research or building new machine tools. COM's consortium effort has established an industry-designed, agile enterprise system. It is a systems approach that incorporates computer-aided concurrent engineering techniques to intelligently integrate the activities essential in optical design, engineering, manufacturing and distribution.



Opticim - Optics Computer Integrated Manufacturing

Opticim® integrates both the operational and production needs of the manufacturing enterprise. The system will replace today's isolated design decisions and enable industry concurrent engineering. It will also accommodate above-the-shop-floor operations, which provide the input and feedback needed to manage the enterprise. Opticim® is DoD CALS (Continuous Acquisition and Life-cycle Support) compatible and increases the effectiveness and capability of flexible automation.

Consortium-developed computer-aided procedures, knowledge-based instruction sets, and software modules have been combined into a system that integrates the activities performed during the design and manufacture of an optical device. A COM-industry created standardized *Optical Design Data File* provides the lens design parameters needed by the system to automate the generation of manufacturing and management data. It supplies all the information needed to process the part, including part set-up, tools, and the NC (numerical control) machining instructions. Parts can be designed, engineered and manufactured in real-time, or process information can be stored in a database to be later retrieved and downloaded to the machine controller.

Integrating Computer-aided design (CAD) and Computer-aided manufacturing (CAM) technology automates the fabrication process. Machine accuracy, on-board metrology and closed loop controls replace labor-intensive processes that previously required skilled hand-operations. Computer numerically controlled machining centers have the capability to produce lenses with levels of surface roughness and subsurface damage so low they do not require final polishing in many non-imaging and infrared optical applications. Lens grinding accuracy and the reduction of grinding damage are improved by a factor of ten over conventional processes.

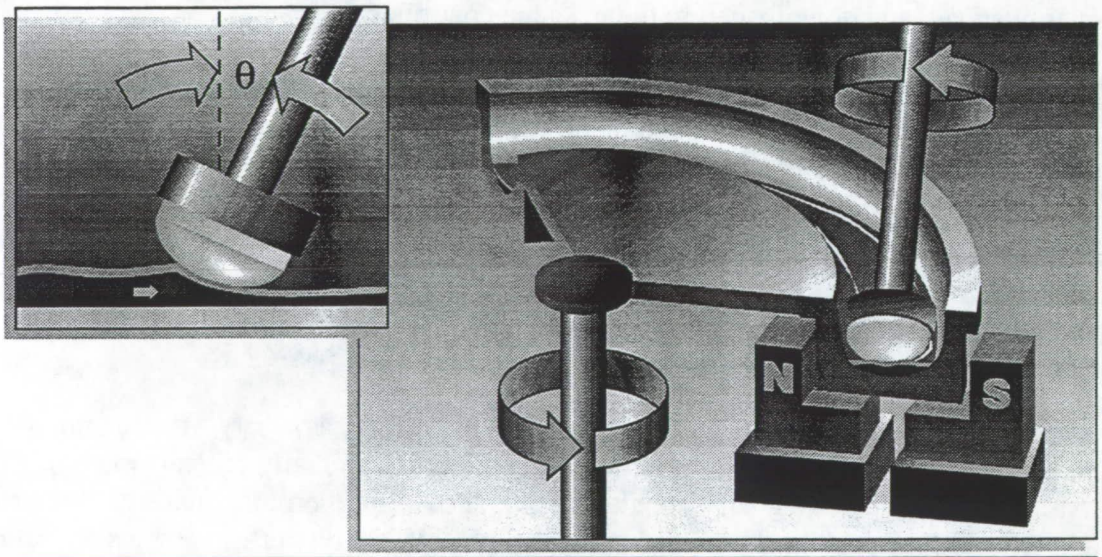
A robotic tool changer provides the flexibility to perform a variety of surfacing, edging, and centering operations. This additional flexibility is especially critical when the optical axis and outside part geometry are closely toleranced. Furthermore, automated metrology and closed-loop feedback control allow in-process correction and statistical process control is used to automatically flag process drift and reject conditions. The machine resident quality management system can also provide hardcopy inspection data without additional off-line labor. *Opticam®* machines eliminate the need for specialized optician skills and dedicated tooling, permitting inexpensive and rapid prototyping.

Deterministic Microgrinding

The expanded capabilities of *Opticam®* equipment result in a repeatable and predictable regime for machining brittle materials. The computer-controlled deterministic microgrinding process demonstrates a very high level of form accuracy ($\lambda/2$ peak-to-valley) and surface finish (<100Å rms). Application of deterministic microgrinding

technology to spherical lens production significantly improves industry competitiveness in the short term. The process takes about ten (10) minutes per surface and is unprecedented in its success.

Continuing COM developments will sustain those gains in the long term as the deterministic microgrinding process is optimized for a wide range of application to other optical and brittle materials. COM intends to build on these breakthroughs and extend deterministic microgrinding to the development of a family of computer numerically controlled machining centers. This will include micro-optics, cylinders, toroids and aspheres. Each machining center will evolve through concept development, design, component subsystem purchase and assembly, system level assembly and prototype machine acceptance testing and feasibility demonstration. These machines will automate the production of the full spectrum of refractive and reflective optics.



Magnetorheological Finishing: Spindle mounted optic is polished in the abrasive slurry supported by the compliant MR "fluid lap."

Magnetorheological Finishing

In parallel, a revolutionary new concept in precision polishing called magnetorheological finishing (MRF), is being developed and integrated with *Opticam*® machining centers. MRF, in concert with deterministic microgrinding, has the potential to finish optical

surfaces in minutes and to eliminate the conventional iterative pitch polishing process that uses skilled-labor and rigid laps.

MRF technology provides a deterministic surface finishing solution with computer control and an independence from specialized tooling. The process is best understood by thinking of the MR (magnetorheological) fluid as a compliant replacement for the conventional rigid lap in the loose abrasive grinding or polishing process. The magnetorheological fluid's shape and stiffness can be manipulated and controlled in real-time, allowing its application to any optical element geometry. It has notable application advantages when compared to the non-deterministic pitch or synthetic polishing process.

The Center plans to develop magnetorheological finishing techniques for all optical surfaces, regardless of symmetry, geometry, or slope variation. After process variables are determined, a computer algorithm will control the specific radial position and material removal function to accurately dictate the final surface shape. MRF will provide the capability to finalize the microground surface to pristine finish, figure and form. The combination of *Opticam®* deterministic microgrinding and magnetorheological finishing will redefine the manufacturing capabilities and competitive dimension of the precision optics industry.

Manufacturing Sciences

Underpinning these machining and finishing efforts, is a comprehensive manufacturing science and process technology development program designed to develop the tooling, techniques, material science, and metrology required to advance the precision optics industry. These tasks are being addressed by a manufacturing science consortium that includes the collaborative efforts of the University of Rochester and regional development centers at the University of Arizona's Optical Sciences Center, the University of Central Florida's Center for Research in Electro-Optics and Lasers (CREOL), several APOMA industrial members, and includes CRADAs (Cooperative Research and Development Agreements) with Lawrence Livermore and Los Alamos National Laboratories.

The primary objective of the manufacturing science program is to improve and extend industry manufacturing capabilities and optimize productivity. The near term goal is to establish manufacturing science basics that optimize current technology, while several program initiatives offer new approaches and solutions. The newest of these is the ARPA (Advanced Research Projects Agency) managed Active Vibration Cancellation program that will improve surface microroughness by applying active vibration cancellation technology that can sense and compensate for undesired vibration-induced errors at the cutting tool-part interface.

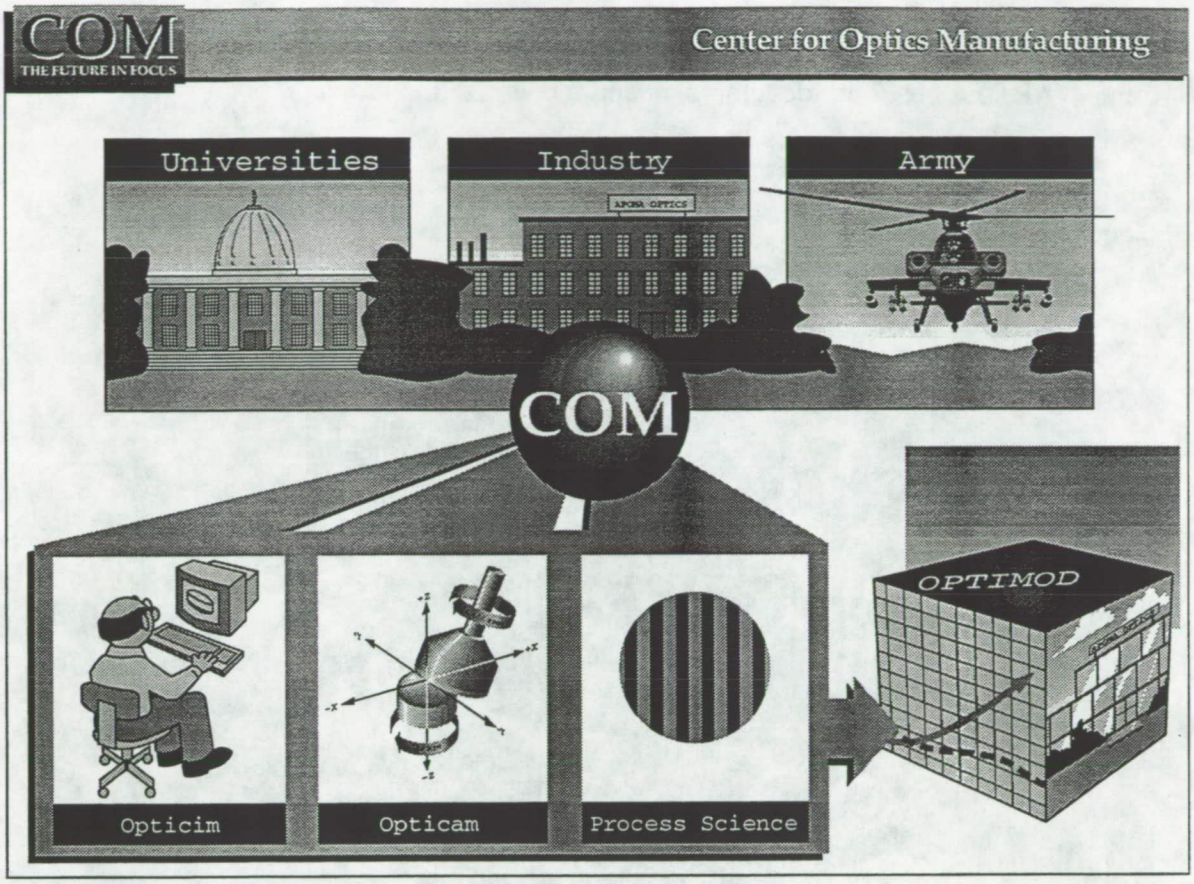
Another proposed program will develop an optical/electrical/mechanical fabrication and assembly system that can efficiently mass produce products that incorporate elements from each of several now independent engineering disciplines. Integration of the current stand-alone optical, mechanical, and electronic system and manufacturing infrastructures will provide the springboard for the creation of a new optomechatronics industry. Special challenges will include fabricating small, fragile components that require complex automated assembly to achieve the desired optical performance. This new assembly concept will include advanced manufacturing system design tools and state-of-the-art flexible assembly techniques that require new concepts of automation and control.

COM manufacturing science focus areas relate to the exploration of the complex interactions between material characteristics, properties, environmental conditions, and machining parameters for a wide range of brittle materials. To realize the full advantage of the computer integrated manufacturing environment, a database must be generated to enable real-time, rule-based process control for these more efficient machines. The long term goal is to constantly develop new optics manufacturing technology and maximize industry potential - a continuous improvement program for the optics industry.

Enabling Technology Implementation

COM's focused cooperative alliance is forging the optics manufacturer's course for modernization and continuous improvement. COM enabling technologies and modernization support will facilitate efforts to establish a competitive advantage in the global market and extend commercial opportunities. Computer-aided manufacturing technology offers significant economic advantage over current methods and will ensure the availability of competitive production processes that support the introduction of new commercial and military optical products. COM will continue to develop breakthrough advances to extend the state-of-the-art and bring new capabilities to the factory floor.

Industry implementation is the only hallmark of development success. To promote implementation, COM's Optimod program provides industry participants with the opportunity to determine firsthand how business challenges can be cost-effectively managed through the adoption of COM technologies. Industrial participants tryout newly developed manufacturing technology at the Center's User Laboratory, using their own parts and technical staff (who receive hands-on training during the visit). This unprecedented approach reduces industry's implementation risk and allows industry participants to actually project savings before capital commitment.

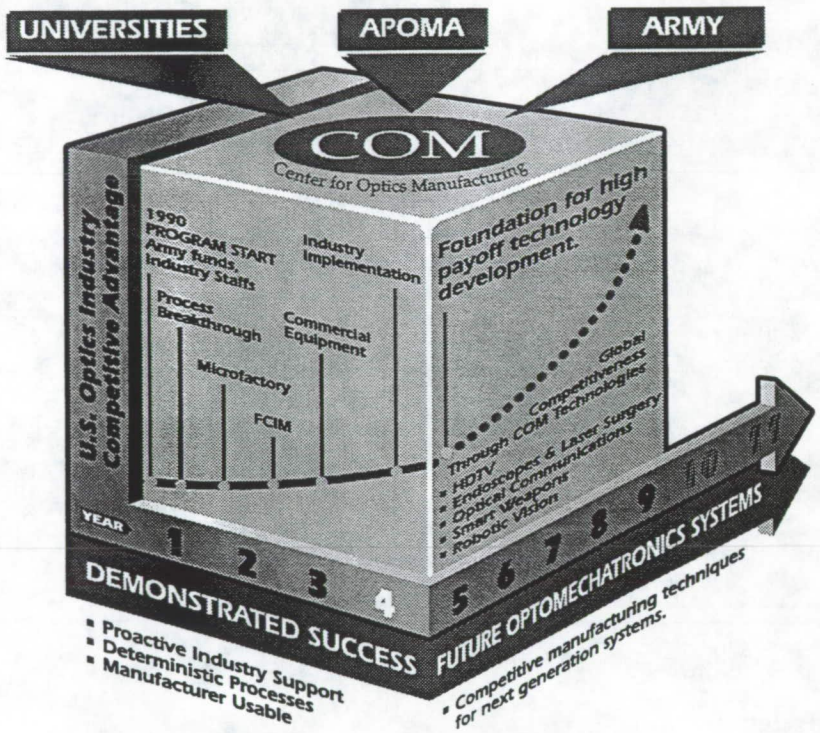


Industry Modernization

COM projects are governed by two mandatory rules: the need for and development of the technology project must be supported by industry; and the continuous and rapid transfer of technology to industry must be assured. The Center's most effective asset in meeting these objectives is its consortium alliance of experts from industrial, academic, and government organizations. These experts are proven leaders in optics manufacturing, and bring extensive insight and experience to the modernization process. They include scientists, engineers and practitioners - all APOMA industry and academic members - that accelerate the process of technology development, transfer, and implementation.

COM's model approach has been extremely successful at establishing a strategic partnership between industry, universities, and government that is providing the teamwork necessary to accelerate industry modernization. Drawn from this expert group is COM's Manufacturing Advisory Board (MAB), a cross-section of industry managers

that are the ultimate technology end-users. The MAB helps define industry needs and ensures that developments are cost effective and manufacturer usable. The U.S. Army provides substantial support through the Army Research, Development and Engineering Center (ARDEC) for both development and modernization projects.



COM has provided a framework for defining manufacturer usable research, development, and education that will maximize manufacturing efficiencies, expand product possibilities, and redefine business opportunities. The consortium fosters cooperation and understanding among APOMA industry members, universities, and government. COM's revolutionary systems approach is determining the optics industry's course for continuous improvement by providing the infrastructure that will speed implementation. These advantages and opportunities will propel the optics industrial base into the 21st century.

9-16-94

Lithographic News Ink Consisting of 100% Vegetable Oil and Pigments

Marvin O. Bagby and Sevim Z. Erhan
Oil Chemical Research, USDA, ARS, NCAUR, Peoria, Illinois 61604

ABSTRACT

A technology has been developed for preparing ink vehicles using vegetable oils in the complete absence of petroleum. From these vehicles, lithographic news-inks have been formulated over the complete range of viscosities required by the industry by adding appropriate pigments. The inks were evaluated for performance and quality characteristics and compared with the Newspaper Association of America (NAA) soyoil-petroleum formulation. Our vehicles are exceptionally light colored, thus inks may be formulated by using lesser amounts of pigments. Most ink formulations have low ruboff characteristics. Those tested on NAA pilot plant offset commercial press showed the desirable hydrophobic and emulsion characteristics, less than normal buildup on the various contact points and easy press cleanup. This technology should be cost competitive with petroleum-based inks with similar quality factors. Inks prepared by the technology have been evaluated for their potential biodegradation. Commercial news inks consisting of vehicles prepared with petroleum resin base and either mineral oil or soybean oil solvents were used for comparison. Results showed that pigments slowed the degradation of the ink vehicles. However, the USDA inks degraded faster and more completely than either of the petroleum resin-based commercial inks. Degradation averaged 80, 30 and 16%, respectively, for soy, partial soy and mineral oil vehicles.

United States agriculture produces over 16 billion pounds of vegetable oils each year. These domestic oils are extracted from the seeds of soybean, corn, cotton, sunflower, flax and rapeseed. Although more than 12 billion pounds of these oils are used for food products such as shortenings, salad and cooking oils and margarines, large quantities serve feed and industrial applications. The latter, representing about 300 million pounds of soybean oil, include applications such as plasticizers, emulsifiers, surfactants, plastics, resins and the recent new market for soybean oil in lithographic inks (40 million pounds annually, Private communication, American Soybean Association). Research and development approaches frequently take advantage of natural physical or chemical properties of the oils or their major constituent fatty acids. However, more often it is advantageous to modify those properties for specific applications. The latter approach has provided a variety of products, obtained by microbial and chemical modifications [1]. One example is the preparation of ink vehicles using vegetable oils in the complete absence of petroleum. From these vehicles, lithographic news-inks have been formulated over the complete range of viscosities required by the industry by adding appropriate pigments. The technology has been demonstrated with several vegetable oils to include a broad range of Iodine Values and representative fatty acids. This technology is cost competitive with petroleum-based inks with similar quality factors.

The following describes the preparation and properties of the vehicles and the lithographic newsinks.

SOYBEAN OIL INKS

Soy inks, alternatives to conventional petrochemical based inks, were developed by the Newspaper Association of America (NAA, formerly the American Newspaper Publishers Association, ANPA) and introduced to the marketplace in 1987 when the Cedar Rapids Gazette (Iowa) printed the first press run. This soy-petroleum-hybrid technology, consisting of about 35 to 50% degummed soybean oil, 20 to 25% petroleum resin and pigments [2], has enjoyed rapid acceptance by the newspaper publishers industry, especially so for the colored inks. However black inks formulated by the NAA technology were not cost competitive with typical offset news inks. Because the technology consists of a direct substitution of soybean oil for the mineral oil portion of the vehicle (entraining and dispersing agent for the pigments and other solid substances), other oils of similar fatty acid composition should be directly interchangeable. In fact, some formulators have prepared inks containing mixtures consisting of soybean and corn oils. Economic considerations and marketing strategies govern the final selection for applying that technology. Since then at the request of NAA and the American Soybean Association, USDA developed a technology in which the vehicle is totally derived from vegetable oils (3-6). Although soybean oil was emphasized because of dependable supply and economic factors, this new technology was demonstrated with several commodity oils. Those oils selected provide a broad range of iodine values and representative fatty acids consisting of saturated, mono-unsaturated, di-unsaturated and tri-unsaturated acids. Besides elimination of petroleum, this technology permits formulation of inks over a desirable, broader range of viscosity and is cost competitive with conventional

offset news inks. Further, inks formulated with this technology have low rub-off characteristics equal to those formulated and marketed as low rub inks (4). The following discusses the development and the characteristics of that technology.

We selected alkali refined canola, cottonseed, soybean, sunflower and safflower oils to demonstrate the technology [3,6]. Alkali refining removes the gums, waxes and free fatty acids. The presence of anyone of these materials will interfere with the desirable hydrophobic characteristics of the vehicle and the ultimate ink formulation.

This hydrophobic characteristic deserves further comment. The off-set printer plate or cylinder consists of two distinct areas. One area has been rendered hydrophobic (image area) while the non-image area is hydrophilic. Thus, the off-set printing process involves a two phase system consisting of an oil phase (the ink) and aqueous phase (the fountain solution). During the printing process, these phases must not form stable emulsions, or they will not separate properly on the printer plates. Poorly separated phases lead to smudged or ill-defined print. Understanding of this characteristic directed our attention toward techniques for modifying vegetable oils that would provide relatively non-polar products, i.e., low oxygen content polymers.

The vehicles were prepared from vegetable oils by two methods [3]. In the first method, vegetable oils were heat polymerized at a constant temperature in nitrogen atmosphere to a desired viscosity. In the second method, the heat polymerization reaction was permitted to proceed to a gel point, and then the gel was mixed with vegetable oils to obtain a desired viscosity. The apparent weight average molecular weights (Mw) of both the heat-bodied vehicles and the gels were determined by Gel Permeation Chromatography [5].

INK VEHICLE METHODS

Vehicles were prepared in a four-necked reaction flask equipped with a mechanical stirrer. Two major methods were used in preparation of polymers. (A) Alkali-refined vegetable oil was polymerized with stirring at $330\pm3°C$ under nitrogen atmosphere to the desired viscosity. Some polymers prepared by this method were used directly as vehicles, others having Gardner-Holdt viscosities as high as Z_8-Z_9, were admixed with low viscosity polymers and/or unmodified, alkali-refined vegetable oil at 65-75°C in a reaction flask equipped with a mechanical stirrer. (B) Heat bodying was continued until the oil gelled. The reaction was discontinued at the transition point when clumps of gel began to climb up the shaft of the mechanical stirrer. The gel was blended in various ratios with unmodified alkali-refined vegetable oil at $330\pm3°C$. The heating softened the gel and promoted blending. Agitation was continued until a smooth vehicle was obtained. The proportions of the gel and unmodified oil determined the resultant vehicle viscosities.

The viscosities and color of the vehicles were established by ASTM D-1545-63 (Gardner-Holdt Bubble) and ASTM D-1544-63 (Gardner Color Scale). The apparent Mw was determined by Gel Permeation Chromatography as tetrahydrofuran solutions [5]. For more detailed description of experimental approaches see References 3-6.

INK RESULTS AND DISCUSSION

The vehicles that we prepared typically had viscosity values in the range of G-Y on the Gardner-Holdt Viscometer Scale or about 1.6-18 poises [7]). These viscosities correspond to apparent weight average molecular weights (Mw) of about 2600-8900 [5]. As the oils are heated, they undergo polymerization and isomerization reactions. Thus, molecular weights and viscosities increase. The more highly unsaturated oils containing greater amounts of linoleic and linolenic and having the higher I.V., of course, react more rapidly.

Triglyceride, consisting of three fatty acids at which addition may occur, introduces the possibility of forming very complex structures and very large molecules. Thus, the reaction time necessary to reach a desired viscosity depends on mass, structure of the reactants, rate of heat transfer and agitation (8). Gelling times for safflower (I.V. = 143.1), soybean (I.V. = 127.7), sunflower (I.V. = 133.4), cottonseed (I.V. = 112.9) and canola (I.V. = 110.2) oils were 110, 255, 265, 390 and 540 min, respectively. Their respective polymerization rates ($K \times 10^{-3}$) at $330 \pm 3°C$ were 17.23, 11.02, 7.07, 3.21 and 2.14 [8]. The use of catalyst shortens the heating time by about 25 to 50% or can lower the polymerization temperature by 25-30°C. Although iodine values of cottonseed and canola oil are similar, canola oil with its high oleic acid content reacts more slowly and requires a longer time to gel. For comparison purposes, corn oil, with an I.V. of about 127 and an abundance of linoleic (62%), reactivity should closely parallel those responses shown for soybean and sunflower oils. Heat-bodied oils of different viscosities can be blended to produce a desired vehicle viscosity. Gels can, also, be blended with unmodified oils to give appropriate characteristics.

Vehicles, prepared by these technologies, are compatible with pigments for producing the four colors commonly used in the newspaper printing industry; namely, black, cyan, magenta and yellow. These vehicles are characterized by an exceedingly light coloration. Except for canola, they have values on the Gardner Color Scale of about 6 or less and typically are in the range of about 2-4. This property permits some reduction in the amount of pigment required for colored inks as compared to the pigment levels required by the much darker commercial vehicles [2,3] having Gardner Color Scale values of about 14 and greater.

Properties of ink formulated with soybean, cottonseed, canola, safflower and sunflower oils are characterized by viscosities in the range of 5 to 46 poises and by tack values of about 2 to 7 gram-meter (g-m). The typical viscosity for a black lithographic newsink is in the range of about 13 to 24 poises and about 5 to 12 poises for a black letterpress newsink. Tack values for lithographic inks are about 3.5 to 4.8 g-m and about 2.6 to 3.4 g-m for letterpress. The thickening effect of the pigment on the base vehicle should be considered when preselecting a vehicle viscosity. Because of the vehicle system we use [3], it is fairly easy to tailor the viscosity and tack values of the formulated inks. These inks, with a large range of viscosities and tack values, are suitable for both letterpress and lithographic applications. A variety of additives may be formulated into the inks including driers, lubricants, antioxidants and the like.

The water take up for lithographic ink performances was also tested, and the range of 20-30% water take up was well within the acceptable range of 20-36% [9]. Inks having these properties were also characterized by acceptable or superior ruboff values. The majority showed blackness of less than 6% after 2 hr, thus they demonstrate good rub resistance. With one exception, all formulations tested had ruboff values lower than that formulated by NAA's technology.

Properties of yellow, blue and red inks also meet the industrial standards. The addition of up to 5% thickening agent and/or optical brightener is an option but not a necessity for our color ink formulations. Elimination of the hydrocarbon resin from the vehicle results in significant savings for both black and colored inks. In addition, some reduction of pigment for colored inks due to the light coloring of vehicle lowers the price of the ink further.

Although the NAA soy-hybrid ink formulations and many presumably similar formulations have been advertised as biodegradable, until recently data regarding degradation were unavailable. Erhan and Bagby [10] reported biodegradation data obtained by evaluating the USDA 100% soybean oil vehicle and commercial vehicles consisting of petroleum resins and either soybean oil (NAA formulation) or mineral oil solvents. Both monocultures and mixed cultures of common soilborne organisms Aspergillus fumigatus, Penicillium citrinum and Mucor racemosus during a 25-day fermentation degraded the vehicles by about 86, 63 and 21%, respectively. An ANOVA model showed that the main effect of variation was due to duration of the fermentation and type of ink vehicle [11]. Subsequent biodegradation studies of the same vehicles in the presence of pigments as formulated inks gave similar results. However, the degradation was somewhat slower and less complete showing about 80, 30 and 16% degradation, respectively, in 25 days [12]. Thus, it is clearly demonstrated that, in terms of biodegradation, the USDA soy vehicles are far superior to the NAA soy-hybrid which, in turn, is superior to the 100% petroleum vehicle.

In 1987, 550 million pounds of letterpress and lithographic newsink were produced according to Census of Manufacturer. If a total conversion to our soy-oil ink occurs, about 2.5 billion pounds of soybean or 500 million pounds of soy oil will be used to supply the newsink market. That would be a significant increase over the current 300 million pounds of soybean oil now used for nonfood commercial products. One nonexclusive license has been awarded.

In summary, the USDA lithographic newsink technology readily satisfies the initial requests of NAA and the American Soybean Association by being (a) cost competitive with petroleum based inks of comparable quality characteristics, (b) completely free of petroleum solvent or resin in the vehicle and (c) usable over a broad range of viscosity and tack by exceeding the industry needs at both the high and low extremes. Further the ink is formulated in the absence of volatile organic compounds and is highly biodegradable, and the technology provides inks with the characteristics of low rub-off. During commercial printing tests, minimal accumulation appeared on machine rollers, and these residues are soft and easy to clean.

REFERENCES

1. Bagby, M.O. and K.D. Carlson, Chemical and Biological Conversion of Soybean Oil for Industrial Products, in Fats for the Future, edited by R.C. Cambie, published for International Union of Pure and Applied Chemistry by Ellis Horwood Limited, Chichester, 1989, Chapter 20, pp. 301-317.

2. Anon., American Newspaper Publishers Association, Manufacturing directions for soybean oil-based ANPA-ink, 1988, p. 1.

3. Erhan, S.Z., and M.O. Bagby, J. Am. Oil Chem. Soc. 68(9):635-638 (1991).

4. Erhan, S.Z., M.O. Bagby and H.W. Cunningham, Ibid. 69(5):251-256 (1992).

5. Erhan, S.Z., and M.O. Bagby, J. Appl. Polym. Sci. 46:1859-1862 (1992).

6. Erhan, Sevim Z., and Marvin O. Bagby, Vegetable Oil-Based Printing Ink, U.S. Patent 5,122,188 (1992).

7. Mattil, K.F., F.A. Norris, A.J. Stirton and D. Swern, Bailey's Ind. Oil Fat Prod., 1964, pp. 513.

8. Erhan, S.Z., and M.O. Bagby, J. Am. Oil Chem. Soc., 1994, in press.

9. Surland, A., A Laboratory Test Method for Prediction of Lithographic Ink Performance, Sun Chemical Corporation, 1980, pp. 5.

10. Erhan, S.Z., and M.O. Bagby, TAGA Proceedings, 1993, pp. 314-326.

11. Erhan, S.Z., M.O. Bagby and T.C. Nelson, TAGA Journal, 1994, in press.

12. Erhan, S.Z., and M.O. Bagby, Taga Proceedings, 1994, in press.

ADVANCED MANUFACTURING BY SPRAY FORMING: ALUMINUM STRIP AND MICROELECTROMECHANICAL SYSTEMS

Kevin M. McHugh
Idaho National Engineering Laboratory
Idaho Falls, ID 83415-2050

ABSTRACT

Spray forming is an advanced materials processing technology that converts a bulk liquid metal to a near-net-shape solid by depositing atomized droplets onto a suitably shaped substrate. By combining rapid solidification processing with product shape control, spray forming can reduce manufacturing costs while improving product quality. The Idaho National Engineering Laboratory is developing a unique spray-forming method based on de Laval (converging/diverging) nozzle designs to produce near-net-shape solids and coatings of metals, polymers, and composite materials. Properties of the spray-formed material are tailored by controlling the characteristics of the spray plume and substrate. Two examples are described: high-volume production of aluminum alloy strip, and the replication of micron-scale features in micropatterned polymers during the production of microelectromechanical systems.

INTRODUCTION

In spray forming, liquid metal is converted to a spray of fine droplets and deposited onto a substrate or pattern to form a near-net-shape solid. Researchers at the Idaho National Engineering Laboratory (INEL) are developing a unique approach for producing near-net-shape metals, polymers, metal matrix composites, and polymer matrix composites [1-7]. A liquid is aspirated or pressure-fed into a de Laval (converging/diverging) nozzle. There it contacts a high velocity, high temperature inert gas that disintegrates the liquid into very small (~ 20 μm) droplets and entrains the droplets in a highly directed spray.

Spray deposition with de Laval nozzles typically involves transonic gas-particle flow through the nozzle and subsonic free jet flow from the nozzle to the substrate. Momentum transfer from the atomizing gas results in droplet velocities of about 50 m/s. Despite their short transit time to the substrate, droplets arrive in semi-solid, solid, and undercooled states. This is due to the significant heat sink effect of the spray jet as it entrains large volumes of ambient gas. Droplets cool in-flight by convection at rates of 10^2 to 10^5°C/s, as determined from measurements of secondary dendrite arm spacings in solidified powders. Upon impacting the substrate, the droplets weld together to form a coherent solid while releasing the remaining enthalpy by conduction through the substrate. After impact the cooling rate drops to about 1°C/s. Droplet dendrite fragmentation at impact supplies a high concentration of nuclei that aid in structure refinement during equiaxed grain formation. Rapid solidification also reduces segregation compared to cast materials and can offer other property improvements by extending solid solubility. Economic advantages are inherent to the process because the spray-formed product is near-net-shape and fewer unit operations, such as machining, forging, or rolling, are required.

On-line diagnostics are used to characterize spray properties. An in-flight particle diagnostics system is used to simultaneously measure droplet size, velocity, and temperature in the atomized plume. This system measures particle diameters between 5 and 1,000 μm using an absolute magnitude of scattered light technique. Velocities of 10 to 100 m/s are measured with a dual-beam laser Doppler velocimeter; particle temperature is measured with a high-speed two-color pyrometry technique. Modeling the multiphase flow, heat transfer, and solidification phenomena provides guidance for component design and process control.

Properties of the deposit are tailored by controlling the characteristics of the spray plume (droplet size distribution, velocity, heat content, flux, and flow pattern) and substrate (material properties, surface finish, and temperature). Two examples are described: high-volume production of aluminum alloy strip, and the replication of micron-scale features in micropatterned polymers during the production of microelectromechanical systems.

ALUMINUM ALLOY STRIP

Currently, nearly all commercial aluminum sheet is produced by conventional ingot metallurgy (I/M) hot mill processing. Rapid solidification/near-net-shape processing by spray forming offers technical and economic advantages. Cost savings will largely result from process simplification and the need for fewer unit operations. In a recent technoeconomic analysis [8], industry savings of $15MM per year (10% overspray) or $27MM per year (no overspray) were estimated for an annual production rate of 2.1 billion pounds of spray-formed aluminum sheet. Energy savings for high volume strip/sheet alloys are substantial and result from eliminating intermediate, energy intensive, hot rolling steps necessary with conventional I/M processing. For example, each ton of spray-formed aluminum sheet saves 4.2×10^6 Btu compared to I/M processing. This amounts to a savings of 4.4×10^{12} Btu/year for the U.S. aluminum industry if 25% of current U.S. aluminum sheet production is converted to spray forming. Technical benefits include improved metallurgical homogeneity, elimination of macrosegregation, refined equiaxed grain structures, and improved impurity tolerance.

Strip Preparation and Properties

The approach used to produce aluminum strip on a continuous basis, depicted schematically in Figure 1, has also been used to spray-form steel strip and metal-matrix-composite strip. The alloy to be sprayed is melted under an inert atmosphere, superheated about 150°C, and pressure-fed into a de Laval (converging/diverging) spray nozzle of our own design. A high-temperature, high-energy inert gas flow rapidly disintegrates the molten aluminum into fine droplets that are entrained by the jet in a directed flow, and deposited onto a grit-blasted, water-cooled, mild steel drum. The resultant strip is detached and hot rolled to full density. A purged nitrogen atmosphere within the spray apparatus minimizes oxidation of the melt, surface oxidation of the strip, and in-flight oxidation of the atomized droplets.

Nitrogen is used as the atomizing gas because it does not react with aluminum under spray conditions. A gas-to-metal mass flow ratio (G/M) of about 0.3 is typically used, and metal mass flow rates are in the range 500 to 3000 lb/h per inch of nozzle width transverse to the flow direction. This nozzle dimension is scaled for a desired sheet width or mass throughput. For example, a 24 in. wide nozzle produces strip at a rate of approximately 12,000 to 72,000 lb/h. Overspray losses, defined as unconsolidated particulate and thin edge trimmings, are about 9% for a bench-scale (~3/4 in. wide) nozzle, and decrease as the nozzle is scaled.

The cooling rate of droplets in 6061 aluminum spray jets was estimated by measuring the dendrite cell size in polished/etched powders. Powder was partitioned into size bands using sieves of 300, 212, 177, 149, 125, 75, 63, 45, 38, 25, 20, 15, 10, and 5 μm. In general, dendrite cell size increased with increasing powder size, consistent with previously published results on gas atomized aluminum alloys [9-13]. For example, the cell size increased from about 1.8 μm for a 20 μm particle, to about 9 μm for a 200 μm particle. Cell size was found to follow a power law relationship with powder size. Measured cooling rates (Figure 2) ranged from about 10^2 to

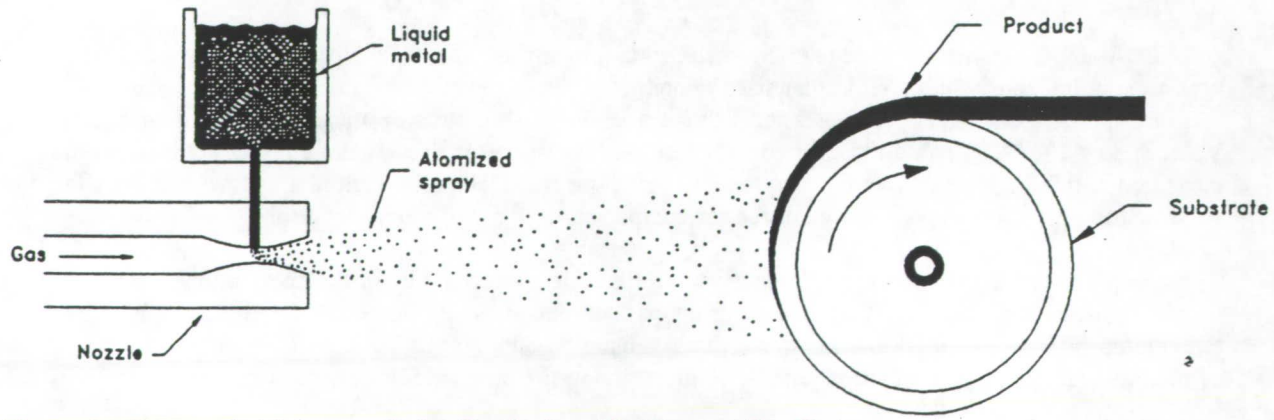


Figure 1. Schematic of Approach for Producing Aluminum Alloy Strip.

10^4 K/s depending upon powder size, which is well within the range of rapid solidification and at least one order of magnitude higher than the average cooling rate of the bulk spray deposit.

After a transit time on the order of milliseconds, the droplets impacted the substrate, positioned about 12 in. from the nozzle, producing a strip of metal > 0.020 in. thick. Strip thickness is dictated by the rate at which the substrate is swept through the spray, the metal mass flowrate, and, to a lesser extent, the thermal conductivity of the substrate. The transverse cross section of 0.25 in. thick 6061 aluminum strip shown in Figure 3 was spray formed with a bench-scale de Laval nozzle.

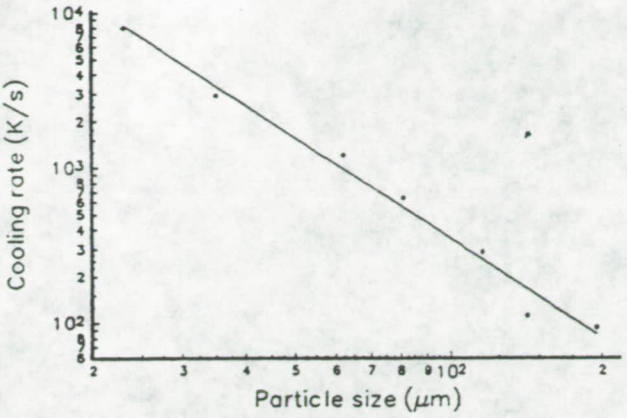


Figure 2. Cooling Rate of 6061 Aluminum Droplets.

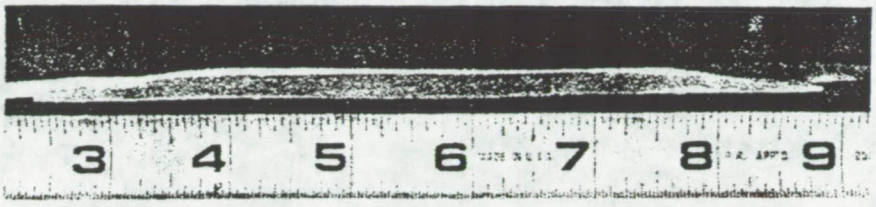


Figure 3. Transverse Cross Section of Spray-Formed 6061 Aluminum Strip.

The as-deposited density of spray-formed aluminum strip, measured by water displacement using Archimedes' principle, is 95 to 99% of theoretical. Photomicrographs of etched as-deposited 6061 (Figure 4a) and 3003 aluminum (Figure 4b) indicate a refined equiaxed microstructure with good constituent dispersion and no macrosegregation. Depending upon conditions, average grain size is 15 to 50 μm for these alloys. Figure 4c is a transverse photomicrograph of the strip in Figure 4b after hot rolling at 450°C to 44% thickness reduction. The average grain size was reduced by 25%.

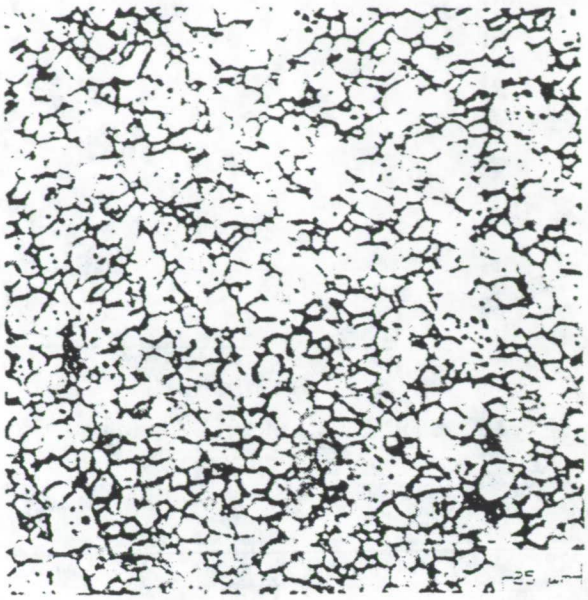
Room temperature tensile properties were determined for spray-formed 6061 aluminum after the samples had been solution heat treated and precipitation hardened to yield a -T6 temper. Results are summarized in Table 1.

The size distribution of overspray powders was analyzed using a Microtrac Full Range Particle Analyzer. Typical frequency and mass distributions are given in Figure 5. The mass median diameter, area mean diameter, and volume mean diameter of the powder were, respectively, 61 μm, 51 μm, and 71 μm. The geometric standard deviation ($\sigma_g = (d_{84}/d_{16})^{1/2}$) was determined from the cumulative mass distribution plot to be 1.7, indicating a narrow droplet size distribution in the spray plume. SEM analysis revealed that nearly all powder particles were spherical. These results compare favorably with commercial aluminum powders produced by other means.

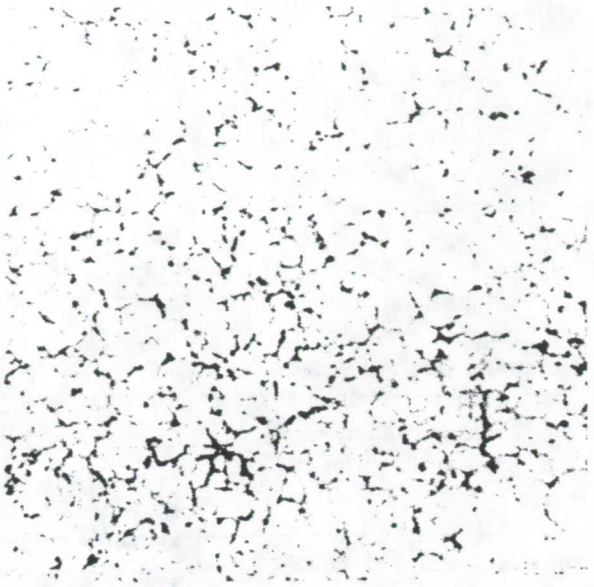
Table 1. Tensile Properties of 6061 Al Strip.

Sample	Yield Strength, 0.2% Offset, MPa (ksi)	Ultimate Strength, MPa (ksi)	Elongation in 50 mm, %	Elastic Modulus, GPa
Commercial 6061-T6	277.2 (40.2)	307.5 (44.6)	12.1	70
Spray Formed and Hot Rolled* 6061-T6	306.1 (44.4)	320.6 (46.5)	7.4	77

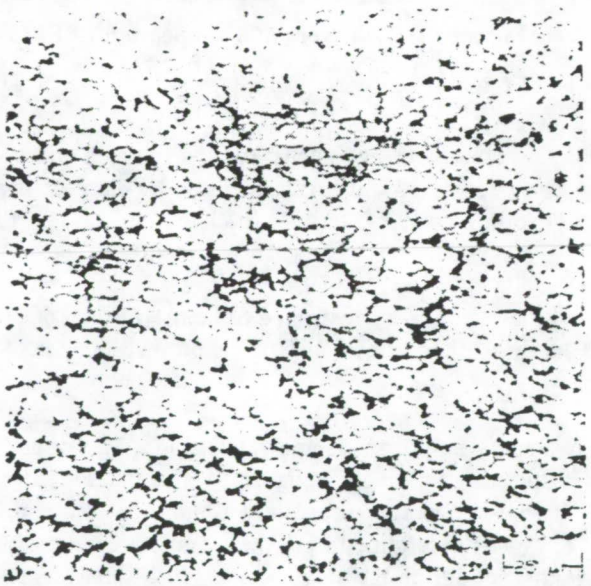
*80% thickness reduction



(a) As-Deposited 6061 Al



(b) As-Deposited 3003 Al



(c) Material of (b) after Hot Rolling at 450°C to 44% Thickness Reduction.

Figure 4. Microstructure of Spray-Formed Aluminum Sheet Alloys.

SPRAY FORMING MICROELECTROMECHANICAL SYSTEMS

The field of microelectromechanical systems (MEMS) is a rapidly growing, multidisciplinary research area. A long-term goal of this research is to merge microsystems of electrical, mechanical, and magnetic components into low-cost, highly-sophisticated control systems. Ultimately, it is believed that "smart" structures, such as embedded sensors, actuators, and control mechanisms, will be fully integrated into a structure, giving it the ability to sense stimuli and to take an appropriate response in a predetermined and controllable fashion [14]. Many of the potential applications of these systems in health care, biotechnology, industrial automation, automotive systems and many consumer products depend on the availability of a materials processing technology that will enable microminiature

parts to be produced from hard magnetic and other specialty alloys at low cost and at high production rates. At present, no such materials processing technology exists.

High aspect ratio micropatterned polymers (MPPs) consisting of three-dimensional microstructures can be generated by deep etch uv or x-ray lithography of a suitable polymer such as polyimide (PI) or poly (methyl)methacrylate (PMMA). Microstructures, such as miniature gears, turbines, actuators, and filters (or miniature molds of these microstructures) are produced with overall dimensions of about 10 to 500 μm. Deep etch lithography can be used in combination with electroforming to make metal microstructures, as exemplified by the LIGA method [14]. Alternatively, 3-D microstructures can be generated in silicon using etching techniques born in the 1960s [15]. MEMS production using etching or electroforming methods, however, places limitations on material choices and fabrication rates.

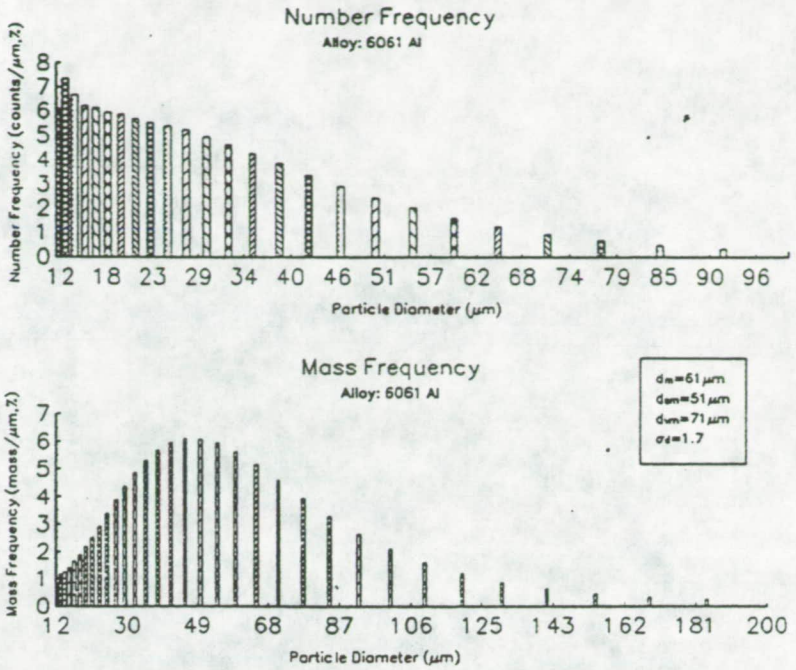


Figure 5. Frequency Distributions of 6061 Aluminum Overspray Powder.

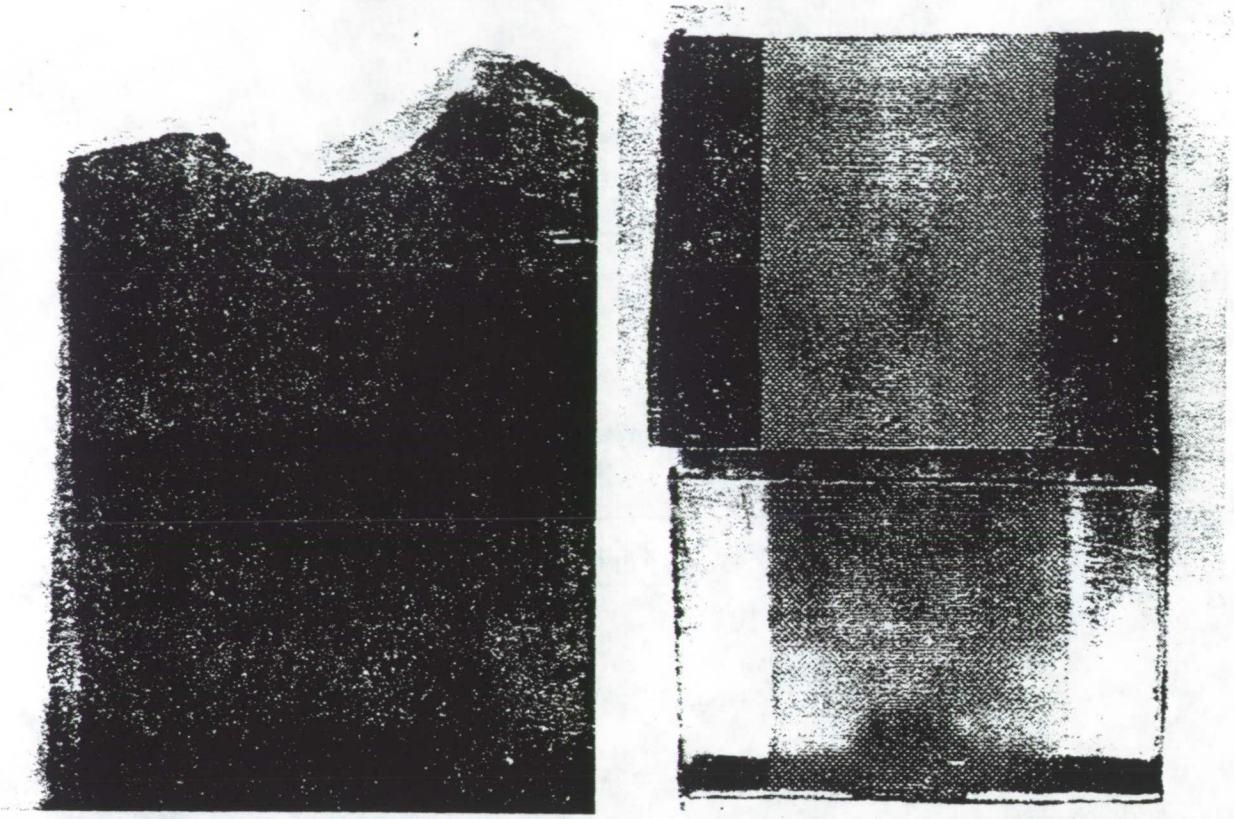
A study was conducted at INEL to examine the feasibility of producing MEMS by spray forming. In this approach, atomized metal droplets were rapidly quenched in-flight and deposited onto MPPs to produce complementary microstructures in metal. One benefit of this approach is that it is applicable to most metals of interest to this field, including hard magnetic alloys. Another benefit is that thin (~ 75 μm) films can be deposited onto MPPs at high rates using de Laval nozzles designed to produce wide spray patterns.

Approach and Results

To simulate a continuous network of microstructures, an ultra-fine wire (64 μm diameter) mesh screen and a miniature threaded rod (90 threads per inch) were hot pressed into plastic and detached. A variety of commodity thermoplastics and advanced polymers have been used as substrate materials including PMMA, polyetherimide (PEI), polypropylene (PP), polycarbonate (PC), and PI. Particular attention was paid to both PMMA and PI since these materials are commonly used in x-ray and ultraviolet lithography. Glass materials (soda-lime glass, annealed pyrex, and fused silica) were micropatterned by scratching the surface with micro-scribing tools.

Tin and Zn-3Al-11Cu were spray formed using bench-scale spray nozzles operated at G/M as high as 10. Deposits were separated from the micropatterned substrates by peeling or by dissolution of the substrate material. Deposit microstructure (grain structure and porosity) and surface finish were evaluated using optical microscopy. Conformality was evaluated using both SEM and optical microscopy.

With tin, melt temperatures of 400°C and atomizing gas temperatures of 300°C have yielded good results with all substrate materials tested. Replication of fine detail on surfaces has been excellent. For example, the image of a fingerprint that was left on a PMMA surface was clearly reproduced in a spray-formed metal deposit (Figure 6a). Likewise, the image of a 64 μm diameter wire mesh in PMMA was reproduced by spray forming tin onto the micropatterned polymer (Figure 6b). SEM photographs of the pattern and spray-formed metal deposit of Figure 6b



(a) Metal Fingerprint.

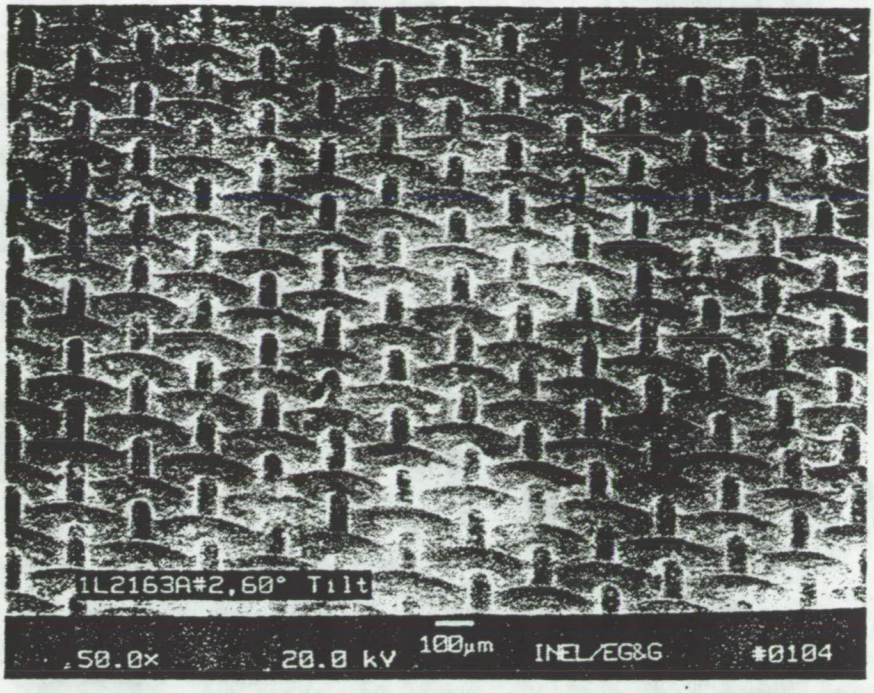
(b) Image of Fine Wire (64 μm diameter) Mesh in PMMA (lower area) Reproduced in Tin (upper area).

Figure 6. Replication in Metal of Micropatterned Polymers.

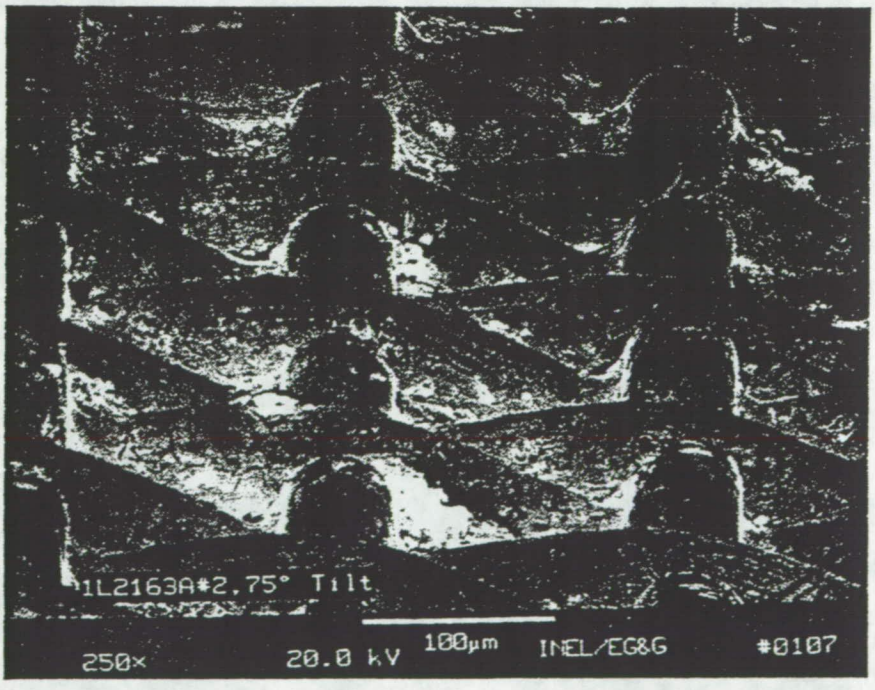
are shown in Figure 7. In Figure 8, SEM photographs show the replication in metal of the image of a 90 threads/inch threaded rod that was left on a PMMA surface.

At low metal throughputs (< 100 lb/h), similar results were obtained with the zinc-base alloy superheated to 600°C and atomized with argon or nitrogen heated to as high as 630°C. General purpose polymers gave good results despite maximum continuous use temperatures that typically do not exceed about 80°C [16], provided the droplets were lost sufficient heat in flight to avoid damaging the substrate.

At low metal throughput, the microstructure of sprayed deposits was not particularly sensitive to substrate material but the deposit surface finish and conformality were. Surface finish ranking for tin was typically found to follow the order glass > PMMA > PEI > PP ~ PC > PI, while conformality was found to follow the order PMMA > glass > PEI > PP > PC > PI. This order was also followed for the zinc-base alloy except that glass substrates gave poor results unless the substrate was preheated. For example, peak-to-valley surface roughness, measured with a stylus profilometer, dropped from about 60 μin. to about 4 μin. when glass substrates were preheated to about 100°C. This is presumably due to higher viscous energy dissipation for the zinc-base alloy, compounded by a higher surface tension which resulted in very rapid arrest of the spreading droplets on the cold substrate. Droplet impacts with polymer substrates were less sensitive to substrate surface temperature because of the low thermal conductivity of polymers. For a given metal, the order of these rankings depended on experimental conditions, but the order was reproducible for a given set of conditions.

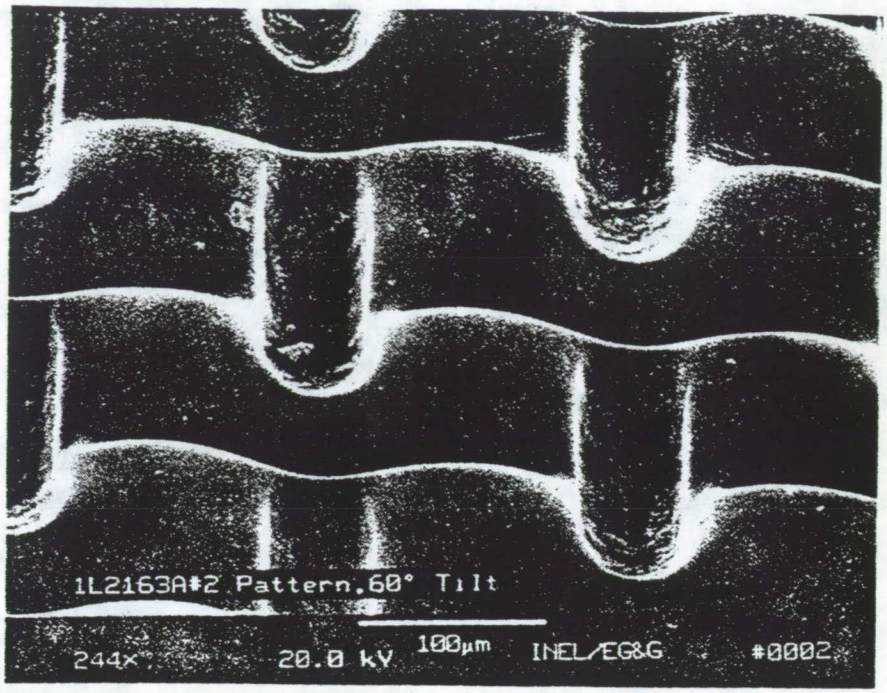


(a) Metal Deposit, 50X.



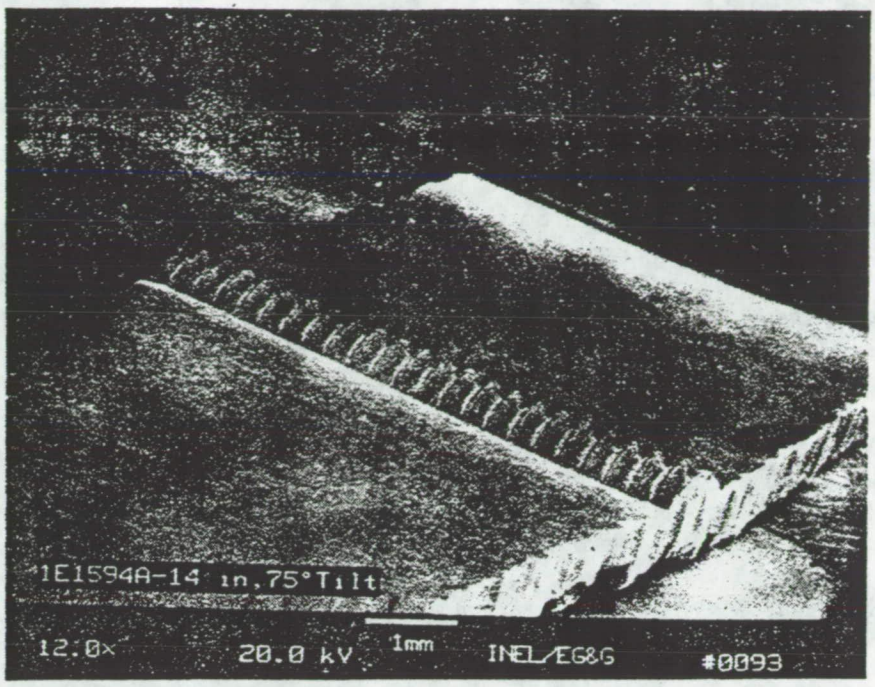
(b) Metal Deposit, 250X.

Figure 7. SEM Photographs of Metal Deposit and Micropatterned Polymer of Figure 3 (b).

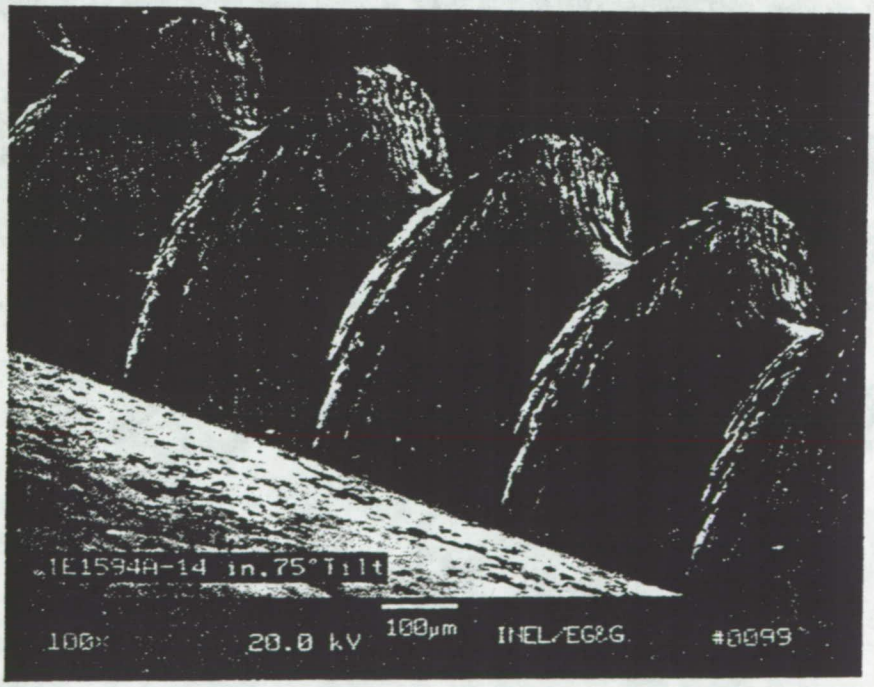


(c) Micropatterned Polymer, 244X.

Figure 7 continued.



(a) 12X.



(b) 100X.

Figure 8. SEM Photographs Showing Replication in Metal of the Image of a Miniature Threaded Rod Left in a PMMA Surface.

ACKNOWLEDGMENTS

The author gratefully acknowledges significant contributions of James Key, Woody Russell, Linda Wallace, and Bruce Wickham in this research. This work was supported by the U.S. Department of Energy, Office of Conservation and Renewable Energy, Office of Industrial Technology, and by the INEL Laboratory Directed Research & Development Program under DOE Idaho Operations Office Contract DE-AC07-94ID13223.

REFERENCES

1. K. M. McHugh and J. F. Key, "INEL Spray-Forming Research," *Proceedings of the Third National Technology Transfer Conference*, NASA Conference Publication 3189, 1, p. 15, 1992.
2. K. M. McHugh and J. F. Key, "Recent INEL Spray-Forming Developments," *P/M in Aerospace, Defense, and Demanding Applications - 1993*, ed. F. H. Froes, MPIF, Princeton, NJ, 177 (1993).
3. K. M. McHugh and J. F. Key, "Use of De Laval Nozzles in Spray Forming," *Thermal Spray Coatings: Research, Design and Applications*, ed. C. C. Berndt and T. F. Bernecki, ASM International, Materials Park, Ohio, p. 75, 1993.
4. K. M. McHugh and J. F. Key, "Performance Aspects of De Laval Spray-Forming Nozzles," *Proceedings of the ASME Fluids Engineering Conference*, June 1993.
5. K. M. McHugh, "Near-Net-Shape Manufacturing: Spray-Formed Metal Matrix Composites and Tooling," *Proceedings of the Fourth National Technology Transfer Conference*, NASA Conference Publication 3249, 1, p. 64, 1993.
6. K. M. McHugh, "Materials Processing with De Laval Spray-Forming Nozzles: Net-Shape Applications," *Thermal Spray Industrial Applications*, ed. C. C. Berndt and S. Sampath, ASM International, Materials Park, Ohio, p. 477, 1994.
7. K. M. McHugh and J. F. Key, "Use of De Laval Nozzles in Spray Forming," *J. Thermal Spray Technol.*, 3 (2), 191 (1994).
8. Spray Forming Aluminum: Phase II Technical Proposal.
9. I. C. Stone and P. Tsakiropoulos, *Internat. J. Rapid Solid.* 7, 177 (1992).
10. J. P. Lyle and W. S. Cebulak, *Metall. Trans.* **6A**, 685 (1975).
11. S. A. Moir and H. Jones, *Mater. Sci. Eng.* A173, 161 (1993).
12. C. G. Levi and R. Mehrabian, *Metall. Trans.* **13A**, 13 (1982).
13. H. Jones, *J. Mat. Sci.* **19**, 1043 (1984).
14. W. Ehrfeld and D. Munchmeyer *Nucl. Instrum. Methods Phys. Res.*, A303, 523 (1991).
15. K. D. Wise and K. Najafi, *Science*, 254, 1335 (1991).
16. R. B. Seymour, "Polymers for Engineering Applications," ASM International, Metals Park, OH, (1987).

ADVANCED CONTROL OF CUPOLA MELTING

E.D. Larsen, D.E. Clark, H.B. Smartt
Idaho National Engineering Laboratory
Mail Stop 2210
P.O. Box 1625
Idaho Falls, ID 83415-2210

K.L. Moore
College of Engineering
Idaho State University
Campus Box 8060
Pocatello ID 83209-8060

Paul King
US Bureau of Mines
Albany Research Center
1450 Queen Ave SW
Albany OR 97321

ABSTRACT

The cupola is a furnace used for melting pig iron, scrap steel, cast iron scrap, foundry return scrap, and ferroalloys to a specified tapping temperature and chemical composition. Its energy source is coke (processed coal); it has historically been the primary method for melting iron because of its low cost and simplicity. Cupolas can be quite large, up to 13 feet inside diameter, and a large foundry may operate several, but they can also be much smaller (as small as 18 in. inside diameter), to suit the needs of smaller foundries.

Although cupola melting as a process has a long history, automatic control has been elusive because the process has been poorly understood. The purpose of this work, which has been underway for about six months of an anticipated four year program, is to develop a controller for the cupola process using intelligent and conventional control methods. Improved control of the cupola process will lower material and processing costs, reduce scrap, and improve product quality. This will reduce foundry costs and improve the competitive position of U.S. foundries in the world market.

The Cupola Process

The cupola furnace [1,2] is one of the oldest processes for making cast iron, and is still the dominant method. It takes the form of a water-cooled vertical cylinder, as large as 4m in diameter and 20m tall (Figure 1). The charge consists of scrap metal, alloying ingredients, limestone, and coke for fuel; it is weighed and fed through doors near the top of the cupola. Air, usually heated and often enriched with oxygen, is blown in near the bottom through tuyeres extending a short distance into the interior of the cylinder. As the coke is consumed, the charge drops and melts. A continuous flow of iron emerges from the bottom of the furnace, at rates as high as 100 tons per hour. This molten iron is then transferred to casting lines in the foundry. The exhaust gases emerge from the top of the cupola, where they are processed to meet pollution standards.

The interior environment of the cupola furnace is very harsh. Extreme heat, chemical reactivity, along with the large mass flow and the presence of liquid slags and metals, tend to either plug or destroy sensors that might be introduced to measure operating conditions. Partly for this reason, and partly because traditional operating methods have evolved over

the centuries, the cupola is operated in open loop mode, with the expertise of experienced operators applied to obtain the desired melt rate, metal composition, and metal temperature.

Although cupola melting as a process has a long history, automatic control has been elusive because the process has been poorly understood. The purpose of this four-year program, which has been underway for about six months, is to develop a controller for the cupola process using intelligent and conventional control methods. Improved control of the cupola process will lower material and processing costs, reduce scrap, and improve product quality. This will reduce foundry costs and improve the competitive position of U.S. foundries in the world market.

TECHNOLOGY TRANSFER POTENTIAL

There are economic reasons why this is a good time to pursue the advanced control of cupola melting. Beginning in the 1950s, the need for pollution control devices complicated cupola melting and altered its economics. Stiff foreign competition is also a factor; lower labor and material costs, foreign government assistance (import restrictions and export incentives), less stringent emission control requirements, lower capital costs, and government subsidies have resulted in 30-50% lower import prices to the U.S. Nonetheless, the cupola remains the primary tool for producing liquid iron for castings, currently accounting for 70% of cast iron melting in the U.S., or 14 million tons per year. [3] Hundreds of cupolas are in operation, and the potential for quality improvements and energy savings is substantial. A comparison between the domestic and foreign foundry industries showed no significant differences in technology; it was noted, however, that the foreign industry typically had better trained operators, resulting in better control of the process, which would indicate that automatic process control could benefit the domestic foundry industry.

The time is also ripe for application of several recent technical developments in intelligent control to cupola melting. Neural networks, fuzzy logic, and expert systems have been independently developed to the point where their application to a variety of processes is easier and better understood than it was only a few years ago. A finite difference model of the cupola that can compute chemical as well as thermal conditions has been developed by the American Foundrymen's Society. This model can be used to simulate the process for training neural networks and tuning intelligent controllers. Because of the complexity of the chemical and physical processes in cupola operation, there are major inefficiencies, e.g., only 35% energy efficiency and oxidation losses of 10-50% of the valuable alloying elements (Si, Mn) in the charge. These inefficiencies can be reduced by better control of the process.

Current Operating Practice

Since cupolas have been used for centuries, and some individual cupolas for many decades, a great deal of empirical data about their operation has been gathered, both locally and in published handbooks, and they are operated according to experience and tradition.

Cupolas are typically operated for extended periods of time, called "campaigns." A typical campaign might run a week, to coincide with the molten metal requirements of a foundry running one or more shifts, but shut down for the weekend. Longer campaigns up to several weeks are sometimes used. In any case, the operating cycle consists of starting the cupola (igniting the initial coke charge), then gradually bringing the charging components (coke, limestone, scrap steel, scrap iron, and alloying elements) to their desired continuous values. Blast rate and temperature and oxygen addition, if any, are similarly adjusted. At steady state, the charge falls more or less continuously as the coke is consumed and the metal is melted. The molten metal falls in drops through the coke bed,

collects in the bottom of the cupola, and eventually runs out a tap hole. From there it runs to a holding furnace, or into crane-borne buckets directly to the casting line. Metal composition is important, because the solidification characteristics and the mechanical properties of the final casting depend on maintaining composition within rather narrow limits. Metal temperature is important because the temperature at pouring determines whether the mold will fill properly. Both composition and temperature have an effect on solidification, microstructure, and final casting properties.

At the end of the campaign, the sand-covered bottom of the cupola is dropped, and the entire mass of the charge falls into a pit below. After the cupola cools down, maintenance on the refractory lining and water cooling systems can be performed. The bottom is propped back into place and sand tamped in, the furnace is recharged, and the cycle begins again.

Major variables for cupola control are shown in Table 1. Blast rate is controlled by changing the speed of the mechanical blowers; blast temperature by regulating the input to the heaters; oxygen by regulating the oxygen flow. These variables have short time constants and little delay. The metal charge ratio and the coke ratio, however, are changed by altering what is charged, i.e., loaded into the top of the cupola. It takes some time, up to an hour, for these changes to propagate with the slowly sinking charge into the active zone of the cupola.

Table 1: Cupola Control Variables

Output	Primary Input	Secondary Input
Iron Temperature	Blast Rate Blast Temperature Oxygen Addition	Coke Ratio
Iron composition (Carbon, Silicon, Manganese)	Metal Charge Ratio	Coke ratio Oxygen addition Blast temperature
Melting Rate	Blast rate Coke ratio	Oxygen addition

In current practice, these variables are not automatically controlled; while an independent control loop might be used, for example to maintain constant blast rate, it is up to the operator, based on experience and on procedures developed for a specific cupola and foundry system, to select the blast rate and other inputs to achieve the required outputs. When the casting line needs no metal, because molds are full, cranes are busy, or it is shut down for a shift or longer, the cupola can be "banked," with a reduced blast rate. There are transient effects associated with this event, which are more severe the longer the shutdown. A feedback control system would also be desirable during the startup phase, to reduce time and material consumption, and to avoid the formation of unusable "transition iron."

The task of the present work is to design a feedback controller for the cupola, using a combination of conventional and advanced control techniques. The scheme for this is shown in Figure 2. This control configuration is characterized by the use of both feedforward and cascade control.

AFS Numerical Model

The American Foundrymen's Society, an industry group with a strong interest in cupola control and optimization, has developed with DOE matching funding, a numerical model of the cupola process. It is a one-dimensional, steady-state code that allows for various

material properties (chemical and geometrical), chemical reactions, heat transfer, various charge materials, and various input parameters (blast rate and temperature, oxygen addition, etc.). It is coded in FORTRAN, solved by finite difference techniques, and takes 5-10 minutes to converge on a set of conditions using a typical 33MHz 486 processor. It is expected that improvements in the algorithm and the use of faster microprocessors will eventually reduce this time to under one minute. This model has been verified with experimental data, and is expected on the market within the next year with a user friendly Windows™ interface. This software will prove very useful to cupola operators in assessing the metallurgical and economic effects of parameter changes. By correlating variables that are relatively easy to sense (such as offgas composition and temperature) with others of more immediate interest that are quite difficult to sense (such as melting rate), the model also opens up new feedback control paths.

The calculation time of the model is, however, probably not fast enough for real time feedback control of the cupola, since input conditions such as changes in blast rate may change over a shorter time than it can recalculate. In addition, operator skill levels, the foundry physical environment, and cost considerations preclude the use of fast workstations in the foundry. In practice, if the algorithm cannot operate within a PC/PLC hardware environment, it will probably not be used at all.

NEURAL NETWORKS

Neural networks have recently become more popular as a way of dealing with nonlinear systems with many inputs and outputs [4]. They consist of neurons (Fig. 3), processing elements that do relatively simple calculations, summing inputs and applying a nonlinear function, producing a single output. Many such elements, arrayed in several layers and variously connected (the "neural network"), can give fast answers to complex problems (Fig. 4). The "neural" nomenclature applied to artificial neural networks (ANNs) is a deliberate reference to the way biological nervous systems are perceived to function. ANNs are typically implemented in software.

Like biological nervous systems, artificial neural networks must be trained to solve particular problems. Initial work with ANNs in the 1960's was limited because of the lack of learning techniques, but the development of the backpropagation of errors (or simply backpropagation) algorithm provided a simple, iterative way to train networks. This algorithm can take many iterations, often requiring hours of computer time, using a training set consisting of known inputs and outputs, to set the weights multiplying the input to each neuron. Once the network is trained, however, it operates in the feedforward mode, where only summations and multiplications are required.

Applications for neural networks include predicting the state of systems that show deterministic chaotic behavior, robotics, speech recognition, and (some people will tell you) predicting the stock market. The details of picking the right network topology and training methods for a particular application are, to some extent, a matter of trial and error, though they are also incorporated into commercially available ANN software, and extensive information on this topic has appeared in the open technical literature in recent years.

Neural Network Implementation of the AFS Model

A neural network can be trained to approximate the output of the AFS model by making a number of model runs, varying the inputs over a range of typical operating parameters. The trained network approximates the AFS model quite accurately. As long as the model is operating within the range of inputs covered by the training set, it can be expected that the network will allow accurate interpolation; this has been shown in results obtained thus far. Since the neural network is only an interpolation of a data set, however, and not a valid physical model (neural networks don't do physics) it is not to be trusted outside the

original training limits. Nonetheless, because large data sets can be created easily, and networks trained from them in an automated fashion, fairly complete coverage of the operating space of the cupola can be obtained. Because the neural network operating in the feedforward mode is very fast (only a few hundred addition and multiplication operations are required in this instance), the calculation of outputs from a given set of inputs seems instantaneous to the user, and even a quite modest personal computer can provide a response fast enough for control purposes.

The training sets provided by the AFS model are used, after some preprocessing, in neural networks implemented under NeuralWare,™[5], a commercial neural network product. The network was trained using conventional backpropagation techniques.

Forward and Reverse Models

The operator can use the neural network approximation of the model to play "what-if" games-- changing coke ratios or oxygen enrichment, for example, based on their effects on iron production rate, or factoring in the prices of inputs to do economic analyses. But what the operator really wants to know when it comes to setting the cupola control knobs requires a "backwards" model-- given the desired outputs (melting rate, iron temperature, carbon content, etc.), what input settings (blast rate and temperature, coke ratio, oxygen enrichment, etc.) are required (Fig. 5)? The AFS model, or the neural network approximation to it, would require an iterative approach, which can be time consuming when several inputs and outputs are involved because of the complex interaction of the variables.

By training the neural network from the same training sets, except using the process outputs as inputs to the neural network, a reverse model is obtained that allows the setting of all the outputs, and gives, virtually instantly, the required inputs (Fig. 6). The reverse model is useful for providing control loop set points, allowing the process to come to equilibrium faster, with less "transition iron"-- material that is out of spec because it was produced during transient conditions in the furnace. In control terms, it serves the purpose of linearizing the plant so that simpler PID or fuzzy logic controllers may be used.

LabVIEW User Interface

In addition to making the model more usable for real time control, the model-trained neural networks have been reduced to C-code and implemented in LabView™, [6] a graphical programming environment suitable for sensing and control that creates "virtual instruments." Slide-type indicators and controls allow the user to visually examine the effects of changing the variables, observing which are covariant and which countervariant, and which have minima or maxima as some other variable is changed. The forward and reverse directions can be selected with a switch on the panel. In the "forward" mode, the model mimics the input/output relationships of the cupola: one can test the effect of various input values by moving the sliders and watching the virtually instantaneous results. In the "reverse" mode, the desired output conditions can be set, and the model will give input settings that will achieve them at steady state. It should be emphasized that the AFS model is steady state only, and says nothing about transitions. These are being measured experimentally for incorporation into the control scheme.

Such an interface allows a user to deal with the complex, nonlinear model of an industrial or agricultural process in terms of familiar variables, and, like other neural network applications, it does not require a mathematically rigorous model of the process, but can also be adapted to empirical data. It is a compact, inexpensive, and useful tool, even if automatic control based on the underlying model is not implemented.

Control Approach

A controller is generally developed by: (1) constructing a model of the physical system, (2) analyzing the model to determine the properties and dynamic response of the system, (3) designing a control algorithm which, when coupled with the model of the system, results in the desired closed-loop behavior, and (4) implementing the controller through either hardware or software. The finite difference model of the process provided by the American Foundrymen's Society will be used for controller development. The steady-state conditions and the dynamic response of cupolas will be determined from the model and from experimental data. The Bureau of Mines operates an experimental-sized cupola with an 18in. inside diameter and a melting capacity of 1500 LB/hr that will be used to verify the AFS mathematical model for this cupola and to obtain transient data that, along with industrial data, will be used for tuning and validating the controllers. An artificial neural network trained using the backpropagation algorithm will be used to learn the inverse relationship of the steady-state process for use with the controller and to provide the initial conditions of the process.

OTHER APPLICATIONS

The neural networks use has expanded substantially in the past few years as faster computers, more effective algorithms and more efficient user interfaces have been developed. It is actually quite easy for a relatively naive computer user to produce neural nets that yield useful information about diverse applications, that would have been difficult or impossible to come by via other techniques. A number of small companies with expertise in the mechanics of handling neural nets have sprung up that do just this, but it is also possible for a person with expert knowledge of the system of interest to sit down at a PC with a packaged program and do the same thing.

It helps to have a basic understanding of the process that is being controlled. As an expert, such a person knows which questions to ask, and what kinds of answers are reasonable. The understanding may be mathematically based, using models or statistical inference, or it may be largely experiential, in which case techniques such as fuzzy logic and tools for developing expert systems may be useful.

Neural networks can linearize a nonlinear plant in a control scheme, as used in the cupola control scheme shown in Figure 1. Neural networks can learn and control systems characterized by deterministic chaos. Although there is no indication that the cupola exhibits such behavior, other systems do, such as arc furnaces for steel making, and neural networks have been successfully applied to these control problems.[7]

FUNDING ACKNOLOGIMENT

This work is being performed at the Idaho National Engineering Laboratory (INEL), the U.S. Bureau of Mine Albany Research Center, and Idaho State University under funding from the U.S. Bureau of Mines under Contract No. J0134035, and from the National Technology Transfer Center under Contract No. B17961, through Department of Energy Contract No. DE-AC07-76ID01570. Proposals have been submitted jointly by the INEL and the American Foundrymen's Society for further work under the DOE Metals Casting Competitiveness Research Program.

REFERENCES

[1] American Foundrymen's Society, Cupola Handbook, 5th ed., American Foundrymen's Society, Des Plaines, IL, 1984.

[2] American Society for Metals, Metals Handbook: Melting and Casting of Ferrous Metals, 8th edition, vol. 5, 1970.

[3] Kaiser, Francis T, 1994. "Why Melt Cupola?" Proceedings of International Cupola Conference, March 1-3, 1994, Rosemont, IL, American Foundrymen's Society, Des Plains, IL, 1994.

[4] Hinton, G. E., "How Neural Networks Learn from Experience", Scientific American, September 1992, pp. 145-151.

[5] NeuralWare Incorporated, Penn Center West Building IV, Pittsburgh PA 15276-9910 phone 412.787.8222

[6] National Instruments, 6504 Bridge Point Parkway, Austin TX 78730-9824 phone 800.433.3488

[7] Staib, W. E., N. G. Bliss, and R. B. Staib, "Development in Neural Network applications: The intelligent electric arc furnace", Iron and Steel Engineer, June 1992, pp. 29-32.

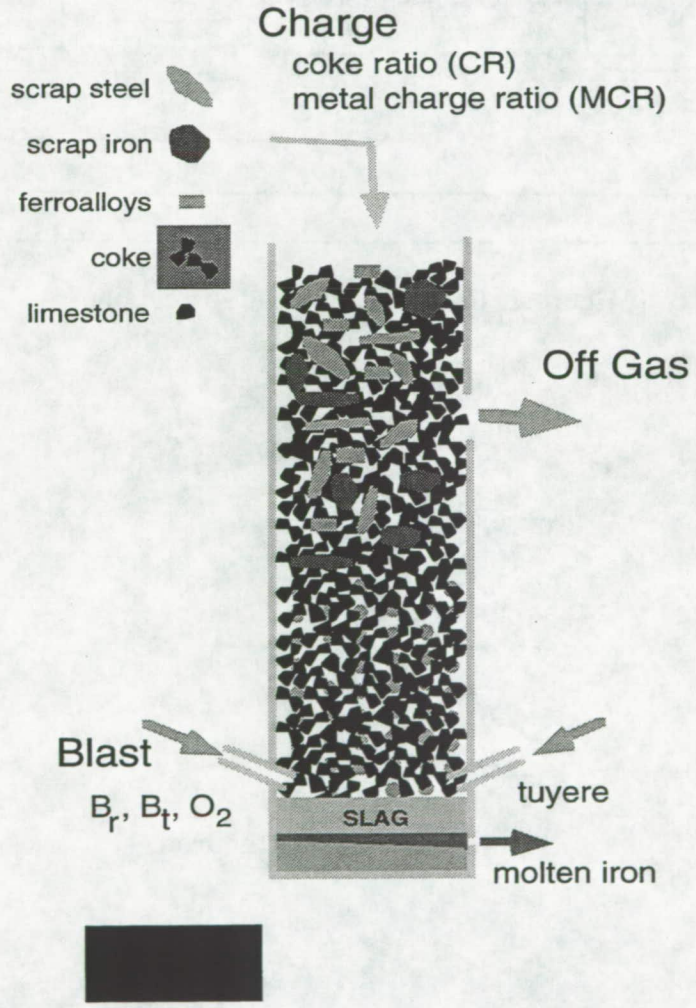


Figure 1. Basic Cupola Schematic

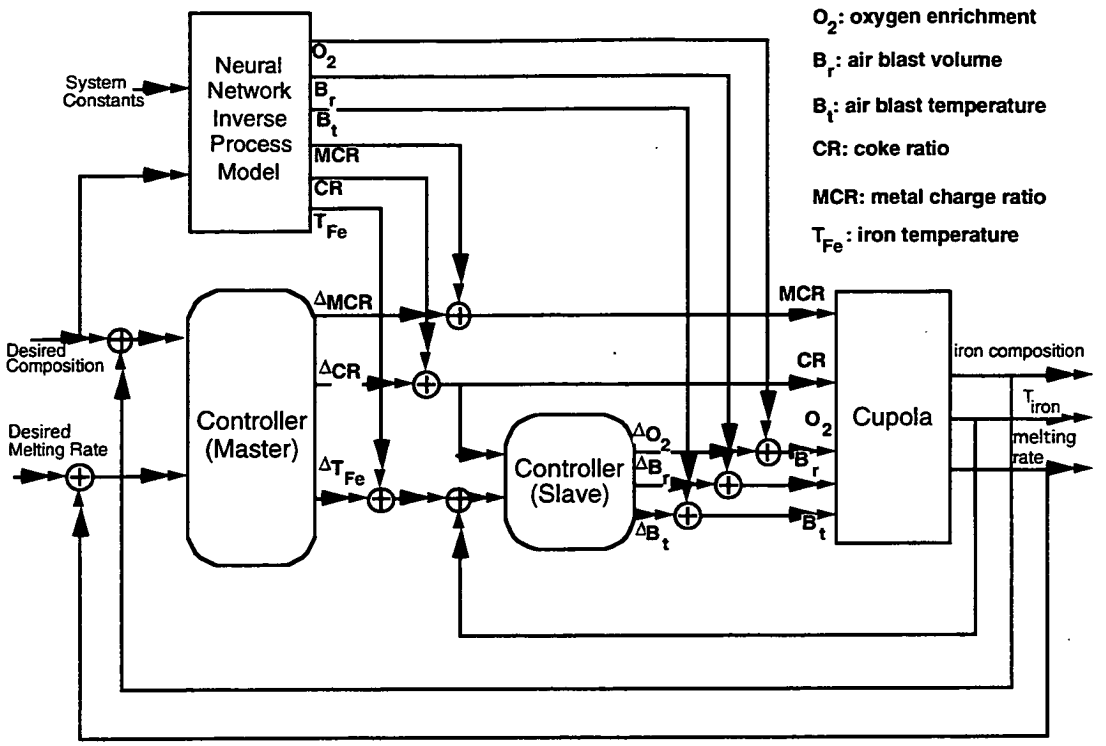


Figure 2. Control schematic for cupola.

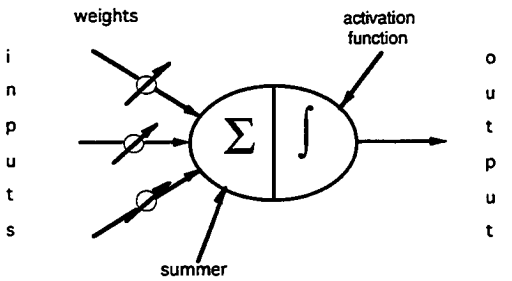


Figure 3. Non linear neuron.

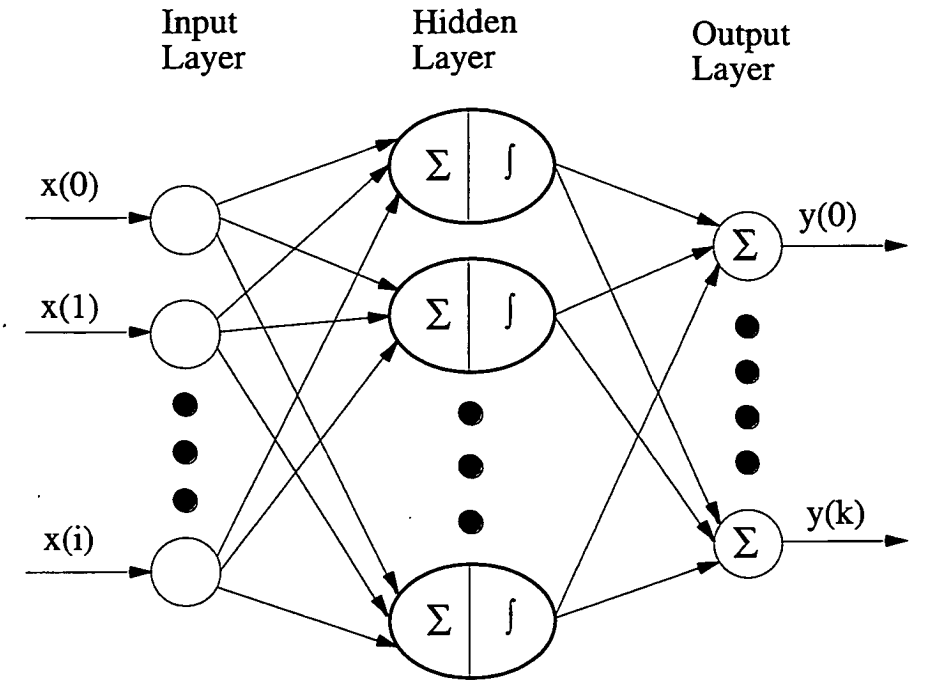


Figure 4. Artificial Neural Network

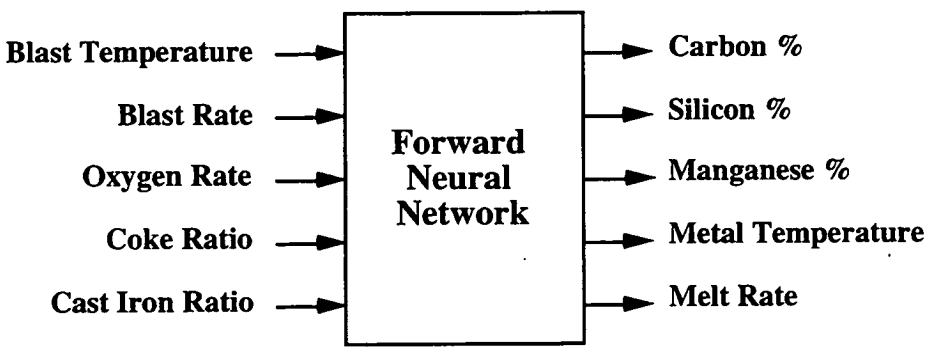


Figure 5. Cupola Neural Network model inputs and outputs in the forward mode.

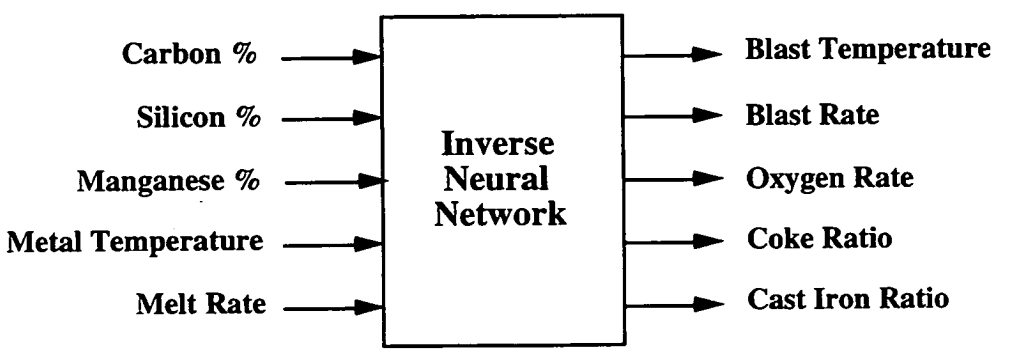


Figure 6. Cupola Neural Network model input and outputs in the reverse mode.

Biotechnology

De Novo Design of Human Immunodeficiency Virus Protease-Cleavage Complexes

David Noever* and Subbiah Baskaran*+

*Biophysics Branch, ES-76, National Aeronautics and Space Administration, George C. Marshall Space Flight Center, Huntsville, AL 35812 USA

+Institut fuer Molekulare Biotechnologie, e.V., Beutenbergerstr. 11, DO-7745, Jena, Germany

Among the 3 retroviral enzymes encoded by HIV (Fig. 1), currently the protease (HIV-PR) possesses the only known high resolution 3-dimensional structure[1-2]. Because hydrolysis of HIV-PR renders the virus non-infectious[3], a search for cleavage complexes carrying this cleavage potential has received attention. Only a fraction of the known sequences now circulate in the open literature, although unofficial counts suggest that many as 160 candidate HIV-PR cleavage complexes[4] have been found. Novel amino acid sequences which can hydrolyze HIV-PR are thus required to locate and classify the physico-chemical properties of each cleavage site and its accompanying cleavage residue pattern. In the search for HIV-PR cleavage complexes, various strategies have been employed using X-ray crystallography (tertiary structure), molecular modelling (graphics overlay and energy minimization) and chemical (Alladin) database searches for non-peptide (small molecule) cleavage complexes. Among these techniques, de novo design has remained relatively unexplored.

Recent researchers[5-6] have proposed a classification scheme for predicting HIV-PR cleavage potential for a given octapeptide. Their predictive success seems robust, since they correctly classify 80 to 92% of the known cleavage complexes and non-cleavage complexes. No

previous method, however, has proven itself capable of generating de novo sequences. The present approach likewise shows greater than 92% accuracy for predicting cleavage but also can generate and select best (and worst) candidates among the 26 billion ($[20]^8$) possible octapeptide combinations which can potentially inhibit HIV-PR.

To span this vast search space of available sequences, the method employs an artificial intelligence method (Fig. 1) called the genetic algorithm[7-8]. First residue frequencies for each site get extracted from the available library of known HIV-PR cleavage complexes, then these frequencies get compared to similar occurrence frequencies for an algorithmically generated test sequence. To construct the predictive procedure for octapeptide generation and evaluation, a given 8-letter sequence (e.g. BYYYCYYY) was compared to the known cleavable peptide databases. Each amino acid sequence can be taken to represent a 160-dimensional vector (e.g., v(x)=(00000010000000000000 01000000000000000000.....00000000000000000010) or

$$v(x)=\sum_{i=1}^{20}\sum_{i=1}^{8} v_j^i(x)$$

where each twenty-letter sequence has only one non-zero element which signifies the presence of that amino acid at that cleavage site point (i.e. C=01000000000000000000, which is the second amino acid of the twenty and found in the second octapeptide site of the above example vector). The comparative distance of this generated vector can be evaluated accroding to standard vector geometry.

Non-cleaving (negative) octapeptide (37) were selected from the lysozyme sequence, none of which (even in denatured form) can cleave HIV-PR. An additional random library of amino acid frequencies was enumerated to establish the information content of the known sequences. Since the random vector signficantly fails to predict cleavable peptides (1 generated sequence with cleavage potential out of its 30 best octapeptides), even after many trials of random search, then the vital core information is concluded to reside within the known

library vector itself and not within the algorithmic procedure of random amino acid substitutions.

After refining these sequences over many generations, the best 5 rank-ordered de novo cleavage complexes are shown in Fig. 2 for HIV-1 and HIV-2. Thirty-seven of the published forty or so experimental sequence can be generated de novo in this way (92.5%). More specific cleavage complexes which cleave HIV-1 protease, but not HIV-2 protease (and vice versa), can equally follow from maximizing the sequences' proximity to the HIV-1-PR cleavage database while minimizing the proximity to the HIV-2-PR cleavage database. For comparison to these de novo designs, the best existing cleavage and its evaluated cleavage potential is indicated. For HIV-1 and HIV-2 (both specific and non-specific) protease cleavage complexes, the de novo designs score an improved cleavage potential compared to the best experimentally known sequences.

A few generic models for cleavage complexes of HIV-PR have been considered from experiments, molecular modelling and X-ray crystallography (Tables 1 and II). Among the models[9-12] put forward to explain site requirements for inhibition, hydrophobicity and β-structure serve as primary qualifiers[5]. Previous work[9] has noted a homology among retroviral proteases which carry the motif of small and hydrophobic residues residing next to a proline 1-cleavage site. The diversity of HIV-PR cleavage complexes, however, supports a multipoint attachment mechanism, wherein residue substitution far (up to 4 sites) from the scissile bond can prevent hydrolysis[11]. The multiplicity of these results support many weak bonds, rather than a single strongly determining attachment site as the HIV-PR pattern. Analysis of 27 physico-chemical properties for the de novo substitution pattern reveals significant polarity and small bulkiness (Fig. 3) among successful site substitutions[8]. The present method should apply equally to a variety of related biochemically significant enzyme-

viral interactions[13] (e.g. Rift Valley Fever virus, simian immunodeficiency virus, etc).

Most promising among the de novo findings is the ability to generate new rules or find rare exceptions which may violate previously proposed site-specific rules. For example, some forbidden substitutions have been proposed, but the limited size of the library database has restricted authors' ability to generalize confidently. At the 1 site for instance, isoleucine (I) and valine (V) have previously been thought to negate cleavability because of its bulkiness and side chain branching. None of the library octapeptides have (I) residues at site 1 and among synthesized sequences, the cleavability of otherwise identical sequences was reversed upon (I) substitution. Kotler[14] have seen earlier that (I) substitution in site 1 of a peptide substrate blocked cleavage by avian myelblastosis virus PR and that this peptide also served as an inhibitor. The de novo search however was able to find an HIV-2 protease cleavage (GQNIAEEF, score=2.98) which not only places I at position near the cleavable bond, but also ranked among the 30 all-time best cleavage complexes. This contrasts with the general de novo finding that a remarkable 83% of the worst HIV-2 protease cleavage complexes contain significant β-branching either as V (80%) or I (3%). Indeed the de novo identification of a promising octapeptide with β-branching at the 1-site is a rare find.

As an additional example, it has been emphatically stated that otherwise abundant lysine residues are highly unlikely to appear close to the cleavage bond (anywhere between 2-2') Tomaselli[15] reported never seeing a lysine within 4 residues of the bond hydrolyzed. However, from the de novo designs, the sequence (AEVFFETK, score=2.87) appears on the list of top 30 HIV-1 cleavage complexes. Although the lysine appears on the 4-site, a strong rule negating cleavage in the presence of lysine seems improbable. To support this de novo finding, a 3-site lysine has been observed in one viral polyprotein cleavage site[16]. In the nucleocapsid protein[17], both a 2 and 2' lysine have been inferred. As conducted successfully by the genetic algorithm, this type of search for previously forbidden substitution motifs

corresponds to the classic problem of finding of 'a needle in a haystack'.

In conclusion, the results show: 1) the method can generate de novo amino acid sequences with optimal predicted functionality (cleavage quality); 2) the vast search space can be intelligently partitioned for identifying optima; 3) amino acid sequences can be rank-ordered according to their comparative quality between no desired functionality (high distance metric) to ideal functionality (low distance metric). In principle, the octapeptides provide a simple probe for PR activity (generation of processed viral proteins) and permit rapid manipulation of the viral (*pol*) gene sequence. These findings underscore the challenge of successful de novo design: how to enumerate and test all possible residue substitutions. For HIV-PR the puzzling multiplicity of substitution patterns demands a systematic way of testing cleavage architecture. By way of analogy, comparison between previously tried ($\sim 10^2$) and possible ($\sim 10^{10}$) HIV-PR cleavage complexes is as the size of a human palm compared to the size of the Earth. The effort aims towards de novo design of cleavage complexes which may serve as natural substrates for binding studies, as well as may offer therapeutic value in AIDS treatment and detection.

1. Wlodawer, A., Miller, M, Jaskolski, M, Sathyanarayan, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J, Kent, S.B.H. (1989) *Science*, **245**, 616-621.

2. Navia, M.A., Fitzgerald, P.M.D., McKeever,B.M., Leu, C.T., Heimbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L., and Springer, J.P. (1989) *Nature*, **337**,617-620

3. Kohl, N.E., Emini, E.A., Schleif, W.A., Davis, L.J., Heimbach, J.C., Dixon, R.A.F., Scolnick, E.M., Sigal, I.S. (1988) *Proc. Natl Acad. Sci.*, **85**, 4686-4690.

4. Sathyanarayana, B.K. and Wlodawer, A. (1993)*Curr. Sci.* , **65**, 835-847.

5. Poorman, R.A, Tomasselli, A.G., Heinrikson, R.L., Kezdy, F.J. (1991) *J. Biol. Chem.*, **266** , 14554-14561.

6. Chou, J.J. (1993) *Biopolymers*, **33**, 1405-1414.

7. Holland, J.H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, MI: University of Michigan Press; Forrester, S. (1993) Genetic Algorithms: Principles of Natural Selection Applied to Computation, *Science*,**261**, 872-878.

8. Noever, D.A. and Baskaran, S. unpublished

9. Pearl, L.H. and Taylor, W.R. (1987)*Nature*, **329**, 351-353.

10. Pearl, L.H. and Taylor, W.R. (1989) *Nature*, **329**, 482.

11. Griffiths, J.T., Phylip,L.H., Konvalinka, J., Strop, P.,Gustchina, A., Wlodawer, A., Davenport, R.J., Briggs, R., Dunn, B.M., Kay, J. (1992) *Biochem.*, **31**, 5193-5200.

12. Pettit, S.C., Simsic, J., Loeb, D.D., Everitt, L., Hutchinson, C.A., and Swanstrom, R. (1991) *J. Biol. Chem.* , **266**, 14539-14547.

13. Hellen, C.U.T., Karusslich, H.G., and Wimmer, E., (1987) *Biochem*,**28**, 9881-9889.

14. Kotler, M. Katz, R. A., Danho, W.., Leis, J., and Skulka, A.M., (1988) *Proc. Natl. Acad. Sci. USA*, **85**: 4185-4189.

15. Tommasselli, A.G., Hui, J.O., Adams,L., Lowery, D., Greenberg, B., Yem, A., Deibel,M.R., Zurcher, Neely, H., Heinrickson, R.L. (1991) *J. Biol Chem.*, **266**, 14548-14553.

16. Henderson, L.E., Benveniste, R.E., Sowder, R.C., Copeland, T.D., Schultz, A.M. and Oroszlan, S. (1988)*J. Virol*, **62**, 2587-2595.

17. Roberts, M.M. and Oroszlan, S. (1989) *Biochem. Biophys. Res. Commun.*, **160**, 486-494.

Figure captions

Fig. 1. Schematic of de novo procedure of generating and evaluating cleavage complexes. Inset right shows the lifecycle of HIV from viral insertion to mature budding. Three enzymes are shown to catalyze maturation and infection. For the retrovirus, RT is required to connect viral RNA to the healthy cells' DNA. IN duplicates this new viral DNA. PR is required for the budding of new viral particles with a condensed core of infectious agent (after Sathanarayan and Wlodawer, 1993). Shown upper right is the gag-pol elements of the viral gene which carry the enzyme's genetic information.

Fig. 2. Best HIV- protease cleavage complexes as found by the de novo method and comparative scores for the best existing cleavage complexes. Results show the best 5 cleavage complexes found for HIV-1, HIV-2, specific HIV-1, and specific HIV-2. Cleavage potential is the sum of three distance metrics (correlation angle, Euclidean and Minkowski distance) from the reference vector of known HIV-PR cleavage complexes. The library or reference norm vector was constructed similarly as a 160-dimensional vector, which represents the relative frequency (or likelihood) for a given amino acid (1-20) to appear in a octapeptide cleavage site (1-8). To underscore the cleavage property, the database frequencies were subtracted from the natural abundance as:

$$v_j^i(k)=V_j^i(k) - \eta^i$$

where $v_j^i(k)$ is the reference component of the cleavage vector, $V_j^i(k)$ is the training set value, and η^i is the natural abundance identified over the 20 amino acids.

Three measures were included to evaluate the distance between the octapeptide sequence (example) and the database frequencies for cleavable HIV-PR:

1) the correlation angle between test sequence and library sequence;

$$\vartheta(k)= \arccos \left\{ \frac{\sum_{i=1}^{160} v_i(x)\, v_i(k)}{\left[\left(\sum_{i=1}^{160} v_i^2(x)\right)\left(\sum_{i=1}^{160} v_i^2(k)\right)\right]^{\frac{1}{2}}} \right\} \qquad (1)$$

2) Minkowski point to point distance (polynomial distance sum);

$$d^M(k)=\left[\sum_{i=1}^{160} \mathrm{abs}(v_i(x)-v_i(k))^p\right]^{\frac{1}{p}} \quad p=1 \qquad (2)$$

3) Euclidean point to point distance (squared distance sum)

$$d^E(k)=\left[\sum_{i=1}^{160} (v_i(x)-v_i(k))^2\right]^{\frac{1}{2}} \qquad (3)$$

A composite measure was constructed from the weighted sum of the three measures with an assignment of one as normalized for each distance metric (normalization condition).

Fig. 3. Physico-chemical properties of the 30 best HIV-1 protease cleavage complexes on a site-specific basis. The values are fractional deviation from a normalized average for all 20 amino acid residues (0=average value); 1=100% deviation above the average; -1=100% deviation below the average). Shown inset are the standard deviation of the site-specific property. High standard deviations indicate a large variability and much tolerance for different substitution residues. The two properties shown relate the polarity and bulkiness of each site in the 30 best de novo HIV-1-protease cleavage complexes.

Table Captions

Table I. Summary of good single site substitution patterns for HIV-1-PR as found in various proposed protease models.

Table II. Summary of poor single site substitution patterns for HIV-1-PR as found in various proposed protease models.

THE ROLE OF IMMUNOGLOBULIN BINDING FACTORS IN IMMUNOTHERAPY

F.M. Cowan
President, Uppsala Inc.
Colora, MD 21917

ABSTRACT

Receptor-ligand reactions are vital to many biological systems and can be simply modeled as "lock and key" with the receptor as the lock and the ligand as the key. Unlocking the gates of Fc receptor (FcR) dependent immunity by the Fc region of antibody is a critical property of immune homeostasis. The use of FcRs or FcR-like microbial immunoglobulin binding factors (IBF) as drugs to modify reactions between FcR and Fc antibody may impact the abnormal FcR dependent immunity that is often observed in autoimmunity, cancer and AIDS.

INTRODUCTION

Any agent that can elicit an immune response is called an antigen. Antigens include infectious agents such as bacteria and viruses, drugs, toxins, foreign tissue transplantation antigens, self-tissue autoimmune antigens, cancer cells and even the new configurations called idiotypes that antibody forms in response to an antigen. The primary functions of the immune response toward antigens are recognition, specific response, clearance and memory. Exposure of a human or an animal to a given antigen activates the immune system to develop circulating antibody, or humoral immunity. Cellular immunity can be developed by formation of sensitized cells that bind to Fc antibody or display receptors against antigen. Such humoral and cellular immunity toward antigens are the pillars of the immune response. Both positive and negative regulation of the immune response involves humoral ligand keys, such as antigen, antibody or cytokines, unlocking their respective receptor locks on the membranes of immune cells. This allows the selective opening or closing of the gates that regulate access to the pathways that lead to immunity to antigen. The receptor locks may also be secreted by cells and these humoral receptors act to further control ligand binding to cellular membrane receptors.

The cellular structures that bind antibody Fc constant regions were designated FcRs by Paraskevas and coworkers in 1972. Like the T and B cell antigen receptors FcRs are members of the immunoblobuiln super family. FcRs reside in the plasma membranes of all types of immune-system cells and are secreted to circulate as humoral IBFs. Different FcR can bind specifically to distinct antibody classes such as immunoglobulin (Ig), IgA, IgG, and IgE as well as to subclasses, for instance, IgG1 and IgG2. FcRs may bind to a single antibody molecule as does the IgG FcRI designated cluster of differentiation 64 (CD64) or be selective for immune complexes, as is IgG FcRII (CD32) or FcRIII (CD16). This group of molecules, collectively referred to as FcRs, forms an important junction between antigen recognition and immune response to antigen, that prompts both effector functions, such as antibody-dependent cellular cytotoxicity (ADCC), phagocytosis and complement activation, and orchestration of immune regulatory signals for T-helper and T-suppressor lymphocytes.

Antibodies are Y-shaped molecules constructed with two identical Fab-fragment "arms" that contain the variable regions that bind antigen, and an Fc constant region "stem" that as either a single antibody molecules or an antigen-antibody immune complex is the ligand for FcR (Figure 1). Multiple antibodies binding to antigen determinants form immune complexes, which can aggregate FcRs and activate FcR-mediated immunity. Cellular FcR-mediated immune response is contingent on the relative concentrations of antibody and antigen-antibody immune

complex and the degree of attachment of these ligands to available cellular FcR. Each antigen is bound to the variable sites on the Fab arms of an antibody molecule specific for that antigen, whereas the antibody Fc constant region engages FcR and activates FcR-mediated immunity independent of the antigen specificity (Figure 2). Thus, the antigen non-specific activation of FcR immunity by a single receptor system influences immunity to the infinity of antigens recognized by the Fab portion of antibody. Systemic FcR/IBF immunotherapy modifies the ratio of FcR locks to ligand keys by administering humoral IBF to alter the opening or closing of the gates of the immune system which provoke abnormal immunity to antigen.

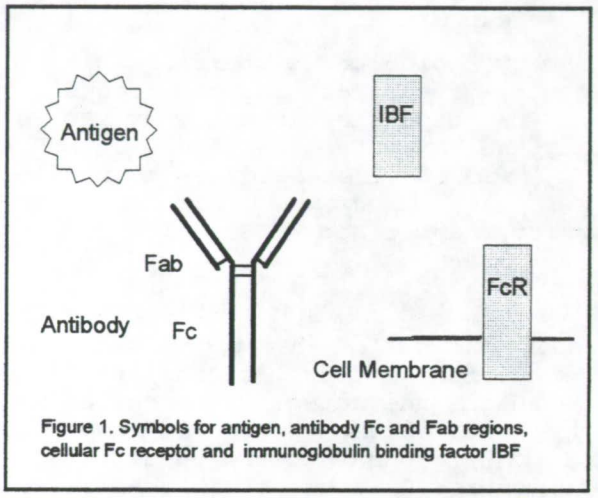


Figure 1. Symbols for antigen, antibody Fc and Fab regions, cellular Fc receptor and immunoglobulin binding factor IBF

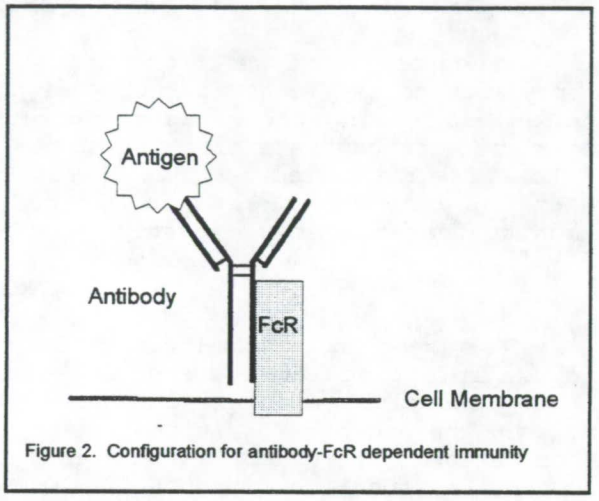


Figure 2. Configuration for antibody-FcR dependent immunity

The physiological immune responses are designed to eliminate antigens, and the formation of antigen-antibody immune complexes is crucial for the clearance of antigens and other immune reactions. However, immune complexes circulating in biological fluids may become pathogenic and cause inappropriate activation or inactivation of humoral or cellular immunity. Antibodies, antigen-antibody complexes and abnormal FcR immunity have long been known to influence the immune system and have been implicated in the immunopathology of infectious agents, immunodeficiency diseases such as AIDS, autoimmune disorders and cancer. Sinclair has reviewed the biological consequences for inflammatory and immune regulatory functions of immune complexes bound to FcRs. B cells are responsible for antibody production and are activated by the cross-linking of their membrane associated antibody, which comprises the B cell antigen receptor (Figure 3). Sinclair & Panoskaltsis have integrated elements of Bretscher and Cohn's associative recognition theory for antigenic signals and of Jerne's hypothesis for immunoregulation by idiotype anti-idiotype networks, and further postulated that FcRs are also important immunoregulatory molecules for production of antibody. In this model, B cell activation is inhibited when either anti-idiotypic or antigen specific humoral antibody binds the activated B cell receptor complex, allowing the unfettered Fc region of the humoral antibody to further engage an FcR on the B cell membrane. This cross-linking of the B cell antigen receptor and an FcR dampens antibody production acting as a "dimmer switch" for humoral immunity (Figure 4). Fridman and coworkers (1986) have expanded the theme of Sinclair's work, hypothesizing that, in addition to Jerne's idiotypic network, an isotypic network of antibody-FcR lock and key reactions contributes to immune regulation. Jerne's idiotypic network for regulation of antigen-specific immunity predicts that the idiotype configurations formed to bind antigen, located at the Fab variable regions of antibodies, can themselves be antigenic causing the production of antibody against these sites, which are called anti-idiotypic antibody. Signals generated by the interactions of idiotypes and anti-idiotypic antibody then control immunity. According to Jerne (1984) this idiotypic network, "reflects first ourselves, then produces a

reflection of this reflection, and ... subsequently . . . reflects the outside world: a hall of mirrors". "Mirrors" of the idiotypic network arrayed in concert with an FcR-isotypic network could react to abnormalities in FcR-dependent immunity by eliciting immunological disorder. Furthermore, therapies designed to correct abnormalities in the antibody-FcR isotypic network might realign a disarrayed idiotypic network and restore order to the immune system.

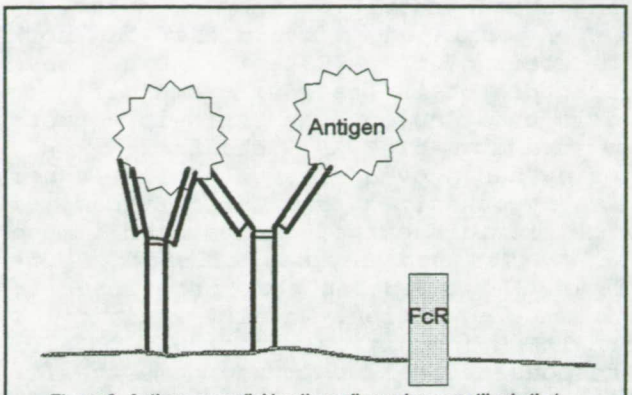


Figure 3. Antigen cross-linking the cell membrane antibody that comprises the B-cell antigen receptor activates antibody synthesis

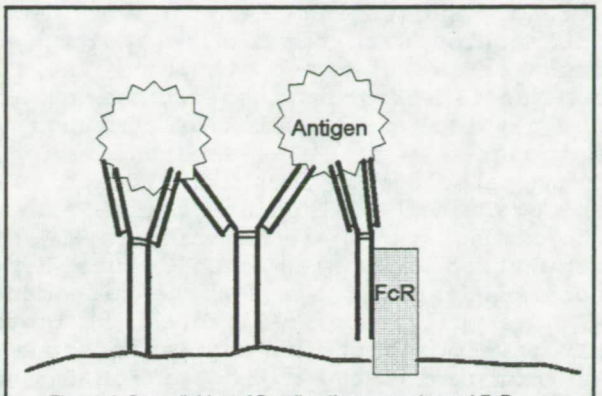


Figure 4. Cross-linking of B-cell antigen receptor and FcR by immune complex terminates signal for antibody production

FcRs and related immune receptors for cytokines have achieved prominence in immunopharmacology. Fernandez-Botran (review 1993) predicts that soluble receptors moderate the interaction of ligands with many receptor systems, and that humoral cytokine receptors may have implications for immunotherapy. Fridman proposes that secreted FcR humoral IBF may act as cytokine-like molecules effecting FcR-dependent immunity and influencing the immunopathology of cancer and infectious disease, and he concludes that IBF should be studied as possible immunotherapeutic agents. Microbes sometimes offset host immune defenses by producing proteins that imitate mammalian receptors and undermine ligand-receptor reactions; such microbial immunomodulators have also been advanced as plausible immunotherapeutic drugs. For example, microbial IBFs are produced both by viruses, such as retroviruses and herpes viruses, and by bacteria, such as Staphylococcus aureus. Microbial IBFs are known to alter FcR-dependent immunity, and the bacterial IBF Staphylococcus aureus protein A (SPA) has been used for over a decade for immunotherapy. SPA has five FcR-like regions; however, this IBF valence of five can be reduced or increased, e.g. one to ten Fc bindng sites. In addition to concentrations, the valences, or number, of Fc antibody molecules in an immune complex or the number of FcR-like sites in an IBF, and their relative configurations and binding affinities toward each other can influence FcR immunity and IBF immunotherapy.

A patent was filed in 1979 and issued to Cowan in 1993 for the use of IBFs such as leukocyte FcRs and SPA as a biological modifiers for immunotherapy of immune complex diseases and associated abnormal immunity to antigen. The use of SPA as an immune receptor to modify immunity to antigen was, to the best of the inventor's knowledge, the first systemic use of a receptor as an immunotherapeutic drug. In preliminary veterinary and human clinical studies by numerous investigators, SPA bacterial IBF immunotherapy has demonstrated efficacy in neoplastic, autoimmune and retroviral diseases, including human mammary adenocarcinoma, AIDS-associated Kaposi's sarcoma and immune thrombocytopenic purpura. SPA is also the only drug reported to effect complete remissions of the feline leukemia virus (FeLV) that causes retroviral disease in cat similar to AIDS.

AIDS AND IBF

The human immunodeficiency virus (HIV) is the causative agent of AIDS, and HIV envelope proteins gp120 and gp41 and SPA all have IBF activity for Fc IgG. Viral IBF may contribute to the pathogenesis of HIV by "jamming" the network of FcR signals that control FcR-dependent immunity. Conversely, the bacterial IBF SPA has anti-retroviral activity possibly involving antagonism of viral IBF, strongly suggesting that IBF may, depending on the nature of the molecules and interaction with the immune system, act as a double-edged sword that can both cause or avert immunopathology. The first decade of the AIDS epidemic has seen substantial progress in deciphering the structure of the retrovirus in defining the clinical progression of symptoms, and in treating opportunistic infections associated with HIV. Nevertheless, determination of the mechanisms of HIV immunopathology to foster an effective therapy has proven elusive. Zolla-Pazner and Sidhu and Edelman and Zolla-Pazner have hypothesized that the intricacy of the immune system are mirrored in HIV disease, reminiscent of exaggerated immune regulation provoking both immune hyperactivation and immune deficiency. One possible explanation for the pathogenesis of HIV infection that progresses to AIDS is that retroviral protein(s) impact immune regulation. The IBF activity of HIV envelope proteins may cause the abnormal FcR immunity associated with AIDS and compromise the myriad of FcR immune responses vital to immune homeostasis. The limited amount of viral genetic material compels economical coding of viral protein function. The "counterfeiting" of a major immune receptor, such as FcR, by retroviral genes is almost certainly not without function in HIV infection. Furthermore, forgery of immune molecules by HIV may not be restricted to IBF. Matthews and coworkers (1994) have observed similar structures for the HIV P17 matrix protein and the cytokine interferon. IBF is associated with both activation and inhibition of immunity. Helper T cells produce cytokines that can avert suppression of immunity by antibody-FcR, and in the absence of functional helper T cells such antibody-FcR suppression becomes dominant. The production of viral IBF from HIV or other viruses such as cytomegalovirus in a helper T cell compromised host could contribute to immunodeficiency. Cytomegalovirus IBF can inhibit FcR immunity, and HIV GP120 and gp41 IBF blockade of B cell FcR immune regulatory signals is proposed to contribute to B cell polycolonal activation resulting in autoimmune destruction of T cells.

Subversion of FcR immunity by viral IBFs may also influence infection of FcR-positive cells by HIV, cytokine systems and T cell function. The major route of HIV infection is the CD4 receptor on T-helper cells and CD4 has IBF activity for antibody Fab constant regions. In addition to CD4, immune complexes of HIV and anti-HIV antibodies can infect antigen presenting cells (APC), such as macrophages and dendritic cells, by their FcRs. Circulating immune complexes are often elevated in HIV-infected individuals, and immune complexes isolated from the HIV-positive sera infect cells in culture. HIV-IBF envelope proteins might also occupy the Fc antibodies of immune complexes that contain antigens other than HIV. Such immune complex-HIV aggregates when processed by APCs could infect T cell clones responding to antigen. Activation of T-helper cells by APCs requires both the T-cell antigen receptor complex and CD4 (Figure 5). Programmed cell death, or apoptosis, provides normal physiological functions such as clonal deletion, but is also postulated as a pathological mechanism of T cell elimination in AIDS. Activation of T-helper cells in conjunction with a second signal generated by anti-CD4 antibody aggregated by APC FcRs that cross-link CD4 stimulates apoptosis. In HIV disease CD4 might also be cross-linked by HIV gp120-antibody to HIV gp120 immune complex bound to CD4 and aggregated by either cellular FcRs on APC or the IBF envelope proteins on the HIV, precipitate apoptosis and clonal deletion of activated T lymphocytes (Figure 6). Certain cytokines such as interleukin 2 can prevent T cell apoptosis and restore T cell activation. The cytokine interferon can increase FcRs and enhance FcR-dependent immunity, and Fc antibody-FcR ligand-receptor interaction promotes expression of the genes coding for the cytokines interferon and tumor necrosis factor. These observations are compatible with an FcR-cytokine immune regulatory loop controlling FcR-mediated immunity and cytokine production. Disruption of such a

FcR-cytokine immune regulatory loop by viral IBF could precipitate both abnormal FcR-mediated and cytokine-dependent immunity. The CD3 is a component of the T-helper cell antigen receptor, and crosslinking CD3 by anti-CD3 antibodies aggregated by FcRs on APCs causes polyclonal activation of T-helper cells, similar to the antigen specific activation of T cell clones by APC. Anti-CD3 activation of T cells is inhibited when anti-gp120 antibody from AIDS patients binds to HIV gp120 attached to the CD4 receptor, and similar paralysis of T cell function may contribute to the immunopathology of AIDS. The T cell antigen receptor complex shares both structural similarities and common biochemical signaling pathways with FcRs, and anti-CD3 activation of T cells in the presence of APCs is FcR dependent. Unlike other cells with FcRs, T cells express their FcR only during activation by antigen or anti-CD3 antibodies. The selective expression of FcRs on activated T cells underscores the danger of FcR-dependent HIV infection and apoptosis, and provides the opportunity for antibody Fc-FcR lock and key signals to influence activated T cells. T cell FcR signals may emulate B cell immunoregulation by IgFc-FcR signals and afford a common pathway for coordination of humoral and cellular immunity.

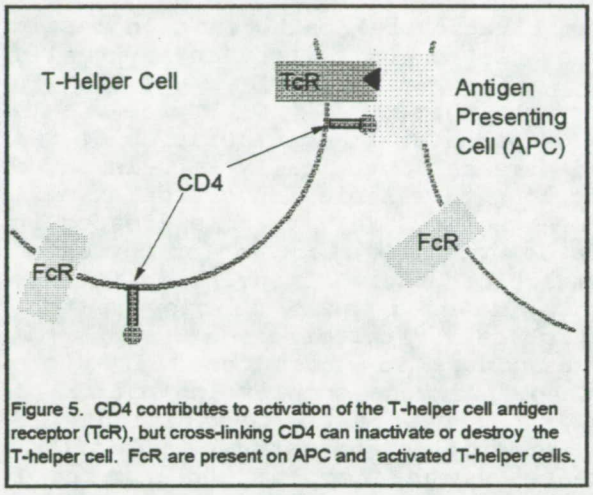


Figure 5. CD4 contributes to activation of the T-helper cell antigen receptor (TcR), but cross-linking CD4 can inactivate or destroy the T-helper cell. FcR are present on APC and activated T-helper cells.

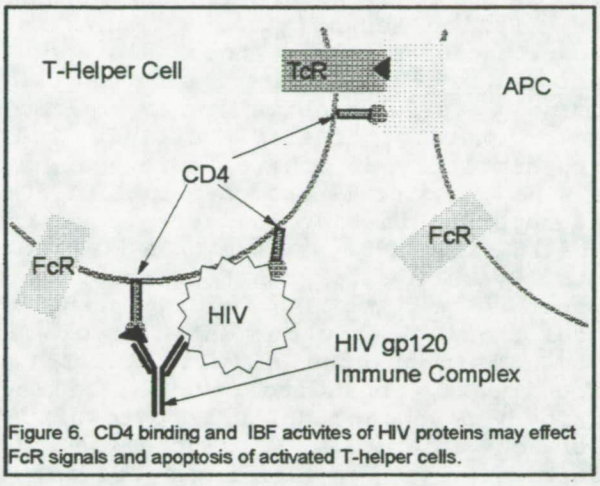


Figure 6. CD4 binding and IBF activites of HIV proteins may effect FcR signals and apoptosis of activated T-helper cells.

HIV antigenemia, decline of antibody to HIV proteins, reduced number of T-helper cells, and loss of circulating IBF produced by immune cells are correlates and prognostic markers for the progression of AIDS. The elevation of humoral HIV-IBF proteins and the absence of the naturally occurring IBF of the immune system may contribute to the defective FcR effector functions associated with morbidity in AIDS patients. These FcR effector defects also imply a broader disorder encompassing FcR-dependent immune regulation. In the early stages of infection HIV-IBF may cause aberrant FcR immunity with a predominance toward autoimmune eradication of T cells. The subsequent erosion of T cell function renders antibody-FcR immune suppression dominant, reinforcing the immune anergy of the terminal phase of AIDS. Paradoxically, HIV may generate both abnormal FcR-dependent autoimmunity and immune deficiency that sequentially and cumulatively contribute to the immunopathology of AIDS. SPA and HIV gp120 and gp41 viral envelope proteins have IBF for human IgG Fc; and both SPA and HIV gp120 also bind IgM Fab antibody V_H3 gene products. The common IBF properties of SPA and HIV pg120 are also reflected in their biological activates, and both molecules can inhibit IgG FcR signals, activate the complement system and cause polyclonal activation of B cell expressing V_H3. Immune molecules that remain elevated in healthy HIV patients and are lost with the advance of the disease have been considered as a possible source of therapeutic drugs. The stage-related decrease of patients' IBF levels observed in AIDS qualifies IBF as a possible therapeutic. The anti-retroviral activity of SPA and the Fc and Fab IBF activity of SPA and HIV envelope proteins may be merely coincidental; however, it is tempting to speculate that the IBF activity of SPA might replace deficient natural IBF and thus antagonize the immunopathogenic action of viral IBF protein(s).

STAPHYLOCOCCUS AUREUS PROTEIN A BACTERIAL IBF IMMUNOTHERAPY

Veterinary and human clinical data indicate that SPA bacterial IBF immunotherapy is effective in neoplastic, autoimmune and retroviral diseases. Unlike the systemic use of SPA, extracorporeal SPA therapy involves removing blood from a patient, separating the cells from the sera, passing the sera over a SPA column as an immunoadsorbent and returning the treated sera to the subject. In patients who had failed conventional therapies, treatment with protein A affinity columns resulted in a 36% response rate for mammary adenocarcinoma, a 47% response rate for AIDS-associated Kaposi's sarcoma, and normalized platelet counts in AIDS patients with immune thrombocytopenic purpura. Extracorporeal treatment of feline leukemia and rat or canine mammary adenocarcinoma had antineoplastic activity similar to that observed with systemic SPA immunotherapy. One explanation for the like antineoplastic actions of extracorpereal and systemic SPA therapies is that SPA may leach from extracorpereal columns. This is not unreasonable since tests for SPA leakage from columns are generally done with saline solutions, contrary to the patient's sera, which contains proteases capable of cleaving SPA from the column matrix.

Feline leukemia virus (FeLV) is a major retroviral pathogen of domestic cats. It causes anemia, neoplasia and FeLV-associated mortality that is usually a consequence of immune impairment similar to the immunodeficiency of AIDS in humans. FeLV is an animal retroviral model for the therapy of AIDS. Twice-weekly use of extracorporeal SPA affinity columns or parenteral administration of SPA were similarly effective in FeLV-associated disease. Chronically FeLV-infected cats had an overall response rate of 50%, and long-term remission of the disease was reported in 6 of 18 cases (33%) after SPA therapy. FeLV viremia cleared in 5 of 18 cats (28%), a more than 30-fold increase over that reported for untreated cats. Overall cats with FeLV-related abnormalities other than overt malignancy responded well to SPA immunotherapy, which achieved greater efficacy in this model than ATZ, the current drug of choice for retroviral disease.

SPA modulates a variety of immune responses in "test tube" in vitro experiments, including reversal of FcR dependent immune complex inhibition of antibody-dependent cellular cytotoxicity (ADCC) and anti-IgM termination of B cell activation. SPA also increases cytokine production and natural killer (NK) activity. In vivo SPA immunotherapy decreases immune complex and number of activated T-suppressor cells; increases cytokine production, NK cell activity, anti-idiotypic antibody, antibodies against viral and tumor antigens, and absolute numbers of activated T-helper cells and B cells; and suppresses tumor growth. The efficacy of SPA for cancer and retroviral disease may encompass many of these immunomodulating activities, since even a single primary site of action could quickly cascade to encompass a myriad of FcR dependent effector and regulatory processes. SPA IBF activity is responsible for the antineoplastic activity of SPA preparations against mouse fibrosarcoma, and the efficacy of extracorporeal SPA therapy for human breast adenocarcinoma is lost after chemical treatment of columns eliminates SPA IBF activity.

These results strongly suggest that SPA IBF activity is required for antineoplastic activity; however, synergy between SPA and other immune factors such as cytokines may also contribute to efficacy. Singh et al. have demonstrated that SPA and interleukin 2 synergistically amplify lymphokine activated killer (LAK) activity and suggest that SPA might augment LAK immunotherapy in cancer patients. Kasparian et al. have shown that incubation of SPA with the defective NK cells found in HIV-seropositive subjects produces a "remarkable enhancement" of NK activity. A dominance toward either humoral or cellular immunity is observed in immune complex mediated FcR-dependent "split tolerance". Such increased humoral immunity to the detriment of cellular immunity may be a factor in AIDS. The reversal of such split tolerance by SPA may contribute to its efficacy in neoplastic and retrovirus diseases. However, the doses of SPA either injected or potentially leached off SPA affinity columns during extracorporeal profusion are inadequate to irreversibly neutralize circulating immune complex. Therefore, SPA IBF sites and Fc antibody in immune complex may by a competitive and reversible interaction alter the locks and keys that govern immunity to

antigen. The fact that the extremely low concentration of humoral IgG IBF (10^{-8}M) in normal human sera and HIV-infected patients could be significantly altered by the doses of SPA used for immunotherapy favors this hypothesis.

The efficacy of systemic SPA immunotherapy might be enhanced and major problems of immunotoxin therapy might also be eliminated by the use of antibody-SPA as an "immunotoxin" (Figure 7). Cytotoxic biological toxins or chemicals can be targeted to cancer cells by monoclonal antibodies specific for tumor antigens. Monoclonals can be conjugated to these toxins by chemical reagents or recombinant DNA techniques and may consist of either intact antibody or antigen binding Fab fragments of the antibody molecule. Their are several major obstacles to the development of immunotoxins: (1) toxins require cellular internalization to kill the cancer cell; (2) obtaining human monoclonal antibodies; (3) intact mouse monoclonals can act as antigens in humans, simulating anti-mouse antibody immune reactions; and (4) mouse antibody poorly activates human Fc receptor immunity. Like standard immunotoxins, the SPA molecule is complexed to a substance such as monoclonal antibody Fab specific for the target molecule or antigen, such as tumor antigen or microbial antigen on a virus or antibiotic resistant bacteria. The SPA of the immunotoxin binds the Fc region and aggregates the patient's circulating antibody. The monoclonal antibody binds to the target antigen, locating the host antibody aggregate that activates Fc receptor mediated immunity, in close proximity to the antigen (Figure 8). It is a critical difference that, unlike conventional immunotoxin therapy, SPA immunotoxin is not directly cytotoxic but rather activates the subject's Fc receptor dependent immunity to eradicate the antigen. Major obstacles to the efficacy of immunotoxin therapy are overcome by the emulation of natural immunity by SPA immunotoxin. For example, antibody-SPA acts extracellularly at the cell membrane and does not require internalization. SPA activates Fc receptor immunity by binding the Fc constant region of the host antibody. This avoids the reduced efficacy and adverse immune reactions associated with the use of intact antibodies from a different species such as mouse monoclonals, in humans. The use of SPA immunotoxin was described in the patent of Cowan that was filed in 1979 and issued in 1993, and such SPA-antibody Fv hybrid molecules were later made by Ito and Kurosawa in 1993 to increase antibody Fab binding. SPA immunotoxins have the potential of eradicating antigens the immune system has not encountered previously or that have fail to stimulate an adequate immune response.

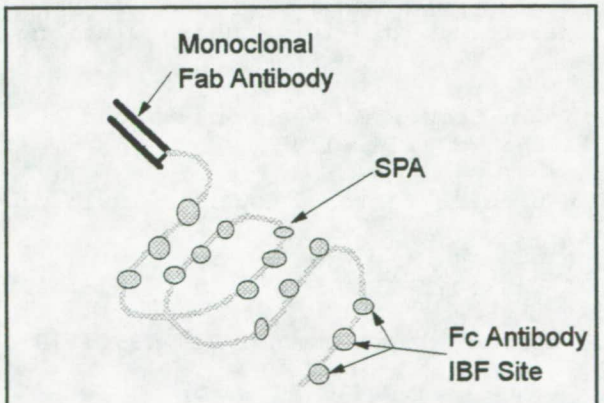


Figure 7. SPA immunotoxin

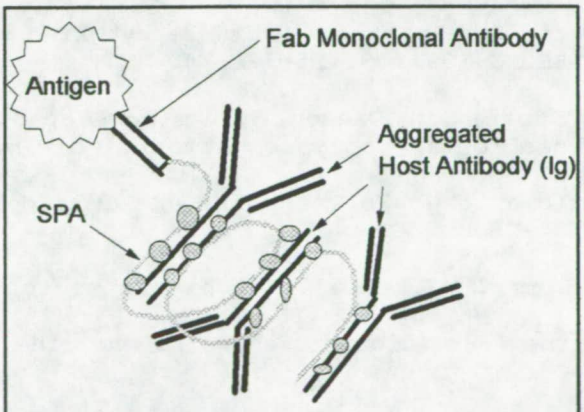


Figure 8. SPA immunotoxin bound to antigen by monoclonal antibody and FcR immunity activated by aggregation of host Ig

The relative poverty of currently available experimental data and the unanswered questions concerning the mechanisms and efficacy of FcR/IBF immunotherapy mandate caution in predicting the possible future of this therapy. Nevertheless, defining the mechanisms by which IBF from mammalian and microbial sources modify FcR immunity may increase our knowledge of the immune response and eventually lead to new immunopharmacologic weapons in the fight against AIDS, cancer, and other disorders with immunopathological components.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. D.D. MacFadden and Mr. Fleet E. Cowan for their help and encouragement during the course of this work. A special acknowledgement is extended to James Ray Cowan and May MacFadden.

REFERENCES

Patent:

Cowan, FM: A method of treating cellular Fc receptor mediated immune disorders. U.S. Patent Number: 5,189,014, Feb. 23, 1993, (original parent application Serial No.99,741 filed December 3, 1979).

Reviews:

AIDS the unanswered questions. Science 1993: 269: 1219-1293.

Boyle, MPD, Reis, KJ. Bacterial Fc Receptors. Bio/Technology 1987:5: 697-703.

Bretscher, P, Cohn, M. A theory of self nonself discrimination. Science 1970: 169: 1042-1049.

Capon, DJ, Ward, RHR. The CD4-gp120 interaction and AIDS pathogenesis. Annu Rev Immunol 1991: 9: 649-78.

Cowan, FM, Klein, DL, Armstrong, GR, Stylos, WA, Pearson, JW. Fc receptor mediated immune regulation and gene expression. Biomedicine 1980: 32: 108-110.

Cowan, FM, Madsen, JM. The role of immunoglobulin binding factors in the pathogensis and therapy of AIDS. Medical Hypothesis (in Press)

Edelman, AS, Zolla-Pazer, S. AIDS: a syndrome of immune dysregulation, dysfunction, and deficiency. FASEB J 1989: 3: 22-30.

Fernandez-Botran, R. Soluble cytokine receptors: their role in immunoregulation. FASEB J 1991: 5: 2567-2574.

Fridman, WH, Daeron, M, Amigorena, S, Rabourdin-Combe, C, Neauport-Sautes, C. Bases for an isotypic network. Mol Immunol 1986: 23: 1141-1148.

Fridman, WH. Fc receptors and immunoglobulin binding factors. FASEB J 1991: 5: 2684-2690.

Fridman, WH, et al. Soluble Fc receptors. J. Leukocyte Bio. 1993: 54: 504-512.

Jerne, NK. Idiotypic networks and other preconceived ideas. Immunol Rev 1984: 79: 5-24.

Marrack, P, Kappler, J. Subversion of the immune system by pathogens. Cell 1994: 76: 323-332.

Ravetch, JV. Fc Receptors: Rubor Redux. Cell 1994: 78: 553-560.

Sinclair, NR StC. Antibody-antigen complexes: Biological Consequences. In Encyclopedia of Human Biology. Vol.1, New York: Academic Press Inc., 1991: 297-305.

Sinclair, NR StC, Panoskaltsis, A. Immunoregulation by Fc signals. Immunol Today 1987: 8: 76-79.

Sandor, M, Lynch, RG. Lymphocyte Fc receptors: the special case of T cells. Immunol. Today 1993: 14: 227-231.

Sandor, M, Lynch, RG. The biology and pathology of Fc receptors. J. Clin. Immunol. 1993: 4: 237-246.

Theofilopoulos, AN, Dixon, FJ. The biology and detection of immune complexes. Adv. Immunol. 1979: 28: 89-220.

Articles:

Berberian, L, Goodglick, L, Kipps, TJ, Braun. Immunoglobulin V_H3 gene products: Natural ligands for HIV gp120. Science 1993: 261: 1588-1591.

Choy, EHS, Adjaye, J, Forrest, L, Kingsley, GH, Panayi, GS. Chimaeric anti-CD4 monoclonal antibogy cross-linked by monocyte Fc IgG receptor mediates apoptosis of human CD4 lymphocytes. Eur J Immunol 1993: 23: 2676-2681.

Cowan, FM, Klein, DL, Armstrong, GR, Pearson, JW, Ablashi, DV, Bensinger, WI. Inhibition of rat adenocarcinoma metastases by Staphylococcus aureus protein A. Pharmacotherapy 1982: 36: 29-31.

Engelman, RW, Good, RA, Day, NK. Clearance of retroviremia and regression of malignancy in cats with leukemia-lymphoma during treatment with Staphylococcal protein A. Cancer Detect Prev 1987: 10: 435-444.

Ito, W, Kurosawa, Y. Development of an artificial antibody system with multiple valency using an Fv fragment fused to a fragment of protein A. J. Biological Chem. 1993: 268,: 29668-20675.

Kasparian, S, Stefanescu, I, Rivard, GE, Menezes, J. Differential Staphylococcal protein A-induced enhancement of natural killer cell activity of lymphocytes from HIV-seropositive individuals. Int J Immunopharmac 1991: 13: 865-873.

Khayat D, et al. Changes of soluble CD16 levels in serum of HIV-infected patients: Correlation with clinical and biologic prognostic factors. J Infect Dis 1990: 161: 430-435.

Matthews, S, et al. Structural similarity between the p17 matrix protein of HIV-1 and interferon. Nature 1994: 370: 666-668.

Messerschmidt, GL, et al. Protein A immunoadsorption in the treatment of malignant disease. J Clin Oncol 1988: 6: 203-212.

Rahimpour, R, Anderson, CC, Sinclair, NR StC. Blockade of immunoregulatory Fc-signalling by HIV peptides: oligopeptides from HIV gp120 and gp41 bind the Fc portion of IgG and increase the in vitro anti-ssDNA response. Clin Exp Immunol 1993: 94: 1-6.

Singh, KP, Shau, H, Gupta, RK, Kopald, K, Ray, PK. Protein A potentiates lymphokine-activated killer cell induction in normal and melanoma patient lymphocytes. Immunopharmacol. Immunotoxicol. 1992: 14: 79-103.

Snyder, HW, Cochran, SK, Balint, JP, et al. Experience with Protein A-Immunoadsorption in treatment-resistent adult immune thrombocytopenic purpura. Blood 1992: 79: 2237-2245.

Susal, C, Kirschfink, M, Kropelin, M, Daniel, V, Complement activation by recombinant HIV-1 glycoprotein gp120. J. Immunol. 1994: 152: 6028-6034.

A MONOCLONAL ANTIBODY TO DISTINGUISH EGGS OF TWO COTTON INSECT PESTS: ITS POTENTIAL AS A TOOL FOR PEST CONTROL AND INSECTICIDE RESISTANCE-MANAGEMENT[1]

Matthew H. Greenstone
U.S. Department of Agriculture - Agricultural Research Service
Biological Control of Insects Research Laboratory
P.O. Box 7629, Columbia, Missouri 65205

ABSTRACT

The cotton bollworm and tobacco budworm cost Southern cotton growers over three hundred million dollars annually. The problem is compounded by the fact that the budworm becomes resistant to pyrethroids, the most widely used insecticides. Although either or both species may infest fields simultaneously, and the eggs are virtually indistinguishable, growers usually spray if egg numbers exceed the economic threshold. This leads to costly and ineffective spraying of budworm populations, exacerbation of pyrethroid resistance, loss of biological control through destruction of beneficial insects, and environmental pollution. A monocolonal antibody-based immunodot assay identified over 2,400 insectary-reared eggs with false negative and false positive rates of 0.4% and 0.1%, respectively; 345 of 346 eggs collected from cotton fields in four Southern states were correctly identified. The antibody could be packaged into a kit which would enable a consultant or grower to determine budworm and bollworm egg numbers in less than an hour. Armed with this knowledge, the grower could make an informed decision either to spray at a stage requiring less insecticide than one made later, or to employ cost-effective alternatives immediately. High costs of insecticides and insecticide application, absence of inexpensive alternative chemicals, renewed interest in biological control, and societal pressures to reduce environmental pollution combine to make this technology very attractive to cotton growers and consultants.

MANAGEMENT OF COTTON INSECTS AND OF PESTICIDE-RESISTANCE

Pyrethroids have been the chemicals of choice for control of the tobacco budworm (TBW), *Heliothis virescens* (F.), and the cotton bollworm (CBW), *Helicoverpa zea* Boddie, in cotton. Pyrethroid resistance was first detected in TBW in cotton in California in the early 1980's and in the midsouth and Texas in the mid 1980's (1) and has increased in intensity in several regions during subsequent years (2, 3, 4). Levels of resistance vary temporally and geographically, tending to start out low and increase during the growing season, and to be higher in cotton growing than in non-cotton growing areas (4, 5, 6). These trends indicate not only that the evolution of resistance is promoted by pyrethroid use, but that when pyrethroid use is reduced, pyrethroid resistant individuals are less successful than susceptible ones. This interpretation is supported by laboratory studies demonstrating reduced female attractiveness, reduced fecundity, and increased development times in resistant populations (7,8). Hence a reduction of pyrethroid use can be expected to retard the development of resistance and extend the useful life of these chemicals (9).

In Texas, midsouth, and Mexican cotton, TBW management becomes more challenging due to the co-occurrence of susceptible populations of CBW. It is important to be able to distinguish these two close relatives as early as possible, preferably at the egg stage, in order to avoid overexposing TBW to pyrethroids. Although subtle morphological differences between eggs of the two species have been described (10), they have not proven practical for wide scale use in field identification. Here I describe a simple, rapid immunoassay which can be scored by eye and which distinguishes TBW and CBW eggs unequivocally.

DEVELOPMENT OF A PROTOTYPE IMMUNOASSAY

Insects

Except where otherwise stated, all eggs used in this research were from colonies maintained in continuous culture at the Biological Control of Insects Research Laboratory (BCIRL) in Columbia, Missouri, using established rearing methods (11, 12, 13).

Antibody Production

Monoclonal antibodies (MABs) were produced according to a previously published protocol (14), using the soluble fraction of CBW whole egg homogenate (hereinafter egg homogenate) as immunogen. Hybridoma supernatants were screened for activity by indirect ELISA, using whole egg homogenate at 100 μg/ml as plate coating antigen. These and all other protein determinations were made by the method of Bradford (15), using a commercial dye reagent concentrate (Bio-Rad Laboratories, Hercules, CA). Hybridomas of interest were cloned twice by limiting dilution (16), using 50% SP2/0-conditioned medium in place of thymocytes. Antibodies from selected clones were produced by mass tissue culture in Nutridoma serum-free medium (Boehringer Mannheim Corp., Indianapolis, IN), concentrated by ultrafiltration and treatment with Aquacide (Calbiochem, La Jolla, CA) to at least 10.0 mg/ml, and stored at -80^0 C prior to use.

Seven fusions were performed, yielding a total of 340 hybridomas, nine of which produced MABs recognizing the immunizing antigen at an ELISA absorbance level at least three times above background. Three of these nine were successfully cloned, and all were specific for CBW egg homogenate only. One MAB, which was designated HZE-1, gave ELISA absorbances more than five times above background, and was chosen for further characterization. HZE-1 belongs to antibody subclass IgG1, as determined by the Sigma ImmunoType Mouse Moncolonal Antibody Isotyping Kit (Sigma Chemical Co., St. Louis, MO).

Immunoassay Protocol

Assays were performed on reinforced nitrocellulose membranes (Hybond-C Super, Amersham Life Sciences, Arlington, IL). For each assay, a membrane was cut to appropriate size and taped over a vinyl-jacketed matrix of darkly photocopied dots; the dots, visible through the translucent membrane, provide reference for the placement of individual

1.0 μl spots of egg homogenate or individual eggs. After a minimum of fifteen minutes for drying, the membrane was blocked for 1 h in 5.0 % BLOTTO (17) containing 5.0 mg/ml protease IV (Sigma Chemical Co., St. Louis, MO) to inhibit endogenous peroxidases. The remaining steps followed the immunodot protocol developed for egg predator gut analysis (18). Total assay time is about 3 h.

Assay sensitivity was determined by two-fold serial dilutions, from 1.0 μg to 1.0 ng protein, of whole egg homogenate. After 15 min. for drying, the membrane was assayed using MAB HZE-1 at 25, 50 and 100 μg/ml. The assay detected the same amount of CBW egg homogenate protein at all three MAB concentrations tested, 125 ng (Fig. 1). In order to minimize reagent costs, MAB concentration was set at the minimum, 25 μg/ml, for egg squashblot assays.

Prototype Egg Squashblot Assay

Individual eggs from long-term BCIRL moth colonies were obtained from insectary egg sheets (paper towels) or from soybean or cotton leaves exposed to adult moths in a rearing chamber. Infertile CBW eggs were obtained by exposing soybean leaves to virgin females. Eggs from egg sheets were loosened by immersion in tap water, while those from leaves were removed with an artist's brush. The brush was used to position individual eggs over the spots seen through the nitrocellulose membrane. Each egg was then squashed in place with a 2.5 mm diameter wooden applicator stick (Abco Dealers, Inc., Milwaukee, WI). The completed membrane was allowed to dry and then assayed.

Eggs squashed on membranes sometimes left a permanent brownish residue regardless of species. However only CBW eggs lead to development of a large pink spot in the assay (Fig. 2). HZE-1 recognizes an antigenic determinant which is found in CBW egg homogenate but is missing from TBW egg homogenate (C.L. Goodman & M.H.G., unpublished data). Nevertheless TBW eggs sometimes appeared weakly positive when assayed on membranes blocked with phenylhydrazine as peroxidase inhibitor (18). When increases in phenylhydrazine concentration and incubation time failed to reduce background and improve the signal-to-noise ratio, a commercial laundry whitener containing protease (19) was tested (data not shown). The laundry whitener improved the ratio but could not be made to go completely into solution at effective concentrations and had to be removed from the membranes by hand rubbing under running water. Blocking with protease IV at 5.0 mg/ml for 1 h reduces TBW non-specific reactions without degrading signal strength of CBW eggs. A 15.0 mg/ml (3X) stock may be prepared ahead of time and stored at -20^0 C, then thawed and diluted for use on the day of the assay. Residual protease causes casein coagulation in the MAB incubation step. Nevertheless CBW and TBW eggs are unambiguously distinguishable on the developed membrane (Fig. 2).

More than 2,500 insectary-reared eggs were assayed (Table 1). Five of 1,269 CBW eggs were negative, giving a false negative rate of 0.4%; one of 1,157 TBW eggs was positive, for a false positive rate of 0.1%. These rates are conservative, since females occasionally escape and may gain access to egg sheets of the other species in the insectary. TBW and CBW eggs oviposited on cotton and soybean leaves gave reliable results, and all

infertile CBW eggs were positive.

Assay of Eggs from Field Populations

Although long-term continuous colonies of CBW and TBW are maintained in several laboratories in different geographic areas of the United States, they were probably all derived from just a few field acquisitions. In order to be sure that the species-specificity of MAB HZ5-1 is not an artifact of long-term culture, an attempt was made to collect eggs from widely distributed field populations of both species. In late summer and fall of 1993, CBW and TBW larvae or adults were identified to species and collected from cotton and corn fields in Tift Co., Georgia, Washington Co., MS, Fayette Co., TN, and Burleson and Hidalgo counties, TX. The adults were mated and the resulting eggs were sent to the author's laboratory by air freight; larvae were allowed to pupate and the pupae were sent to the author's laboratory, where they were permitted to eclose, mate, and oviposit. These eggs were placed on membranes along with control eggs from the BCIRL insectary, and assayed as above.

Data from field these assays are presented in Table 2. Both species were collected at the Tift Co., GA site; all but one CBW egg were positive and all TBW eggs were negative (see Fig. 4). At the other field sites, only TBW were collected. The eggs resulting from these collections were all negative in the assay.

Assay of Additional Species

Other close relatives of the CBW and TBW may occasionally oviposit on cotton. In order to determine whether there is significant risk of false positives from such incidents, additional membranes were prepared with eggs of the following species maintained in the BCIRL Insectary: the ground cherry fruitworm (GCFW)*Heliothis subflexa* Grote, the fall armyworm (FAW) *Spodoptera frugiperda* (J.E. Smith), the cabbage looper (CL) *Trichoplusia ni* (Hübner), and the velvetbean caterpillar (VBC) *Anticarsia gemmatalis* (Hübner). All tested eggs of these four species were negative. FAW eggs tended to melanize intensely, but the resulting inky black dots were easily distinguished from the large pink spots of CBW positives (Fig. 3).

NEED AND PROSPECTS FOR OF A QUICK EGG IDENTIFICATION KIT

We in the United States and Mexico are fortunate that pyrethroid resistance has not become fixed in the TBW. The egg squashblot assay described here could be the cornerstone of a pyrethroid resistance management strategy which would reduce the frequency of pyrethroid sprays while prolonging their useful life against this pest. Additional benefits would be cost savings to growers, reduced environmental contamination, and enhancement of biological control through conservation of natural enemies. The total costs to the Southern cotton industry sue to yield loss costs of control is over three hundred million dollars (20), so there should be a substantial market for this technology.

Analysis of an analogous resistance problem involving two Old World *Helicoverpa* species is instructive. In Australian cotton, the Australian bollworm (ABW) *H. punctigera* Wallengren is fully susceptible to pyrethroids, while the Old World Bollworm (OWB) *H. armigera* (Hübner) exhibits high levels of resistance. Furthermore there does not appear to be a significant cost in success due to resistance in the OWB so that relaxing pyrethroid use does not lead to reduced levels of resistance. Nevertheless a decade-long insecticide resistance management (IRM) strategy has prolonged the usefulness of pyrethroids against this pest (21).

During the 1993-1994 field season, the use of a monoclonal antibody-based field identification kit to distinguish the two species was added to the IRM strategy (22). The assay is analogous to the one reported here, but packaged with proprietary technology into a faster format. Guidelines for the use of the kit (23) could be easily adapted to our situation. Briefly, scouts would collect individual eggs, at the rate of perhaps 100 per 400 ha, while performing their normal scouting duties; the eggs could be squashed and assayed upon return from the field. A prediction of whether or not pyrethroids would effect satisfactory control would depend upon the total TBW+CBW density, the proportion of eggs belonging to each species, and levels of resistance in TBW populations. Precise guidelines would need to be worked out, but generally speaking, spraying with pyrethroids when the proportion of TBW is high would not be effective unless both total densities and TBW resistance levels were low. Furthermore as the new strategy was implemented and natural enemy populations rebounded, it might be possible to avoid spraying pyrethroids at higher total densities and higher proportions of TBW, due to the increased mortality attributable to biological control. In situations where pyrethroid spraying was deemed inappropriate, biopesticides such as *Bacillus thuringiensis*, or insecticides with other chemistries, could be employed. However the latter course would require very careful choice and judicious use of chemicals, since pyrethroids are not the only class of insecticides to which the TBW has evolved resistance (24).

Insecticides will continue to play a role in insect pest management for the foreseeable future. As the number of approved and effective chemicals dwindles, emphasis must be placed on management programs which retard the evolution of resistance by susceptible pest populations. Added benefits of such programs include cost savings from avoiding unnecessary or ineffective spraying, reduced environmental contamination, and enhancement of pest population suppression by the preservation of biological control agents.

TABLES

Table 1. Results of squashblot assays on insectary-reared eggs

Species	CBW		TBW		GCFW		FAW		CL		VBC	
	+	-	+	-	+	-	+	-	+	-	+	-
SOURCE												
Egg sheets	1,096	5	1	1,052	0	84	0	84	0	84	0	84
Cotton leaves	24	0	0	24								
Soy leaves	80	0	0	80								
Infertile	64	0										

Table 2. Results of squashblot assays on field-collected eggs

Species	CBW		TBW	
	+	-	+	-
LOCALITY				
GA	135	1	0	84
MS			0	42
TN			0	24
TX			0	60

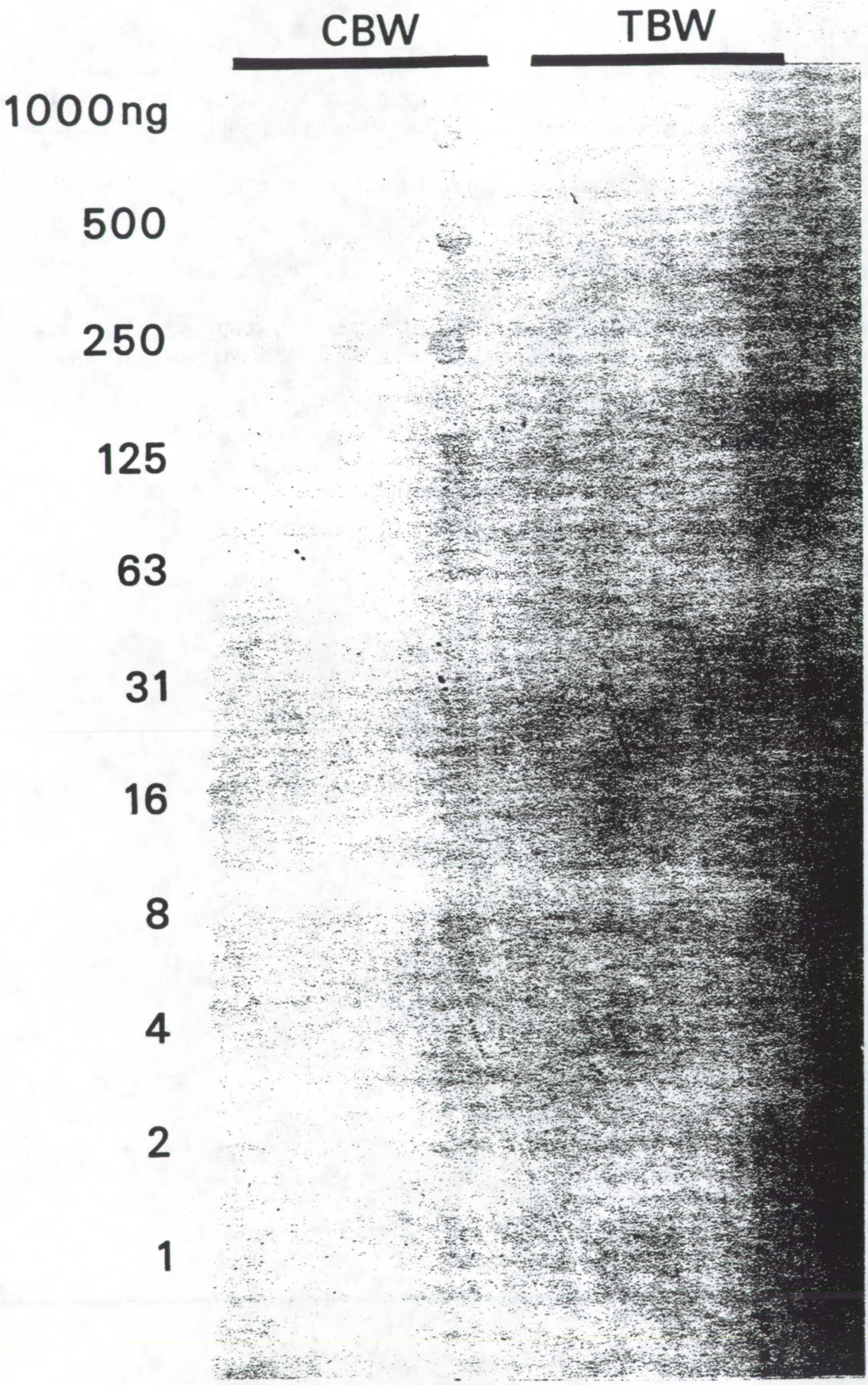
FIGURE LEGENDS

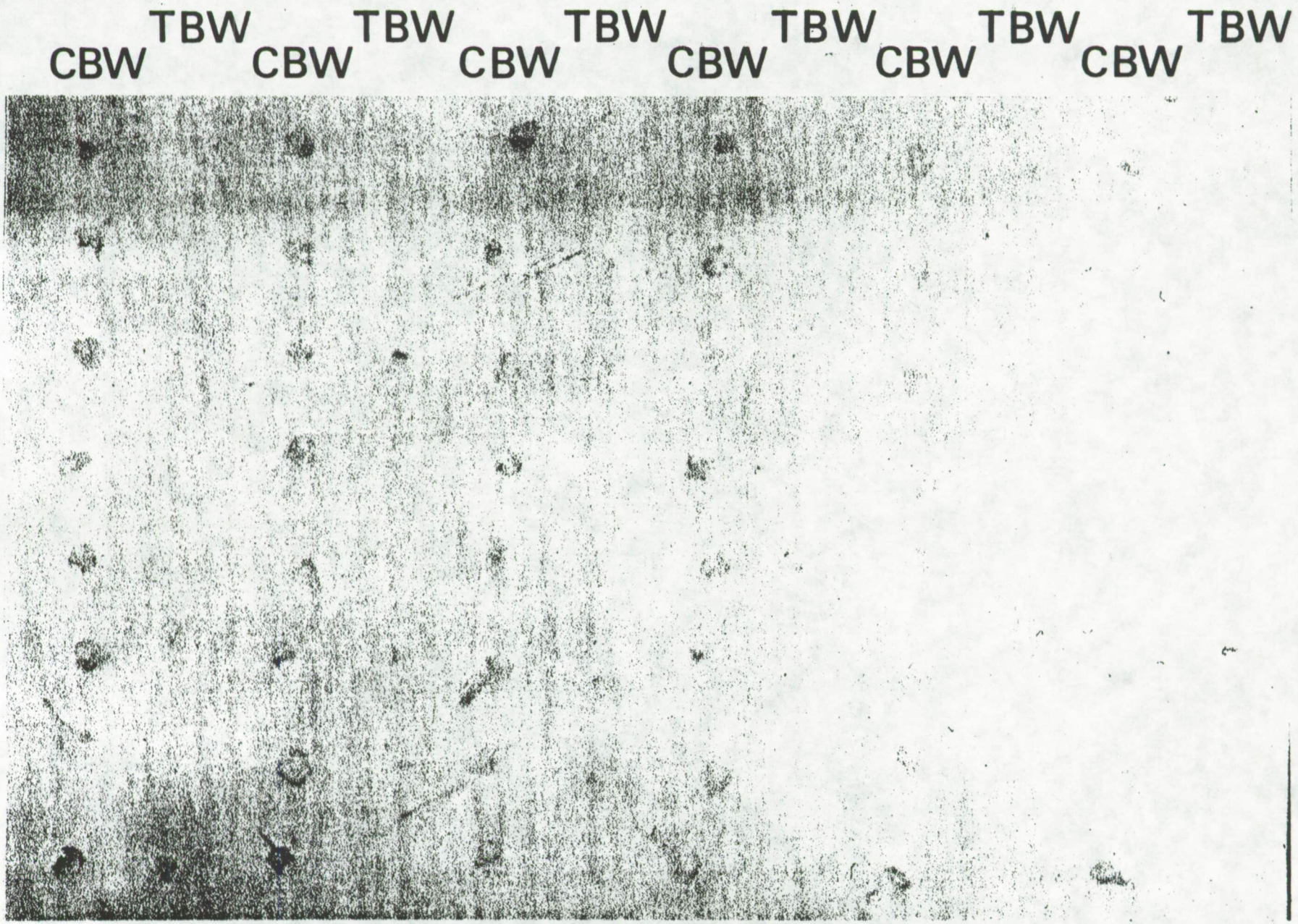
Fig. 1. Determination of assay sensitivity by titration (by triplicate serial two-fold dilutions) of egg homogenate from 1000 to 1 ng, against MAB HZE-1 at 25 μg/ml; identical results were obtained with the MAB at 50 and 100 μg/ml. Abbreviations per text.

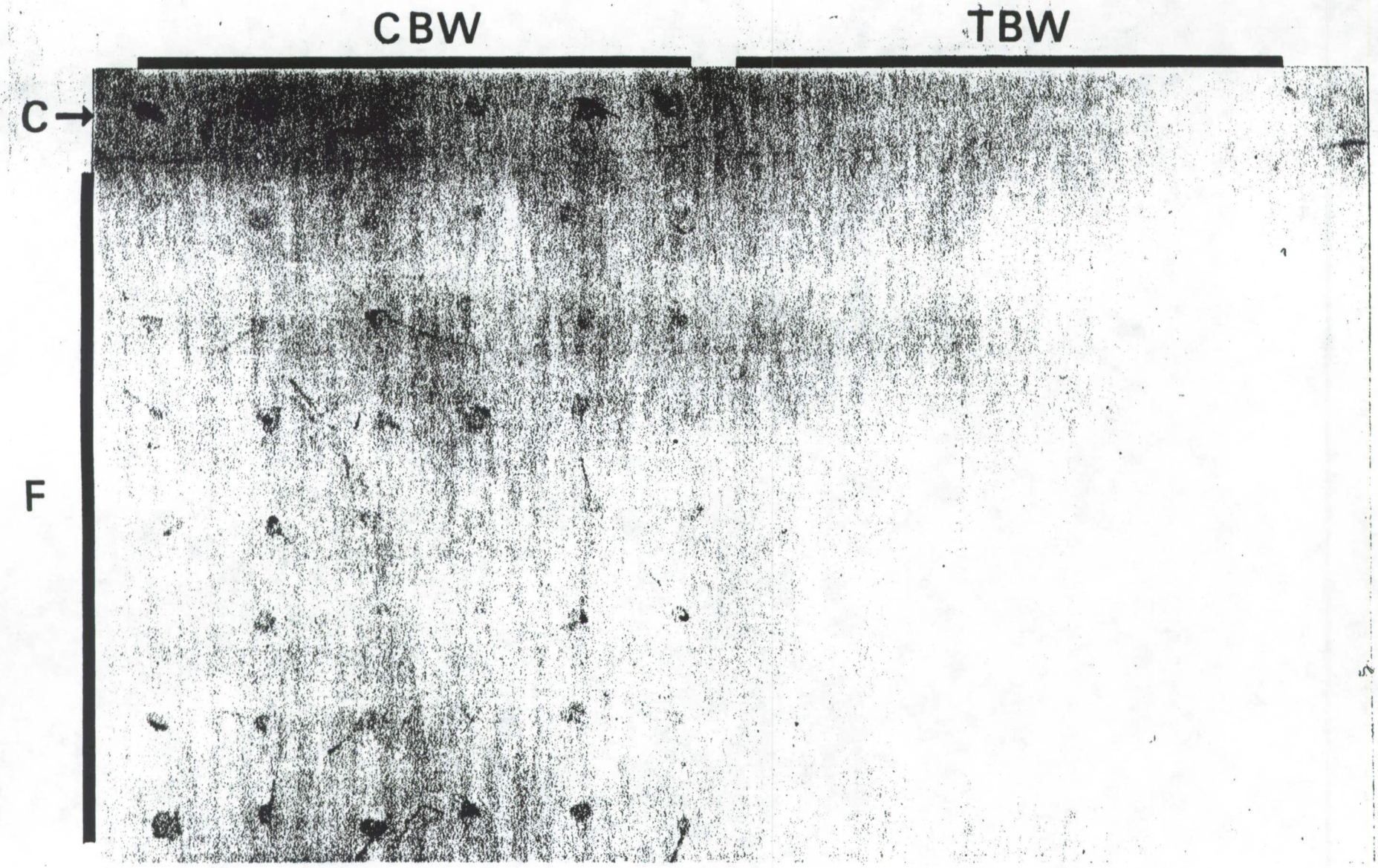
Fig 2. Species-specific identification of individual eggs by squashblot immunoassay. Abbreviations per text.

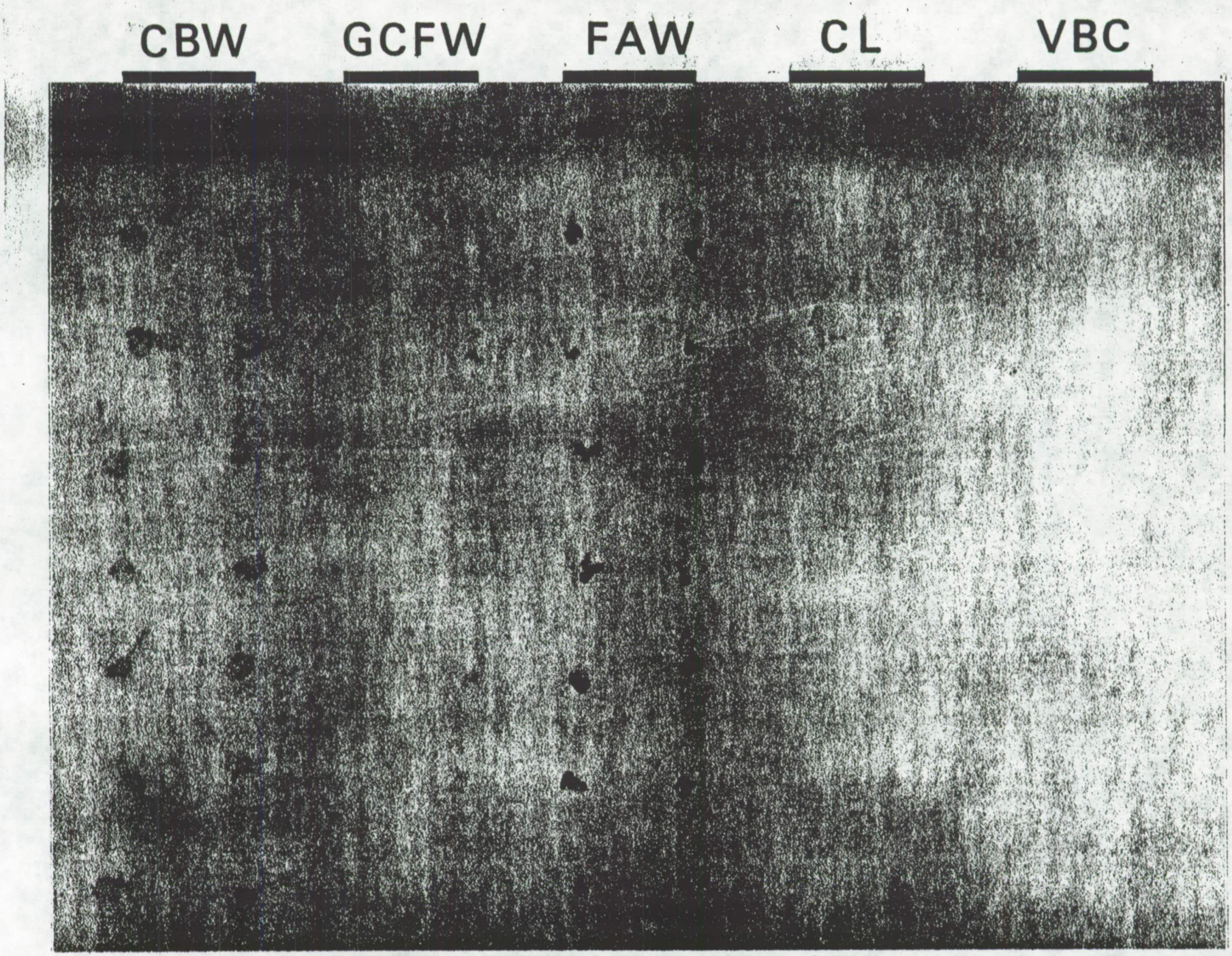
Fig. 3. Assay of eggs of additional species. Abbreviations per text.

Fig. 4. Assay of field collected eggs from Tifton, GA. C, control eggs from BCIRL insectary colonies; F, field collected eggs; other abbreviations per text.


CBW
TBW
1000ng
500
250
125
63
31
16
8
4
2
1


CBW
TBW
CBW
TBW
CBW
TBW
CBW
TBW
CBW
TBW
CBW
TBW


CBW
TBW
C
F


CBW
GCFW
FAW
CL
VBC

ACKNOWLEDGMENT

I thank Gary Elzen, Harry Gross, Mike O'Neill, Phillip Roberts and Dale Spurgeon for providing collected insects, thank Paula Peters for rearing insects in culture, Clyde Morgan and Laurel Shulse for technical assistance, and Fred Gould, David Ragsdale and Melissa Stuart for comments on an earlier version of the manuscript. Laurel Shulse was supported by a and USDA-ARS Research Apprenticeship.

REFERENCES CITED

1. **Gage, E.V., L.D. Hatfield & D.A. Wolfenbarger. 1991.** Introduction. Southwest. Entomol. Suppl. 15: 1-3.

2. **Carillo, J.L.M. 1991.** Monitoreo de resistencia a piretroides de gusano tabacalero *Heliothis virescens* en el noroeste de Mexico, pp. 59-67. *In* J.M. Brown & Deborah A. Richter [eds.], Proceedings Beltwide Cotton Production Research Conference, Nashville, TN, 2-7 January 1989. National Cotton Council of America, Memphis, TN.

3. **Elzen, G.W. 1991.** Pyrethroid resistance and carbamate tolerance in a field population of tobacco budworm in the Mississippi Delta. Southwest Entomol. Suppl. 15: 27-31.

4. **Graves, J.B., B.R. Leonard, S. Micinski & E. Burris. 1991.** A three year study of yrethoid resistance in tobacco budworm in Louisiana: resistance management implications. Southwest. Entomol. Suppl. 15: 33-41.

5. **Luttrell, R.G., K. Knighten, W.F. Kitten, G.L. Andrews, F.A. Harris & J. Reed. 1991.** Monitoring pyrethroid resistance in the tobacco budworm in Mississippi: implications for resistance management. Southwest. Entomol. Suppl. 15: 5-26.

6. **Plapp, F.W., Jr. 1991.** Pyrethroid resistance in the tobacco budworm: five years later. Southwest Entomol. Suppl. 15: 69-73.

7. **Campanhola, C. & F.W. Plapp, Jr. 1989.** Managing pyrethroid resistance in the tobacco budworm: bioassays, resistance mechanisms and biological constraints, pp. 352-359. *In* J.M. Brown & Deborah A. Richter [eds.], Proceedings Beltwide Cotton Production Research Conference, Nashville, TN, 2-7 January 1989. National Cotton Council of America, Memphis, TN.

8. **McCutcheon, B.F., F.W. Plapp, Jr. & H.J. Williams. 1989.** Reproductive deficiencies associated with pyrethroid resistance in the tobacco budworm, pp. 364-366. *In* J.M. Brown & Deborah A. Richter [eds.], Proceedings Beltwide Cotton Production Research Conference, Nashville, TN, 2-7 January 1989. National Cotton

Council of America, Memphis, TN.

9. **Mallett, J. & R. Luttrell. 1991.** A model of insecticidal control failure: the example of *Heliothis virescens* on cotton. Southwest. Entomol. Suppl. 15: 201-212.

10. **Bernhardt, J.L. & J.R. Phillips. 1985.** Identification of eggs of the bollworm, *Heliothis zea* (Boddie), and the tobacco budworm, *Heliothis virescens* (F.). Southwest. Entomol. 10: 236-238.

11. **Berger, R.S. 1963.** Laboratory techniques for rearing *Heliothis* species on artificial medium. USDA Agr. Res. Svc. ARS-33-84: 1-4.

12. **Ignoffo, C.M. 1965.** The nuclear-polyhedrosis virus of *Heliothis zea* (Boddie) and *Heliothis virescens* (Fabricius). II. Biology and propagation of diet-reared *Heliothis*. J. Invertebr. Pathol. 7:217-226.

13. **Ignoffo, C.M. & O.P. Boening. 1970.** Compartmented disposable plastic trays for rearing insects. J. Econ. Entomol. 63: 1696-2697.

14. **Lenz, C. J. & M. H. Greenstone. 1988.** Production of a monoclonal antibody to the arylphorin of *Heliothis zea*._Arch. Ins. Biochem. & Physiol. **9**:167-177.

15. **Bradford, M.M. 1976.** A rapid and sensitive method for quantitation of microgram quantities of protein utilizing the principle of dye binding. Anal. Biochem. 72: 248-254.

16. **McKearn, T.J. 1980.** Cloning of hybridoma cells by limiting dilution in fluid phase, p. 374. *In* Kennett, R.H., T.J. McKearn & K.B. Bechtol [eds.], Monoclonal antibodies, Plenum Press, New York.

17. **Johnson, D.A., W. Gautsch, J.R. Sportsman & J.H. Elder. 1984.** Improved technique utilizing non-fat dry milk for analysis of proteins and nucleic acids transferred to nitrocellulose. Gene Anal. Tech. 1: 3-8.

18. **Greenstone, M.H. and S.C. Trowell. 1994.** Arthropod predation: A simplified immunodot format for predator gut analysis. Ann. Entomol. Soc. Amer. 87: 214-217.

19. **Johnson, D.W., A.F. Cockburn & J.A. Seawright. 1992.** Quick blots and nonradioactive detection of DNA probes for the identification of mosquitoes. J. Amer. Mosq. Contr. Assoc. 8: 231-236.

20.

21. **Forrester, N.W., M. Cahill, L.J. Bird & J.K. Layland. 1993.** Management of pyrethroid and endosulfan resistance in *Helicoverpa armigera* (Lepidoptera:

Noctuidae) in Australia. Bull. Ent. Res. Suppl. 1: 1-132.

22. **Forrester, N.W., I. Anderson & B. Pyke. 1993.** Insecticide resistance management strategy for Heliothis 1993-1994, pp. 8-11. *In* A.J. Shaw [ed.], Cotton Pesticides Guide 1993-94, New South Wales Agriculture, Narrabri, NSW, Australia.

23. **Forrester, N.W. 1993.** Guidelines for using the new Heliothis identification kit, pp. 11-13. *In* A.J. Shaw [ed.], Cotton Pesticides Guide 1993-94, New South Wales Agriculture, Narrabri, NSW, Australia.

24. **Elzen, G.W., B.R. Leonard, J.B. Graves, E. Burris & S. Micinski. 1992.** Resistance to pyrethroid, carbamate, and organophosphate insecticides in field populations of tobacco budworm (Lepidoptera: Noctuidae) in 1990. J. Econ. Entomol. 85: 2064-2072.

FOOTNOTE

[1]This article reports the results of research only. Mention of a proprietary product does not constitute an endorsement or a recommendation for it use by the USDA. All programs and services of the USDA are offered on a nondiscriminatory basis without regard to race, color, national origin, religion, sex, age, marital status, or handicap.

SUPER FAST MOLECULAR DYNAMICS - A SPACECRAFT TECHNOLOGY TRANSFER

Hon M. Chun, Chief Technical Officer
Moldyn, Inc., Cambridge, Massachusetts 02138

Carlos E. Padilla, Senior Staff Scientist
Moldyn, Inc., Cambridge, Massachusetts 02138

Kim B. Blair, Senior Staff Scientist
Moldyn, Inc., Cambridge, Massachusetts 02138

Jun H. Li, Senior Research Assistant
Moldyn, Inc., Cambridge, Massachusetts 02138

David N. Haney, Consultant
Haney Associates, San Diego, California 92131

Howard E. Alper, Computational Chemist
Moldyn, Inc., Cambridge, Massachusetts 02138

ABSTRACT

Molecular modeling is beginning to play an increasingly important role in the drug discovery and materials design process. Classical molecular dynamics models, which treat each atom as a three degree-of-freedom particle, have been widely used for such purposes. However, as researchers became interested in larger molecules and longer time scales, such modeling methods have become severely limited. This is due to the large number of degrees of freedom, as well as the high frequency content of the dynamics, which requires small integration step sizes. To address this problem, the authors have developed a modeling approach that is based on spacecraft dynamics modeling techniques. A molecule is substructured into a set of interconnected rigid and flexible bodies and atomistic regions. The atoms that are aggregated into bodies are those that exhibit collective motion, with small relative motions among each body's constituent atoms. High frequency content is reduced by retaining only low frequency modes for flexible bodies. The substructured description of the system has a much smaller number of degrees of freedom than the all-atom model. Further development of the method is proceeding, through collaborations with researchers in academia as well as in industry.

MOLECULAR MODELING

Molecules are modeled with computers based on the cartesian coordinates of every atom. The ability to visualize every atom and its movements has provided valuable information for chemists studying structure-function relationships. This type of molecular modeling tool has been made possible by contributions in four major areas: (1) X-ray crystallography and the publication of structures in the Brookhaven Protein Data Bank [1], (2) molecular mechanics and dynamics software and forcefields developed primarily by Karplus et al. [2], and Allinger [3], (3) molecular graphics hardware and software, and (4) raw computational power.

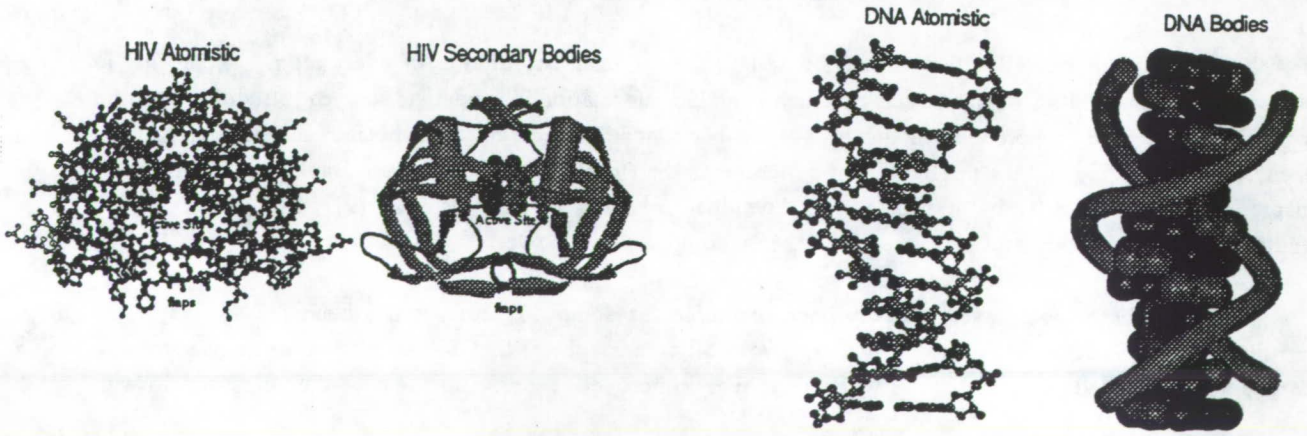
Although these tools have generally been available for about 10 years, it has been the advent of the powerful graphics UNIX workstation (particularly from Silicon Graphics) that has caused an explosive growth. Researchers have studied many molecules with this tool and improved its performance over the years to deal with

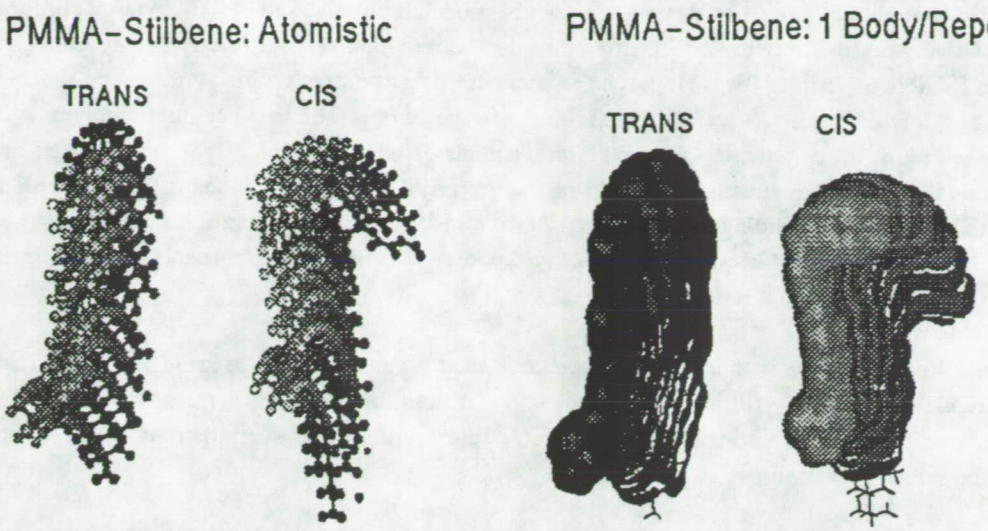
larger and larger systems. However, since the focus of these systems is based on classical molecular mechanics, the tools continue to focus on individual atoms as the structural unit. This has two primary drawbacks (a) limiting the size (or length of simulation) of the molecular system (to less than 25,000 atoms or 100 picoseconds) and (b) difficult analysis of structure-function (collective atom architectures) relationships.

In addition to providing insight to structure-based mechanisms, a key goal to molecular simulations is to reproduce and predict laboratory experiments that are based on molecular events that occur in the micro- to milli-second time frame. Many large molecules are organized into aggregates of atoms that function collectively. Such collections of atoms could be things like monomer residues, alpha helices, beta sheets, molecular domains, or whole subunits (chains). In order to simplify the vast amounts of information in large biological and synthetic molecules, and focus on molecular substructures responsible for function, it is necessary to be able to aggregate atoms that function with collective motions into larger objects or bodies. Not only will this provide us with the proper perspective to structure-function relationships, it will allow us to avoid calculating the time-consuming and uninteresting high frequency motions of these molecular systems.

Software developed in the 1980's for solving large aerospace simulation problems [4] has been adapted to molecular systems [5] in order to provide this atom aggregation (molecular substructuring) and body-based simulation technique. The resulting simulation software is called MBO(N)D (Multi-Body Order (N) Dynamics) This method enables the multigranular treatment of systems which contain particles (atoms), flexible bodies and rigid bodies. Examples of molecular structures and substructures with bodies are shown in Figure 1. In this figure, the protein is substructured into secondary bodies, beta sheets (represented as flat ribbons) and alpha helices (represented as cylinders); the nucleic acid is substructured into two backbone bodies and base bodies; and the polymer is substructured into monomer bodies. These substructuring schemes represent only one scheme (for each type of molecule) of many possible schemes available to the substructuring methodology. For example, the protein in Figure 1A could be substructured as 2 bodies representing the 2 subunits; the nucleic acid could be substructured to include base-pairs or several base-pairs into a single body; and the polymer could be substructured to include one whole chain in a body, allowing one to simulate more chains within the system. By using these methods it is possible to reduce the number of objects computed from hundreds-of-thousands of atoms to tens or hundreds of bodies, and increase the simulation time step from one femtosecond to tens- or hundreds- of femtoseconds.

FIGURE 1: (A) HIV protease, a protein, (B) 1 turn of DNA, a nucleotide, (C) PMMA-stilbene, a polymer





Organizing these large molecular systems into bodies not only improves the computational efficiency, but it also allows a more simplified analysis of the collective atom (substructure)-function relationships. Examples of molecular events whose study will be enabled by this methodology include (a) molecular docking events, (b) molecular conformational changes such as HIV flap movement, subunit shift in hemoglobin between oxy and deoxy states, (c) protein/polymer folding and diffusional conformational changes, and (d) free energy changes. As the molecular structure of large systems becomes more well defined this need to substructure molecules into architectural units and to simulate molecular bodies will become more important even on the next generation of supercomputers.

SUBSTRUCTURED MODELING APPROACH

This paper presents a new approach to the modeling of macromolecular systems. The method is based on the idea that the essential dynamics of such systems are captured by the low frequency modes of the system [6-11]. There are many biological processes that take place in the nano- to milli-second time frames that cannot be modeled with current methods because of the large size of the systems and small time steps needed for numerical integration. Mechanical properties of polymers are difficult to predict unless large time-scale simulations are performed on large enough models of the material. By considering only the low frequency behaviors, much larger integration time-steps can be used. Modal approaches also reduce the number of degrees of freedom that need to be modeled. Modal methods have been used in the past to study macromolecular systems. However,because of the linearization that is introduced in order to obtain the mode shapes and frequencies, such models are valid only for a small region near the conformation about which the linearization was performed.

The approach that we have developed is based on the substructuring of a large molecular system into bodies and particles. The substructuring is determined by the amount of motion to be expected between atoms. For regions where motions are expected to be very small, or small enough to be unimportant for the purposes of the simulation, the atoms can be grouped together into rigid bodies. Regions where there are moderate amounts of motion can be modeled as flexible bodies. Regions where large conformational changes are expected can remain atomistic. Since this modeling approach allows large motions between bodies, as well as between individually modeled atoms, it is expected to be valid for a much larger region in conformational space.

Secondary structures of proteins are primary candidates for grouping into bodies. Alpha helices and beta sheets are naturally thought of as having collective motions. In fact, analysis reported in the paper by Swaminathan

et al. [12] has demonstrated this. Loop and turn regions can be modeled atomistically to allow large conformational changes. Parts of these regions might also be grouped into flexible bodies. Although synthetic polymers are composed of more flexible molecules than biopolymers, there are substructures within synthetic polymers that can be rigid and act as collective groups as well. In addition, since the function of some synthetic polymers is due in part to the close packing of many polymer chains, the movements of whole chains of polymers can be correlated and provide a means for molecular substructuring. Aromatic groups, side chains residues and even whole polymer fibers can be treated as bodies depending on the study. Because of the distinct separation of the system into high frequency atomistic regions and low frequency flexible/rigid bodies, this modeling approach is highly amenable to treatment by multiple time scale integration techniques.

The eigensolution process is more tractable for this substructured model because the mode shapes and frequencies are calculated separately for individual bodies, rather than for the entire system. It is computationally less expensive to compute eigensolutions of component bodies within a system than it is to compute the eigensolution of the entire system.

There are several advantages of this substructured modeling technique. The elimination of high frequency content from aggregated groups of atoms allows larger time-steps, and hence longer time frame simulations or larger molecular systems. The ability to manipulate groups of atoms collectively as bodies allows the modeling and simulation of events that cannot be treated by all-atom models. Examples include flap motion of HIV protease, sub-unit interactions in hemoglobin, and dynamic docking of ligand and substrate. The exact constraint formulation allows the relative motion between bodies to be optionally specified as a function of time, thereby allowing inverse dynamics simulations to be performed based on hypothetical or experimentally observed molecular motions.

Within this substructured modeling approach, there are several alternatives for the calculation of bond and non-bond interactions. Conventional all-atom calculations [2] can be performed. Body forces and torques are obtained by summing up atomistic forces and moments over the atoms that make up each body. Modal forces are computed by multiplying atomistic forces by the mode shapes. This projects the physical forces into the low frequency subspace of the body. A modified approach replaces the body internal interactions by modal stiffness terms. The bond and non-bond pairlists are reduced in size by eliminating explicit interaction calculations between atoms that reside on the same body, resulting in a more efficient calculation. Fast multipole algorithms [13-15] could be applied to speed up non-bond calculations for large systems. As noted above, we have developed a body-based multipole algorithm for better computational efficiency at high levels of aggregation.

The outputs from the substructured modeling approach are of the same type as that from all-atom simulations. This is because the coordinates and velocities of every atom are known once the translations, rotations, and modal amplitudes of the atom's parent body are found. Thus, conventional post-simulation analysis algorithms can be directly applied to these simulations.

PERTINENT SPACECRAFT MODELING TECHNOLOGIES

Many elements of this substructured modeling approach have been adapted from spacecraft and mechanical dynamics modeling techniques. In such systems, large relative motions are allowed between articulated bodies which may be either rigid or flexible. Elastic behavior of the individual bodies are modeled by component modes. Constraints at hinges/joints are handled in an exact manner, and there have been recent developments [16-19] that have resulted in fast algorithms with Order (n) computational complexity.

MBO(N)D traces its roots to NASA's multibody dynamics simulation software called DISCOS (Dynamics Interaction Simulation of COntrols and Structures). The authors have had extensive experience with the use of

DISCOS for the modeling of spacecraft dynamics and controls for a variety of NASA of Department of Defense missions [20-23]. More recently, Chun, Turner, and Frisch [10] replaced the original dynamics algorithm with a highly efficient Order (n) recursive algorithm that greatly reduces the computational requirements for large systems. This resulted in a new version of the code, called N-DISCOS. In addition, a rigorous formulation for topological closed loops [24] and an efficient gear reduction model were implemented [25] in support of the NASA Flight Telerobotic Servicer. Modifications were made to enable the modeling of human biomechanics for the purpose of task analysis, man-machine interface, and human factors engineering. The resulting software, HMT-CAD (Human-Machine-Task - Computer Aided Design) [26], was streamlined to run on Personal Computers. That work was supported through the NASA SBIR Program. The MBO(N)D molecular dynamics modeling software builds upon these algorithmic and software developments, and demonstrates a significant leap in the transfer of technology from spacecraft applications to the biotechnology and materials engineering arena.

DEVELOPMENT OF SUPER-FAST MOLECULAR DYNAMICS

There were several important modifications to N-DISCOS that were made to create the molecular modeling capability embodied in MBO(N)D. A three degree-of-freedom particle element was introduced to handle atomistically modeled regions of the molecular system. Exact and rigorous bond length and bond angle constraints were implemented to eliminate the high frequency vibrations due to bond stretching and angle bending. These constraints can be optionally specified. A hierarchical body-based multipole algorithm was implemented for electrostatic interactions, so as to take advantage of the fact that invariant multipole coefficients can be computed if they are based on body-fixed coordinate frames. The MBO(N)D code was interfaced to conventional computational chemistry software to utilize their user and data interface modules, as well as to access their force field calculations.

Test Results

Many tests were conducted to demonstrate the various elements of the substructured modeling approach. Constrained dynamics of glycine dipeptide was used to verify the bond-length constraints within MBO(N)D. Comparison with AMBER results showed good agreement in the dihedral angle dynamics. Figure 3 shows the similarities in the dihedral angle time histories between the AMBER/MBO(N)D simulation and the AMBER/SHAKE simulation. (SHAKE is a conventional iterative algorithm for maintaining constraints.) Analysis of the numerical results showed that the mean values of the dihedrals exhibited good agreement (three significant digits of agreement) with AMBER results.

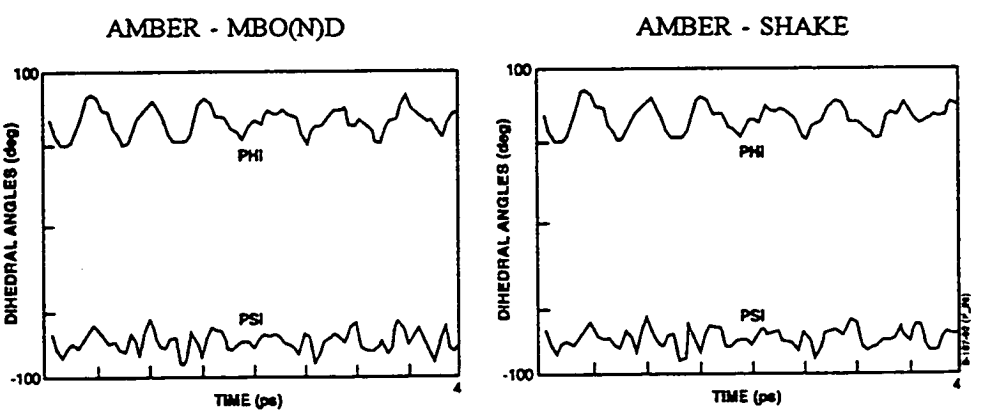


Figure 3: Comparison of dihedral angle time histories for bond-length constrained simulations using MBO(N)D vs. using SHAKE.

Information on modal modeling was obtained through modal simulation of decaglycine, which indicated that 20 modes was sufficient for capturing the essential dynamics of the molecule. In this study, an error criterion was developed which required that the rms position deviations agreed to within 0.1 Å of AMBER results. All-atom AMBER results were compared to MBO(N)D simulations using various numbers of modes. Table 1 shows that the error criterion was satisfied at 20 modes and beyond. The table also illustrates the reduction in frequency content possible with the different number of modes. For the 20 mode model, a time step on the order of around 100 fs is sufficient to adequately simulate the dynamics of this 79 atom molecule, compared to 1-2 fs needed for a fully atomistic simulation.

Table 1: Exact (AMBER) vs. Modal (MBO(N)D) Approximation --
20ps Decaglycine Dynamics

NUMBER OF MODES	MEAN DEVIATION FROM AMBER SIMULATION (Å)	HIGHEST FREQUENCY RETAINED (cm^{-1})	PERIOD OF HIGHEST MODE (fs)
1	4.32	18	1899
5	0.26	32	1049
10	0.14	57	587
20	0.08	101	329
27	0.07	146	228
33	0.04	197	168
94	0.02	595	56

Several substructuring strategies were investigated for modeling a portion of bacteriorhodopsin. The bacteriorhodopsin system could logically be treated as either one single body or two alpha helix bodies connected via a proline ring. The proline ring could be modeled as either a body or as a group of individual particles. The results are summarized in Table 2, where the numbers in parentheses indicate the number of modes used. This study demonstrated that a substructuring strategy involving flexible body models with eight modes for the alpha helices and atomistic models for the proline ring was able to capture the essential bending motion of bacteriorhodopsin, as indicated by comparison with an all-atom AMBER simulation.

Table 2: Comparison of several substructuring strategies for bacteriorhodopsin particles, flexible bodies, or rigid bodies for the helix-proline-helix model.

TYPE OF RUN	RMS FLUCTUATION (DEGREES)
Amber	17.55
One Flex Body	8.05
Three Rigid	1.5
Flex (8) Flex (4) Flex (8)	10.79
Flex (2) Particle Flex (2)	12.01
Flex (8) Particle Flex (8)	18.18

* Numbers in parentheses indicate the number of flexible modes retained.

Accuracy tests of the body-based multipole algorithm were performed on crambin and several complete turns of the B-DNA Dickerson dodecamer. Electrostatic interactions obtained from the body-based multipole algorithm were found to agree with those obtained atomistically using X-PLOR and AMBER by about four digits of accuracy, which is considered a good agreement.

Preliminary timing runs using various substructuring schemes for B-DNA (with up to around 2,000 atoms) showed that even at low levels of aggregation (about 30-60 atoms per body), the MBO(N)D code executes about 10-20 times faster. (See Table 3.) The stability of the 10 fs time-step used by MBO(N)D (compared to 1-2 fs needed by atomistic calculations) is demonstrated in Figure 4, which shows the conservation of energy for a 3-turn DNA simulation.

Table 3: Timing Results from B-DNA Simulations

STRUCTURE	NO. BODIES	NO. ATOMS	AMBER CPU TIME/ SIM TIME (s/ps)	MBO(N)D CPU TIME/SIM TIME (s/ps)	SPEEDUP
1 TURN DNA	12	758	1790	146	12
				74 (dt=20 fs)	24
3 TURN DNA	72	2278	7115	470	15

AMBER runs performed at 1 fs timesteps
MBO(N)D runs performed at 10 fs timesteps, except where indicated otherwise

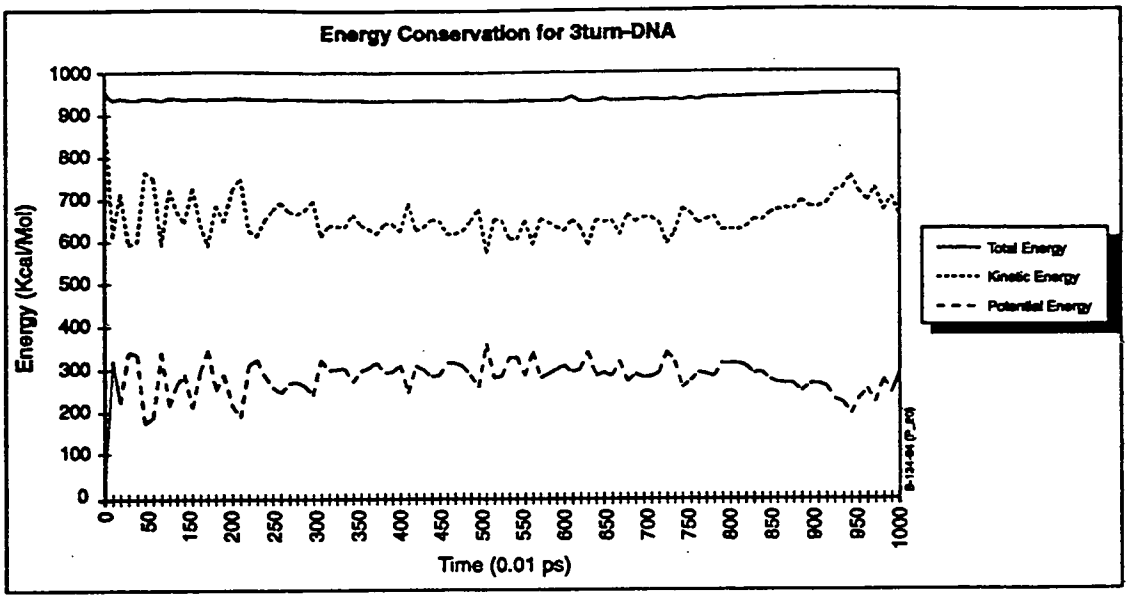


Figure 4: Stability of MOB(N)D 10 fsec Time Step Simulation for a 3-turn DNA

Future Work

Additional research is planned for investigating the substructured modeling approach to assess its speedup potential and modeling accuracy, and to gather information on appropriate substructuring strategies for various molecular systems. Collaborations for this purpose have been set up with Harvard University, Yale University, and Princeton University, as well as with several pharmaceutical companies. The MBO(N)D code is currently at the prototype stage. It is at preliminary levels of integration with AMBER 4.0, CHARMM, and X-PLOR.

CONCLUSION

The substructured modeling approach provides the framework for dealing with collective motions, models with varying degrees of fidelity (higher fidelity near active site, lower fidelity elsewhere), and a way of treating large systems and long time-frame systems in a computationally tractable manner. It is our belief that this modeling technique will provide the enabling technology for the scientific study of a wide range of biomolecular and materials phenomena.

BIBLIOGRAPHY

1. Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, T., Shimanouchi, T., and Tasumi, M., *J. Mol. Biol.*, **112**, 535-542 (1977)
2. Brooks, B.R., R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*,, 4, 187-217 (1983)
3. Allinger, N.L., *J. Am. Chem. Soc.*, 99, 8127-8134 (1977)
4. Bodley, C.S., Devers, A.D., Park, A.C., and Frisch, H.P., NASA Technical Paper 1219 (1978)
5. Turner, J. D., Weiner, P. K., Chun, H. M., Lupi, V., Gallion, S., and Singh, U. C., Ch. 24., Computer

Simulation of Biomolecular Systems: Theoretical and Experimental Applications, **2**, Gunsteren, W.F., Weiner, P.K., and Wilkinson, A. J., eds., ESCOM, Leiden (1993)

6. Levy, R. M., Karplus, M., Kushick, J., and Perahia, D., *Macromolecules*, **17**, 1370 (1984)
7. Ichiye, T., and Karplus, M., *Proteins*, **11**, 205 (1991)
8. Horiuchi, T., and Go, N., *Proteins*, **10**, 106 (1991)
9. Space, B., Rabitz, H., and Askar, A., *J. Chem. Phys.*, **99**, 9070, (1993)
10. Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C., *Proteins*, **17**, 412 (1993)
11. Mizuguchi, K., Kidera, A., and Go, N., *Proteins*, **18**, 34 (1994)
12. Swaminathan, S., Harte, W. E., and Beveridge, D. L., *J. Am. Chem. Soc.*, **113**, 2717 (1991)
13. Greengard, L., and Rokhlin, V., *J. Comp. Phys.*, **73**, 325 (1987)
14. Greengard, L., and Rokhlin, V., *Chem. Scr.*, **29A**, 139 (1989)
15. Ding, H.-Q., Karasawa, N., and Goddard III, W.A., *J. Chem. Phys.*, **97**, 4309, (1992)
16. Bae, D.-S., and Haug, E.J., *Mechanics of Structures and Machines*, **15**, 359-382 (1987)
17. Chun, H. M., Turner, J. D., and Frisch, H. P., Paper AAS-89-457, AAS/AIAA Conf., Stowe, VT (1989)
18. Rodriguez, G., Jain, A., and Kreutz, K. *International Journal of Robotics Research*, **10**, 371-381 (1991)
19. Jain, A., Vaidehi, N., and Rodriguez, G., *J. Comput. Phys.*, **106**, 258-268 (1993)
20. Chun, H. M., et al, Talon Gold Engineering Simulation Program - Fourth Interim Technical Report, CSDL-R-1729, Contract No. F04701-82-C-0062, Section 2, The Charles Stark Draper Laboratory, Inc., Cambridge, MA, 5-91 (1984)
21. Turner, J.D., Chun, H.M., and Soosaar, K., N-ROSS Satellite Dynamic Stability Analysis: Final Report, Cambridge Research, A Division of Photon Research Associates, Inc., Cambridge, MA (1986)
22. Chun, H.M., Turne, J.D., and Yu, C., Control and Structural Optimization for Maneuvering Large Flexible Spacecraft, Task 6: Final Report, NASA Contract No. NAS1-18763, Subcontract No. 89521160H, Photon Research Associates, Inc., Cambridge, MA (1990)
23. Padilla, C.E., MACE - Nonlinear Modeling, Simulation , and Preliminary Contorl of the Baseline Test Article, MIT Space Engineering Research Center Report No. 13-91R (1990)
24. Chun, H.M., Turner, J.D., and Frisch, H.P., Paper AIAA-90-3385-CP AIAA Guid. Nav. & Control Conf., Portland, OR (1990)
25. Chun, H.M., Turner, J.D., and Frisch, H.P., Paper AAS-91-455 AAS/AIAA Astrodynamics Specialist Conference, Durango, CO (1991)
26. Blair, K.B., Kondraske, G. et al., Human-Machine-Task CAD Software Development: Final Report. Contract #NAS5-31493 (1993)

Computer Aided Design and Engineering

Feature-based Modeling for Rapid Functional Prototyping: The Quick Turnaround Cell

by
John F. Montgomery, US Army Missile Command
Dr. David C. Anderson, Purdue University Engineering Research Center
Dr. Tien Chien Chang, Purdue University Engineering Research Center

Abstract

The US Army Missile Command (MICOM), The National Science Foundation Engineering Research Center at Purdue University, Loral Vought Missile Systems, and several industrial partners are currently developing a Quick Turnaround Cell (QTC) System that provides for rapid functional prototyping of machined parts utilizing feature-based modeling and generative process planning. The goals of the QTC project are to develop an integrated "art to part" system that assists both defense and commercial developers in prototyping concepts while helping eliminate many producibility / manufacturability issues during early development phases. The intent of the QTC system is to provide designers with the capability to easily design components for a new concept and rapidly produce functional components for concept testing.

The benefits gained from the QTC include both a large reduction in time/cost required to go from idea to physical parts and the ability to address manufacturing concerns during early design activities. The philosophy behind the QTC is consistent with emerging themes of Concurrent Engineering (CE) and Integrated Product/Process Development (IPPD). While the QTC is being developed to meet the needs of MICOM and the ERC industry partners, the QTC technologies offer the potential for significant benefit to the industrial base and national competitiveness.

The QTC provides an integrated system that includes manufacturing-feature-based modeling, automatic generative process planning, automated numerical control (NC) code generation, simulation of machining processes for verification of process plan, fixturing and numerical control codes, digital numerical code communication with selected NC machines, inspection/validation of machining process, and geometry portability to a commercial computer aided design system. The QTC domain consists of machinable parts that are typically produced using conventional milling and turning processes. The significant accomplishments of this project are the results of a tight coupling between design and manufacturing planning afforded by the feature-based modeling approach.

INTRODUCTION

As the Department of Defense undergoes a scaled down and as industry struggles to remain competitive in a global economy, it is becoming more evident that the technical community must develop a better set of engineering tools that allow manufacturing issues to be addressed earlier in the development cycle thereby reducing the overall time and cost of new products [44]. This philosophy is the driving force behind the development of the Quick Turnaround Cell. The QTC provides the capability to bring manufacturing issues related to the machining of components to the attention of the designer while the "design intent" is

still fresh on the designers mind and while alternative concepts have minimum impact on the overall system. The QTC capabilities are the result of a tight coupling between design and manufacturing made possible by a modular systems architecture and technical approach based on feature-based design and generative process planning.

SYSTEM ARCHITECTURE

The QTC system architecture, as shown in Figure 1, consists of several tightly coupled modules including Feature-based Design, Feature Refinement, Generative Process Planning, Automated Numerical Control (NC) Code Generation, Machining Simulation and a Cell Controller for Direct Numerical Control code download to selected NC machines [37].

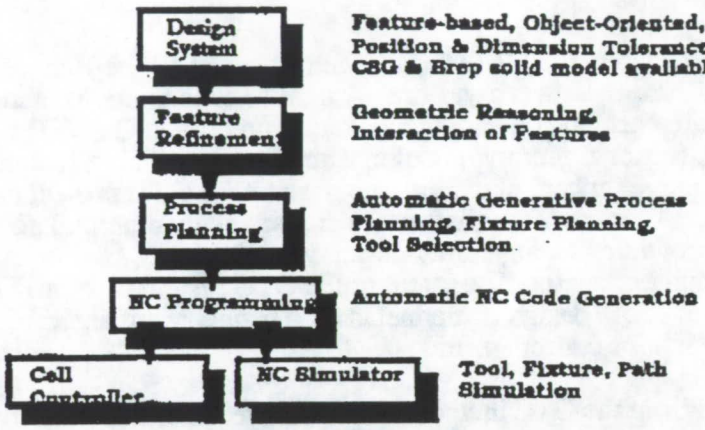


Figure 1. System Architecture

The intelligence of the QTC system is based on the use of features that allow for both a geometric representation required to support concept definition while providing a topology and taxonomy used to conduct geometric reasoning and generative process planning. The design module provides an object oriented, graphical, three-dimensional design based on a small number of features. Utilizing a relatively small number of features, the designer can combine features to form complex geometric shapes. The Feature Refinement module [1] utilizes geometric reasoning to create a global precedence graph of feature relationships. This precedence graph represents significant geometric relationships between features and is used to determine subsequent process sequencing. The Process Planning module [20] then determines the process plan for producing the required features and advises the designer of manufacturing problems in machining of the part. Once the process planning is completed, the designer can choose to machine the part using the Cell Controller module, graphically simulate the machining of the part using the NC Simulator module or edit the part further based on the machining information now available.

The QTC system currently executes on a Silicon Graphics Inc. (SGI) workstation utilizing the Graphics Library functions [37]. However, the QTC software is written in C, C++ and common LISP and can be ported to other platforms that support the OPENGL graphics library. An SGI graphics emulator has been developed for the Apple Macintosh platform and the QTC Design module ported to the Macintosh platform as a demonstration of system portability.

FEATURE-BASED DESIGN

In general terms, a feature is a representation of shape with geometry and attributes that are recognizable and meaningful to humans while providing an information topology that can be understood by analysis routines. In the case of the QTC system, features consists of prismatic entities such as rectangular stock feature, holes, slots, pockets, counter-bore, NURB surfaces, etc. and cylindrical entities such as cylindrical stock feature, round-groove, rectangular groove, V-groove, and axisymmetric taper. These features have both a geometric meaning from a concept design perspective and a process meaning from a manufacturing perspective.

The Feature-based Design Graphical User Interface, shown in figure 2, is a multi-window, highly graphical, interactive environment that allows the designer to conceive the form of a component part and easily use features to define the part while viewing both a wireframe representation and a shaded image or hidden line representation. This environment closely resembles the conventional CAD environment familiar to most concept designers.

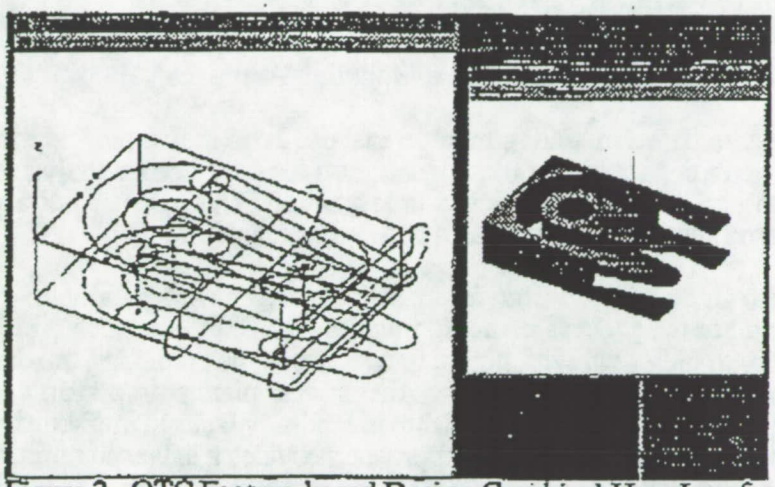


Figure 2. QTC Feature-based Design Graphical User Interface

Each feature is easily positioned, oriented and sized on the work face of the stock by utilizing graphical icons called "handles". Features are positioned with respect to other features by specifying position vectors between handles. Both position vectors and dimensional vectors contain tolerances. The Design Module allows feature to be placed on any of the faces of the stock feature. The resulting part model is a high-level representation of the design that contains all feature information including position and dimensional tolerances [1]. Figure 3 illustrates the QTC Feature-based design approach. The unique ability of the QTC system to represent a three dimensional geometric model and a three dimensional tolerance model is essential to conducting generative process planning.

The QTC Design environment is implemented as a modular, object-oriented approach that allows new features to be added by defining the graphical representation, position and dimensional attributes (handles) and process planning rules. This modular approach allows the QTC system to be expanded over time to address a wider range on manufacturing processes.

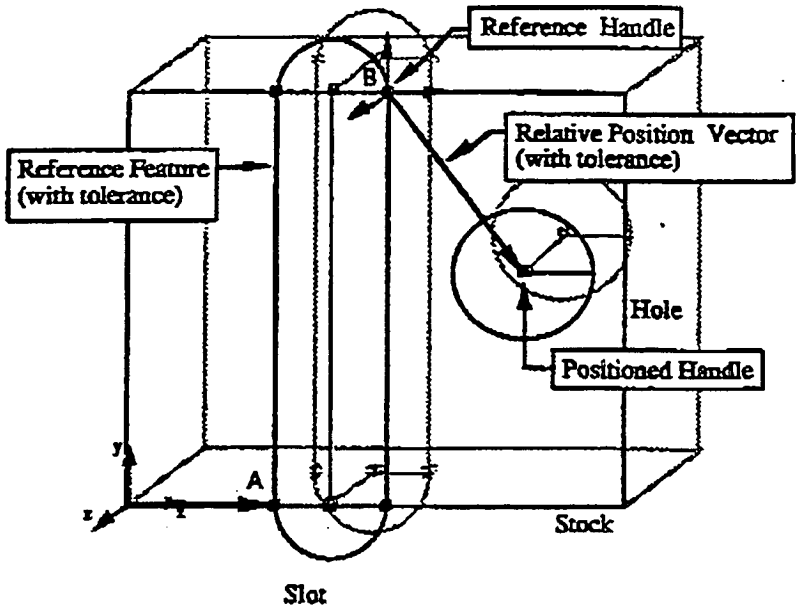


Figure 3. Feature-based Modeling Approach

The QTC feature-based modeling module has also been utilized in support of several engineering activities including mesh generation for engineering analysis [40], surface modeling [25], assembly [6], non-contact inspection [7], geometry portability between geometry based systems, and mill-turn and 3-axis machining [17].

Approaches to automating "design for manufacturing" have typically favored either the design function or the process planning function [1]. Such approaches either restrict the design activity to only deal with planning operations thus reducing the designer's flexibility, or on the other extreme, force the process planner to perform complex and lengthy deductions from a design model that generally lacks manufacturing content. The QTC system feature-based modeling approach provides a balanced approach that allows the designer sufficient flexibility to represent geometrically complex parts while providing a model that contains sufficient manufacturing content to conduct automated process planning activities.

GEOMETRIC REASONING

In recent years there have been several research and commercial efforts to automate process planning for manufacturing [4, 11, 13]. One of the major technical obstacles identified by these efforts is the large number of combinations and complexity that occur due to the interaction of features. A typical approach to resolve this interaction problem has been to expand the number of features to account for each possible combination. This approach results in a very large set of features and overly complicated planning algorithms. In some efforts the developers have attempted to conduct "feature extraction" from a conventional geometric model and discovered that this approach becomes extremely difficult as parts become geometrically complex.

The QTC system has overcome this technical obstacle by developing a CAD Interface module that utilizes geometric reasoning to resolve the relationships between features. For example, Figure 4. illustrates the interaction of a single stock feature with a single slot feature and the resulting combinations such as a through-slot, a blind-slot, pocket, or step.

By utilizing the CAD Interface to resolve these relationships, only two features are required to represent this relationship. This approach greatly simplifies both the design module and process planning module algorithms and greatly reduces the number of feanires required to represent geometrically complex parts. The CAD Interface module also resolves precedence relationships where one feature must be removed before removing the second. Figure 4 also illustrates a case where the hole is located on the bottom face of the slot. The CAD Interface resolves this case and provides a precedence graph to the process planning module for proper sequencing of machining operations.

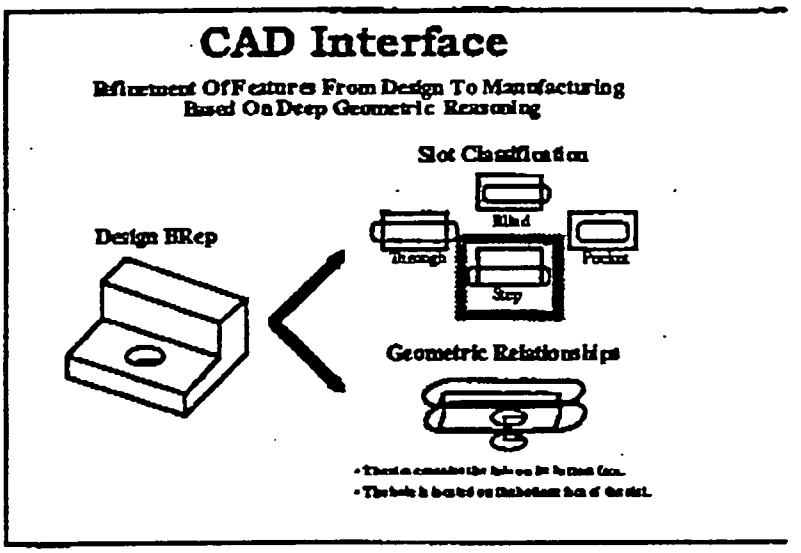


Figure 4. CAD Interface, Refinement of Features

GENERATIVE PROCESS PLANNING

The function of the Generative Process Planning module, shown in Figure 5, is to generate a process plan for the part consisting of process selection [6, 7, 8], sequencing of operations, set-up planning, fixture planning [19], tools selection, cutter path generation, and estimation of machining time. The planning module utilizes several user definable databases including a tool database, fixturing database, machinability database and machine configuration. These databases allow the QTC system to be customized for a specific machine shop. The planning module outputs a planning document that contains instructions for part set-up, fixturing and machining operations and a cutter location data file. Cutter location data is stored in APT-like format for subsequent translation to NC machine format. Multiple part setups are generated depending on the orientation of the part required for machining. The Process planning and cutter location data are stored within the QTC system as text files thus allowing easy interface to an Material Requirement Planning (MRP) system where these files may be used in shop loading applications.

Machining problems detected by the planning module are identified to the designer thus providing manufacturing feedback to the designer in a timely manner. Typical problems encountered include such items as non-standard hole sizes that would require special size tooling or dimensional tolerances too tight for a machining process. This timely identification of manufacturing issues allows the designer to consider alternative design

MACHINE CELL CONTROLLER

The Machine Cell Controller module provides a direct network link between the QTC system and numerical control (NC) machines and provides setup information to the NC machine operators. Information provided to the machine operator includes raw stock size, part offset position (for machines not equipped with edge finding probes), and machine tools required for the setup. Prior to downloading of the cutter location data to the NC machine, a translation routine (i.e. post processor) is executed to convert the QTC cutter location data to the part program format required by the specific machine. Post processors, written in C language routines, have been developed for Cincinnati Milacron T-10 controllers and Emco Mier T-10 controllers [37].

A Machine Cell Controller Graphical User Interface, shown in Figure 7., shows the information discussed above and illustrates how the raw stock should be fixtured for machining. Multiple setups, previously generated by the planning module, are displayed by a simple "PREVIOUS or NEXT" selection.

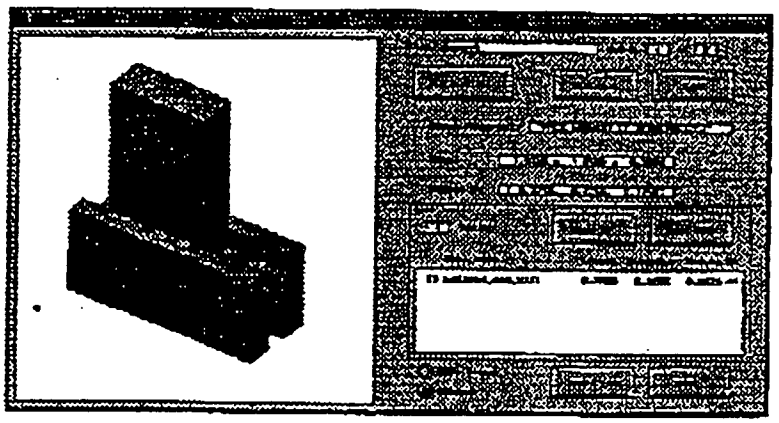


Figure 7. Machine Cell Controller Graphical User Interface

CONCLUSIONS

The QTC system is an evolving effort by the Purdue ERC under the sponsorship of the National Science Foundation, Industry partners of the Purdue Engineering Research Center, and the Army Missile Command. This alliance of academia, government and industry partners is essential to insure relevant research at an academic level, development of practical engineering tools from that research, and formulation of new research areas made evident by practical application of the engineering tools. The QTC system testbeds are currently implemented at the Purdue ERC, at the Army Missile Command in Huntsville Alabama, Loral Vought Systems in Dallas Texas, and the Automation and Robotics Institute in Arlington Texas. There are a number of research activities currently underway at the Purdue ERC that will contribute to the future advancements of the QTC system.

The QTC system represents a leading edge technical accomplishment made possible by the feature-based modeling approach. The technical approach used in the QTC system is being disseminated by several means including use of test-bed systems, location of experienced graduate students from the ERC into commercial CAD software development companies and through publication of technical results. Several of the members of the development team are now working for commercial CAD vendors where they are utilizing the skills developed at the Purdue ERC to advance commercial CAD products to better address

manufacturing issues. Commercial CAD users are beginning to see several commercial products that utilize "geometric form features" to augment the design process. In the future we can expect the trend toward feature based modeling and an emphasis on manufacturing to continue.

REFERENCES

[1] Anderson, D. C., and Chang, T. C., "Geometric Reasoning in Feature-Based Design and Process Planning," Computers and Graphics, Vol. 14, No. 2, 1990.

[2] ANSI Dimensioning and Tolerancing, ANSI Standard Y14.5, ASME, 1983.

[3] ANSI/EIA, Axis and Motion Nomenclature for Numerically Controlled Machines, Electronic Industries Association, EIA-267-C, 1990.

[4] Brown, D. R., Cutkosky, M. R. and Tenenbaum, J. M., "Next-Cut: A Computational Framework for Concurrent Engineering," Proc. 2nd International Symposium on Concurrent Engineering Research, Morgantown, WV, 1990.

[5] Chamberlain, M. A., Joneja, A. and Chang, T. C., "Protrusion-Features Handling in Design and Manufacturing Planning," Computer-Aided Design, Vol. 25, No. 1, pp. 19-28, 1993.

[6] Chang, T.C., *Expert Process Planning Systems*, Addison Wesley, Reading Massachusetts, 1990.

[7] Chang, T.C. Anderson, D.C., Mitchell, O.R., "QTC - An Integrated Design/Manufacturing/Inspection System for Prismatic Parts", Proceedings of the ASME Computers in Engineering Conference 1988, SF, Cal., 1,pp 417-26, 1988.

[8] Chang, T.C., Wysk, R.A., and Wang H.P., Computer Integrated Manufacturing, Printice Hall, 1991.

[9] Choi, B. K., Barash, M. M. and Anderson, D. C., "Automatic Recognition of Machined Surfaces from a 3D Solid Model," Computer-Aided Design, Vol. 16, No. 2, pp. 81-86, 1984.

[10] Cutkosky, M. R., "Features in Process-Based Design," Proc. Computers in Engineering, ASME, San Francisco, CA, Vol. 1, pp. 557-562, 1988.

[11] Cutkosky, M. R., Tenenbaum, J. M. and Brown, D. R., "Working with Multiple Representations in a Concurrent Design System," Journal of Mechanical Design, Vol. 114, No. 3, pp. 515-524, 1992.

[12] Desai, V. S. and Pande, S. S., "GFM-An Interactive Feature Modeler for CAPP of Rotational Components," Computer-Aided Engineering Journal, Vol. 8, No. 5, pp. 217-220, 1991.

[13] Dixon, J. R., "Designing with Features: Building Manufacturing Knowledge into More Intelligent CAD Systems," Proc. Manufacturing International-88, Atlanta, GA, Vol. 1, pp. 51-57, 1988.

[14] Fields, M. C., Computational Problems in Feature Research, Ph.D. Thesis, Purdue University, 1993.

[15] Finger, S. and Dixon, J. R., "A Review of Research in Mechanical Engineering Design. Part I: Descriptive, Prescriptive, and Computer-Based Models of Design Processes," Research in Engineering Design, Vol. 1, No. 1, pp. 51-68, 1989.

[16] Finger, S. and Dixon, J. R., "A Review of Research in Mechanical Engineering Design. Part II: Representations, Analysis, and Design for the Life Cycle," Research in Engineering Design, Vol. 1, No. 2, pp. 121-137, 1989.

[17] Henderson, M. R., Extraction of Feature Information from Three Dimensional CAD Data, Ph.D. Thesis, Purdue University, 1984.

[18] Johnson, R. H., Dimensioning and Tolerancing Final Report, CAM-I, Inc., R-84-GM-02.2, 1985.

[19] Joneja, A. and Chang, T.C., "A Generalized Framework For Automatic Planning of Fixture Configurations", *Advances in Manufacturing Systems Engineering*, Winter Annual Meeting of the ASME, San Francisco 1989.

[20] Joneja, A. and Chang, T.C., "Search Anatomy in Feature-based Automated Process Planning", Journal of Design and Manufacturing, Vol 1, 7-15, 1991.

[20] Kanumury, M. and Chang, T.C., "Process Planning in a Completely Automated Manufacturing Environmnet", Journal of Manufacturing Systems, pp 67-78, 1991.

[21] Kanumury, M., Shah, J., and Chang, T.C., "An Automatic Process Planning System for a Quick Turnaround Cell - An Integrated CAD and CAM System" *Proceedings of the USA-JAPAN Symposium on Flexible Automation*, Vol 1, ASME July 1988.

[22] Kim, J. Y., O'Grady, P. and Young, R. E., A Feature Taxonomy for Rotational Parts, Department of Industrial Engineering, North Carolina State University, Technical Report, 91-01, 1991.

[23] Libardi, E. C., Dixon, J. R. and Simmons, M. K., "Designing with Features: Design and Analysis of Extrusions as an Example," Proc. Spring National Design and Conference, ASME, Chicago, IL, 1986.

[24] Luby, S. C., Dixon, J. R. and Simmons, M. K., "Designing with Features: Creating and Using a Features Data Base for Evaluation of Manufacturability of Castings," Proc. International Computers in Engineering Conference, ASME, Chicago, IL, pp. 285-292, 1986.

[25] Merrifield, M. C., A Feature-Based Design Environment for Free-Form Surfaces, M. S. Thesis, Purdue University, W. Lafayette, IN, 1991.

[26] Mouleeswaran, C. B., and Fischer, H. G., "A Knowledge Based Environment for Process Planning," Proc. Applications of Artificial Intelligence in Engineering Problems, Southampton, Engl., Vol. 2, pp. 1013-1027, 1986.

[27] Mullins, S. H., Automatic Tolerance Synthesis and Analysis Based on Assembly Constraints, M. S. Thesis, Purdue University, W. Lafayette, IN, 1991.

[28] Mullins S.H. and Anderson D.C., "Feature-Based Tolerance Representation for Design and Analysis," Journal of Design and Manufacturing, 1(2), pp. 107-118, December 1991.

[29] Pande, S. S., and Palsule, N. H., "GCAPPS - A Computer Assisted Generative Process Planning System for Turned Components," Computer-Aided Engineering Journal, Vol. 5, No. 4, 1988.

[30] Pratt, M. J., :Solid Modeling and the Interface Between Design and Manufacture," IEEE Computer Graphics and Applications, Vol. 4, No. 7, 1984.

[31] Requicha, A. and Chan, S., "Representation of Geometric Features, Tolerances, and Attributes in Solid Modelers Based on Constructive Geometry," IEEE Journal of Robotics and Automation, Vol. RA-2, No. 3, pp. 156-165, 1986.

[32] Salomons, O. W., van Houten, F. J. A. M. and Kals, H J. J., "Review of Research in Feature-Based Design," Journal of Manufacturing Systems, Vol. 12, No. 2, pp. 113-132, 1993.

[33] Salvendy, G., ed., "Handbook of Industrial Engineering," 2nd Ed., John Wiley and Sons, 1992.

[34] Shah, J. J., Rogers, M. T., et al., "The A.S.U. Features Testbed: An Overview," Proc. Computers in Engineering 1990, ASME, Boston, Massachusetts, Vol. !, pp. 233-241, 1990.

[35] Shah, J. J., "Conceptual Development of Form Features and Feature Modelers," Research in Engineering Design, Vol. 2, No. 2, pp. 93-108, 1991.

[37] The Quick Turnaround Cell User Guide, Purdue University Engineering Research Center for Intelligent Manufacturing Systems, 1986-1993.

[38] Turner, G., "An Object-Oriented Approach to Interactive, Feature-Based Design with Tolerancing," M.S. Thesis, Purdue University, W. Lafayette, IN, 1988.

[39] Turner, G.and Anderson, D.C., "An Object-Oriented Approach to Interactive, Feature-Based Design for Quick Turnaround Manufacturing", Proceedings of the 1988 ASME Computers in Engineering Conference, San Francisco, CA, July 1988, pp.551-556.

[40] Unruh, V. and Anderson, D.C., "Feature-Based Modeling for Automatic Mesh Generation" Engineering and Computers, 8, 1-12, 1992.

[41] Vaghul, M., Dixon, J. R., Zinsmeister, G. E. and Simmons, M. K., "Expert Systems in a CAD Environment: Injection Molding Part Design as an Example," Proc. Computers in Engineering, ASME, Boston, MA, Vol. 2, 1985.

[42] Wang, H. P. and Wysk, R. A., "Intelligent Reasoning for Process Planning," Computers in Industry, Vol. 8, No. 4, pp. 293-309, 1987.

[43] Wilson, P. R. and Pratt, M. J., "A Taxonomy of Form Features for Solid Modeling," Geometric Modeling for CAD Applications, Elsevier Science Publishers B. V., North Holland, pp. 125-135, 1988.

[44] Manufacturing Systems Strategic Plan, Report of The Manufacturing System Committee, Department of Defense, #N0014-92-C-7009, 1993

concepts that would eliminate machining problems early in product development rather than waiting until volume production is planned.

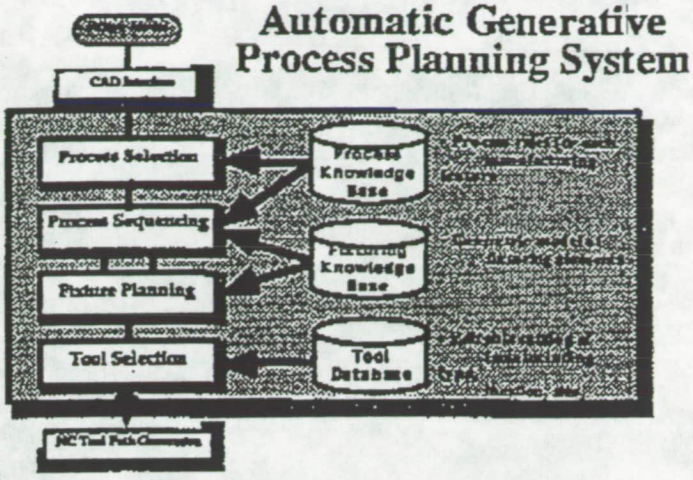


Figure 5. Automatic Generative Process Planning

MACHINING SIMULATION

The machining simulation module, also referred to as Numerical Control Verification (NCV), provides for a graphical simulation of the machining operations for the purpose of cutter path verification, tool and fixture collision avoidance and detection of tool material gouging [25]. The NCV utilizes a Graphic User Interface as shown in figure 6. to display and simulate the machining operations.

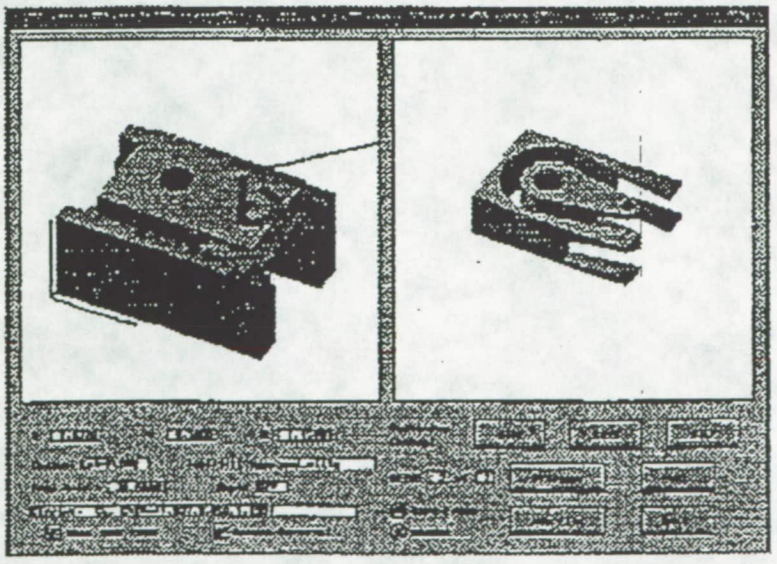


Figure 6. Machining Simulation Graphic User Interface

FAULT TREE BASED DIAGNOSIS WITH OPTIMAL TEST SEQUENCING FOR FIELD SERVICE ENGINEERS

David L. Iverson
NASA Ames Research Center
Moffett Field, CA 94035-1000

Dr. Laurence L. George
Problem Solving Tools
Livermore, CA 94550-5852

Dr. F. A. Patterson-Hine
NASA Ames Research Center
Moffett Field, CA 94035-1000

ABSTRACT

When field service engineers service equipment, they want to diagnose and repair failures quickly and cost effectively. Symptoms exhibited by failed equipment frequently suggest several possible causes. This can lead an engineer to perform unnecessary tests before finding the actual failure. Test sequences presented in service manuals can help with this problem, but are limited to an inflexible routine based on general analysis. We have developed the Fault Tree Diagnosis and Optimal Test Sequence (FTDOTS) software system to find possible causes and recommend an optimal test sequence *conditional on observed symptoms.* FTDOTS uses a fault tree as a diagnostic knowledge base to find sets of possible failures that explain symptoms. The program then uses an operations research technique known as Mitten's Rule to rank the hypothesized failure sets based on how likely they are and how much it would cost, in time or money, to perform tests to confirm each hypothesis. The field service engineer is presented an optimal test sequence that minimizes the time or cost required to find and repair the failures. This paper describes the FTDOTS system and presents an example diagnosis session.

INTRODUCTION

Failed equipment often exhibits symptoms that could be caused by many different failures. Customers and field service engineers want to minimize the time or cost to find a failure and make the repair. Labor and logistics time determines repair time for a given failure, unless the repair includes preventative or opportunistic maintenance [1]. Diagnostic time, the time it takes to determine which repair to make, depends on the diagnostic sequence chosen to test the suspect components. In the words of a veteran field service engineer, diagnose first what gives "the most bang for the buck." In other words, check the components that most likely caused the failure and take the least time to diagnose first. For instance, the first test performed by many field service engineers is to check if the failed equipment has power. Checking if the machine is plugged in or has blown a fuse is quick, cheap, and sometimes the only action necessary.

There are many factors that can lead to unnecessary expense for equipment service. Field service engineers form their own opinions based on experience that may not accurately represent the population of machines they maintain. This often causes them to go through inefficient diagnostic sequences before they find the true failures. Customer complaints frequently suggest many possible causes that can mislead the engineer. Unnecessary diagnostic tests requiring expensive equipment are sometimes performed. Diagnosis occasionally consists solely of replacement and test, also known as shotgun repair. Organizing and using the experience gained with a population of machines can lead to cost and time savings by avoiding these pitfalls and providing an efficient, symptom based diagnostic test sequence.

The Fault Tree Diagnosis and Optimal Test Sequence (FTDOTS) software system was developed to help field service engineers use data available on a machine population to minimize the time or cost required to isolate and repair failures. FTDOTS uses an automated diagnosis system that reasons from a fault tree to hypothesize possible causes for exhibited symptoms. Tests to determine the validity of these hypotheses are then sorted into an optimal test sequence using a technique known as Mitten's Rule [2]. If the engineer tests the suspected components in the order suggested by FTDOTS, the actual failures will be revealed with minimal test cost. Depending on the parameters used with FTDOTS, minimal test cost can be measured in terms of expense, time, or other values.

To develop an FTDOTS system for a piece of equipment a fault tree model of the equipment must be built. Fault trees are common reliability models that are built for many devices during system design and development. A fault tree contains failure events or symptoms that can be observed in the modeled system and shows how system component failures can propagate to cause these higher level observable symptoms. Failure probabilities and test costs can usually be determined for each low level system component modeled in the fault tree. Once the fault tree, failure probabilities, and test costs are obtained for a system, a diagnostic knowledge base for use with FTDOTS can be easily constructed. Building a diagnostic knowledge base using a fault tree usually requires much less effort than building a traditional expert system knowledge base.

FTDOTS and the knowledge base can be loaded on a portable computer and carried by the field service engineer to any equipment site. The engineer can enter observed symptoms into the computer and FTDOTS will suggest possible failures to investigate and provide an optimal test sequence to specifically address those symptoms. FTDOTS could also be installed in the equipment as an on board diagnosis system or used at a phone-in customer support center to recommend customer actions or provide initial information for a field service call.

FAULT TREE MODELS

Fault tree analysis can be described as an analytical technique where an undesired state of a system is specified and the system is analyzed, in the context of its environment, to find all tenable ways this undesired state could occur. The resulting information can be represented as a tree structure with the original undesired event (failure) at the root, the possible causes of that event as the root's children (inputs), the causes of those events as their children, and so on. Each lower level of the fault tree represents a more specific failure. Analysis stops for each branch of the tree when the failure event described by the leaf (bottom) node is fine-grained enough to satisfy the needs of the analysis. These leaf nodes are called basic events and usually correspond to basic system component failures or replaceable system modules.

All non-leaf nodes of a basic fault tree can be thought of as logic gates representing a logical AND or OR. An AND gate signifies that all the child failure events of the node must occur before the event represented by the (parent) node will occur. An OR gate means if at least one of the child failure events occurs, the parent event will occur. Sometimes a NOP gate is used when a node has only one child. A NOP gate is just like an OR gate with only one child event. See Fig. 1 for illustrations of fault tree AND and OR gates. Fig. 2 shows a complete fault tree.

Other information can be associated with each fault tree node. For instance, a node might also contain the probability of occurrence of its associated failure event, or the time interval between the occurrence of a child event and the occurrence of its parent event. When additional information is included in a fault tree it is called an augmented fault tree [4]. Fault trees used by the FTDOTS system can include failure propagation time intervals for each failure event. The failure propagation time interval of an event under an OR gate is an estimate of how much time will elapse from the moment that failure event occurs until its parent failure event occurs. In the case of a child event under an AND gate, the time interval measures the time between the

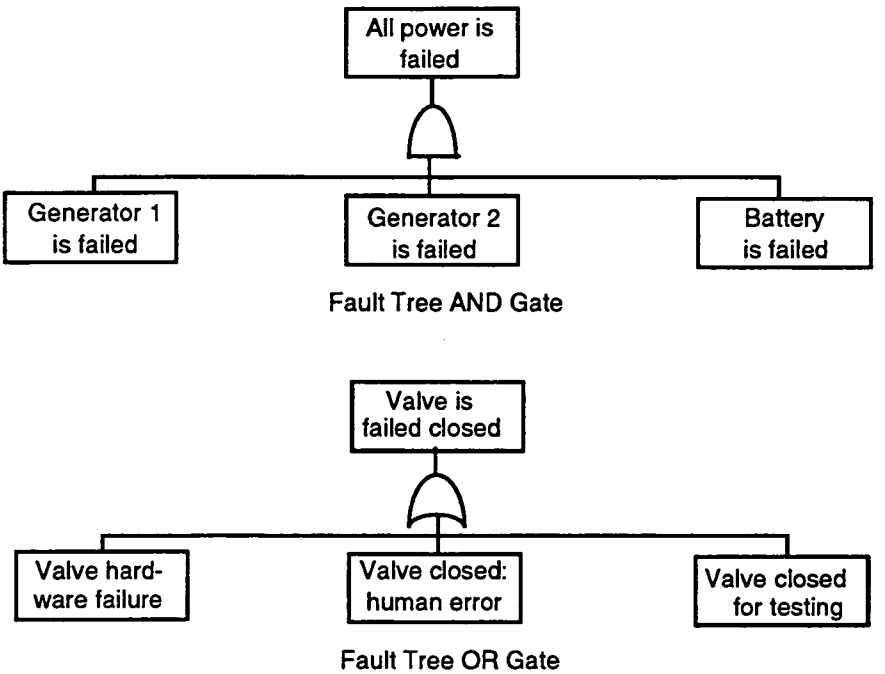


Figure 1: Examples of Fault Tree Gates [3]

moment when all the child events have occurred and the occurrence of the parent event. The FTDOTS system can use time intervals to obtain more accurate diagnoses, as explained in the next section. However, failure propagation time interval information, which is sometimes difficult to obtain, is not required by FTDOTS.

FTDOTS DIAGNOSIS ALGORITHM

The FTDOTS system uses the Fault Tree Diagnosis System (FTDS) diagnosis algorithm [5] to generate failure hypothesis sets. The user provides FTDOTS with information about the equipment being diagnosed in the form of normal and abnormal indicators. A *normal indicator* indicates that a given failure event has not occurred. For example, if we can see that the system has electric power we would set "All Power Failed" as a normal indicator since that failure event has obviously not occurred. An *abnormal indicator* indicates that a failure event has occurred or a failure symptom has been observed. Each possible indicator corresponds to a node in the fault tree. If it is known that the failure event represented by a fault tree node has not occurred, that event is placed in the normal indicators set. If it is known that a failure event has occurred, that event is placed in the abnormal indicators set. The most effective diagnosis process is obtained when abnormal indicator nodes are as low in the fault tree as possible (near the basic failure leaf nodes) and normal indicator nodes are as close to the top of the tree as possible.

The diagnoses produced by the FTDOTS system are sets of basic failure events that causally explain the occurrence of the abnormal indicators while maintaining consistency with the normal indicators. The diagnosis process is initiated by specifying the observed normal and abnormal indicators. If temporal reasoning is used, the estimated time of occurrence of each abnormal indicator failure and the time that each normal indicator was last confirmed is also provided.

The diagnosis begins by determining which failure events in the fault tree could not have occurred. This is done by propagating information from the normal indicator nodes. If a normal indicator node appears as a child of an AND gate, we know that the AND gate parent node

failure could not have occurred since that would imply all of its children have failed. If a normal indicator node appears as a child of an OR gate and all the children of that OR gate are normal indicators, then the parent event of the OR gate must also be a normal indicator. Each time a new normal indicator is added to the set, FTDOTS checks to see if its parent node should also be added. This is known as forward chaining. Similarly, if the parent event of an OR gate is a normal indicator, we know that none of its child node failures have occurred so they are added to the normal indicators set. Whenever a node is added to the normal indicators set, FTDOTS checks if its children should also be included as normal indicators. This technique is called backward chaining.

If temporal reasoning is used, associated times are propagated with the normal indicators to indicate when the designated failure event was known to be normal (i.e., a time at which it was known that the failure had not occurred). These times are obtained from the failure propagation times included in the augmented fault tree [4, 5]. The use of temporal reasoning helps to obtain a more accurate diagnosis. Diagnoses performed without temporal reasoning may miss some hypotheses that would be found with the use of temporal reasoning. If temporal reasoning is not used and the first diagnosis does not reveal the true failure, the missed hypotheses can be found by performing a second, less efficient diagnosis without the normal indicator information.

After the normal indicators set is complete, the FTDOTS system looks for dependencies between the abnormal indicator events. It does this by propagating the failures through the fault tree to see if any of the abnormal indicator failures could cascade and cause any of the other indicated failures. If so, only the most basic abnormal indicators are retained in the abnormal indicators set. Any higher level abnormal indicators that are causally connected to indicated failures beneath them in the tree are assumed to be caused by those lower level failures. These higher level failures are removed from the abnormal indicators set so they are not directly considered in the diagnosis process. The failure nodes remaining in the abnormal indicators set after this phase of processing are used as starting points for the next diagnosis step. Temporal reasoning can also be used in this portion of the diagnosis. If a higher level failure was observed before the occurrence time of a lower failure plus the propagation time between the two failures, the failures are considered independent and will both be retained in the abnormal indicators set for separate diagnoses.

The FTDOTS system uses backward chaining reasoning to find sets of basic failure events that causally explain the starting point failures remaining in the abnormal indicators set. Essentially, the system looks at the child nodes of each abnormal indicator node to find possible causes for the abnormal indicator. It then looks at the child nodes of those nodes to find possible causes for those failures. This continues until basic event nodes are reached along each of the branches followed by the system. These basic event failures are returned as possible causes for the abnormal indicator failure. If an AND gate is encountered during backward chaining, the FTDOTS system finds possible basic causes for each of the child events of the AND gate and combines those causes into sets that could cause all of the AND gate child events and thus the AND gate parent failure event. In this case a failure hypothesis may contain more than one basic event. If temporal reasoning is used, estimated failure times for each basic failure event are included in the hypothesis sets.

When a hypothesis set of basic events has been found for a given abnormal indicator, it is checked for consistency with the normal indicator information. This is done by propagating the hypothesized failures upward through the fault tree using forward chaining. If the failure propagation encounters a normal indicator node, we know that that hypothesis is invalid. The normal indicator provides evidence that the hypothesized failure did not occur. Any invalid hypotheses are discarded. This process will perform temporal consistency checks if temporal reasoning is used. A hypothesis will only be discarded if it predicts a failure of the encountered

normal indicator node at a time before the function represented by the node was observed or predicted to be working properly.

When all hypothesis sets have been found for each abnormal indicator in the starting points set, the FTDOTS system combines them into hypothesis sets that could each causally explain all of the original abnormal indicators. This is accomplished by forming the cross product of the hypothesis sets for each of the starting point nodes.

Given a group of failure hypotheses, field service engineers without FTDOTS will choose a test sequence based on experience and service manuals. Although good service manuals attempt to portray the minimum time or cost diagnostic sequence for the most probable circumstances, they are static documents and do not take into account information that the customer, company, or field service engineer may have regarding actual failure rates and diagnosis costs. The manuals also may not cover all the various combinations of failures that could be encountered. The FTDOTS system will have up to date information and will use it to construct an optimal sequence of tests to determine which of the failure hypotheses is correct. Mitten's Rule is used to construct a testing sequence that minimizes the time or cost of finding and repairing failed components.

MITTEN'S RULE

L. G. Mitten developed an algorithm that can be used to find an optimal test sequence based on failure probabilities and testing costs [2]. The FTDOTS system uses Mitten's Rule to provide the field service engineer with a recommended sequence of tests that will find the failed system components in the least amount of time or at the lowest cost.

To use Mitten's Rule, two values are required for each basic failure event included in the FTDOTS fault tree input file. The first is the probability that the failure represented by that basic fault tree node will occur. This value is often available from the manufacturer of the component or subsystem modeled by that basic event node. Field failure probabilities (the observed failure probabilities of machines operating in the field) should be used if available. The second required value is the cost, in time or money, of testing whether that failure has actually occurred. This value will depend on how difficult it is to access the component for testing, the testing equipment required, the length or cost of performing the test, and other factors. Some of this information may be available from the component manufacturer and some may be dependent on equipment design.

Assume that the probability of occurrence of a given basic failure event is R_i, and the cost of testing the component modeled by that failure event is C_i. Mitten's Rule states that the least cost testing sequence can be obtained with the following procedure:

1. For each test i, compute the ratio C_i / R_i;
2. Run the test with the smallest value for the above ratio first, the one with the second smallest ratio second, ..., and the test with the largest ratio last.

The FTDOTS system calculates this ratio for each hypothesis set generated in the initial diagnosis phase. FTDOTS input files provide values for C_i and R_i for each basic failure event in the fault tree. For those hypothesis sets containing only one failure event, the calculation is a straightforward division. If a hypothesis set contains multiple failure events, FTDOTS assumes that the failures are independent and derives values for C and R. R is calculated by finding the product of the failure probabilities of every event in the set. This gives the probability of all of these failures occurring. C is calculated by summing the testing cost of each failure event. This

will give an upper bound on the testing cost for that hypothesis set. If all hypothesis sets contain only one event, which is often the case, the current FTDOTS system provides a least cost test sequencing. If some hypothesis sets contain multiple events, the current FTDOTS system may not provide a least cost test sequence due to assumptions made when dealing with multiple failure hypotheses, but the recommendation will be close to the least cost sequence. Future enhancements of the FTDOTS system may include revising the test sequencing routine to account for tests performed earlier in the sequence and cost dependencies between different tests. This will provide more accurate results for multiple event hypotheses.

Using Failure Rates with Mitten's Rule

Component reliability data is often presented in failures per hour or failures per million hours. Since Mitten's Rule uses failure probability (i.e., component unreliability) for its calculations, these failure rate values must be converted to failure probabilities. This is usually accomplished by assuming an exponential failure distribution function for the components. Other probability distribution functions can be used. The choice of failure probability distribution will depend on the equipment and how accurate a result is desired.

If the failure rate for component i is λ_i failures per hour and an exponential distribution function is used, then the failure probability, R_i, for that component can be calculated with

$$R_i(t) = 1 - e^{-\lambda_i t} \quad (1)$$

where t is the number of hours the component has been in operation [3]. The exponential distribution can be approximated by its first order term and a reasonable R_i value can be obtained with the equation

$$R_i(t) \approx \lambda_i t. \quad (2)$$

If FTDOTS is used with failure rate data and equation (1) or (2), the field service engineer would provide the number of hours the equipment has been in operation when the failure occurred. This information would be used to calculate the current failure probabilities.

DIAGNOSIS EXAMPLE

Figure 2 shows a fault tree for a piece of medical lab equipment called a hemoanalyzer. This machine provides cell counts and other blood sample measurements. Each node in the fault tree represents the failure of the component or system named in the node. The FTDOTS system can use this fault tree along with the data in Table 1 to diagnose problems with the hemoanalyzer and recommend a sequence of tests to isolate the failed component. Table 1 presents the number of hours required to test each hemoanalyzer component to see if it has failed and the number of expected failures per million hours of operation for each component. For this example we'll assume FTDOTS is not using temporal reasoning so no failure propagation time information is required.

A field service engineer called in to repair an ailing hemoanalyzer would arrive at the site, unpack a portable computer running FTDOTS, and load in the fault tree from Fig. 2 and the data from Table 1. She would then study the hemoanalyzer for failure symptoms. Noticing that the analyzer portion of the equipment was malfunctioning, she would enter 'Analyzer' into FTDOTS as an abnormal indicator. If she found that the hemoanalyzer had power and the communications

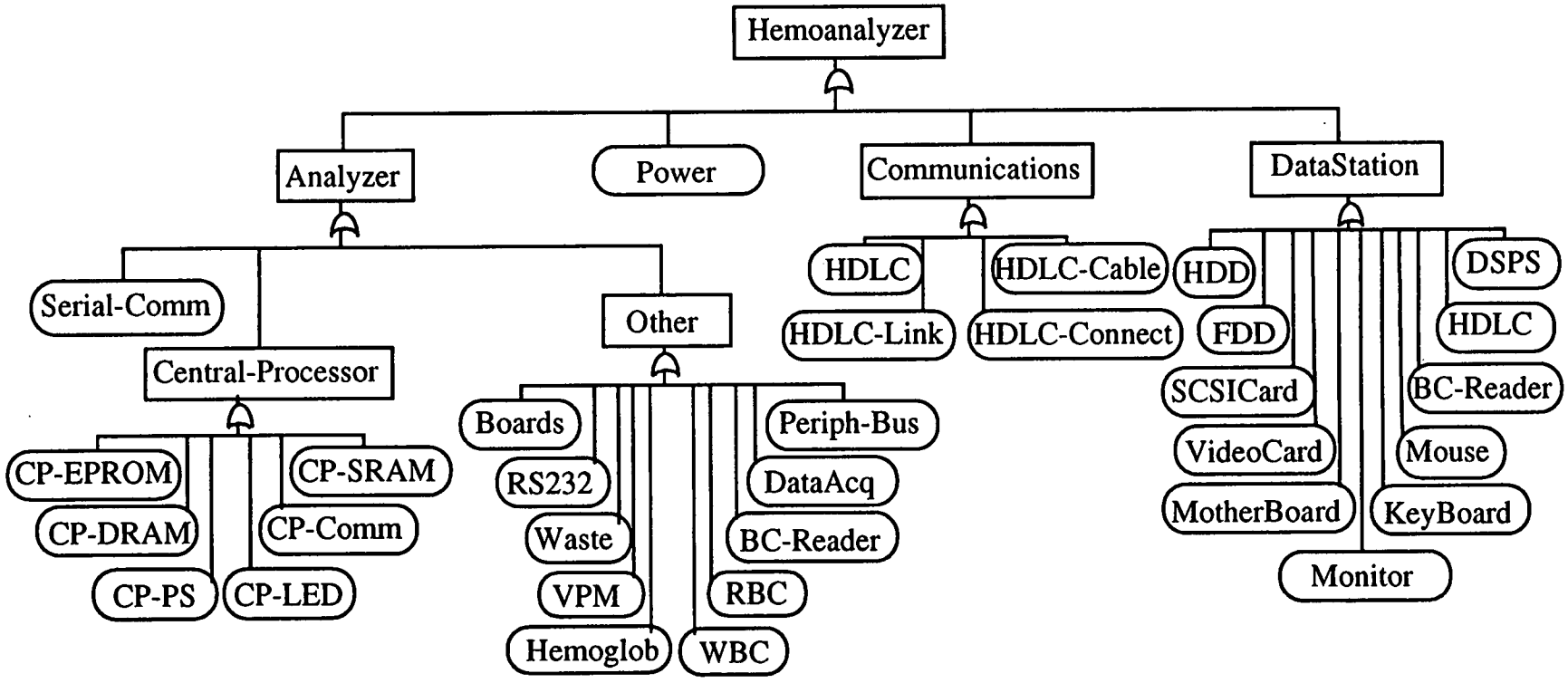


Figure 2: Hemoanalyzer Fault Tree

Component	Hours to Test	Failures / 10^6 hours
BC-Reader	0.50	32.9
Boards	2.00	1000.0
CP-Comm	0.50	416.7
CP-DRAM	0.50	416.7
CP-EPROM	0.50	416.7
CP-LED	0.10	0.1
CP-PS	0.10	1.4
CP-SRAM	0.50	416.7
DataAcq	0.10	2.8
DSPS	0.20	416.7
FDD	0.20	416.7
HDD	0.20	416.7
HDLC	0.10	7.0
HDLC-Cable	0.10	7.0
HDLC-Connect	0.10	7.0
HDLC-Link	0.10	7.0
Hemoglob	1.00	25.0
KeyBoard	0.20	416.7
Monitor	0.20	416.7
MotherBoard	0.20	416.7
Mouse	0.25	416.7
Periph-Bus	1.00	10.0
Power	0.10	100.0
RBC	0.10	246.4
RS232	0.50	416.7
SCSICard	0.50	416.7
Serial-Comm	0.50	25.0
VideoCard	0.50	416.7
VPM	1.00	300.0
WBC	1.00	246.4
Waste	2.00	200.0

Table 1: Hemoanalyzer Component Test Time and Failure Rates

system and data station were functioning properly, she would enter 'Power', 'Communications', and 'DataStation' into FTDOTS as normal indicators. Since the hemoanalyzer data contains failure rates instead of failure probabilities, the number of component operation hours would be entered or retrieved from a database to allow FTDOTS to find current component failure probabilities with equation (1) or (2). A future version of the FTDOTS system may have a cable to connect the portable computer to the failed equipment so relevant information can be automatically transmitted to FTDOTS.

FTDOTS begins the diagnosis by looking at the normal indicators and determining which failure events in the fault tree have not occurred (i.e., which components must be working). Since the power, communications, and data station systems all have normal indicators, FTDOTS concludes that the child events beneath the Communications and DataStation OR gates must also have normal indicators. This means that the HDLC, HDLC-Link, HDLC-Cable, HDLC-Connect, HDD, FDD, SCSICard, VideoCard, MotherBoard, Monitor, DSPS, BC-Reader, Mouse, and Keyboard are all working properly. Note that the BC-Reader node appears twice in the fault tree. Once under the DataStation node and once under the Other node which is a child of the Analyzer node. Both of these nodes represent the same failure.

Since there is only one abnormal indicator, FTDOTS does not look for abnormal indicator dependencies and proceeds directly to backward chaining reasoning to find basic failure event hypothesis sets. The reasoning begins at the Analyzer node and looks at the children of that

node. It finds that the Serial-Comm, Central-Processor, and Other systems may have failed. Serial-Comm is a basic failure event so a hypothesis set containing Serial-Comm is created. Continuing the backward chaining, FTDOTS looks at the children of the Central Processor and Other nodes. The result is 17 single event hypotheses including: Serial-Comm, CP-EPROM, CP-DRAM, CP-PS, CP-LED, CP-Comm, CP-SRAM, Boards, RS232, Waste, VPM, Hemoglob, Periph-Bus, DataAcq, BC-Reader, RBC, and WBC. However, the system notices that the BC-Reader is in the normal indicators set, so BC-Reader is removed from the hypothesis sets since we know that failure has not occurred.

FTDOTS has presented the field service engineer with 16 possible failures that could cause the exhibited symptoms. In the worst case, she must test 16 different components before finding the failure. Using Mitten's Rule to sequence those tests will help to minimize the cost and effort required to isolate the failed component. In most cases only a few tests are required if they are performed in the order suggested by Mitten's Rule. The failed component will probably be found early in the testing sequence making the remaining tests unnecessary.

To find the optimal test sequence FTDOTS calculates the Mitten's Rule value, C / R, for each hypothesis set. Since the hypothesis sets in this example only contain one failure event each, the Mitten's Rule value will be C_i / R_i for the component i in the hypothesis set. For example, we see from Table 1 that the Central Processor Static RAM (CP-SRAM) values are 0.50 hours to test and 416.7 failures per million hours. For this example we will assume all components have been in operation for 1000 hours and obtain the failure probability R using equation (2) to give $R \approx \lambda_i t = .4167$. The testing time C = 0.50 is taken straight from the table. This would give CP-SRAM a Mitten's Rule value of 0.50 / .4167 = 1.199. Calculating the Mitten's Rule value for the remaining 15 failures and sorting from lowest to highest we obtain this testing sequence: RBC, CP-Comm, CP-DRAM, CP-EPROM, CP-SRAM, RS232, Boards, VPM, WBC, Waste, Serial-Comm, DataAcq, Hemoglob, CP-PS, Periph-Bus, CP-LED. This recommends that the field service engineer begin by testing the Red Blood Cell Fluid Circuit (RBC). If the RBC is functioning properly she should continue down the list to test the Central Processor Communications function (CP-Comm), and so on until the failure is found. Once the failure is isolated, the offending component can be replaced or repaired and the equipment should be tested again. If another malfunction is noticed, FTDOTS can be used to help find and repair that problem as well. By following the FTDOTS recommendation instead of using an arbitrary test sequence, the field service engineer will spend significantly less time diagnosing and isolating failures.

CONCLUSIONS

Field service engineers must isolate component failures quickly and efficiently. For each set of symptoms a piece of equipment exhibits, there may be several possible component failures that could cause those symptoms. A field service engineer must find which of those components has failed by performing a series of tests to isolate the failure. A lack of information about the failure history of the equipment being repaired could make this an unnecessarily costly process. The Fault Tree Diagnosis with Optimal Test Sequencing (FTDOTS) system will determine the possible causes of a set of symptoms and recommend an optimal sequence of tests that will isolate the actual failure in the least amount of time or at the least cost.

To use FTDOTS, a fault tree model of the system must be built and probability of failure and the time (or cost) to test each component failure must be found. Fault tree models are sometimes built during the system design phase to check for system failure characteristics. If a model was not built during system design, a reliability engineer or someone else familiar with fault trees can build the model using design data and blueprints. Failure probabilities (or failures per hour) for each component can be obtained by keeping records of failures during testing and

field operations. Sometimes these values are available from component manufacturers. Time or cost for testing a component depends on system configuration and test requirements. These values can be estimated by system designers and service personnel.

Once a fault tree model is built and data is obtained for each component failure, the information can be formatted for FTDOTS use and loaded onto portable computers. When a field service engineer makes a call, she can enter observed symptoms into the portable computer and soon afterwards receive a list of possible failures and an optimal series of tests to perform. Provided the fault tree and input data are correct, the use of the FTDOTS system assures efficient, cost effective troubleshooting and repair.

REFERENCES

1. George, L. L. and Yat H. Lo, "An Opportunistic Look Ahead Replacement Policy," Annals of SOLE, Vol. 14, No. 4, Winter 1980.

2. Mitten, L.G., "An Analytic Solution to the Least Cost Testing Sequence Problem," *The Journal of Industrial Engineering*, pp 16-7, January-February, 1960.

3. Vesely, W.E., F.F. Goldberg, N.H. Roberts, and D.F. Haasl, Fault Tree Handbook, NUREG-0492, U.S. Nuclear Regulatory Commission, Washington, D.C., 1981.

4. Narayanan, N.H., and N. Viswanadham, "A Methodology for Knowledge Acquisition and Reasoning in Failure Analysis of Systems," *IEEE Transactions on Systems, Man, and Cybernetics,* pp. 274-288, March/April 1987.

5. Iverson, D.L. and F.A. Patterson-Hine, "Object-Oriented Fault-tree Models Applied to System Diagnosis," *Proc. of SPIE Applications of Artificial Intelligence VIII,* Orlando, FL, Vol. 1293, pp. 1013-1023, April 1990.

National Aeronautics and
Space Administration
Ames Research Center
Moffett Field, CA 94035-1000

eply to Attn of: IC:269-1

TO: NASA Headquarters
Attn: Code CU/Mr. Leonard Ault

FROM: David L. Iverson, Computer Engineer
Computational Sciences Division

SUBJECT: Technology 2004 paper submission

Please find enclosed two copies of the paper entitled "Fault Tree Based Diagnosis with Optimal Test Sequencing for Field Service Engineers," by D. L. Iverson, L. L. George, and F. A. Patterson-Hine, to be presented at the Technology 2004 Conference, November 8-10, 1994, in Washington, DC.

If you have any questions, I can be reached at 415-604-3115.

David L. Iverson

2 Enclosures:
1) Paper (2)
2) 3.5' Floppy disk (2)

MULTIPROCESSOR SCHEDULING OF DIGITAL SIGNAL PROCESSING ALGORITHMS

Robert L. Jones III
NASA Langley Research Center
Hampton, Virginia

ABSTRACT

A graph-theoretic design process and software tool are presented for selecting a multiprocessor scheduling solution for algorithms typically used to solve digital signal processing and control-law problems. An algorithm must be represented by a dataflow graph, which provides a graphical and mathematical model for determining performance bounds, scheduling constraints, and resource requirements. The software tool, which provides an analytical solution to the dataflow model, is shown to facilitate the application of the model in finding and optimizing a scheduling solution on multiple processors. Commercial potential for this technology is also discussed.

1. INTRODUCTION

For years, digital signal processing (DSP) systems have been used to realize digital filters, compute Fourier transforms, execute data compression algorithms, and a vast amount of other compute-intensive algorithms. Today, both government and industry are finding that computational requirements, especially in real-time systems, are becoming increasingly more challenging. As a result, many users are relying on multiprocessing solutions to meet the needs of these problems. To take advantage of multiprocessor architectures, novel methods are needed to facilitate the mapping of DSP applications onto multiple processors.

A graph-theoretic design process is defined in this paper for selecting a multiprocessor scheduling solution for DSP algorithms. DSP algorithms are characteristically executed repetitively for an infinite input data stream and can be described using dataflow graphs. The dataflow paradigm uses graph nodes to represent DSP algorithm functions and directed edges to describe the data dependencies between the functions. A dataflow graph models the algorithm as a partial ordering of functions, which protects the integrity of the data and ensures correct computational results. Describing an algorithm with a partially-ordered precedence relationship, as opposed to one that is totally (sequentially) ordered, exposes the parallelism inherent in the algorithm (constrained only by the data dependencies) in a way that facilitates the design of a multiprocessor solution [1]. The dataflow approach also provides a general computational model. That is, nodes can represent large-grain functions, e.g., the computation of a Fast Fourier Transform (FFT), or represent primitive functions, such as a single instruction. Dataflow graph analysis and design techniques are introduced and shown to effectively determine bounds on the sampling rate, processor requirements, admissible schedules, and speedup. The model assumes that the algorithm will be executed iteratively on a set of identical processors.

A software program developed to implement the dataflow analysis is shown to facilitate the application of the design process to a given problem, including performance optimization through the inclusion of artificial data dependencies. The tool predicts the steady-state behavior of cyclic (having recurrence loops) as well as acyclic dataflow graphs. This is important since most DSP algorithm realizations require past values of the input, output, and intermediate data values. Thus, there is a practical desire to accommodate cyclic graphs as well as graphs that require delayed samples corresponding to z^{-1} terms in traditional signal flow graph representations. It is shown that such delays on data paths can be modeled with initial tokens.

The modeling of an example DSP algorithm with a dataflow graph is discussed in Section 2. Section 3 provides a demonstration of the software tool implementation of the dataflow model for the example problem. Section 4 discusses some commercial uses of the software tool, and concluding remarks are provided in Section 5.

2. DATAFLOW GRAPH MODELING

A dataflow, graph-theoretic design process is presented in this section. The design process is suitable for selecting a multiprocessor scheduling solution for deterministic computational problems. The dataflow paradigm is very useful in exposing the parallelism inherent in algorithms. It provides a graphical and mathematical model which describes a partial ordering of algorithm functions based on data precedences. The dataflow graph is also a general model in that the functions, the basic unit of scheduling, may represent a single instruction or a complex function composed of many instructions. Ultimately, however, the scheduling granularity is governed by the hardware/software implementation and the associated synchronization/communication overhead.

DSP problems are well suited for dataflow analysis since they are typically described using signal flow graphs, the signals being *data* samples. One can easily translate signal flow graphs to dataflow graphs by locating and representing computations (addition and multiplication) with graph nodes and representing unit delays (inverse-z terms) with initial tokens. Analytical analysis of the dataflow graph is possible for many digital signal processing (DSP) and control law algorithms which are deterministic. For determinism, the time to execute functions are assumed constant from iteration to iteration when executed on a set of identical processors. Also, it is assumed that the dataflow graph is data independent. That is, any decisions present within the computational problem must be contained within the graph nodes rather than described at the graph level. As an example, DSP filter realizations characterized by the difference equation

$$y(n) = \sum_{k=1}^{N} a_k y(n-k) + \sum_{k=0}^{M} b_k x(n-k) \qquad (1)$$

can be represented directly by dataflow graphs. For demonstration purposes, Figure 1 shows the dataflow graph for Equation 1 where N = M = 2. Because the dataflow graph can be constructed directly from the difference equation, it is called a *direct form I* representation of a 2nd-order infinite impulse response (IIR) filter [2]. An alternative representation which reduces the number of delays (initial tokens) required of a sequential processor is shown in Figure 2 and is called a *direct form II* representation [2]. The *direct form II* equivalency to Figure 1 is based on the linear, shift-invarient property of such systems. This property allows one to realize the poles of the transfer function of Equation 1 first, and the zeros second, as opposed to the other way around. The *direct form II* dataflow graph shown in Figure 2 will be used to demonstrate the model and software tool presented in this paper.

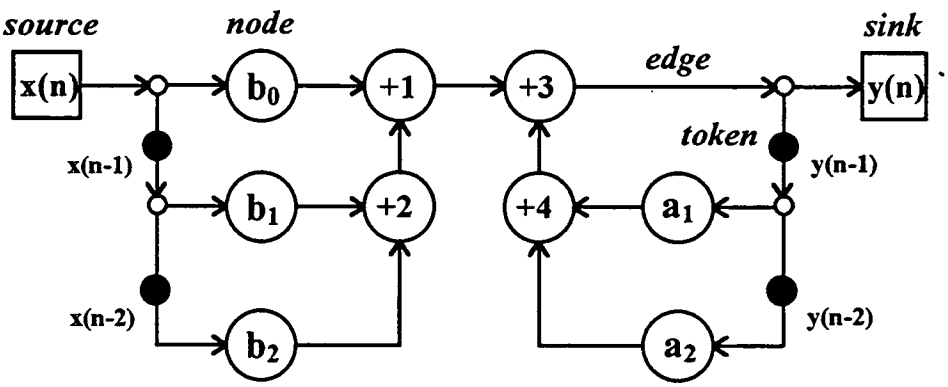


Figure 1. The *direct form I* dataflow graph representation of a second-order IIR filter.

The dataflow model defined in this paper incorporates *node latency* (L_{node}) and *edge latency* (L_{edge}) as model parameters. The node latencies model the expected execution duration of the functions, and edge latencies model the communication cost associated with communicating a result from one node to another. For the IIR filter shown in Figure 2, let's assume that two-input additions (nodes +1, +2, +3, and +4) take 10 *time units*, multiplications with filter coefficients (nodes a_1, a_2, b_0, b_1, and b_2) take 20 *time units*, and all edge latencies are 5 *time units*.

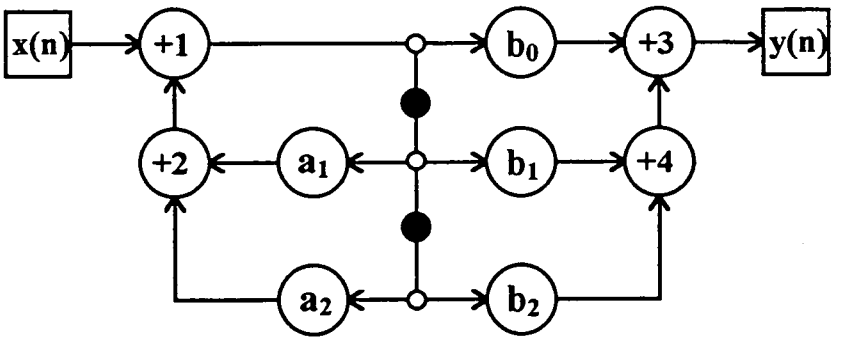


Figure 2. The *direct form II* dataflow graph representation of a second-order IIR filter.

Marking an edge with an initial token effectively introduces a unit delay (a logical delay) between the production and consumption of tokens on the edge. Referring to Figure 2, for example, the edge directed between nodes +1 and a_2, denoted as (+1, a_2), is initialized with 2 tokens to model the logical delay of 2 units required to realize the filter equation. The 2-unit delay between nodes +1 and a_2 implies that the nth firing of node a_2 would be enabled by the (n-2)th token produced by node +1. Since edges imply physical memory implemented by first-in first-out (FIFO) queues or buffers, and tokens imply initial data values (zero values in this case) within the buffers, run-time implementation of *delay* does not incur any overhead. Delays are simply implemented by initializing the buffers. The graph transitions, through markings, as a result of a sequence of node firings. A node is enabled for firing when a token is available on every input edge of the node, indicating that the corresponding function has all of its operands, and space is available on each output edge for the result token. When the node fires, it consumes one token from each of its input edges, delays an amount of time equal to the function latency, and then deposits one token on each of its output edges. Transitions called sources and sinks are also provided to model the input and output data streams of the filter. Sources and sinks have special firing rules in that sources are unconditionally enabled for firing and sinks consume tokens, but do not produce any. In practice, it is assumed that input tokens representing input data samples will arrive from the source at fixed intervals, the sampling period of the A/D converter for instance.

Graph-theoretic performance bounds and metrics are defined in this section. Rather than explain in detail, the proofs and derivations of the equations, a brief explanation of each will be given, and the significance of the equations will be demonstrated using the IIR filter example. The reader is invited to refer to [1] and [3-5] for a detailed discussion of the theory defined here.

There are two mathematical expressions which define the precedence relationships modeled by the dataflow graph. The first expression:

$$t_s \geq t_p + L_p + L_{(p,s)} \tag{2}$$

defines the *forward precedence constraint* imposed on all *intra-iteration* dependent nodes, i.e., associated with the same logical computation. The expression simply assures that given a precedence relationship between a predecessor node (p) and a successor node (s), the time (t_s) in which the successor node is scheduled for execution must be after the predecessor node produces the result token needed for input by the successor node. This start time (t_s) must be greater than or equal to the summation of the predecessor node's start time (t_p), execution latency (L_p), and the edge delay between the two nodes, $L_{(p,s)}$. Unlike the forward precedence constraint presented in [5], this expression also includes the edge delay $L_{(p,s)}$ which provides a simple, but effective, means of modeling the cost of communicating data between nodes. This communication model assumes that nodes with multiple output edges can communicate the data simultaneously. Unless a static scheduling partition is imposed on the graph up-front, the graph-theoretic design process defined in this paper applies this communication model to all edges, thereby assuming that every edge will incur a communication cost. As demonstrated in the next section, once a static assignment of nodes has been made to processors utilizing the dataflow model, intra-processor communications can be removed and the effect observed.

The second expression:

$$t_s + D_{sp} \cdot T \geq t_p + L_p + L_{(p,s)} \tag{3}$$

defines the *backward precedence constraint* imposed on all *inter-iteration* dependent nodes, where D_{sp} is the logical delay between node s and node p (i.e., the number of initial tokens) and T is the iteration period of execution. Inter-iteration dependencies associated with different logical computations result from delaying the availability of intermediate data samples within a DSP algorithm, as modeled by the inverse-z terms in a signal flow graph or initial tokens in a dataflow graph. The precedence constraint of Equation 3 models the fact that at steady state, every node fires at a rate equal to T and given a *logical* delay of D_{sp} (a token generated by node p will be delayed D_{sp} units before being used by node s), the maximum *time* the token will be delayed is equal to the product of D_{sp} and T.

The dataflow graph also has the ability of predicting the algorithm-imposed minimum iteration period referred to as T_{min}, regardless of the number of processors the algorithm is mapped onto. Such a graph-theoretic limit has been widely published, showing that T can be limited by recurrence loops within the algorithm. It has been shown for example in [3] and [5] that T_{min} is defined by the maximum loop time per loop delay. In addition to the node latencies incorporated in such ratios defined in [3] and [5], the edge delays which represent communication costs will also be incorporated here. The maximum ratio defines the critical loop and can be easily determined from Figure 2 to consist of nodes +1, a_1, and +2. Hence, T_{min} for the IIR example problem, given by the following equation:

$$T_{\min} = \frac{L_{loop}}{D_{loop}} = \frac{\sum_{node \in loop} L_{node} + \sum_{edge \in loop} L_{edge}}{D_{loop}} = \frac{55}{1} = 55.\textit{time units} \tag{4}$$

is calculated to be 55 *time units*. Note, however, that T_{min} is an algorithm graph-imposed limit. The minimum iteration period that may be achieved given a finite number of processors (P) is given by:

$$TBO_{LB} = \max\left(T_{\min}, \frac{TCE}{P}\right) \tag{5}$$

where *TCE* is defined as the total computing effort and equal to the sum of all node latencies [1], [5]. This iteration period is denoted as TBO_{LB}, referred to as the lower bound time between outputs. The TBO_{LB} metric not only defines a lower bound on the iteration period, but also conveys the fact that an output will be produced every *TBO* interval—the inverse of *TBO* being throughput [1].

Another metric that is often useful, is the measurement of time between input and output (*TBIO*) of the graph. When considering both the forward and backward precedence constraints, it has been shown in [1] that *TBIO* is defined by the path which maximizes the following expression:

$$TBIO = \sum_{node \in critical\ path} L_{node} + \sum_{edge \in critical\ path} L_{edge} - D_{critical\ path} \cdot TBO \tag{6}$$

for a given iteration period *TBO* and total delay along this *critical path*, $D_{critical\ path}$. Given an iteration period of T_{min} equal to 55 *time units*, *TBIO* is computed for the IIR filter example to be 50 *time units* due to the critical path: +1, b_0, and +3.

The lower limit on the processor requirement for a given iteration period can also be defined and is referred to as the *calculated processor* requirement [5]. This value can be calculated by applying the ceiling function to the ratio of *TCE* to *TBO* [1]. Given a *TBO* of 55 *time units* for the IIR filter example, at least 3 processors would be required for scheduling as calculated by Equation 7.

$$P = \left\lceil \frac{TCE}{TBO} \right\rceil = \left\lceil \frac{140}{55} \right\rceil = 3 \tag{7}$$

The next section will present a software tool developed to solve Equations 2 and 3 for a given dataflow graph and provide automatic calculation of the performance bounds and metrics defined in this section.

3. THE DATAFLOW DESIGN TOOL

A software tool is presented in this section which implements the graph-theoretic analysis concepts discussed in the previous section. The software is formally called the *Dataflow Design Tool*, referred to hereafter as the Design Tool. The Design Tool was written in C++ for Microsoft Windows 3.1 or Windows NT and can be hosted on an 386/486 personal computer or compatible. The tool takes input from a text file, which specifies the topology and attributes of a dataflow graph. The various displays and features are shown to provide an automated and interactive design process, which facilitates the selection of a multiprocessor solution based on dataflow analysis. In addition to providing the numerical performance metrics discussed in the previous section, the tool graphically portrays system behavior using Gantt charts and concurrency envelopes. This section will demonstrate some of the displays and capabilities of the Design Tool while applying it to the analysis of the dataflow graph shown in Figure 2. Besides providing a graphical portrayal of the scheduling criteria, the tool will show the automatic calculation of the same performance and timing characteristics determined by hand in the previous section.

The Design Tool provides a graphical user interface, referred to as the Metrics window shown in Figure 3, containing buttons and menus for the purpose of displaying performance bounds, setting *TBO*, and setting/displaying calculated processors. The time measurements shown by the tool are given in generic *time units*, so that the resolution of the measurement can be user-interpreted. Upon analyzing the dataflow graph of Figure 2, the Design Tool has determined that *TCE* is 140 *time units*, the $TBIO_{lb}$ is 50 *time units*, TBO_{lb} is 55 *time units*, and the calculated processor requirement is 3, as expected from the previous calculations. One other performance metric not previously discussed is *schedule length*, denoted as ω. Schedule length is defined as the minimum time to execute all functions for a given computation. For DSP algorithms, ω is typically greater than *TBIO*. In this example, ω is measured to be 145 *time units*. If one were to change the available processors, the Design Tool would automatically calculate new performance values and bounds as needed. The Metrics window also allows the user to invoke various graphical displays. As an example, clicking on the *Graph Play* button, invokes a window that graphically portrays the precedence constraints for the steady-state computation associated with a single data sample. The display is formally called a Single Graph Play (SGP) diagram and uses shaded bars to indicate the permissible execution (assignment in time) and duration (the length of the bar) in time (the abscissa of the diagram). Unshaded bars indicate slack time (mobility) of each node and the node names are shown along the ordinate axis.

3.1. Exposing the Parallelism

The Design Tool works on the premise that its better to first, expose as much parallelism and performance bounds that are inherently present in the algorithm decomposition and then second, optimize (if necessary, altering the dataflow-derived schedule) to account for finite resources in terms of available processors and memory. Thus, the Design Tool constructs the SGP diagram assuming infinite resources such that all of the parallelism inherent in the algorithm decomposition can be exposed, constrained only by the precedence relationships. Consequently, the execution bars are positioned to indicate the earliest start times of each node. Earliest start time together with the latest finish time, indicated by the slack bars, provide a range in which the nodes (functions) can be scheduled for execution. Cursors (vertical lines) are provided in the display for purposes of measuring time. The left cursor time is shown next to the word TIME at the bottom of the display, and the time between the left and right cursors is shown in parentheses. For instance, the delay between the completion of node b_0 and the start of node +3, as shown in Figure 3, is measured to be 5 *time units* which corresponds to the communication delay modeled on the edge (b_0, +3). Note how the logical delays modeled by initial tokens allow many of the nodes to be scheduled much earlier than when the results are actually needed. Also, as modeled in Equations 2 and 3, the slack time shown in the SGP takes into consideration the communication delay required to provide the output of one node to the input of an immediate successor node. For example, note how the slack of node +4 stops 5 *time units* short of the earliest start of node +3.

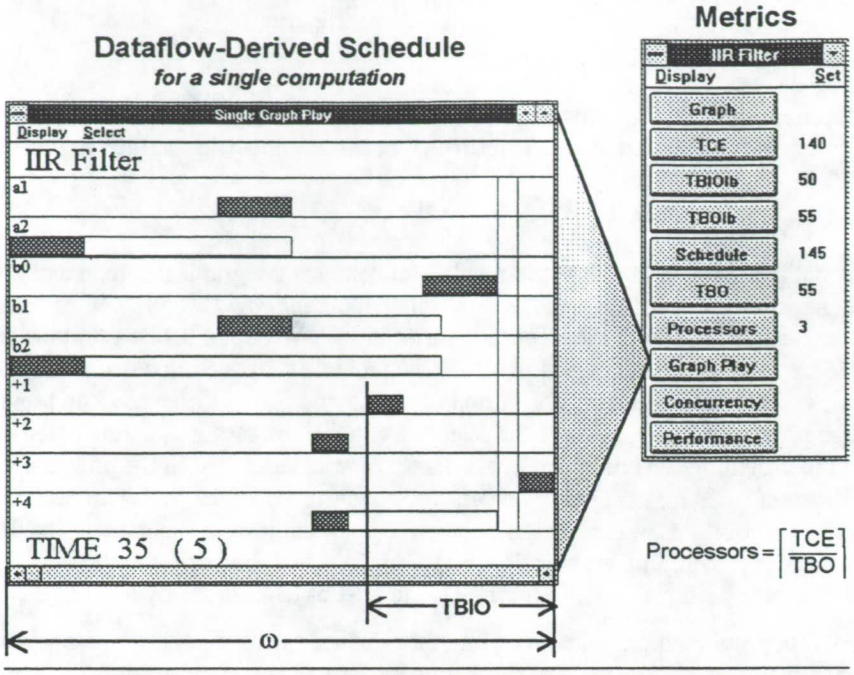


Figure 3. The SGP window exposes the parallelism associated with a single computation.

The SGP shows the steady-state scheduling range associated with computation of a given sample. To determine a permissible schedule (one that is schedulable, consistent, deadlock free and has a bounded resource requirement), a periodic view of the SGP at steady-state is required. The Design Tool provides such a display with another Gantt chart referred to as a Total Graph Play (TGP) diagram. The TGP window shown in Figure 4 portrays, as a snapshot in time, the steady-state, earliest-start schedule over a single periodic interval of length *TBO*. Like the SGP, the TGP represents node assignments in time using bars. The bars are shaded and numbered to indicate the relative iteration associated with a time-scheduled node. That is, if a node bar numbered "1" indicates the nth iteration of that node, then another node bar numbered "3" would indicate the n+3 iteration of that node within the same iteration period. Though not shown in this paper, overlapped bars for a given node indicate that the node function requires multiple instantiations to support the desired level of throughput [6]. Multiple instantiations simply means that the function would be required to execute simultaneously in different processors on different (and sequential) data samples. Construction of the TGP diagram is accomplished by mapping the earliest start times (relative to the SGP diagram) to a time interval of width *TBO* using the mapping function *earliest start time modulo TBO*. As indicated by the shading (and iteration numbers) of the bars, the TGP portrays not only the parallel concurrency (of the same data sample) that is being exploited, but also the pipeline concurrency (multiple sample stream) that results from the simultaneous execution of different data samples within the algorithm graph. The TGP window can also expose the critical loop located by the Design Tool as indicated by the lightest-shaded bars and composed of nodes a_1, +1, and +2, as expected from the Section 2. Note how the edge delays within the critical loop contribute to the minimum iteration period that can be attained, regardless of the number of processors. As a result of the critical loop, the maximum speedup (the ratio of *TCE* to *TBO*) potential that can be expected is 2.5 which peaks at 3 processors as indicated by the Speedup Performance windows also shown in Figure 4. The Design Tool displays the Speedup Performance window for any user-specified processor limit.

Figure 4 provides a good opportunity to briefly discuss the usual discrepancy between the calculated processor requirement (calculated with Equation 7) displayed in the Metrics window and the processor requirement indicated by the TGP window. Whereas, the calculated processor requirement is a lower bound that indicates the optimum number, the TGP indicates the number of processors that are *sufficient*, not necessarily optimum. In this example, the calculated processor requirement is determined to be 3, and the sufficient number is shown to be 5 by counting the maximum overlap of node bars in Figure 4. This may mean that the dataflow-derived, periodic schedule (time assignment) based on the earliest start times is not a good solution, or that the under-utilized idle

time shown in the TGP is due solely to the partial-ordering of nodes and is needed to achieve correct logical results. The nature of this problem stems from the fact that the scheduling problem of partially-ordered nodes is NP-complete [7]. That is, an optimum solution is only assured by an exhaustive search. Since a polynomial-time solution to such problems hasn't been found (and probably doesn't exist), a heuristic (approximate solution) is often employed. The criteria for an optimum, or acceptable, solution also affects the search space. That is, the more parameters such as *TBIO*, *TBO*, and *processors* that are held fixed, the more difficult it is to find a solution. In fact, a solution may not even exist, if both *TBO* and *processors* are held fixed. One of the purposes of the Design Tool is to aid the user in finding an acceptable scheduling solution based on selectable criteria.

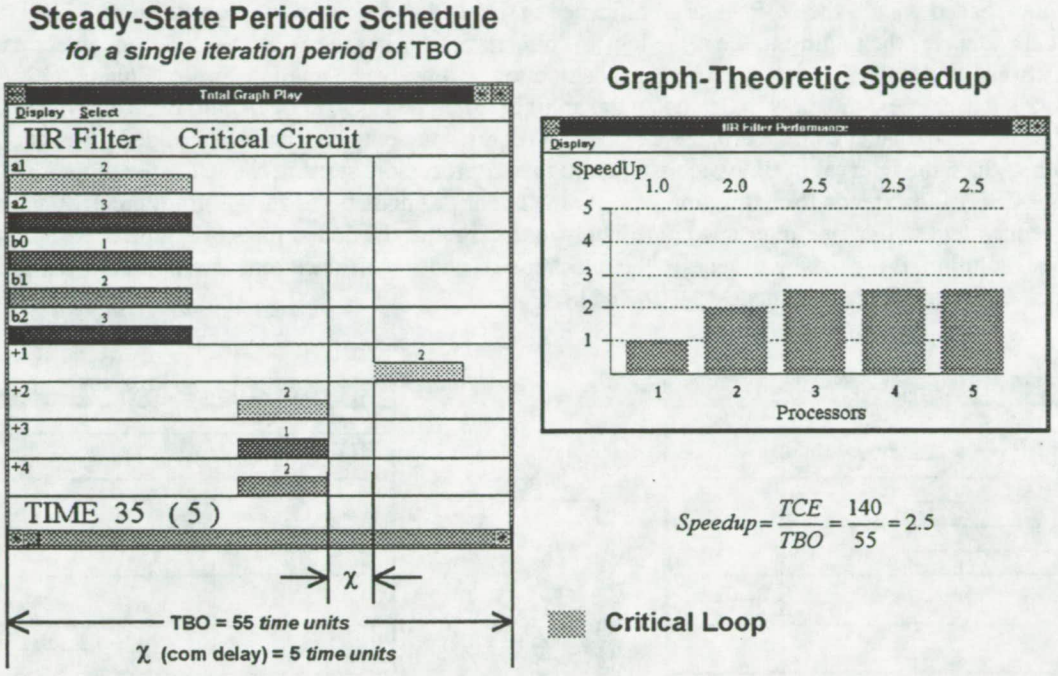


Figure 4. The TGP window exposes the parallelism within an iteration period at steady state.

3.2. Dataflow Optimization

When the sufficient resource requirement exceeds the finite resource availability, and optimization is required, the Design Tool allows one to impose artificial data dependencies between nodes to alter the dataflow schedule. Even though such a re-scheduling mechanism has the disadvantage of quantizing the range in which the nodes can be scheduled, such a reduction in the solution space is often advantageous. For one thing, the solution can still be completely characterized by the dataflow model since the artificial data dependencies are still modeled by edges (called control edges). Another advantage is that the realization of the multiprocessor scheduling solution at run-time requires only a dataflow machine[1] or the scheduling of node functions on a control flow machine based on discrete events (i.e., the completion of other nodes). This section demonstrates the use of the Design Tool in optimizing the dataflow schedule for the dataflow graph of Figure 2 given the criteria that the solution achieves the lower bound iteration period (sampling period) of 55 *times units*, that can be statically scheduled on 3 processors, and minimizes *TBIO* as much as possible (i.e., we are willing to sacrifice *TBIO* performance for *TBO* performance). The static schedule versus a dynamic schedule is advantageous (and often necessary for fine-grain problems like this), since it eliminates the run-time costs associated with processor assignment decisions by making such decisions at compile time.

---

[1] A dataflow machine is defined to be an efficient hardware implemention of the dataflow model of computation or a software implementation of the dataflow model on a von Neumann machine.

Figure 5 demonstrates the use of the Design Tool in re-scheduling a node via the addition of a control edge. The SGP on the left side of Figure 5 shows the result of entering the *add edge* mode and choosing a node (by clicking on a node bar with the mouse) to be moved (by delaying) within its range, or outside (resulting in an increase in *TBIO*) if need be. The left-side SGP also shows a display mode that uses vertical, dotted lines to indicate iteration periods of width *TBO*. The *earliest start modulo TBO* mapping of the SGP to the TGP is actually equivalent to overlapping these *TBO*-width segments. Imposing such a reference onto the SGP allows one to transition more easily between what is viewed in the TGP referenced to a *TBO* period to what's viewed in the SGP referenced to the time interval ω. Another feature of the tool is that after selecting a node to be moved, in this case node b_1 as indicated at the bottom of the SGP window, all nodes independent of node b_1 are highlighted as shown by the lighter shaded bars in the SGP. The significance of this is that the tool allows one to impose only *intra-iteration* dependencies (belonging to the same logical computation) using the SGP. Such dependencies can only be imposed between nodes that are otherwise independent, else the precedence relationships of Equations 2 and 3 would be violated. For example, as indicated by the unhighlighted node +4, node b_1 cannot be delayed behind node +4 for the same logical computation, since node +4 requires the output of node b_1. Node b_1 was chosen for re-scheduling since the interval of time requiring more than 3 processors starts at the left side t of an iteration interval $[t, t + TBO)$ and extends for a duration equal to the latency of node b_1, as shown in Figure 4. This same time interval is equivalent to the time intervals immediately to the right of the dotted lines in Figure 5. Consequently, our optimum solution, if it exists, will require the relocation of nodes with slack that are right-adjacent to these dotted lines. There are three such nodes: a_2, b_1, and b_2.

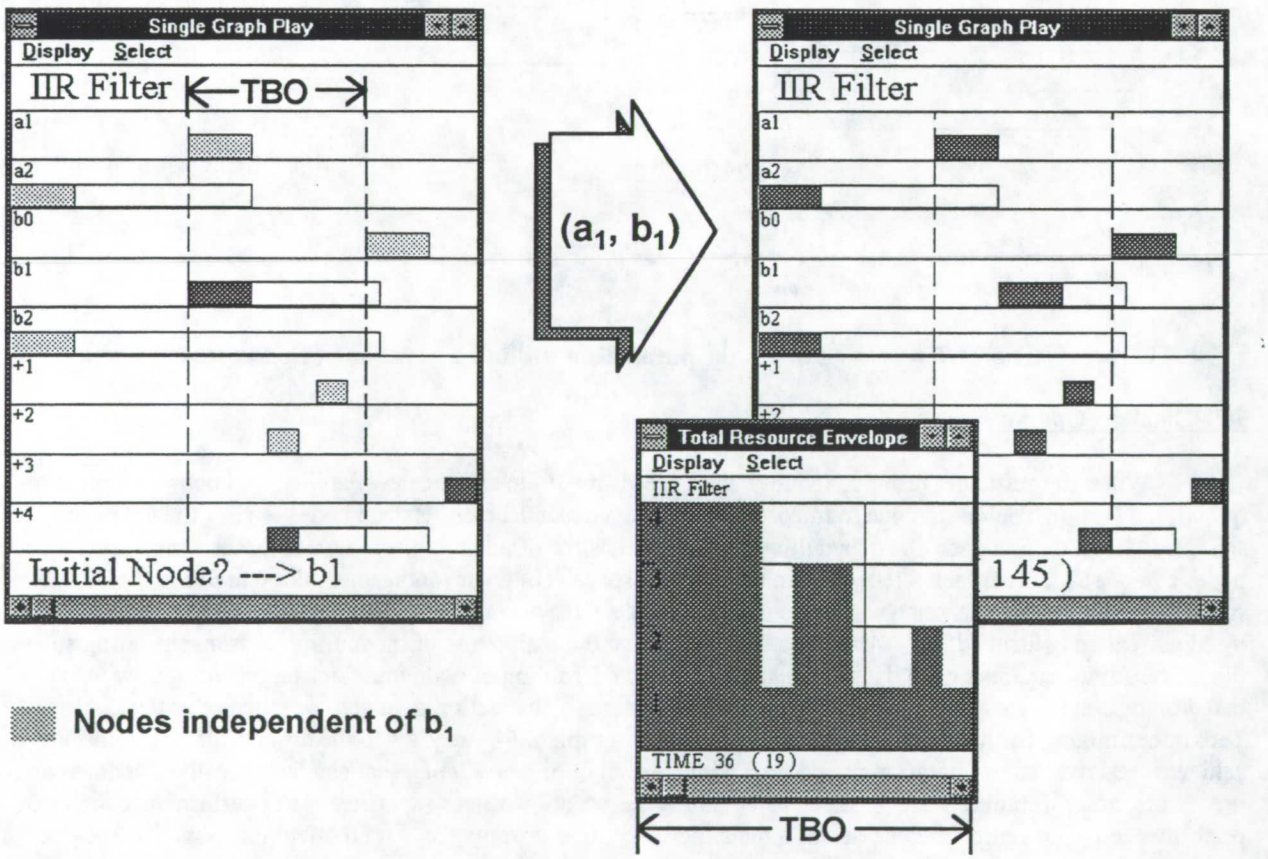


Figure 5. The Design Tool provides the ability to relocate scheduled nodes for optimization.

The effect of relocating node b_1 behind node a_1 is shown by the right-side SGP of Figure 5. The Design Tool accomplishes this task by automatically imposing a control edge between nodes a_1 and b_1, denoted as (a_1, b_1). As a result, the processor requirement has been reduced to 4 as shown by another display shown to the right called

a resource envelope. A Single/Total Resource Envelope is provided for both the Single/Total Graph Play windows and portrays the processor utilization as a function of time. The area under the shaded curve of the Total Resource Envelope (TRE) for a *TBO* interval is equal to the *TCE* metric. Next, we look for additional options that may move the amount of effort (TRE area) requiring 4 processors into the vacant area requiring 1 or 2 processors. For this example, the additional precedence constraints (+2, b_2) and (+2, a_2) provides such an outcome, as shown by Figure 6. The solution not only satisfies the $TBO = T_{min}$ and 3-processor requirements, but also achieves the lower bound *TBIO* since all nodes were relocated within there range defined by the earliest start and slack time metrics. The Design Tool also shows that the schedule length has been reduced to 110 *time units* and that the 3 processors will be utilized 84.8% of the time.

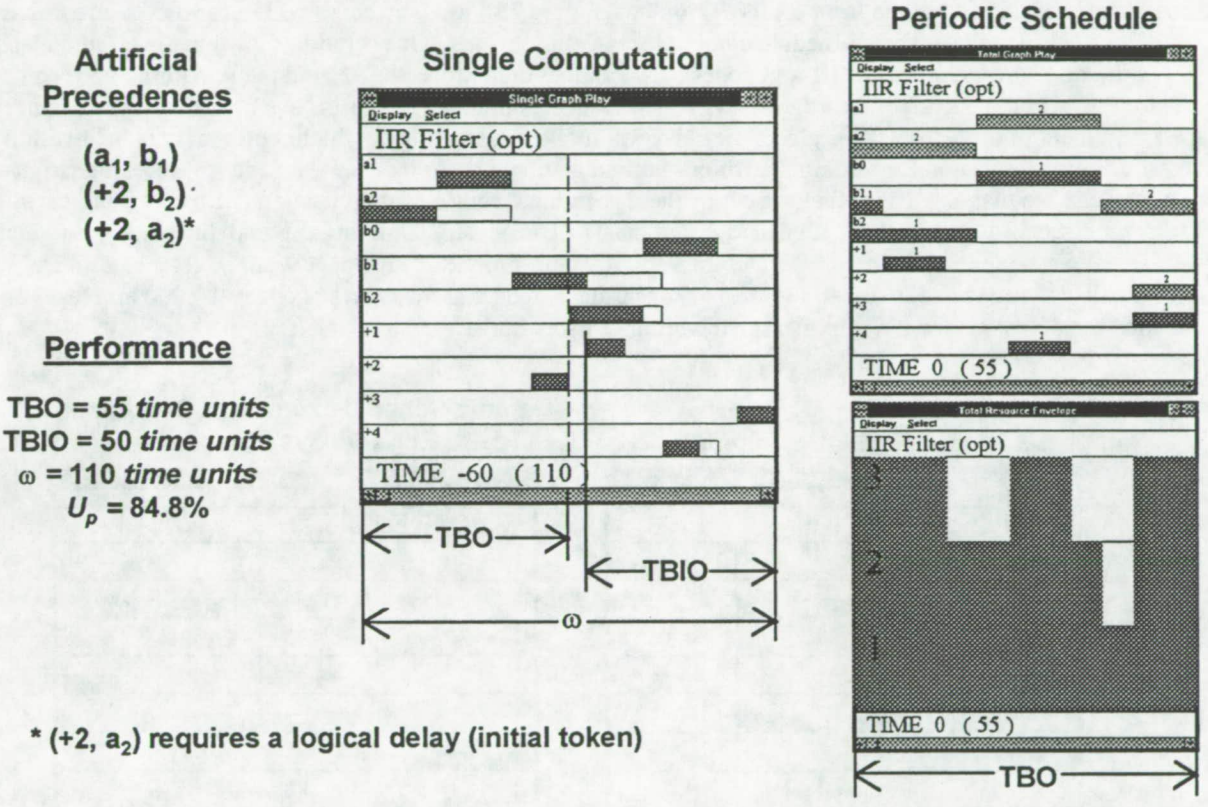


Figure 6. An optimized dataflow schedule for 3 processors.

One last point should be made about this solution before moving on to the static processor assignment. The control edge (+2, a_2) may seem to contradict the earlier discussion about allowing only intra-iteration dependencies. That constraint is only applicable to the SGP window, since it portrays the scheduling range of nodes of a single computation. The TGP window, on the other hand, portrays a permissible schedule of multiple computations related to the iterative execution of the dataflow graph algorithm. As a result, the TGP characterizes both *intra-* and *inter-iteration* dependencies and consequently, allows one to impose both kinds of dependencies by relocating nodes. Whether a control edge results in an intra- or inter-iteration dependency is transparent to the user. Node a_2 must still be scheduled before node +2 for the same logical computation, and it does as shown in the SGP of Figure 6. But, at steady-state and viewed within a single iteration period, node a_2 is synchronized with the completion of node +2 for two different logical computation separated by one iteration period, i.e., a logical delay of one. Therefore, in order for this schedule to be realizable, the Design Tool automatically determines that a unit delay modeled by an initial token is required on the control edge (+2, a_2). The need for this delay can also be foreseen by noting the recurrence loop created by the real data dependency (a_2, +2) and the artificial dependency (+2, a_2), which requires at least a single delay (token) for the algorithm to be schedulable.

3.3. Static Scheduling

The dataflow-derived time schedule discussed previously and shown in Figure 6 is a graph-theoretic prediction of what will occur and the amount of resources (processors and memory) consumed if the algorithm were executed on a dataflow machine. Even though only the processor requirement prediction has been discussed in this paper, the Design Tool can analytically predict the minimum memory requirements of all FIFO buffers that are implied by the dataflow edges to assure proper synchronization of shared data and thus assure correct computational results. Scheduling algorithms (heuristics) have been developed, which consider the scheduling criteria derived from the dataflow analysis provided by this tool in hopes of determining a better solution. The Design Tool tabularizes this criteria for use by such scheduling algorithms and also provides features that allow the user to manually select a static schedule from the TGP window. Figure 7 shows how nodes within the TGP window have been re-shuffled such that a static schedule onto 3 processors is obvious. One solution shown by this figure maps nodes a_1, a_2, and +2 onto processor P1 and nodes b_0, b_2, and +3 onto processor P2, and the remaining nodes onto processor P3. Up to now, all of the edge delays which model communication costs have been considered. Once static assignment of processors takes place, the delays of intra-processor edges, which represent data held within a processor diminish to zero. Envisioning partitions around graph nodes to indicate the node-to-processor assignments, one can see that only edges that are cut by these partitions require communication (with a nonzero cost) between processors. It is only these edge *delays* that must be considered. Utilizing the analytical analysis (which results in the fastest response time) of the Design Tool, the intra-processor edge delays can be reset to zero and the effect quickly determined as shown in Figure 7. For instance, reducing the delays of edges (b_0, +3) and (a_1, +2) eliminated the 5 *time units* delay gap measured with the cursors in the TGP on the left.

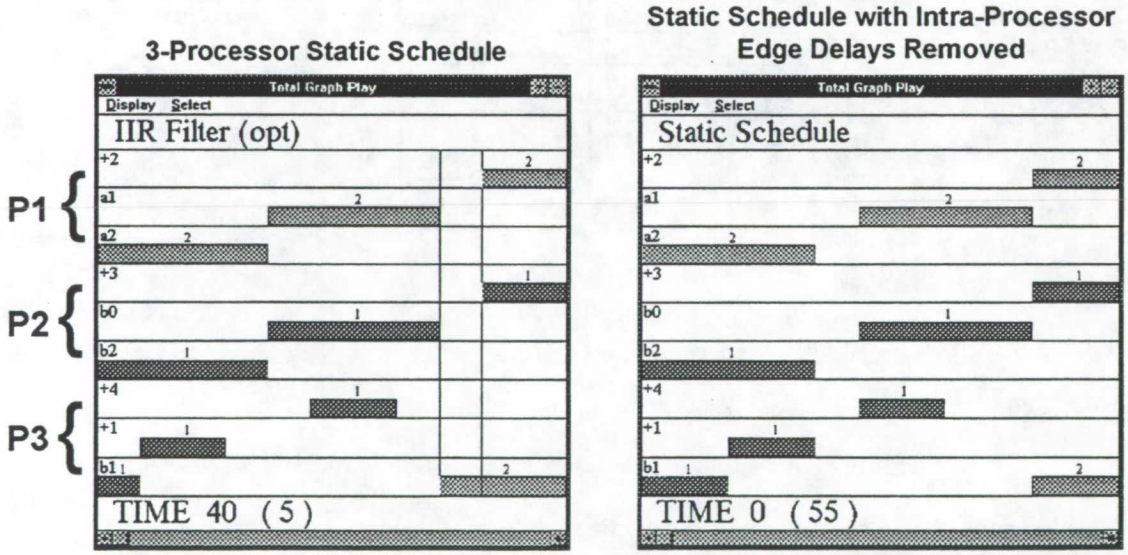


Figure 7. Dataflow scheduling criteria can be used to statically assign functions to processors.

The TGP shown in Figure 7 on the right portrays an optimum, multiprocessor, static-scheduling solution of the IIR filter problem presented in this paper. The 3-processor schedule has considered communication delays between processors and achieves the minimum sampling period possible, while minimizing the input-to-output latency. Though not shown in this paper, the Design Tool also predicts that only a single buffer for each data sample shared between the functions is needed, except for the intermediate values described by edges (+1, a_2) and (+1, b_2). As expected, these two exceptions require 2 buffers each to provide logical delays of 2-units.

4. COMMERCIAL APPLICATIONS

The DSP market has exploded with new and innovative DSP hardware and software architectures which provide mechanisms to efficiently exploit the parallelism inherent in many DSP applications. The dataflow paradigm has also been getting considerable attention in the areas of DSP and real-time systems. The commercial products that are offered today provide the advantages of dataflow only as a graphical programming language. However, the products do not provide an analytical analysis and design of the dataflow graph. While there are many advantages to graphical programming, the full potential of the dataflow representation is lost by not utilizing it analytically as well. In the absence of the analysis/design offered by the Dataflow Design Tool, the commercial toolsets must rely on compile-time approximate solutions (heuristics) or run-time scheduling, which often results in a trial-and-error design approach. It is felt that this software tool has high commercial potential. It could be readily incorporated into existing commercial DSP toolsets to determine a desirable multiprocessor solution at compile time. Other commercial uses of this tool include scheduling of DSP algorithms for real-time applications, including those found in aircraft, automotive, industrial processes, medicine, and remote sensing. The tool could also provide front-end scheduling constraints for other commercial tools utilizing scheduling algorithms, with the potential of finding more efficient solutions.

5. CONCLUSIONS

Dataflow graph analysis concepts were presented in this paper and shown to determine performance bounds inherent in a digital signal processing algorithm. A software program called the Dataflow Design Tool was presented, which implements an analytical dataflow analysis for determining graph-theoretic performance bounds, providing Gantt chart portrayals of scheduling behavior, and plotting processor utilization. The automated and interactive features of the tool were shown to facilitate the selection of a multiprocessor solution for static scheduling. Scheduling ranges, earliest start, and slack time of functions, based on dataflow analysis can also be used as correctness criteria for static-scheduling heuristics.

REFERENCES

[1] Robert L. Jones III, "A Design Tool for Multiprocessing Scheduling and Evaluation of Iterative Dataflow Algorithms," To be published as a NASA Technical Paper-3491, Dec. 1994.

[2] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc., 1975.

[3] S. Som, J. W. Stoughton, and R. R. Mielke, "Strategies for Concurrent Processing of Complex Algorithms," NASA CR 187450, Oct. 1990.

[4] S. Som, R. Obando, R. R. Mielke, and J. W. Stoughton, "ATAMM: A Computational Model for Real-Time Data Flow Architectures," *International Journal of Mini and Microcomputers*, vol. 15, 1993, pp. 11-22.

[5] Sonia M. Heemstra de Groot, Sabih H. Gerez, and Otto E. Herrmann, "Range-Chart-Guided Iterative Data-Flow Graph Scheduling," *IEEE Transactions on Circuits and Systems*, vol. 39, May 1992, pp. 351-364.

[6] R. L. Jones, P. J. Hayes, A. M. Andrews, S. Som, J. W. Stoughton, and R. R. Mielke, "Enhanced ATAMM for Increased Throughput Performance of Multicomputer Data Flow Architectures", *Proceedings of the NAECON '91*, vol. 1, May 20-24, 1991, pp. 238-244.

[7] E. G. Coffman, *Computer and Job-Shop Scheduling Theory*, John Wiley and Sons, Inc., 1976.

Object Oriented Hybrid Network Simulation[1]

John S. Baras[2], Ahsun H. Murad, Kap Do Jang, George C. Atallah[3], Ramesh Karne and S. Joseph Campanella
Center for Satellite and Hybrid Communication Networks
Institute for Systems Research
University of Maryland
College Park, MD 20742

ABSTRACT

The simulation of complex contemporary hybrid communication networks is a daunting task and cannot be handled by the array of conventional network simulation tools available today. This paper discusses the key issues involved in the visualization and simulation of complex hybrid networks and presents the design of an object-oriented hybrid network simulator being developed at the Center for Satellite and Hybrid Communication Networks (a NASA center for the commercial development of space), University of Maryland at College Park. This software incorporates a complex system simulation paradigm, advanced visualization techniques, a flexible interface and an object-oriented database to allow the simulator to be used not only as a network simulation tool, but also as a network design tool, as a network performance evaluation tool, and as part of an intelligent network management software.

INTRODUCTION

As the complexity and diversity of networks have grown, simulation has proved to be an important tool in their design, analysis, testing and performance estimation. A typical hybrid communication scenario is shown in Fig. 1. Hybrid networks involve a variety of network elements—both terrestrial and satellite with their associated protocols, and the services they provide, like commercial video and radio transmissions, voice, data and image transmission services. Because of their complex nature, design and evaluation of hybrid networks is a particularly complicated and challenging task.

A number of tools are available for the simulation of communication networks. Some of these are general-purpose simulation packages like BONeS DESIGNER, GPSS/H, MODSIM II, SIMSCRIPT II.5, SES/*workbench*, SIMAN/Cinema V and SLAMSYSTEM; others like OPNET Modeler provide a communication network simulation language; and yet others are specialized communication network simulators like BONeS PlanNet, COMNET III, NETWORK II.5 and L•NET II.5. For the simulation of communication networks, a dedicated communication network simulator offers advantages over a general purpose simulator, namely, speed of network modeling, a library of predefined network components and perhaps, automatic computation/display and analysis of the key communication-related performance statistics. A general-purpose simulator, on the other hand, has the flexibility to model almost any discrete-event system, though, as the generality of the simulator increases, typically, so does the effort required.

The simulation of a complex hybrid communication network offers a unique challenge to simulation tools. Because systems to be simulated are complex and diverse, constructing a system model using a general-purpose simulator would be too time-consuming. On the other hand, most of the current network-simulation tools are not designed to visualize or simulate such complex networks. The deficiencies in simulating hybrid communication networks with conventional communication network tools are (i) difficulty in modeling the system in sufficient detail and accuracy, (ii) difficulty in capturing the entire communication scenario, (iii) slow speed of execution, (iv) inadequate network performance monitoring/evaluation support both during and at the end of a simulation run.

A number of issues are involved in the simulation of hybrid networks. While being of interest in the simulation of conventional networks, they gain critical status for the simulation of hybrid networks because of the unique demands

---

[1] Work supported in part by NASA grant NAGW-2777S for the establishment of a NASA Center for the Commercial Development of Space, and by grant NSFD CDR 8803012 through the Engineering Research Center Program.

[2] Martin Marietta Chair in Systems Engineering, Rm. 2249 A. V. Williams Building, University of Maryland, College Park, MD 20742. Ph: (301) 405-6606, Fax: (301) 314-9218, email: baras@isr.umd.edu.

[3] Rm. 3117 A.V. Williams Building, University of Maryland, College Park, MD 20742. Ph: (301) 405-7411, Fax: (301) 314-8586, email: atallah@isr.umd.edu.

placed by the complexity and diversity of these networks. They in turn place stringent demands on the simulation tools and motivate the development of new tools tailored to the simulation of these complex systems.

A Complex System Simulation Paradigm — To get an idea of the complexity, let us consider the simulation of Calling's proposed Teledesic satellite network [1]for the provision of global telecommunication services. The proposed system configuration consists of 840–900 satellites in a 21 low-earth orbital plane configuration, with a possibility of ATM support. The capacity of this system is in excess of 2,000,000 simultaneous full duplex 16kbps channels. With the small data-segment size of ATM packets, the simulation of this system operating even at 1% capacity would require the processing of over 10 million packets per second of simulated time. At 100% capacity, a processing rate of over a billion packets a second would be required.

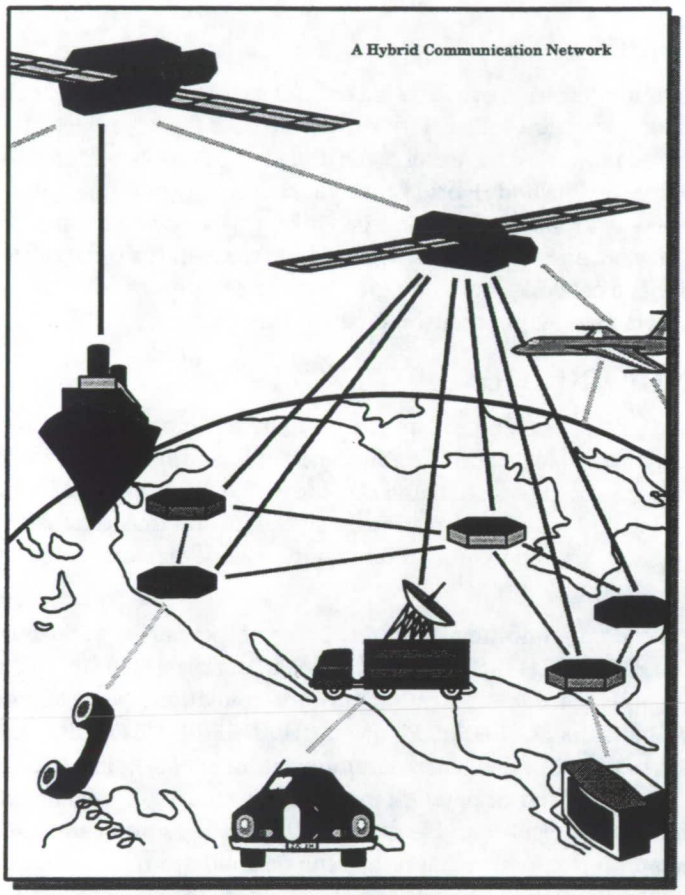


Fig. 1. A Hybrid Communication Scenario.

Assuming the simulation of the life of a packet (generation, end-to-end transfer and consumption) to generate on the order of 100 events (conservative, considering that a typical packet path would involve several satellites), an event based simulator trying to simulate this system would have to process on the order of a billion (1% utilization) to a 100 billion events (100% utilization) for every second of time simulated. With the current technology, even the world most powerful computers would creep through the event-based simulation of such a system—and we haven't even considered the simulation of the terrestrial network that would interact with the Teledesic satellite network, or the work required to simulate the dynamic nature of the satellite/terrestrial network.

Conventional event-driven simulation has been found to be quite efficient for the simulation of conventional networks. However, as the complexity of the system being simulated grows, the vast number of events being generated and required to be individually simulated presents a bottle-neck on the speed of simulation. The back-of-the-envelop calculations we presented for the Teledesic network simulation scenario demonstrate the immensity of the problem. Clearly, a more effective simulation paradigm is required for the simulation of hybrid networks, one that takes advantage of the inherent structure within the sub-systems that make up the complex network. In the hybrid simulation tool, we adapt a multi-tier model based approach– the network system is modeled in terms of a few large functional blocks, which may be recursively defined in terms of smaller functional blocks. The more detailed the model, the more accurate the simulation, and the higher the simulation complexity. This scheme allows a few small, critical blocks to be simulated more accurately, while allowing the other simulation blocks to be replaced by equivalent aggregate parameters, thereby reducing simulation complexity.

Flexibility and Adaptability — A hybrid network, by its very definition, is a diverse system that brings together in a single system, different means of communication, each with its own array of network components, protocols and control strategies. It is not possible for the creator of a simulation tool to provide a model for every possible component that may be required to be simulated. Even if it were, advances in technology and changes in networking ideologies would result in new building blocks that would be required to be simulated, rendering such an effort

useless. It is therefore, very important that a hybrid network simulation tool be designed from the bottom-up to allow easy creation of user-defined components and their efficient integration into the simulation software.

Communication Scenario Visualization and the GUI — Conventional communication networks consist of nodes and links (and specializations of these). While these are the same things that make up the communication aspects of a hybrid network, effective representation and visualization of hybrid networks also involves much more. Hybrid satellite/terrestrial networks often span the globe and interact with the world more directly than conventional networks. For example, correct location of satellite earth-stations and mobile base stations requires knowledge of geographical data, so transformable digital geographical databases need to be incorporated. Accurate prediction of traffic patterns for global networks may also require population modeling in terms of density, economic status etc., requiring the incorporation of other databases. Long-term network performance may also require the incorporation of global climatic data. (For example, rain fade in areas of high rainfall may degrade quality of service to unacceptable levels.)

With the ability to incorporate all this multi-faceted information in a simulation comes the difficult problem of visualization and display of this information to the user and providing him with controls to easily manipulate it as desired. The complexity of the network itself may present problems, and require that visualization of the network be done at a higher level with selective presentation of the most relevant data at the top level.

This points towards the development of a sophisticated graphical user interface (GUI)—one that appears clean and simple, and yet utilizes all possible means (like color, size and shape coding) to convey multidimensional information. New network visualization paradigms are also required to display complex network structures.

Data Management, Databases and Network Management — During a simulation run, vast quantities of data may be generated. Typically, this data would be simply dumped into a file, and accessed later for post-simulation analysis. A more efficient approach is to structure the data on the fly by storing it in a database. This allows simulation data to be manipulated and utilized much more easily during the post-simulation phase. More importantly, it allows access to this data while the simulation is being performed, allowing dynamic computation and display of network performance statistics. The database also serves as an interface point for intelligent network management tools that could use the simulated network as a model to predict the long-term behavior of a real network, and use the performance data generated by the simulator to formulate long-term network management policies.

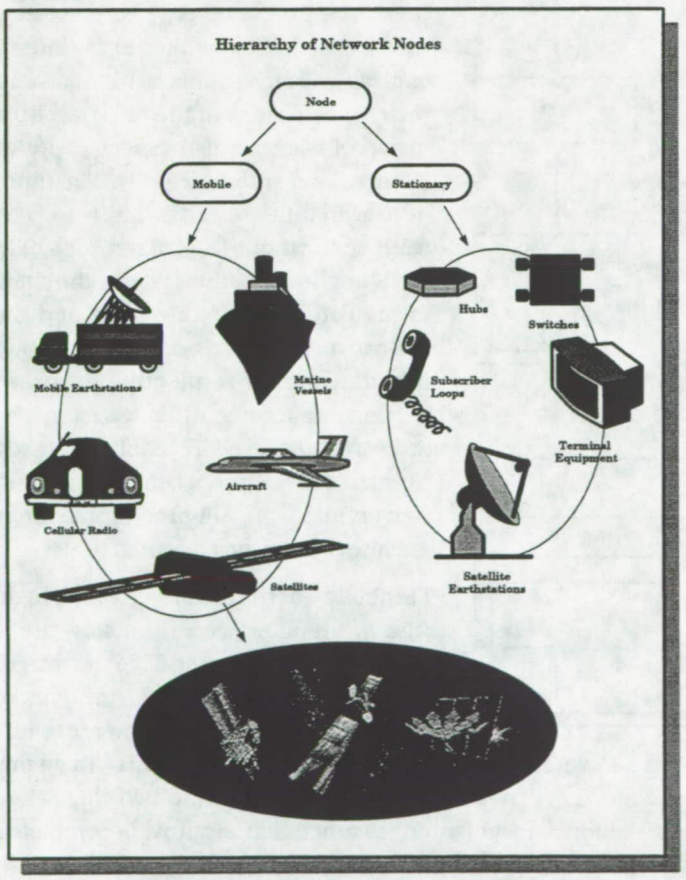


Fig. 2. Object-Oriented Classification of Network Nodes.

Object-Oriented Programming — Object-oriented programming is an advanced approach to structured programming. It is ideally suited to hybrid network simulation for all the reasons that makes it ideal for most complex software efforts: a clean software structure, software and effort reuse, flexibility and adaptability, ease of documentation, and with the availability of modern object-oriented programming languages, efficiency. For example, Fig. 2 illustrates how object-oriented programming may be used to define network nodes; a top-level generic node captures those features common to all nodes, and

more specialized node structures are constructed through inheritance and specialization from more general class structures.

THE HYBRID NETWORK SIMULATOR

This section describes the design and implementation of an object-oriented hybrid simulation network being developed here at the Center for Satellite and Hybrid Communication Networks, University of Maryland at College Park. Keeping in view the motivation behind the development of this software, the proposed system is geared towards the simulation of complex hybrid network simulation systems. The basic structure of the software is shown in Fig. 3. The software is divided into three main parts: the interface, the simulation kernel, and the expansion module library. An object-oriented database serves as a medium for storage and exchange of information between the different components.

The Interface

Three categories of users are expected to make use of the hybrid network simulator. The first of these is the human user who wishes to simulate a communications system. To do this, the user sits in front of a terminal, invokes the software, and then interacts with the **Graphical User Interface** (GUI) to create, simulate and evaluate the performance of a simulated network. The second of these is a programmer, who also sits at a terminal, but this time, wishes to create and incorporate a new module into the software, or perhaps to customize an existing one. The programmer (who is expected to have a deep understanding of the internal organization of the simulator) is allowed to directly access many of the internal structures of the software through a special **Application Programmer Interface** (API). The third kind of user that the software caters to is a software application, (e.g., a network management application) that is using the simulator to simulate a selected network, and is using the network-performance data generated by the simulator for some other purpose (e.g., network management). A lean and efficient **Software Interface** (SI) capability is provided for such a user. In short, the GUI is visualization oriented, the API programmer oriented, and the SI efficiency oriented.

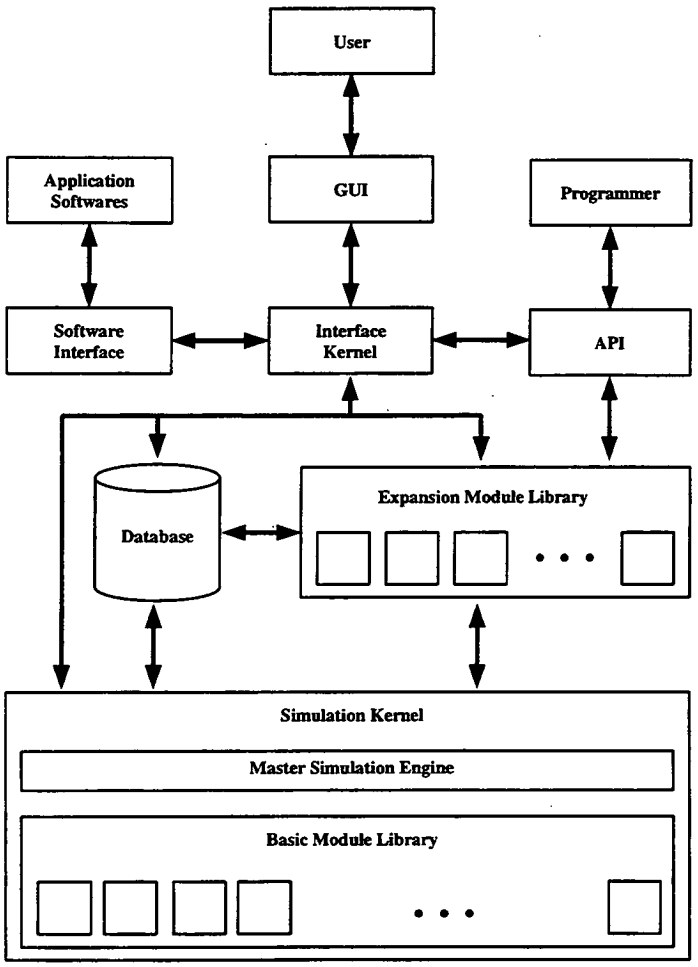


Fig. 3. Block-Diagram of the Hybrid Network Simulator.

Flexibility was the guiding criterion in the design of the interfaces. The software is designed to be easily extendible to many different kinds of graphics platforms, windowing systems, and to be able to interact with a variety of software applications. To do this, the interface is divided into two stages. A platform/user independent **Interface Kernel** provides a canonical set of interface functions that provide complete basic control of the software. The GUI, the API and the software interface are all blocks that use this set of functions. This interface kernel acts as a buffer between the interface and the internal program; changes in the interface do not propagate to the rest of the simulation software. For example, support for a new windowing system can be provided by building a new GUI that still uses the same canonical function set. (Since the API requires greater access to the simulation software, direct interaction also exists between the API module and the simulation software. The API is

however expected to be much more simple to adapt to different platforms, as it is basically a text-based programming interface.)

The Simulation Kernel

The heart of the simulation software is the simulation kernel which consists of the **simulation engine** and the **basic module library**. The simulation engine is an integral part of the control mechanism that implements the complex system simulation paradigm which is described below.

The Complex System Simulation Paradigm

Simulation of complex networks requires a rethinking of the technique of simulation. When systems required to be simulated were simple, a time-based simulation paradigm worked well. With more complex systems, an event-based simulation produces significant reduction in complexity, especially if the system is not very active, temporally, i.e., only a few of all the possible events occur at a time. However, as noted earlier, when system complexity increases, even event-based simulation does not work well. The vast number of events being generated, and required to be individually processed creates a bottleneck which slows down the simulation to unacceptable levels.

For simulation of complex systems, a model based approach is often employed. That is to say, the simulation does not try to emulate what would happen in a real system on an event-by-event basis. Rather, a mathematical model of the system is created that tries to approximate the behavior of the entire system. (For example, a queue might be mathematically modeled as a delay element—the delay modeled as a stochastic process—rather than through the enqueue, dequeue and job service operations on an event-by-event basis.) This is however possible only for very well structured networks, and even then, involves a tradeoff between complexity and accuracy of simulation. We must realize, however, that for simulation of a complex system within an acceptable time frame, such a tradeoff between accuracy and speed has to be made. In light of this, a model based simulation is a promising compromise.

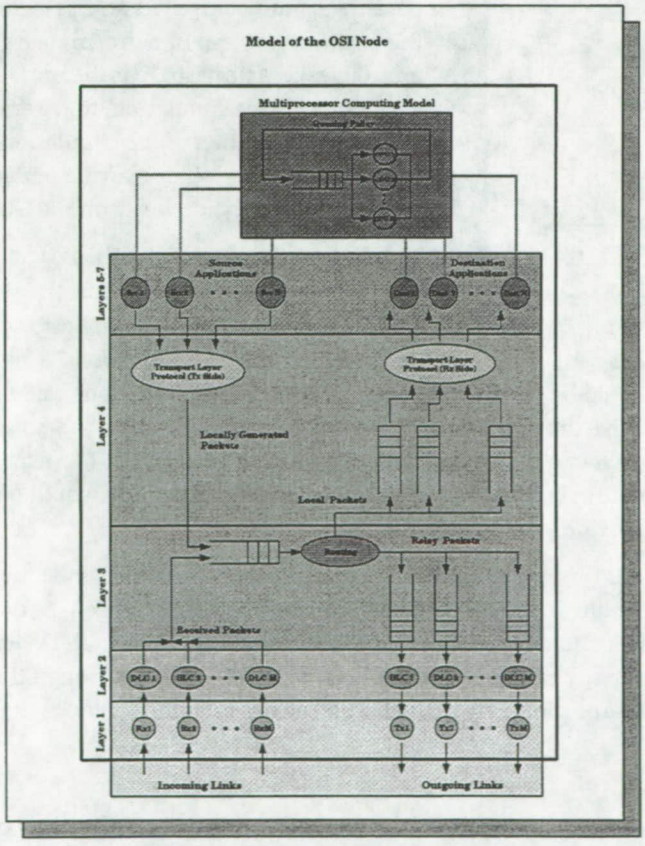


Fig. 4. Structure of the generic OSI Network Node.

Further, it may not be necessary to model the entire network to the same level of accuracy. Non-critical network components may be modeled only coarsely, e.g., a geostationary satellite transponder based up-down link pair may be modeled simply as a broadcast link with a statistical delay and error injection model (rather than as two links and a satellite node). On the other hand, a critical component may be modeled to a much higher level of detail and accuracy, of course, at the cost of an increase in complexity. Often, this modeling may be done in terms of sub-components, which are in turn modeled similarly.

The Basic Modules Library

Also present within the kernel is a **basic modules library**. This is a collection of generic building blocks like queues, switches, transmitters, receivers, timers, e.t.c., and generic nodes and links that can be used to build more complex network components. The modules within the basic module library are however not expected to directly be used as network components. For example, a generic OSI node with a structure as in Fig. 4 may be used as a template to create specialized node structures. The OSI node template is a detailed layer-by-layer model of an OSI structured node with capabilities for multiple incoming and outgoing links with separate transmitters/receivers and queues. Flexibility is provided

to implement various Data Link Control (DLC – Layer-2) protocols, routing algorithms (Layer-3), end-to-end transport layer (Layer-4) protocols and other higher level protocols. Also provided is a model of a node processing resource with provisions for modeling multiple CPUs and I/O resources and different job-queuing strategies like First-Come-First-Served, Round-Robin, Shortest-Job-First, e.t.c., and prioritized versions of these.

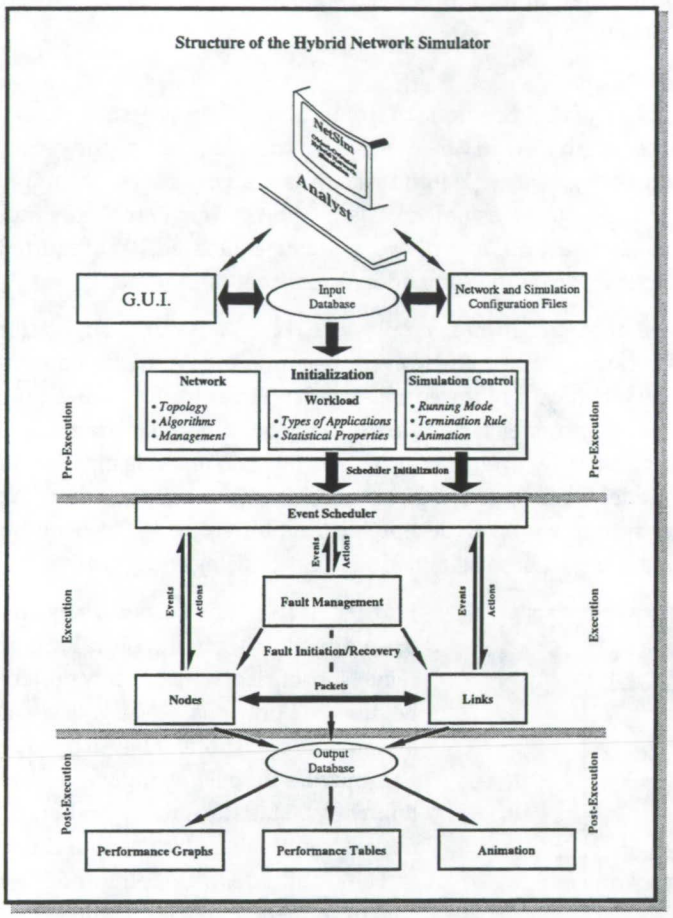


Fig. 5. Temporal View of the Simulation.

Conceptually, the course of a simulation proceeds in three temporal phases, (a) the pre-execution phase, (b) the execution phase, and (c) the post-execution phase. (See Fig. 5.) During the pre-execution phase, the user interacts with the simulation software (either through the GUI, or through files) to create a network scenario to be simulated. During the execution phase, the network is simulated: source applications generate packets of data, which are transported across the network (through links and nodes) and delivered to sink applications. As the simulation proceeds, various network performance statistics are generated and stored in an object-oriented database. As part of the post-execution phase, this database is accessed, and the desired network performance parameters are computed and presented to the user (in the form of performance graphs and tables). As a visual verification of the correctness of the simulation, the user is able to view an animation of the simulation, with the key transactions (like transfer of packets) being displayed dynamically on the GUI.

The Expansion Module Library

The basic module library within the simulation kernel consists of a set of basic components that may be used to build a network component. For example, a specific kind of node could be constructed starting from the generic node and adding transmitters, receivers, queues e.t.c., as required. Unlike the modules within the basic module library, the expansion module library consists of a collection of modules that can directly be used to construct networks. The expansion modules also provide additional functionality. For example, a digital geographical database module would be used to provide map functions; a population module to provide a model of the global population, e.t.c..

Another major difference between the basic module library and the expansion modules library is that while the basic module library is fixed, modules in the expansion module library can be added, deleted, or even changed. At the time of dispatch, the simulation is preconfigured with a collection of modules in the expansion module library. Depending on the needs of users, an application programmer may customize the expansion module library through the API. Users who then wish to simulate a particular communication network will find the exact building blocks that they need in the expansion module library.

The Database

The inclusion of an efficient object-oriented database as an integral component of the simulation software is of vital significance. Vast quantities of information are required to be stored and manipulated as part of the simulation. Online documentation, help messages, test network scenario storage and performance data generated during the course of a simulation run all need to be stored and accessed with speed. The programming effort is greatly reduced through the use of an object-oriented database that specializes in the task of data handling. Equally important is

the role of the database as an interface between the different components of the simulation. As a test scenario is created, the information is organized and stored in the database. When simulation is initiated, this information is retrieved to construct a network and simulate it. As simulation proceeds, performance data is again stored in the database. A network-management software accessing the simulator through the software interface would retrieve this information and use it as input to come up with network management strategies, and perhaps dynamically change the network configuration.

SOME TYPICAL APPLICATIONS

In this section, we describe two typical applications of the object-oriented distributed hybrid network simulation software. The first application tries to highlight some of the network visualization capabilities, which have been covered only sketchily in previous sections. The second application is an illustration of how the software may be used as a tool for network management.

Satellite Constellation Network Simulation

With proposals to provide global communication coverage (telephone/cellular) through a constellation of low-earth orbit (LEO) and medium-earth orbit (MEO) satellites arranged in multiple orbits around the globe, e.g., Iridium, Teledesic and Globalstar, the simulation of satellite networks has suddenly gained prominence. Because of the dynamic nature of satellite networks, both visualization and simulation of such networks presents unique problems. Recognizing their importance, the hybrid network simulator provides a special interface to handle the simulation and visualization of satellites and satellite constellations and their relationship with the earth. The specialized satellite constellation GUI provided by the software incorporates (i) definition of multiple orbits and the positioning of multiple satellites within each orbit; (ii) definition of satellite and earth-station communications parameters; (iii) visualization of the satellite constellation – logically (showing connectivity) and spatially against the globe and on a flat map; (iv) selecting the visual context with respect to a satellite, a point on the globe, or some inertial frame of reference; (v) visualization of end-to-end connection paths, and associated path metrics; (vi) definition and visualization of cell patterns and antenna beam patterns associated with a satellite as it moves around the earth; and (vii) incorporation and visualization of climatic activity and population behavior models.

Fig. 6 shows how some of this information is presented on the GUI. Multiple visualization models are required to effectively visualize the network; for every piece of information, there is a model of visualization that represents it best. In Fig. 6, only the flat-map visualization is shown. The figure shows a communication path between Australia and California over the proposed Teledesic network, and illustrates how the communication path crosses the constellation seam. The path itself is depicted as a bold line, and the shaded ovals represent each satellite's computed footprint (assuming a minimum antenna elevation angle of 40°) projected on the flat map. Other information like beam patterns, connectivity and relative satellite positions can also be effectively visualized on the flat map.

Other visualizations of the satellite constellation network, i.e., a globe-based 3D visualization and a logical view showing communication path and satellite performance related statistics are also available (not shown) to more effectively visualize the network.

Additionally, Fig. 6 also illustrates symbolically, how a network may be assembled using nodes and links and also shows how performance-related data may be viewed dynamically. Finally, the figure also shows an actual distributed image application running on the simulated network (using the software interface feature).

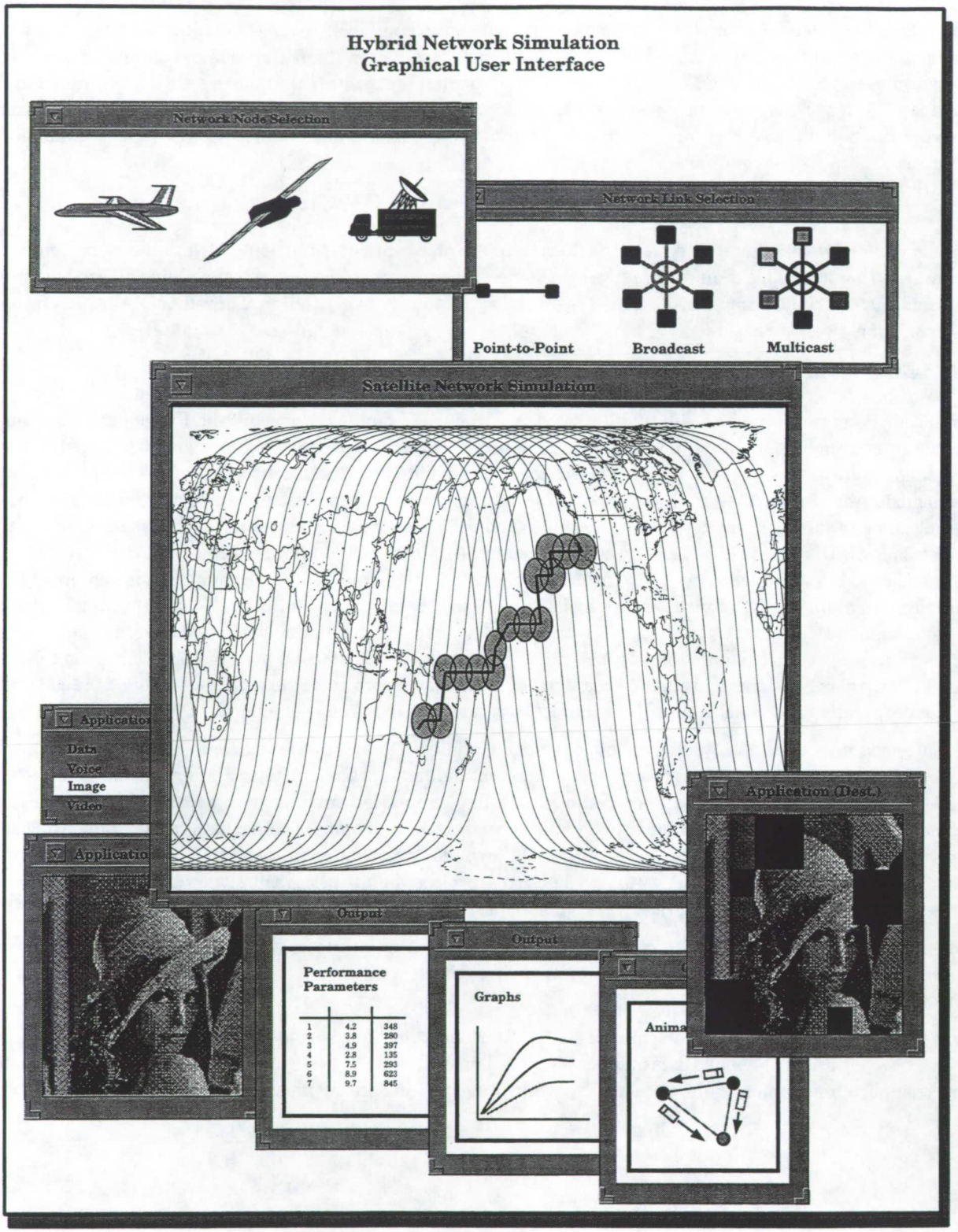
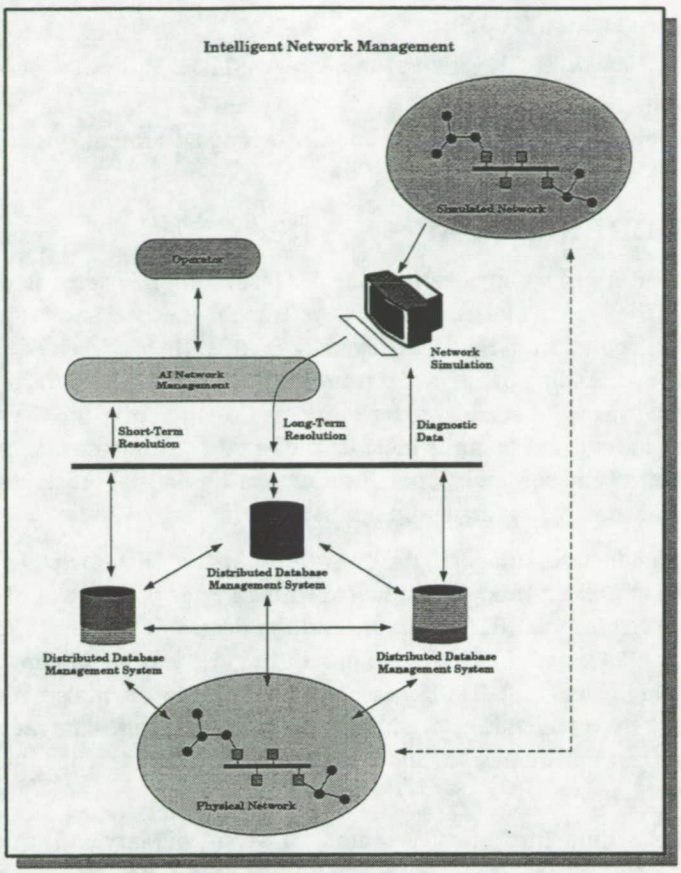


Fig. 6. Visualization of a Satellite Constellation and related parameters.



Intelligent Network Management

The flexible interface and the incorporation of a database prepare the hybrid network simulation software as a tool for network management in a closed loop with a network management software. (See Fig. 7.) In a typical network, network monitor functions gather data on network performance, and store it in distributed databases. Some of this data is used for short-term network management and fault resolution directly, using dedicated network management functions. Additionally, as part of the intelligent management of the network, this data is used to drive the simulation (which runs in parallel with the actual network and mimics its behavior). The long-term behavior of the simulated network now acts as a predictor of the long-term behavior of the network, and may be used to better manage the long-tem network performance. The effect of particular network policies can be safely judged by testing them on the simulated network, before they are applied to the running network. For example, in current work, we are developing fast versions of the simulation to provide online feedback advise for hybrid network configuration and fault management.

Fig. 7. Hybrid Network Simulation Software as a Network Management Tool

SUMMARY

This paper identified and discussed key issues involves in the simulation of complex hybrid communication networks, and presented the design features of an object-oriented hybrid communication software being developed at the University of Maryland, Institute for Systems Research. It was shown that conventional simulation paradigms: time-based simulation and event-based simulation, are unsuitable for complex system simulation; rather a model based simulation approach is adapted. It is also shown how the flexible interface design and incorporation of a database enable the software to be used as part of an intelligent network management tool.

BIBLIOGRAPHY

[1] Application of Teledesic Corporation for a Low Earth Orbit Satellite System in the Fixed Satellite Service, FCC File No. _DSS-P/L-94, Teledesic Corp., 5220 Carillon Point, Kirkland, Washington 98033, March 21, 1994.

[2] BONeS DESIGNER User's Guide, Version 2.5, Comdisco System Inc., FosterCity, CA, 1992

[3] GPSS/H User's Manual, Third Edition, Wolverine Software Corp., Annandale, VA, 1989.

[4] MODSIM II Reference Manual, CACI Products Co., La Jolla, CA, 1993.

[5] SIMSCRIPT II.5 Reference Handbook, CACI Products Co., La Jolla, CA, 1993.

[6] SES/workbench User's Manual, Release 2.1, Scientific and Engineering Software, Inc., Austin, TX, 1992.

[7] SIMAN V Reference Guide, Systems Modeling Corp., Sewickly, PA, 1993.

[8] SLAMSYSTEM User's Guide, Version 4.0, Pritsker Corp., Indiannapolis, IN, 1992.

[9] BONeS PlanNet User's Manual, Version 1.0, Comdisco System Inc., FosterCity, CA, 1992

[10] COMNET III User's Manual, Release 1.0, CACI Products Co., La Jolla, CA, 1993.

[11] Network II.5 User's Manual, Release 9.0, CACI Products Co., La Jolla, CA, 1993.

[12] L•NET II.5 User's Manual, CACI Products Co., La Jolla, CA, 1993.

[13] Simulation Software for Communications Networks: The State of the Art, Averill M. Law and Michael G. McComas, IEEE Communications Magazine, March 1994.

[14] Object Oriented Hybrid Network Simulation, A Functional Specification, Working Document, Institute for Systems Research, University of Maryland, College Park, MD.

AUTHOR BIOGRAPHIES

John S. Baras is a Professor in the Department of Electrical Engineering and the Institute for Systems Research, University of Maryland at College Park. He holds the Martin Marietta Chair in Systems Engineering and is a Fellow of the IEEE. He received his Ph.D. in Applied Mathematics from Harvard University in 1973. He has received several research awards for his work including the 1980 outstanding paper award from the IEEE Control Systems Society. He was the Founding Director of the Institute for Systems Research and the Founding Co-Director of the Center for Satellite and Hybrid Communication Networks. His current interests include intelligent control systems, control and systems theory, computer-aided design of control and communication systems, real-time signal processing and understanding, and computer/communication network management and simulation.

Ahsun H. Murad is a Faculty Research Assistant and doctoral student in the Institute for Systems Research/Electrical Engineering Department, University of Maryland at College Park. He graduated with a degree of Bachelor of Technology from the Indian Institute of Technology, Delhi, India (IITD) in 1989, and the degree of Master of Science from The University of Maryland, College Park in 1992, both in Electrical Engineering. He was a recipient of the Bambawala Award for the best project in Electrical Engineering, IITD for his design and hardware implementation of a high-speed HF modem in 1989. In addition to his research in error-control coding for magnetic data recording and distributed error-control schemes, his current research interests include the design and development of tools for hybrid communication network simulation and management.

Kap Do Jang is a Faculty Research Assistant in the Institute for Systems Research, University of Maryland at College Park. He graduated with a Bachelor of Arts in Economics from Sogang University, Seoul, Korea in 1987 and a Bachelor of Science in Computer Science from the University of Maryland at College Park in 1992. His current interests include object oriented design and programming, graphical user interfaces, and object-oriented database design.

George C. Atallah is a Senior Member of Technical Staff in the Institute for Systems Research, University of Maryland at College Park and a former Westinghouse and IBM research and development scientist. He received his Ph.D. in Systems Science and Applied Mathematics from the State University of New York at Binghamton. His current interests include advanced computer applications and architectures, neural networks, complex systems modeling and analysis, hybrid communication network simulation and management, nonlinear systems dynamics and control, systems reliability, maintainability and availability analysis, and systems/concurrent engineering methodology and management.

Ramesh Karne is a Senior Member of Technical Staff in the Institute for Systems Research, University of Maryland at College Park since October, 1993. Before joining the Institute, he was at IBM corporation for over 15 years and was involved with large computer system design, expert system applications and software development. He received the M.S. degree from the University of Wisconsin in 1978, and a Ph.D. in Computer Science from George Mason University in 1992. His research interests include object-oriented technology, artificial intelligence applications, distributed operating systems, communication systems, parallel and distributed simulation and application oriented architectures.

S. Joseph Campanella is an Adjunct Professor, Electrical Engineering, and Senior Research Scientist in the Institute for Systems Research at the University of Maryland. He retired from COMSAT Laboratories after 26 years of research and development in all aspects of commercial satellite communications. He had held the positions of Executive Director of the Communication Laboratory, V.P. and Assistant Director of COMSAT Laboratories and V.P. and Chief Scientist of COMSAT Laboratories. His current research interests include all aspects of satellite communications, hybrid network architectures and new satellites with on-board processing capabilities.

Materials Science

Low Cost Fabrication of Silicon Carbide Based Ceramics and Fiber Reinforced Composites

M. Singh*
NYMA, Inc.
2001 Aerospace Parkway
Brookpark, Ohio 44142

S.R. Levine
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, OH 44135

ABSTRACT

A low cost processing technique called reaction forming for the fabrication of near-net and complex shaped components of silicon carbide based ceramics and composites is presented. This process consists of the production of a microporous carbon preform and subsequent infiltration with liquid silicon or silicon-refractory metal alloys. The microporous preforms are made by the pyrolysis of a polymerized resin mixture with very good control of pore volume and pore size thereby yielding materials with tailorable microstructure and composition. Mechanical properties (elastic modulus, flexural strength, and fracture toughness) of reaction-formed silicon carbide ceramics are presented. This processing approach is suitable for various kinds of reinforcements such as whiskers, particulates, fibers (tows, weaves, and filaments) and 3-D architectures. This approach has also been used to fabricate continuous silicon carbide fiber reinforced ceramic composites (CFCCs) with silicon carbide based matrices. Strong and tough composites with tailorable matrix microstructure and composition have been obtained. Microstructure and thermomechanical properties of a silicon carbide (SCS-6) fiber reinforced reaction-formed silicon carbide matrix composites are discussed.

INTRODUCTION

Silicon carbide based advanced ceramics and fiber reinforced composites are leading candidate materials for a number of applications in aeronautics, energy, electronics, nuclear, and transportation industries [1-3]. In the aeronautical arena, these materials are being considered for applications in hot sections of jet engines such as the combustor liner of the high speed civil transport (HSCT). Rocket nozzles and rentry thermal protection systems are among other potential aerospace applications. Applications in the energy industries include radiant heater tubes, heat exchangers, heat recuperators, and components for land based turbines for power generation. These materials are also being considered for use in the first wall and blanket components of fusion reactors, in furnace linings, and bricks, and are being used as components for diffusion furnitures (boats, tubes) in the microelectronics industry.

There are a number of critical issues related to commercially available silicon carbide

* Mailing address: MS 106-5, NASA Lewis Research Center, Cleveland, OH 44135

ceramics and fiber reinforced composite materials. Fabrication by current processing techniques of complex shape components of near net shape is very expensive, and even then many desirable properties are not achievable. Typical problems associated with reaction bonding are the excessive amount of free silicon (~15-40 vol%) and presence of sintering aids in some sintered materials. The coarsely distributed free silicon and sintering aids in commercially available silicon carbides degrade high temperature properties. On the other hand, fiber reinforced silicon carbide matrix composites available commercially at present are produced by the chemical vapor infiltration (CVI) process. The CVI processing time is typically hundreds of hours, with multiple process interruptions to machine off the surface case in order to expose the interior to further infiltration. The final material contains about 10-15% residual porosity. The long processing time and intermediate machining steps contribute to the high cost of components made by the CVI process.

At NASA Lewis, a low cost processing technique for the fabrication of silicon carbide-based advanced ceramics and composites has been developed which requires significantly less fabrication times and/or processing temperatures than current processing techniques. The products produced by this process are fully dense, have tailorable microstructure and mechanical properties, and can incorporate second phases without the detrimental effects of unwanted sintering aids [4-10]. This process consists of the production of a microporous carbon preform and its subsequent infiltration with liquid silicon or silicon-refractory metal alloys. The microporous preforms are made by the pyrolysis of a polymerized resin mixture. Pore volume and pore size are carefully tailored to control the size and distribution of the final constituents [5-8]. This process produces silicon carbide based ceramics at low cost and has near-net and complex shape capabilities.

This processing approach has been used to fabricate ceramic matrix composites with silicon carbide based matrices and is suitable for various kinds of reinforcements such as whiskers, particulates and continuous fibers (tows, weaves, and filaments). Various types of fiber reinforcements i.e., carbon, alumina and silicon carbide has been used as reinforcements in this process. It can also be used to fabricate composites with 3-D architectures. Key properties such as strength and toughness, creep, and environmental and thermal shock resistance of these materials can be tailored.

The work reported here is part of a program aimed at developing low cost, near-net and complex shape components of reaction-formed silicon carbide based ceramics and composites with tailorable microstructure and thermomechanical properties. In this paper, fabrication process and advantages of using these materials in various applications will be presented. Mechanical properties (elastic modulus, flexural strength, and fracture toughness) of three reaction formed silicon carbide materials will be discussed. The microstructure and mechanical behavior of silicon carbide (SCS-6) fiber reinforced composites fabricated by this process will also be presented.

PROCESS DESCRIPTION

A schematic description of the reaction forming of silicon carbide based ceramics and composites is given in Fig. 1. This process consists of the production of a microporous carbon preform and its subsequent infiltration with liquid silicon or silicon-refractory metal alloys [4-8]. The microporous preforms are made by the pyrolysis of a polymerized resin mixture. The pyrolysis process affords excellent control of pore volume, pore size and distribution. The silicon or silicon-refractory metal alloy infiltration of the preforms and reaction of liquid silicon with carbon, yields silicon carbide based ceramics with tailorable microstructure and composition. Alloying silicon with

refractory metals reduces the amount of residual silicon phase. Materials with controllable amount of silicon or other phases (4-40%) can be obtained.

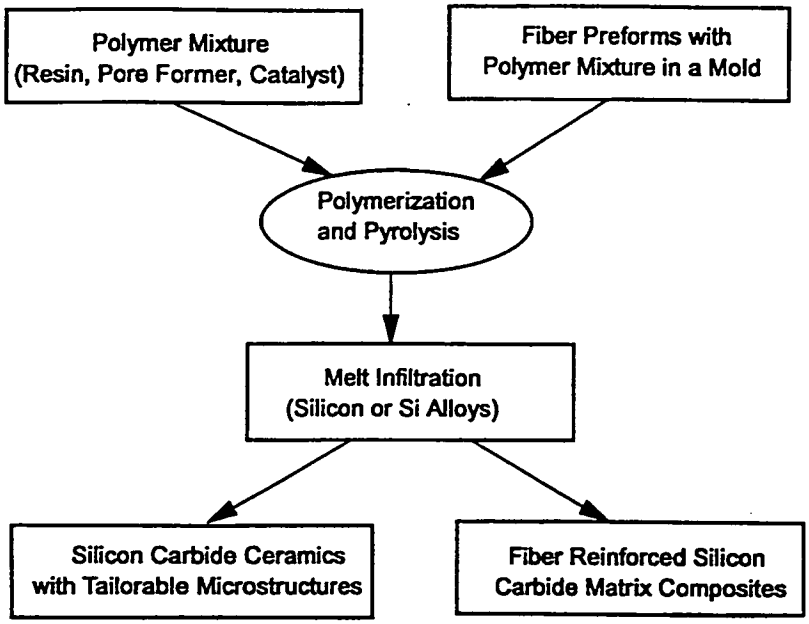


Fig. 1 : Process flow chart for the fabrication of silicon carbide based ceramics and fiber reinforced composites.

The precursor materials (resin, pore formers, etc.) for microporous carbon preforms are enexpensive ($2/lb) and can be handled under normal laboratory conditions without any special equipment. Silicon in lump form is also enexpensive e.g., 6N pure electronic grade silicon lumps can be bought for $10-12/lb. If metallurgical grade or other type of silicon is used for infiltration, the final material cost will be further reduced.

All the mechanical property data reported in this paper are for ceramics formed by the reactive infiltration in to identical microporous carbon preforms of molten silicon (NSC-20), silicon-1.7 at% molybdenum (NSC-21) and silicon-5 at% niobium (NSC-22) alloys. The details of microporous carbon preform fabrication and other infiltration conditions have been described elsewhere [5-8]. The control of processing conditions is critical to avoid microstructural coarsening, silicon vein and lake formations, and cracking due to thermal expansion mismatch and volume change. After infiltration, samples were cross-sectioned and polished for metallographic studies. Powder x-ray diffraction analysis was used to identify different phases in the reaction-formed materials.

TAILORABLE SILICON CARBIDE CERAMICS

Microstructure

An optical micrograph of the NSC-20 material, fabricated by silicon infiltration of carbon preforms, is given in Fig. 2. Complete conversion of carbon to silicon carbide and the uniform distribution of a silicon phase (white) throughout a silicon carbide matrix (grey) are evident in Fig. 2. Similar microstructures were obtained for the silicon-1.7 at% molybdenum and silicon-5 at%

niobium alloy infiltrated materials. The free silicon content of the NSC-20 material was ~ 10 %, significantly lower than the silicon content (~ 20-30 %) of commercially available reaction-bonded silicon carbide materials. The amount of residual silicon in the silicon-alloy infiltrated materials is less than that in the material made using silicon infiltration due to the formation of refractory disilicides.

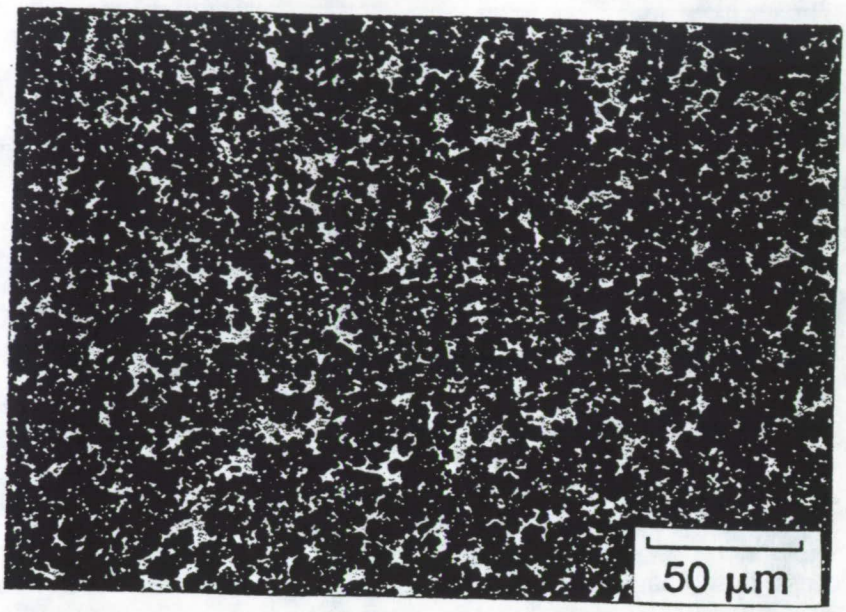


Fig. 2 : Microstructure of reaction-formed silicon carbide (NSC-20) (Si is white and SiC is grey)

Mechanical Properties

Elastic Modulus

Elastic moduli data for reaction-formed silicon carbide ceramics along with some commercial silicon carbides are given in Fig. 3. Moduli data for the commercial silicon carbides were taken from the literature [12-13]. The elastic modulus of silicon containing silicon carbide material (NSC-20) is 360 GPa which is comparable to REFEL-1 and Coors SC-2, but lower than NC-203 (450 GPa). There is no significant difference in the modulus of NSC-20 (360 GPa) and NSC-21 (347 GPa), which may be due to the low amount of molybdenum disilicide in the NSC-21 material. In addition, the modulus of molybdenum disilicide is comparable to that of silicon carbide (427 GPa). The modulus value of NSC-22 is 292 GPa, somewhat lower, probably due to presence of niobium disilicide (300 GPa) and silicon.

Flexural Strengths

The flexural strengths of silicon infiltrated reaction-formed material (NSC-20) along with two commercial silicon carbides as a function of temperature are presented in Fig. 4. The average room temperature strength of silicon infiltrated material (NSC-20) is 371 $\pm$ 28 MPa. The data points represent mean results for ten tests at room temperature and five tests at elevated temperatures. In general, flexural strengths of NSC-20 are higher than two commercial silicon carbides. The apparent increase in strength at 1100°C may be due to healing of machining damage. Further experimental

studies are in progress to confirm this hypothesis.

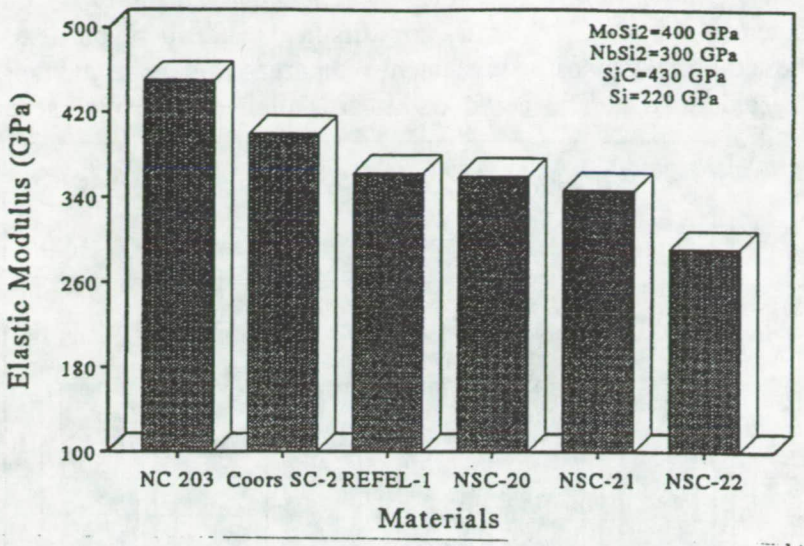


Fig. 3 : Elastic properties of reaction-formed silicon carbide ceramics and other commercial materials.

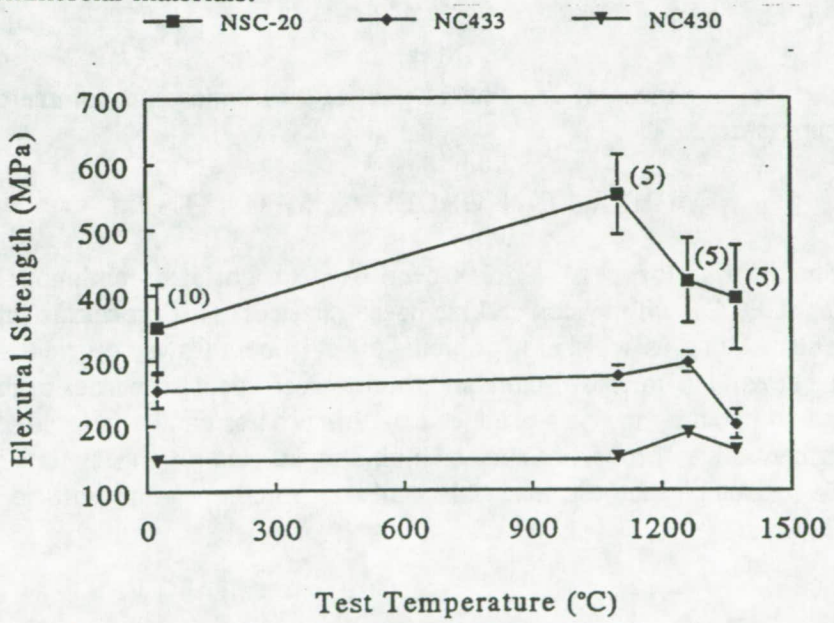


Fig. 4 : Flexural strengths of reaction-formed SiC (NSC-20) and two commercial SiC as a function of temperature.

Fracture Toughness

The room temperature fracture toughness of these materials, determined using the single-edge-precracked-beam (SEPB) method, is given in Fig. 5. The fracture toughness of silicon-5 at% niobium alloy infiltrated material (NSC-22) is 3.7 $\pm$ 0.3 MPa $m^{1/2}$ while for the silicon infiltrated material (NSC-20) it is 2.5 $\pm$ 0.2 MPa $m^{1/2}$. This difference in the fracture toughness of two ceramics can be explained on the basis of the thermal expansion coefficients of the different phases. The toughening mechanism in the silicon-alloy infiltrated material may be crack deflection by thermal

residual stress field which develops around refractory disilicide and silicon particles upon cooling due to thermal expansion mismatch [12]. The thermal expansion coefficients of molybdenum disilicide (8.2 x 10^{-6}/°C), niobium disilicide (11.7 x 10^{-6}/°C) and silicon (7.63 x 10^{-6}/°C) are higher than β-silicon carbide (4.4 x 10^{-6}/°C), the refractory disilicide and silicon particles are under isostatic tensile stresses. These thermal expansion mismatches impart radial tensile and tangential compressive stresses to silicon carbide matrix. This could lead to crack deflection around second phase particles.

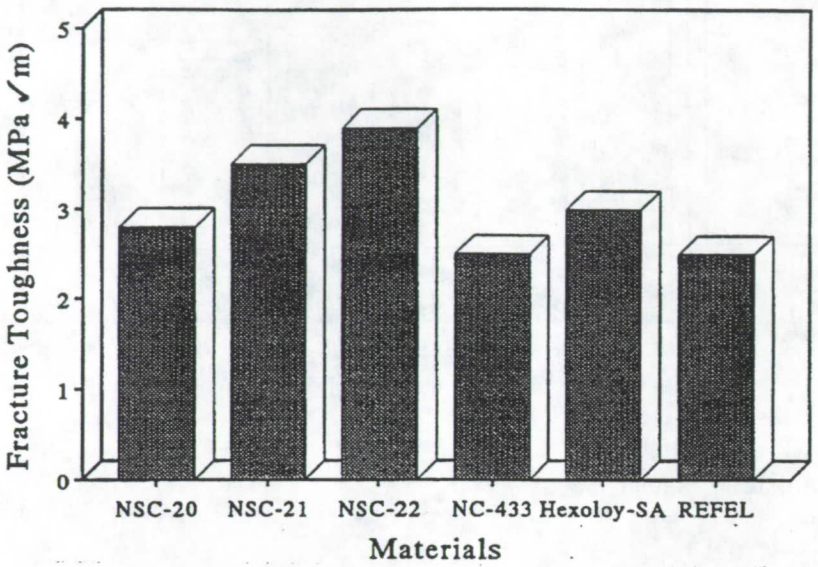


Fig. 5 : Fracture toughness of reaction-formed SiC ceramics and commercial SiC at room temperature.

FIBER REINFORCED COMPOSITES

The reaction forming process has also been used to fabricate continuous fiber reinforced ceramic composites (CFCCs) with silicon carbide based matrices. This process is suitable for various kinds of reinforcements such as whiskers, particulates and fibers (tows, weaves, and filaments) as well as 3-D architectures. The microstructure and thermomechanical properties of the silicon carbide based ceramics and composite materials produced by this process can be tailored to suit design and application requirements. Key properties such as high strength and toughness, creep, environmental and thermal shock resistance of these materials can also be tailored and will be discussed in the following sections.

Microstructure

The silicon carbide fiber-reinforced silicon carbide matrix composites produced by the reaction forming process are fully dense. The liquid silicon or silicon alloy infiltration time is relatively short (15 minutes to one hour) in contast to the long times (several days) required for the chemical vapor infiltration process. An optical micrograph of silicon carbide (SCS-6) fiber reinforced -silicon carbide matrix composites is given in Fig. 6. This micrograph shows uniform microstructure with no unreacted carbon or porosity. One important point to note here is the protection afforded by the coating on as received SCS-6 fibers. Various investigators have used 4-5 μm barrier coatings to prevent molten silicon attack on these fibers. These additional coating applications complicate processing and handling, and add extra cost to composite manufacture. Dense composites have also been obtained with small diameter woven fiber composites.

Mechanical Properties

The stress-displacement behavior of a SiC(SCS-6) fiber-SiC matrix composite is given in Fig. 7. This composite has been tested in four-point bending at room temperature. The unidirectional composites with ~17 percent by volume of fibers have a first matrix cracking stress above 262 MPa (38 ksi). The load-displacement curve shows a non- linear behavior with a graceful failure. The interfacial shear strength of this composite is in the range of 30-60 MPa depending on the processing time, temperature and the infiltrant composition. Detailed thermomechanical and thermochemical characterization of these composites under the hostile environments to be encountered in engine applications is underway.

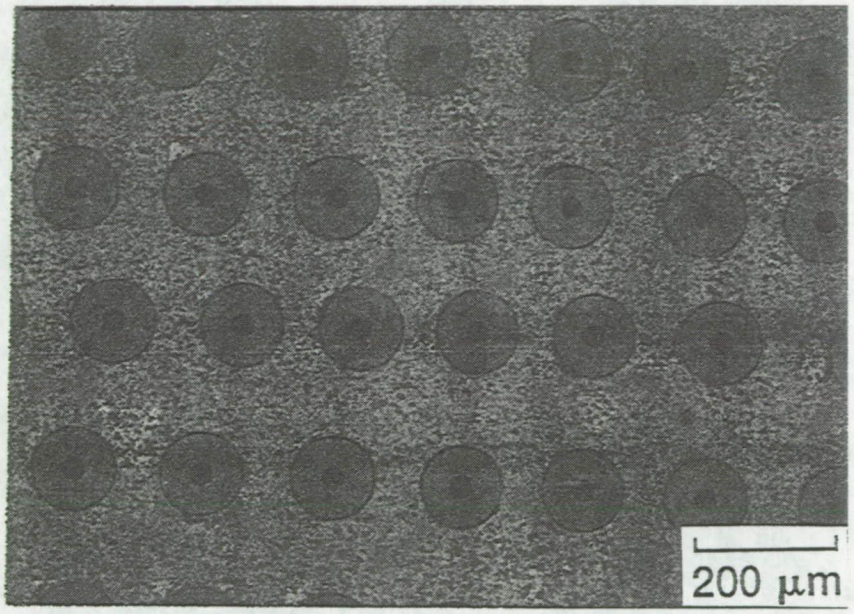


Fig. 6 : Optical Micrograph of a SiC (SCS-6) /SiC matrix composite.

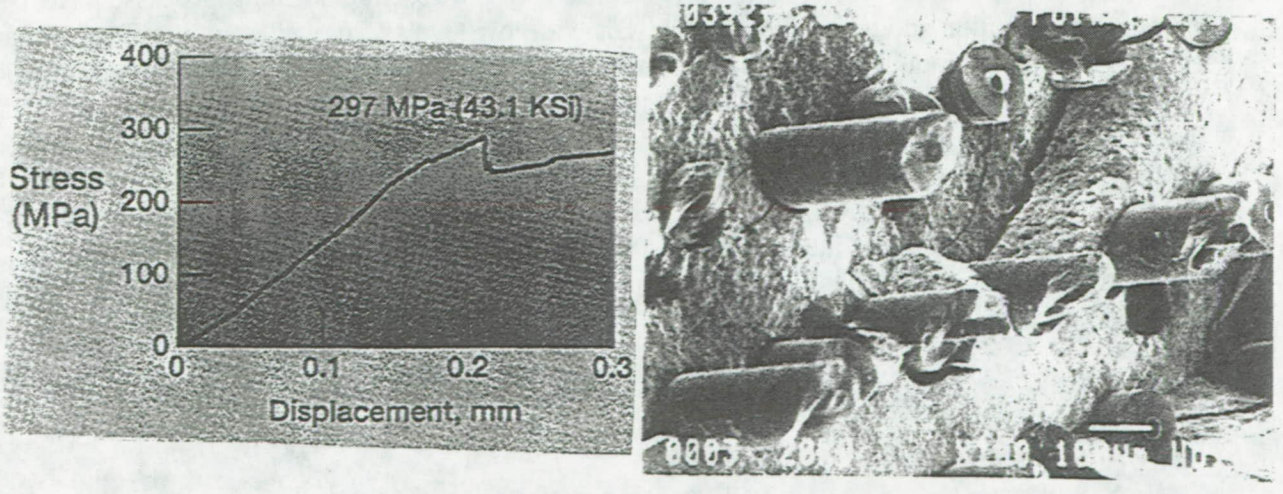


Fig. 7 : Stress-displacement behavior and fracture surface of SiC(SCS-6) fiber-SiC matrix composites.

Thermal Properties

Unidirectional composites made with SiC(SCS-6) fibers have also been tested in burner-rig thermal shock tests. In this test, a hot flame is impinged on the surface of a 2"X3" composite and the temperature at the back surface is measured by pyrometer. The data for a variety of materials and the temperature difference between the front and back surface is given in Fig. 8. The NCC-20 material made by the process described above has the lowest ΔT indicating highest thermal conductivity.

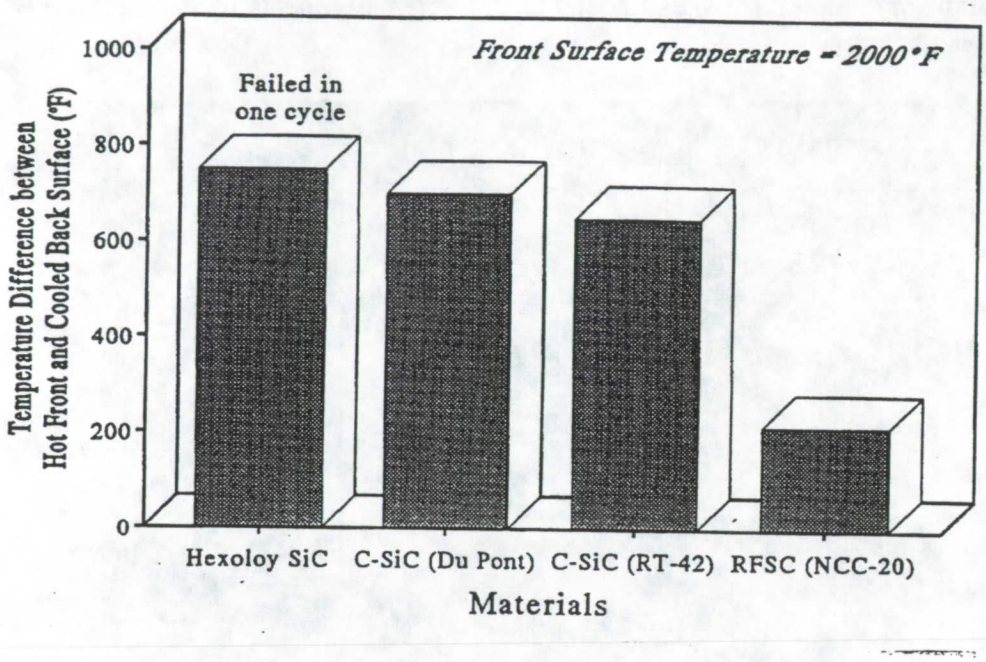


Fig. 8 : Thermal properties of SiC(SCS-6)/SiC composites under burner rig tests.

ADVANTAGES

The silicon carbide based ceramics, produced by the reaction forming process, have very good phase distribution and morphology. Materials with low levels of free silicon (<4-5 vol%) have been produced. The amount of free silicon can be further reduced by silicon-refractory metal alloy (molybdenum or niobium) infiltrations. This process is low cost with near-net and complex shape capabilities.

The silicon carbide fiber-reinforced silicon carbide matrix composites produced by this process are fully dense. The liquid silicon or silicon alloy infiltration time is relatively short (15 minutes to one hour) rather than the long times (several days) required for the chemical vapor infiltration process.

NEAR TERM AND FUTURE APPLICATIONS

Aerospace Applications

The principal application of this technique is to fabricate near-net and complex shaped silicon carbide based ceramic and composite components for aeronautics and space applications. It is

expected that these materials will find applications in the hot section of jet engines; for example, the combustor liner of the proposed high speed civil transport (HSCT). The combustor liner requires a material which is non-porous, has high thermal conductivity and is resistant to degradation in both rich- and lean-burn combustion environments. Another area of application for these silicon carbide based materials is in the thermal protection systems of reentry vehicles and high speed aircraft, where materials that can withstand the high temperatures and thermal gradients generated by the high speeds are required.

The silicon carbide based multiphase ceramics containing different phases such as $MoSi_2$, $NbSi_2$ and others are candidate material for space shielding applications. These materials can outperform conventional armor materials. There is considerable interest in light weight shields to protect spacecraft and bases from debris and micrometeorites for the Space Station and other proposed manned missions back to Moon and on to Mars.

Terrestrial Applications

Silicon carbide based ceramic and composite materials fabricated using this technique will have a variety of applications in the energy industries. These applications include radiant heater tubes, heat exchangers, heat recuperators, various components for land based turbines for power generation, transportation (automobiles and trucks), and energy conversion devices. The excellent oxidation resistance of silicon carbide based materials offer alternative possibilities in the low temperature range as condensing heat exchangers. The main stumbling block for these applications has been the high cost of components fabricated to near-net and complex shape.

In nuclear industries, these materials are under consideration for use in the first wall and blanket components of a fusion reactor. Conventionally sintered silicon carbide materials, apart from being costly, contain sintering aids such as boron which lead to volume swelling upon irradiation. On the other hand, commercially available SiC/SiC composites, produced by CVI, contain 15-20% of residual porosity and are quite expensive. The materials produced by the process reported here contain no sintering aids. SiC/SiC composites have been produced with virtually no porosity (theoretically dense). The application of reaction-formed silicon carbide materials in the high radiation environment of fusion reactors can provide a very low activation system. This will significantly reduce waste disposal and maintenance problems and environmental and safety concerns.

Conventional monolithic silicon carbide ceramics need high sintering temperatures which drives their cost up. For their toughness improvement, various phases have to be hot pressed together. In this technique, second phases for toughening have been produced in-situ. The improved fracture toughness along with other properties makes these materials a strong candidate for armor applications. In addition, the near-net shape capabilities of this process can be used to produce materials for cutting tool applications.

Due to its high strength, excellent thermal shock resistance, and good oxidation resistance, silicon carbide based ceramics are being considered for furnace linings, bricks and other components in furnaces. Another area where low cost silicon carbide materials are desired is in the aluminum industry for reduction cells. The major requirements for the refractories for these applications are their structural and chemical stability over the entire cathode temperature range during cell start-up and operations. Silicon carbide based ceramics are ideal materials due to their excellent resistance

to air oxidation, thermal shock tolerance and good resistance to molten salts.

CONCLUSIONS

A method to produce silicon carbide-based high performance ceramics and composites which requires fabrication times and/or processing temperatures significantly less than other processing techniques such as chemical vapor infiltration, reaction sintering, pressureless sintering, hot pressing, and hot isostatic pressing has been developed at NASA Lewis. The products produced by this process are fully dense, with excellent control of microstructure, second phases and mechanical properties and without the detrimental effects of sintering aids. In addition, the materials produced by this process are low cost.

The near-net and complex shape capabilities, low processing time and temperature, and low cost will make the materials produced by this process affordable in a variety of applications. The advantages and benefits of using these materials for various components in key industries such as aeronautics, energy, automotive, and nuclear can have a significant impact on U.S. competitiveness in the world.

REFERENCES

[1] D.C.. Larsen , J. Adams, L. Johnson, A. Teotia, and L. Hill, "Ceramic Materials for Heat Engines", Noyes Publications, NJ (1985).

[2] E. Fitzer and R. Gadow, Am. Ceram. Soc. Bull., 65, 2 (1986) 325-35.

[3] E.E. Hucke, AMMRC Report TR-83-5, January 1983.

[4] D.R. Behrendt, M. Singh and R.F. Dacek, 16th Conf. on Metal, Carbon and Ceramic Matrix Composites, NASA CP- 3175 (1992) 695-705.

[5] M. Singh and D.R. Behrendt, NASA-TM- 105860 (1992).

[6] M. Singh and D.R. Behrendt, J. Mater. Res., 9, 7 (1994) 1701-1708.

[7] M. Singh and D.R. Behrendt, Mater. Sci. Engg. (1994) in press.

[8] M. Singh and D.R. Behrendt, Mater. Sci. and Engg. (1994) in press.

[9] D.R. Behrendt and M. Singh, J. Mater. Synth. Process., 2, 2 (1994) 133-139.

[10] D.R. Behrendt and M. Singh, NASA-TM-106414 (1993).

[11] Y-M. Chiang, R.P. Messner, C. Terwilliger, and D.R. Behrendt, Mater. Sci. Engg. A 144, (1991) 63-74.

[12] M. Singh, R. Pawlik, J.A. Salem and D.R. Behrendt, Ceramic Trans. Vol 38, 'Advances in Ceramic Matrix Composites', Edited by N.P. Bansal, The American Ceramic Society, Westerville, OH (1993) 349-360.

[13] Ceramic Source, Vol. 7, American Ceramic Society, Westerville, OH (1991) 313.

HIGH TEMPERATURE PHTHALONITRILE RESINS/COMPOSITES

Teddy M. Keller*, James P. Armistead and Satya B. Sastri

Materials Chemistry Branch
Naval Research Laboratory, Code 6127
Washington, DC 20375-5320

ABSTRACT

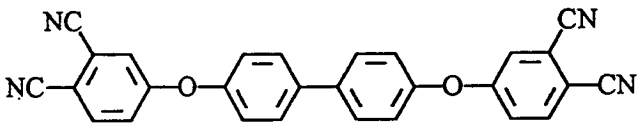
A major technological opportunity lies in the usage of phthalonitrile resins for composite and electronic systems. The phthalonitrile resins are easy to process into void-free components, exhibit high thermal and oxidative stability approaching 372°C (700°F) in air, have low equilibrium moisture absorptivity, and can be made to exhibit controlled electrical conductivity for potential radar absorbing material (RAM) applications and as anti-static materials. Moreover, phthalonitrile prepreg shows indefinite stability at room temperature, which enhances their importance in the fabrication of composite components. Additionally, the cure exotherm is easily controlled for thick composite fabrication. The resin and composite display outstanding mechanical properties with property retention after extended aging at 372°C.

INTRODUCTION

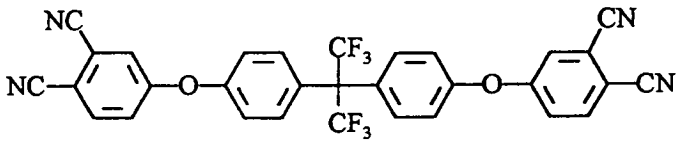
Phthalonitrile-based polymers under development at the Naval Research Laboratory offer promise of bridging the gap between currently used high temperature polymers and ceramics/metals in weight critical aerospace and advanced marine applications. High performance organic polymers, which can withstand temperatures in excess of 300°C (572°F), could bridge the gap between currently used high temperature polymers and metals/ceramics. The high density of steel severely limits its use in aerospace applications, and while titanium is less dense, it is expensive and costly to process. Thus, replacement of these metals by organic polymeric composites would effect substantial weight, cost, and energy savings, and improve performance. Presently, epoxies and polyimides are being heavily used for composite applications but each has disadvantages. Conventional epoxy-based composites and adhesives are limited to 120°C maximum service temperature. Other problems associated with these polymers include their brittleness and water absorptivity. A major problem of the polyimide systems is the inability to process void- and blister-free components in high yield. These process problems, attributed to the evolution of volatile components formed during the polymerization condensation reaction, must be alleviated before these materials will be widely accepted for high temperature applications. In addition to their limited temperature capability, conventional composite matrices suffer such shortcomings as complicated logistics of handling due to low temperature storage requirements of the prepregs, poor processability, brittleness, significant water absorption problems causing a reduction in the glass transition temperature (T_g), and delamination resulting from water penetration into the interface between the matrix material and the reinforcing fiber.

Considerable advances have been made in recent years concerning the synthesis and development of thermally stable polymers. The PMR polymers[1,2] have found numerous applications in the aerospace industry. The aliphatic cross-linking moiety, which is the weak linkage for thermal degradation in these resins, limits the utility of these polymers for high temperature applications. The initial weight loss in the thermo-oxidative degradation of PMR-15 has been attributed to both the nadic- and diphenylmethane-components with the nadic cross-linking unit ultimately becoming the more significant contributor to overall resin loss.[3]

The chemical principles used in the successful design of the phthalonitrile resins appear to have yielded a breakthrough in the desired performance, processing, and shelf-life characteristics, setting the stage for detailed engineering and optimization studies. Most of our research studies have been concerned with 4,4'-bis(3,4-dicyanophenoxy)biphenyl 1A (biphenol phthalonitrile), which shows excellent high temperature, mechanical and flammability properties. This resins shows superb properties to current state-of-the-art (SOA) high temperature thermosetting polymers. Current research activities at NRL concerned the development of 1A for composite applications. Daychem Laboratories, Inc. offers the monomer 1A in research quantities for testing.



1A



1B

RESULTS AND DISCUSSION

Phthalonitrile Technology. Phthalonitrile-based polymers are a class of high temperature polymers having a wide range of potential uses such as composite matrices[4], adhesives[5], and electrical conductors[6-9]. Aromatic ether[6,7,10-12], thioether[13], sulfone[14,15], and imide[16,17] linkages have been successfully incorporated between terminal phthalonitrile units. Polymerization occurs through the cyano groups to afford heterocyclic cross-linked products[18,19]. Neat polymerization is extremely sluggish requiring extended heat treatment at elevated temperatures before gelation occurs. Curing agents that have been used to cure phthalonitrile resins include organic amines[20], strong organic acids[21], strong organic acid/amine salts[22], metallic salts[23], and metals[24]. Only a minute quantity of curing additive is needed to initiate the polymerization reaction. The polymerization reaction occurs through the cyano groups of the phthalonitrile units by an addition mechanism and appears to propagate through multiple reaction pathways involving polytriazine, polyimine, and polyphthalocyanine formations. Recent FTIR and UV studies in our laboratory indicate that phthalocyanine formation is probably a minor product and occurs early in the polymerization reaction before gelation. These heterocyclic cross-linked

products are known to exhibit good thermal and oxidative stability. The preferred reaction pathway is probably dependent on the curing agent used. The polymerization rate is easily controlled as a function of the concentration of curing additive and curing temperature, which enhances the importance of the phthalonitrile-based resins as a matrix material for thick composite applications and as a candidate for resin transfer molding (RTM).

Synthesis and Polymerization. Phthalonitrile polymers, prepared from polymerization of 4,4'-bis(3,4-dicyanophenoxy)biphenyl 1A[7] and 2,2-bis[4-(3,4-dicyanophenoxy)phenyl]-hexafluoropropane 1B[12], show excellent thermal and oxidative stability at temperatures approaching 372°C (700°F). The phthalonitrile monomers 1A and 1B are readily converted to highly cross-linked thermosetting polymers in the presence of thermally stable organic amines, phenols, or strong organic or mineral acids. These curing additives are preferred due to the homogeneous nature of the polymerization medium. Polymerization occurs by an addition reaction through the terminal phthalonitrile moieties to afford void-free polymeric materials. The properties are controlled as a function of the amount of curing additive used and the curing temperature (>200°C).

Processability. The phthalonitrile monomer 1 is easily processed into shaped components or films in a controlled manner. After the addition of curing additive to the melt of 1, the polymerization reaction can be performed in one step to gelation or can be advanced to any viscosity desired (B-staged prepolymer), quenched, and stored at room temperature. The B-staged prepolymer is stable indefinitely under ambient conditions and is soluble in numerous common organic solvents such as methylene chloride, chloroform, and the dipolar aprotic solvents. Shaped components are easily processed by heating the polymerization mixture or B-staged prepolymer above the melting point until gelation occurs. When postcured at 375°C, the resulting polymers do not show any evidence of glass/rubber transition upon analysis to 375°C. Due to the solubility properties and indefinite stability at ambient temperatures, the B-staged prepolymer formed from 1 is an excellent candidate for the preparation of stable carbon prepreg for usage in the fabrication of fiber reinforced composite components.

Thermal Stability. The thermal and oxidative stabilities have been assessed under isothermal (see Table 1) and dynamic conditions. Isothermal measurements in an air flow at 100 cc/min. indicate that the polymers can be expected to perform well for relatively long exposures at

Table 1. Isothermal Oxidative Weight Loss Studies in Air for 100 Hours

	Percent Weight Loss	
Temp. °C	1A	1B
316	--	--
343	1.8	1.3
357	7.2	4.3
372	--	28.4

moderate temperatures between 300-357°C (572-675°F). The thermal and oxidative properties were found to be enhanced as the polymers were postcured at elevated temperatures. Catastrophic failure as determined by thermogravimetric analysis (TGA) in air consistently occurred between 525-600°C (980-1110°F).

Mechanical Properties. The phthalonitrile-based polymers have been shown to exhibit excellent mechanical properties. Compared with other state-of-the-art thermosetting polymers, the phthalonitrile polymers display superior tensile strength values and retain these values when aged at 315°C (600°F) for extended periods in an oxidizing environment (see Table 2). The phthalonitrile polymers have also been found to exhibit a fracture toughness similar to the values of unmodified epoxy resins.

Table 2. Effect of Cure Conditions on the Room Temperature Tensile Strength of Phthalonitrile Polymer 1A

Cure/Postcure Conditions[a]			
(°C)	(Hrs)	Atmosphere	Tensile Strength at Break (MPa)
315	24	Air	94 ± 17
350	12	Argon	94 ± 21
375	12	Argon	80 ± 7
315	100	Air	72 ± 5

[a]All materials were initially cured at 240°C for 6 Hr. and at 280°C for 16 Hr. Tabulated conditions are successive thermal and oxidative treatments.

Adhesive Properties. Phthalonitrile polymers or resins are proving to be superior in physical and chemical properties relative to epoxies, polyimide, and other resins used as high temperature adhesives. Epoxies cannot be used as adhesives for prolonged temperature exposure in excess of about 140°C due to thermal and mechanical instabilities. Polyimides and polybenzimidazoles are difficult to process and volatiles are evolved during the curing or polymerization reaction. A major advantage of phthalonitrile resins is their ability to withstand temperatures in excess of 300°C for extended periods without permanent damage. Table 3 lists the shear strength (lbs./in) for the various metals and the evaluated temperatures. The number in parentheses indicates the number of samples tested.

Electrical Conductivity[6]. The excellent thermal stability displayed by the phthalonitrile polymers contributed to further pyrolytic studies. These studies were performed on the resin prepared from polymerization of 4,4'-bis(3,4-dicyanophenoxy)biphenyl 1A. The electrical behavior can be systematically changed from an insulator to a semiconductor and made to approach metallic regions by controlling the thermal processing temperature (see Table 4). For pyrolytic temperatures up to 500°C, the conductive nature changes from an insulator and approaches the semiconductive region. At 900°C, the conductivity of the pyrolysate 2A approaches the metallic regions. Although there was some shrinkage and a weight loss of 28 percent at 900°C, the pyrolytic material retained its structural integrity and appears visually to

be void-free. However, when viewed through a microscope, microvoids are apparent and probably due to the evolution of gaseous by-products.[25]

The electrical conductivity of 2A is environmentally stable. For example, after pyrolysis at 700°C for 24 hours, the room temperature conductivity of 2A did no change significantly on exposure to air for 1 year. Another sample showed little change in its room temperature conductivity when heated in air at 140°C. The conductivity of 2A is also hydrolytically stable. After being pyrolyzed at 900°C, 2A exhibited no observable change in the electrical conductivity upon immersion in water for 80 days, even though an 1.0% weight increase was observed.

Table 3. Shear Strength (lbs./in) of Phthalonitrile Resin 1A

Metal	Room Temperature	200°C
Aluminum[1]	(7) 1450	(6) 1070
Brass[1]	(6) 1860	(7) 1450
Steel[1]	(2) 1940	-----
Steel[2]	(2) 1610	-----
Titanium[3]	(5) 1480	-----
Titanium[4]	(6) 2215	-----
Titanium[5]	(4) 1250	-----

[1]Cure cycle was 240°C for 16-18 hours, 280°C for 8 hours, and 315°C for 16 hours.

[2]Epon 828 cured with 14.5 pph of m-phenylenediamine at 75°C for 2 hours and at 125°C for 2 hours was used as standard. The literature value for the shear strength of Epon 828 is 7500 lbs./in. Thus, the measured shear strength values for the cured phthalonitrile resin samples are probably too low.

[3]Cure cycle was 240°C for 16 hours, 280°C for 6 hours, and 315°C for 16 hours.
Adherend preparation by filing followed by Pasa Jell treatment; scrim used to control thickness

[4]Cure cycle was 240°C for 16 hours, 280°C for 6 hours, and 315°C for 16 hours.
Adherend preparation by filing and scotchbrite-type sandpaper; no scrim.

[5]Cure cycle was 240°C for 16 hours, 280°C for 6 hours, and 315°C for 16 hours and postcured at 350°C for 4 hours. Adherend preparation by filing and scotchbrite-type sandpaper; no scrim.

Prepreg Processability. Prepregs, which are stable indefinitely under ambient conditions, have been fabricated from the prepolymer of 4,4'-bis(3,4-dicyanophenoxy)biphenyl 1A using phthalonitrile-sized and unsized AS-4 and IM-7 unidirectional carbon fibers. The prepolymer is formed by adding 1.5-2.5% by weight of 1,3-bis(3-aminophenoxy)benzene, obtained from National Starch and Chemical Corporation, to the melt of 1A. The phthalonitrile prepreg is being used to optimize processing parameters and in the fabrication of unidirectional composite panels for mechanical property determinations. A short cure cycle has been developed for

Table 4. Effect of Heat Treatment on Room Temperature Conductivity of 1A and 2A

Temp. °C	Atm.	Conductivity (S/cm)	Time (Hr)
300	Air	$<10^{-8}$	24
400	Inert	$<10^{-8}$	24
500	Inert	$<10^{-8}$	24
600	Inert	$2.7x10^{-2}$	24
700	Inert	$1.5x10^{-1}$	24
800	Inert	$6.7x10^{+1}$	24
900	Inert	$2.7x10^{+2}$	24

fabricating quality composite panels using an autoclave at the National Institute of Standards and Technology (NIST). For example, a composite panel fabricated at 250°C for 1 hour and at 325°C for 3 hours exhibits mechanical properties comparable-to-superior relative to PMR-15 and has a glass transition temperature (T_g) in excess of 350°C. In addition, a composite sample postcured at 375°C for a short time does not exhibit a T_g. Moreover, little changes in the mechanical properties occur at elevated heat treatment temperatures. For comparison, PMR-15 based composites are processed at approximately 232°C for 3 hours and at 315°C for 8-16 hours.

Flame Resistance. The use of composite materials for shipboard and submarine applications has been recognized as offering potential benefit to the Navy in a variety of areas ranging from improvements in weight, maintenance, and corrosion control, to lower thermal conductivity, reduced magnetic and acoustic signatures, and greater versatility in design. Marine applications of composites have tended to rely on glass reinforced polyester matrices. As component designs have increased, organic polymeric composites may be the only option available if the flammability obstacles of polymeric matrices can be overcome. High performance composites that fit the Navy's flammability requirements of MIL-STD-2031[26]: Fire and Toxicity Test Methods and Qualification Procedure for Composite Material Systems Used in Hull, Machinery, and Structural Applications Inside Naval Submarines, have not been available. When compared with other high temperature, high performance composites, preliminary cone calorimetric studies in collaboration with Usman Sorathia at CDNSWC-Annapolis have shown that a phthalonitrile-based composite fabricated from 1A is the only polymeric material that satisfies the Navy's flame retardant specification (T_{IG} and PHR) for lightweight, polymeric composites for ship applications (see Table 5). To meet the Navy's requirement, the time to ignition (T_{IG}) should be greater than 60 seconds and the peak heat release (PHR) at 100-kW/m^2 irradiance should be less than 150 during heat treatment for 300 seconds. Moreover, the low water uptake and ease of processability have enhanced the importance of phthalonitrile composites for both ship and submarine applications. The phthalonitrile composite also meets the FAA requirement (no flammability at 35kW/m^2 irradiance) for composite usage in an aircraft cabin.

Table 5. Phthalonitrile Composite Meets Navy Specification For Flammability As Determined In Cone Calorimeter[a]

Composites	T_{IG}	PHR
Carbon/Phthalonitrile	75	118
Carbon/PMR-15	55	85
Carbon/BMI	22	270
Carbon/PEEK	42	85
Carbon/Epoxy	28	232
Carbon/Phenolic	25	96

[a] Preliminary flammability tests performed by Usman Sorathia and Tom Juska, Fire Resistant Materials Group, CDNSWC, Annapolis.

Technology Impact. The phthalonitrile-based composites provide advanced lightweight structural and electronic materials to meet numerous aerospace and marine needs such as flame resistant polymeric composite materials for ship and submarine applications; high-performance aircraft, rockets, and missiles; engine component applications; heat resistant and flame resistant composite components for usage in the vicinity of an aircraft engine; intrinsic electrical composite conductors as electromagnetic interference (EMI) shields and RAM for aircraft and marine applications; high temperature adhesive; and high temperature molding materials for the fabrication of electronic devices.

CONCLUSION

Due to the unique properties, monomers 1A, and 1B are potential candidates as matrices for numerous composite applications. The monomers are easily fabricated into shaped, void-free components, possess high thermal and oxidative stability, and have high char yields. When postcured at high temperatures, the polymers do not exhibit a T_g. Amorphous prepolymers are stable indefinitely under ambient conditions, which enhance their importance in the fabrication of laminates for composite applications. The polymers retain their properties at elevated temperatures. Upon pyrolysis to 900°C, 1A and has been shown to exhibit electrical conductivity ranging from an insulator to semiconductor and approaching metallic behavior. Potential applications of the polymers include their usage as a matrix material for advanced composites and as a molding material for the fabrication of electronic devices due to the thermal stability at the soldering temperatures required to mount components on a circuit board surface.[27]

ACKNOWLEDGEMENT

The authors wish to thank the Office of Naval Research for financial support of this work. The authors are also grateful to Usman Sorathia, CDNSWC-Annapolis, MD, for his flammability studies.

REFERENCES

1. Serafini, T. T., Delvigs, P., and Lightsey, G. R., **J. Appl. Polym. Sci.**, 16, 905 (1972).
2. Serafini, T. T., **ACS Organic Coatings and Plastics Chemistry**, 40, 469 (1979).
3. Alton, W. B., **SAMPE Proc.**, 12, 121 (1980).
4. Keller, T. M. and Moonay, D. J., **SAMPE Symp.**, 34, 941 (1989).
5. Keller, T. M. and Roland, C. M., **"High Temperature Adhesives"**, US patent 5,242,755 (September 7, 1993).
6. Keller, T. M., **CHEMTECH**, 18, 635 (1988).
7. Keller, T. M., **J. Polym. Sci., Part A: Polym. Chem.**, 26, 3199 (1988).
8. Keller, T. M., **SAMPE Symp.**, 31, 528 (1986).
9. Giuliani, J. F. and Keller, T. M., **Sensors and Materials**, 1, 2247 (1989).
10. Keller, T. M., **Encyclopedia of Composites**, Volume 4, 111 ,1991; VCH Publishers, New York, N. Y.
11. Keller, T. M., **J. Polym. Sci., Part C: Polym. Lett.**, 24, 211 (1986).
12. Keller, T. M., **Polym. Commun.**, 28, 337 (1987).
13. Keller, T. M. and Gratz, R. F., **Polym. Commun.**, 28, 334 (1987).
14. Keller, T. M. and Price, T. R., **Polym. Commun.**, 25, 42 (1984).
15. Keller, T. M. and Price, T. R., **Polym. Commun.**, 26, 48 (1985).
16. Keller, T. M., **Polym. Commun.**, 32, 2 (1991).
17. Keller, T. M., **Polymer**, 34(5), 952 (1993).
18. Burchill, P. J., **J. Polym. Sci.: Part A: Polym. Chem.**, 32, 1 (1994).
19. Snow, A. W. and Griffith, J. R., **Macromolecules**, 17, 1614 (1984).
20. Keller, T. M. and Price, T. R. **J. Macromol. Sci. - Chem.**, A18(6), 931 (1982).
21. Keller, T. M., **Polym. Prep.**, 33(1), 422 (1992).
22. Burchill, P. J., **IUPAC Int. Symp. Polym. Mater. Prep.**, 186 (1991).
23. Keller, T. M. and Griffith, J. R., **ACS Symp. Series: Resins For Aerospace**, 132, 25 (1980).
24. Walton, T. R., Griffith, J. R., and O'Rear, J. G., **Adhesion Sci. and Tech., 1975**, 9B, 665 (1975).
25. Bürger, A., Fitzer, E., Heym, M., and Terwiesch, B., **Carbon**, 13, 149 (1975).
26. Fisher, J. F., **Advanced Composites**, 20 (May/June 1993).
27. Frisch, D. and Ciccarone, R., "Thermal Analysis for Evaluating Laminates," **Circuits Manufacturing**, Benwill Publishing Corporation, July 1977.

INORGANIC-ORGANIC HYBRID POLYMERS FOR HIGH TEMPERATURE APPLICATIONS

Teddy M. Keller* and David Y. Son

Materials Chemistry Branch
Naval Research Laboratory, Code 6127
Washington, DC 20375-5320

ABSTRACT

An emerging technology that holds promise for extending the temperature stability of polymers is inorganic-organic hybrid polymers. The polymeric hybrid combines the desirable features of inorganics and organics within the same polymeric system such as thermal and oxidative stability and processability.

Linear siloxane-acetylenic and carborane-siloxane-acetylenic hybrid polymers have been synthesized and characterized. The poly (carborane-siloxane-acetylene)s 1 and 3 are viscous liquids at room temperature, which can be easily processed to high temperature polymeric thermosets under either thermal or photochemical conditions. The poly (siloxane-acetylene)s 2 are either liquids or low melting solids. Thermal curing of the linear polymers to the thermoset is accomplished above 150°C. The novel thermosetting polymers are readily converted into ceramic materials by heating to 1000°C under inert or oxidative conditions. The thermosets and ceramic materials are stable in air at elevated temperatures.

Various multi-secondary acetylene substituted aromatic hydrocarbons were synthesized and evaluated as to their ease of homopolymerization and their ease of carbonization after being polymerized. Out of our studies, 1,2,4,5-tetrakis(phenylethynyl)benzene 4 has been found to exhibit outstanding properties as a carbon precursor material. Copolymers from 1 and 4 have been shown to exhibit outstanding high temperature oxidative stability.

INTRODUCTION

The development of carborane-siloxane polymers in the early 1960s was a major breakthrough in the search for high temperature elastomers[1]. Poly(carborane-siloxane) elastomers show superior thermal and oxidative properties and retarded depolymerization[2,3] at elevated temperatures relative to poly(siloxanes). The carborane unit is incorporated into a polymeric chain to impart either high temperature and/or specialized chemical resistance. Patents have been issued for thermally stable elastomers consisting of meta- or para-carborane and either silyl or siloxyl units. Some have been produced commercially. However, they lack acetylenic groups or any other functional group for cross-linking purposes to produce thermosetting polymers. Researchers have successfully incorporated carboranes into the backbone of most of the common types of addition and condensation polymers and as pendant groups or side chains[4]. It was recognized that the attachment of a carborane side group onto

the backbone of a polymer did not enhance the thermal properties. However, if the carborane becomes a part of the polymeric backbone, appreciable improvements in thermal stability were achieved.

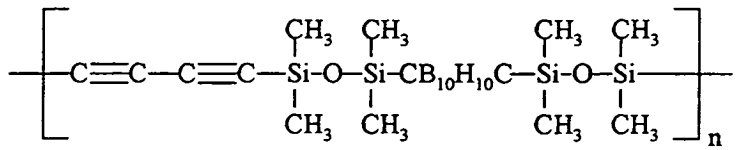
Considerable interest has been shown in the uses of polyfunctional arylacetylenes as precursors to carbon. Carbon erodes rapidly in air at temperatures as low as 400°C. Great effort is thus being devoted at developing techniques for protecting carbon-carbon composites against oxidation at elevated temperatures.[5-8] Much of this interest has arisen from recent plans of the U.S. Government to build a National Aerospace Plane (NASP) which would require lightweight structural materials stable up to 1500°C in air. Phenolic resin systems and petroleum and coal tar pitches are currently used as the carbon matrix precursor material. There are numerous problems associated with these carbon precursor materials. We are engaged in the synthesis of compounds containing three or more phenylethynyl groups substituted on the benzene ring. Copolymers of a multi-phenylethynylbenzene and an inorganic-organic hybrid polymer show outstanding flame resistance or oxidative stability. Both compounds contain acetylenic units for conversion to the copolymer. The resistance to oxidation is a function of the amount of the hybrid polymer present in the copolymer.

In our continuing investigations of high temperature polymers at the Naval Research Laboratory, multi-acetylenic substituted aromatic compounds, linear siloxane-acetylenic polymers, and carborane-siloxane-acetylenic polymers have been synthesized and are being evaluated as high temperature matrix materials for composites and as precursor materials for carbon and ceramics for applications above 400°C and in excess of 1000°C in air. The major advantage of our approach is that the desirable features of inorganics and organics such as high thermal and oxidative stability and processability are incorporated into the same polymeric chain. The siloxane units provide thermal and chain flexibility to polymeric materials. The chemistry involved in synthesizing poly(siloxane) and poly(carborane-siloxane) has been modified to accommodate the inclusion of an acetylenic unit in the backbone. The novel linear polymers have the advantage of being extremely easy to process and convert into thermosets or ceramics since they are either liquids at room temperature or low melting solids and are soluble in most organic solvents. They are designed as thermoset polymeric precursors. The cross-linked density of the thermosets is easily controlled as a function of the quantity of reactants used in the synthesis. The acetylenic functionality provides many attractive advantages relative to other cross-linking centers. The acetylene group remains inactive during processing at lower temperatures and reacts either thermally or photochemically to form conjugated polymeric cross-links without the evolution of volatiles.

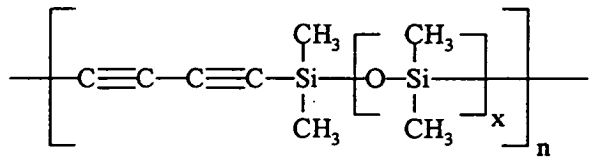
RESULTS AND DISCUSSION

Synthesis. The synthesis of 1, 2, and 3 is a one pot, two stage reaction. A typical procedure is as follows: Dilithiobutadiyne was prepared by the method of Ijadi-Magshoodi and Barton[9,10]. Dilithiobutadiyne was not isolated and was reacted with an equimolar amount of the 1,7-bis(chlorotetramethyldisiloxyl)-m-carborane 5, disiloxyl dichlorides 6, or combinations of 5 and 6 to afford dark-brown viscous polymers 1, 2, or 3, respectively, in high yield (90-97%). GPC analysis of 1 indicated the presence of low molecular weight species (≈500) as well as higher

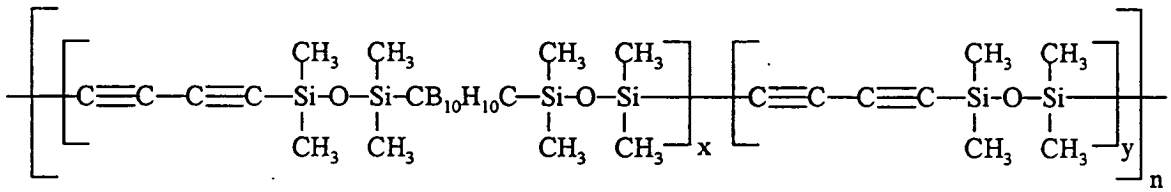
average molecular weight polymers (Mw≈4900, Mn≈2400). Heating under vacuum at 150°C removed the lower molecular weight components leaving a 92-95% overall yield.



1

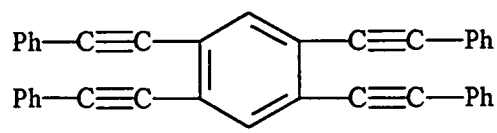


2a, x = 1
2b, x = 2



3

Several aromatic compounds containing multiple phenylethynyl substituents have been prepared and polymerized to thermosets through the acetylenic units. Pyrolysis of the polymers afforded high carbon yields. The acetylenic-based compounds were synthesized from the palladium catalyzed coupling reaction of phenylacetylene with multiple brominated benzenes. The simplicity of the synthetic procedure and the mildness of the reaction conditions enhance the importance of the coupling reaction for the preparation of secondary acetylenes. Out of our studies, 1,2,4,5-tetrakis(phenylethynyl)benzene 4 has been found to exhibit outstanding properties as a carbon precursor material.[11]



4

Polymerization. The acetylenic-based compounds were cured thermally and converted into thermosets at temperatures above 200°C. The poly(carborane-siloxane-acetylene)s 1 and 3, poly(siloxane-acetylene)s 2, and copolymers of 1 and 4 have the advantage of being extremely easy to process since they are either liquids at ambient temperature or low melting solids and are soluble in most organic solvents. Linear polymers 1, 2, and 3 are designed as thermoset polymeric precursors. Cross-linking of 1, 2, and 3 can occur by thermal or photochemical

means through the triple-bonds of the acetylenic units to afford thermosetting polymers[12-14]. A shiny void-free dark brown solid was produced by thermally curing the linear polymers at 200, 300, and 400°C, consecutively, for 2 hours at each temperature either under inert conditions or in air. Gelation occurred during the initial heat treatment at 200°C. Thin films can be cured photochemically in either an inert or oxidizing environment. The copolymers formed from 1 and 4 were mixed in various portions and cured by heating above 200°C in a similar manner to the above.

Conversion Into Carbon/Ceramics. The linear inorganic-organic hybrid polymers 1-3 and various mixtures of 1 and 4 were converted into carbon/ceramics by heating to 1000°C under inert and oxidizing conditions. Pyrolysis under inert conditions afforded chars from 70 to 85%. The chars exhibited outstanding oxidative stability.

Poly(Carborane-Siloxane-Acetylene)s. Carborane-siloxane-acetylenic-based thermosetting polymers are designed to take advantage of the inorganic-organic hybrid approach to produce high temperature polymers and ceramic-based materials.[15] The major advantage of the hybrid high temperature polymeric approach is that the desirable features of inorganics (high thermal and oxidative stability) and organics (processability) are incorporated into a novel polymeric material. Polymers 1 and 3 possess exceptional thermal and oxidative stability. They are viscous liquids that are soluble in most organic solvents. These linear polymers are easy to process thermally into structural components. Thermal cross-linking occurs through the triple bonds to produce the thermoset. These thermosets, in turn, produce a ceramic material in high yield (85% and 92%, respectively) upon pyrolysis in nonoxidizing and oxidizing environments. When heated in an oxidizing environment, these ceramics initially gained weight. The potential utility of polymers 1 and 3 as matrix materials for advanced composites and further conversion into a shaped ceramic component is due to the ease of processability, high ceramic yield, and oxidative stability at elevated temperatures. Surface analysis studies of the thermoset and the ceramic formed from 1 exposed to air indicate that the surfaces are enriched with oxide forms of boron and silicon, which apparently protect the interior part against further oxidation.

Poly(Siloxane-Acetylene)s.[16] Most conventional organic polymers lack the ability to withstand temperatures in excess of 200-250°C for extended periods in an oxidizing environment. In addition, these materials tend to have poor hot-wet mechanical properties. Combining inorganic elements with the organic elements of a polymer should enhance the thermal and oxidative stability of the polymers. Siloxyl groups are a logical choice for inclusion in these polymers as they possess good thermal and oxidative stability and high hydrophobicity.[17] Furthermore, their flexibility should contribute favorably to the processability of the resulting polymers. The polymer $(\text{-C}\equiv\text{C-C}\equiv\text{C-SiMe}_2\text{OSiMe}_2\text{-})_n$ 2a, obtained via the oxidative coupling of 1,3-diethynyltetramethyldisiloxane, was previously reported by Parnell and Macaione.[17] The authors reported obtaining only low molecular weight products (n=2-16) in addition to some insoluble material. Herein, we describe an alternative high-yield synthesis of 2a and the related polymer $(\text{-C}\equiv\text{C-C}\equiv\text{C-(SiMe}_2\text{O)}_2\text{SiMe}_2\text{-})_n$ 2b. GPC analysis indicated broad molecular weight distributions, with peak maxima occurring at ~10,000 (relative to polystyrene) for both polymers. The presence of lower molecular weight species can be attributed to early chain termination and formation of cyclic species during polymerization. Reaction of excess n-butyllithium with a Si-Cl bond can terminate a polymer chain. Also, cleavage of Si-O-Si

bonds by alkyllithium species can result in chain breakage, resulting in lower molecular weights. Elemental analysis data were in general agreement with calculated values. These polymers can be converted into thermosets thermally or photochemically, and can be pyrolyzed to form ceramic materials.

The thermal and oxidative properties of the linear poly(siloxane-acetylene)s were determined to 1000°C. As observed previously[18] for linear polysiloxanes, the thermal stability is diminished as the size of the siloxyl moiety increases. For example, the ceramic yields for 2a and 2b are 74% and 58%, respectively. Upon heating the chars to 1000°C in a flow of air, similar oxidative stabilities were observed. After the oxidative heat treatment, the overall weight loss was about 47%.

Copolymers From 1 and 4. On exposure of the copolymers formed from 1 and 4 to an oxidizing environment, a protective film initially develops that deters or alleviates further oxidation at a given temperature. The formation of the oxidized film and any weight loss associated with the exposure was accelerated by heat treatment of the carbonaceous/ceramic mass to 1000°C in air. Such samples were prepared from 50/50 and 65/35 weight percent blends of 1 and 4 heated to 1000°C, consecutively, in nitrogen and air and then isothermally aged in air in sequence for 10 hours each at 600 and 700°C. The chars from the 50/50 and 65/35 weight percent blends gained and lost about 0.1% and 18% weight, respectively, at 600°C. During the heat treatment at 700°C, the samples lost about 0.4% and 4% weight. The superior performance of the 50/50 blend shows that critical amounts of boron and silicon are necessary to protect a carbon-based material against oxidation.

Surface Analysis. The composition of the thermoset 7 and the ceramic 8 obtained from 1 of the oxidized surface and bulk were examined by XPS (X-ray photoelectron spectroscopy). The bulk compositions of 7 and 8 remained unchanged upon oxidation treatments. After curing and aging in air at 400°C, 7 developed a bilayer inorganic surface, which mainly consisted of boron oxide and silicon oxide (see Table 1). Prolonged oxidation of 8 at 500°C resulted in the volatilization of boron oxide leaving an almost pure silicon oxide layer, which blended into the bulk boron and silicon concentrations at the interface between the surface layer and the bulk.

Table 1. Surface Compositions Obtained From XPS Sputter Profiles After >120 Sputtering Time

Sample	C(Atomic) %	O(Atomic)%	Si(Atomic)%	B(Atomic)%
7[a]	45	6	8	31
7[b]	17	48	24	11
7[c]	5	66	20	10
8[d]	3	58	30	9

[a] Cured at maximum temperature of 400°C for 2 hours. [b] Cured at 300°C for 4 hours in argon and at 400°C for 4 hours in air. [c] Cured in air at 320°C, 350°C, and 400°C in sequence for 2 hours and aged for 100 hours in air at 400°C. [d] Cured and heated at 900°C for 4 hours and aged in air at 500°C for 100 hours.

Technology Impact. The linear siloxane-acetylenic and carborane-siloxane-acetylenic polymers provide advanced structural materials (thermosets and ceramics) for applications above 400°C and in excess of 1000°C in an oxidizing environment. These advanced materials could also be used in the fabrication of advanced ceramic fibers and as high temperature, high performance structural materials for nuclear and fusion power generators, high performance heat engines, gas turbine engines, heat exchangers, cutting tools, coatings to protect carbon/carbon composites against oxidation, and aerospace applications.

CONCLUSION.

Poly(carborane-siloxane-acetylene)s 1 and 3 and poly(siloxane-acetylene)s 2 are precursors for high temperature thermosetting polymers and ceramic-based materials that exhibit outstanding thermal and oxidative properties. The desirable features of inorganics (high thermal and oxidative stability) and organics (processability) are incorporated into novel polymeric materials.

Copolymers of 1 and 4 show outstanding oxidative stability. Both compounds contain acetylenic units for conversion to the copolymer. The resistance to oxidation was a function of the amount of 1 present in the copolymer. The studies show that carbon can be protected from oxidation at various temperatures by proper incorporation of silicon and boron units into a carbon precursor material.

Further studies are underway to evaluate and exploit polymers 1, 2, and 3 and the copolymers from 1 and 4 as matrix materials for high temperature composites and carbon/ceramic composites.

ACKNOWLEDGEMENT

The authors wish to thank the Office of Naval Research for financial support of this work. The authors are grateful to Dr. Pehr Pehrsson of the Surface Branch of the Chemistry Division for the XPS studies, Dr. Tai Ho of George Mason University for his assistance with the GPC measurements, and Dr. Kenneth M. Jones and Dr. Leslie J. Henderson for the original synthesis of 1,2,4,5-tetrakis(phenylethynyl)benzene and poly(butadiyne-1,7-bis(tetramethyldisiloxyl)-*m*-carborane, respectively.

REFERENCES

1. Schroeder, H. A., **Inorg. Macromol. Rev.**, 1, 45 (1970)
2. Papetti, S.; Schaeffer, B. B.; Gray, A. P. and Heying, T. L., **J. Polym. Sci.**, 4 (A-1), 1623 (1966)
3. Critchley, J. P.; Knight, G. J. and Wright, W. W., **Heat-Resistant Polymers**, Plenum Press: New York, 1983 ed., pp. 363-387
4. Williams, R. E., **J. Pure Appl. Chem.**, 29, 569 (1972)
5. Strife, J. R. and Sheehan, J. E., **Bull. Am. Ceram. Soc.**, 67, 369 (1988)
6. Sheehan, J. E., **Carbon**, 27, 709 (1989)
7. McKee, D. W., **Carbon**, 25, 551 (1987)
8. McKee, D. W.; Spiro, C. L. and Lamby, E. J., **Carbon**, 22, 507 (1984).
9. Ijadi-Maghsoodi, S.; Pang, Y. and Barton, T. J., **J. Polym. Sci.: Part A: Polym. Chem.**, 28, 955 (1990)
10. Ijadi-Maghsoodi, S. and Barton, T. J., **Macromolecules**, 23, 4485 (1990)
11. Jones, K. M. and Keller, T. M., **Polym. Mat. Sci. & Eng.**, 68, 97 (1993)
12. Rutherford, D. R.; Stille, J. K., **Macromolecules**, 21, 3530 (1988)
13. Neenana, T. X.; Callstrom, M. R.; Scarmoutzos, L. M.; Stewart, K. R.; Whitesides, G. M.; and Howes, V. R., **Macromolecules**, 21, 3525 (1988)
14. Callstrom, M. R.; Neenan, T. X.; and Whitesides, G. M., **Macromolecules**, 21, 3528 (1988)
15. Henderson, L. J. and Keller, T. M., **Macromolecules**, 27, 1660 (1994)
16. Son, D. Y. and Keller, T. M., **Polym. Mat. Sci. & Eng.**, 71, 305 (1994)
17. Parnell, D. R. and Macaione, D. P., **J. Polym. Sci., Polym. Chem. Ed.**, 11, 1107 (1973)
18. Dvornic, P. R. and Lenz, R. W., **High Temperature Siloxane Elastomers**, Hüthig & Wepf: Heidelberg, 1990

ADVANCED FIBER PLACEMENT PROGRAM FOR AIRCRAFT STRUCTURES

Bruce Brailsford
Sr. Program Manager
GreatLakes Composites Consortium (GLCC)
8400 LakeView Parkway Suite 800
Kenosha, WI 53142-7403

ABSTRACT

One Of The Key Factors driving next-generation fighter and attack aircraft is affordability. With the majority of the composite structures used on today's fighter aircraft being manufactured by hand collation of individual plies of composite materials, new processes must be developed that will reduce costs and improve repeatability. A technology that meets this criteria is fiber placement. Through the U.S. Navy Center of Excellence for Composites Manufacturing Technology (CECMT), funding has been provided to establish and demonstrate the production readiness of Advanced Fiber Placement technology for Navy/DOD applications. Specifically, this includes demonstrating this technology on F/A-18E/F Engine Inlet Ducts, Horizontal Stabilator Skins and Aft Fuselage Panels and on future low observable applications. This paper will discuss the multi-disciplinary and multi-company approach, planned program milestones, and the benefits to the composites fabrication industry and to future weapon system programs.

DISCUSSION

In 1990 The GreatLakes Composites Consortium (GLCC) a member-owned not-for-profit organization was formed to manage the U.S. Navy Center of Excellence for Composites Manufacturing Technology (CECMT). This center was created to develop manufacturing processes for composite materials in both defense and commercial applications. Since its inception the consortium has grown to a membership of over eighty members.

There are three levels of membership within the consortium: Principal, Supporting and Associate. Also, there is an additional category for educational institutions to become involved as partners. Some of the larger companies that are involved include Bell Helicopter-Textron, Lockheed, McDonnell Douglas, Northrop/Grumman, and Rockwell International. Medium sized companies include Cincinnati Milacron, Hercules Composite Products and ICI Fiberite. There is a very large group of smaller companies that have also joined in this effort, while a number of educational institutions bring an academic perspective to the consortium.

With the defense budget being reduced, affordability of composite structures for future Navy aircraft has become a large concern. Most of the composite structures used on today's fighter aircraft are fabricated by hand layup of individual plies of composite materials. Studies have indicated that the largest recurring cost drivers of hand layup manufacturing processes are fabrication and assembly. In light of the reduced defense budgets, the Navy decided to investigate Advanced Fiber Placement technology as a possible new process technology that would meet their demands of more complex designs and could reduce costs associated with component fabrication. In October of 1991, the GreatLakes Composites Consortium was chartered with the task of putting together a program with the objective of establishing and demonstrating production readiness of the fiber placement process on Navy/DOD applications.

Previous to the formation of GLCC, Northrop Aircraft Division (NAD), McDonnell Douglas Aerospace (MDA) and Hercules had entered into a technology exchange that was termed as the Tri-Party agreement in 1987. This agreement allowed NAD and MDA to apply and refine

existing Hercules fiber placement technology for the purpose of aircraft components manufacturing. Although the Fiber Placement process had matured, there had been no production implementation. Certain critical technologies and implementation issues remained to be addressed.

In 1991 a Technical Direction Letter (TDL) was released by the Navy to put together a program that would address fiber placement technology and would involve members of the consortium. A program was structured and given the title of "Advanced Fiber Placement" (AFP) in which three components are addressed as follows: Task One - F/A-18E/F Horizontal Stabilator; Task Two - F/A-18E/F Engine Inlet Duct; and Task Three - Low Observable Technology for future aircraft applications. During March of 1993, another technical direction letter was released from the Navy that added Task 4 to address F/A-18E/F Aft Fuselage Panels. The tasks are represented in Figure 1 and indicate the players that are associated with each of them.

In the fiber placement process, material in the form of continuous prepreg tow is drawn from storage creels to an application head. The head is mounted on a multi-axis machine that is programmed to apply the tow material to the nip point. The application head compacts the material and allows complex contoured surfaces to be fiber placed. The head controls a band of collimated tows that can be individually cut or added. The ability to cut and add tows enables the fiber placement process to manufacture complex ply shapes very effectively using minimum weight designs. Reference Figure 2. The following describes the work effort within each of the tasks:

McDonnell Douglas Corporation is leading Task 1 as they are the contractor of this component for the F/A-18E/F program. Hercules Composite Products is teamed with them in support of this activity. This task is structured into a seven subtask building block approach for the development of the full scale article that will design, build, and test specimens at the coupon, element, and subcomponent levels. In addition, it will also prepare material and process specifications for the fiber placement process. The subtasks consist of : Integrated Product Concept development, Specification Generation, Fiber Placement Flat Panel Demonstration, Subcomponent Demonstration, Full-Scale Demonstration, Technology Transfer, and Production Transition Plan.

Northrop/Grumman Corporation is leading Task 2 as they are the contractor of this component for the F/A-18E/F program. Hercules Composite Products is teamed with them in support of this effort. The scope of this activity is to demonstrate the production readiness of the fiber placement process for this component. This task is structured into a six subtask building block approach for the development of the full scale engine inlet ducts. The subtasks consist of: Duct Design for Fiber Placement, Process Generation and Validation, Quality Assurance, Cost Data, Fiber Placement Cell Readiness, and Technology Transfer.

Northrop/Grumman Corporation, Rockwell International, Cincinnati Milacron and Hercules are teamed together in Task 3 to demonstrate the use of fiber placement techniques to fabricate honeycomb sandwich inlet duct designs using low observable materials. This task consists of a six step building block approach as follows: Establishment of Baseline Design Requirements, Specimen Test Matrix Development, Physical Placement Tests, Structural and Low Observable Properties Testing, Technology Transfer, and a Scale-Up Plan.

Northrop/Grumman Corporation is heading up Task 4 as they are the contractor for this component. The scope of activity is aimed at demonstrating the production readiness of the fiber placement process for the F/A-18E/F aft fuselage skins. This task is structured into a eight subtask building block approach as follows: Part Selection and Design for Fiber Placement,

Element/Subcomponent Fab and Test, Tool Design and Fabrication, Full-Scale Fabrication, Quality Assurance, Cost Data, Production Readiness and Technology Transfer.

In Task 1 an Integrated Product Concept Development approach will be used that will identify and evaluate a number of fiber placement concepts for additional components and identify the most promising for development and production demonstration. The evaluation will consider the parameters of weight efficiency, cost savings potential, suitability for the fabrication using the fiber placement process, and sufficient design flexibility to meet the requirements for an advanced fighter application.

Material and process specifications will be generated within the program to ensure a quality product is being produced for this program and for production implementation. This will ensure the final specifications will not set unrealistic goals as requirements. Also, this will develop a better understanding of critical process variables prior to finalizing the specifications.

A test matrix of flat panels will demonstrate the mechanical properties of laminates fabricated using the fiber placement process are equivalent to those laminates fabricated by conventional processes (hand layup of prepreg broadgoods).

Subcomponent and full-scale articles will investigate critical details of the design and manufacturing method and reduced risk for the program. Several types of test specimens will be obtained from these articles and tested.

The method of technology transfer in this program is through quarterly reviews and reports as well as a final review at the completion of each task. This is a formal way of information exchange and progress evaluation by the team and the customer. Results from low cost manufacturing methods, lessons learned, as well as cost analysis will be reported. In addition, a final report and video tape will be made of the program detailing the significant progress made and the lessons learned.

A production transition plan will be generated in each of the tasks to describe the transition of the development work into the F/A-18E/F production environment. This plan will outline items necessary for implementation such as process control, quality assurance, equipment, and specifications.

PROGRESS

As was indicated in the introduction portion of this paper, this activity is a multi-disciplinary multi-company approach to addressing the development and implementation of fiber placement technology on Navy/DOD applications. The program has had several Quarterly Reviews at various members' locations with the members of the team to discuss the progress that has been made and to see the components that have been fabricated using fiber placement processing.

Task 1 has evaluated several composite parts on the F/A-18E/F aircraft to select high payoff candidates. The evaluation identified design and manufacturing approaches required to allow for fiber placement of the composite parts. Cost, weight, and risk impacts of the proposed changes to fiber placement were assessed. The results were correlated and reviewed and recommendations made. The conclusions indicated that not all composite parts lend themselves to be fiber placed since this technology is very geometry dependent. Also, efficient hand collated components do not show a major savings for fiber placement.

Common material and process specifications have been generated and reviewed by the team and additions and corrections made for fiber placement. These have been approved by the

customer. This effort included coordination with McDonnell Douglas, Northrop, Hercules, and Fiberite to establish common fiber placement specifications for the F/A-18E/F program. Several issues were addressed and resolved such as sampling methods, identification of material, etc.

The objective of the structural equivalency test matrix was to ensure that the properties of laminates fabricated using the fiber placement process have equivalent properties to those using manual layup of broadgoods. This was a large effort requiring many test panels. The mechanical properties test data is being compared to the F/A-18E/F composite structures data base. Analysis of the mechanical properties indicates that fiber placed laminates are equivalent to manually layed up laminates.

The subcomponent article demonstrations investigated the critical design/manufacturing details reduced risk for the full-scale articles. Component designs were created along with tooling being designed and fabricated. The proof of concept article fabricated in Task 1 was successfully fabricated and demonstrated the feasibility of the basic design/manufacturing approach. Non-destructive inspection of the article indicated no significant defects detected.

The full-scale article demonstrations in Task 1 were focused on ensuring that the F/A-18E/F horizontal stabilator is producible using the fiber placement process. The primary objective of this activity was to increase the level of confidence that the design/manufacturing approach of the F/A-18E/F stabilator skin would be successful and that the projected cost savings would be achieved. Articles were fabricated by McDonnell Douglas and Hercules. Detailed fabrication data was acquired during the fiber placement process. This data is being used to review the cost effectiveness of this process and in the fabrication of the particular ply sizes and shapes. One of each of the full-scale parts fabricated at McDonnell Douglas and Hercules were destructively evaluated and the other used for demonstration. Fiber placement has been selected as the baseline approach process for F/A-18E/F horizontal stabilator skins.

Task 2 generated a producible design for a tow placed inlet duct which is representative of the F/A-18E/F inlet duct. This design takes into account the capabilities and limitations of fiber placement. In addition, generation of the equivalence and structural test plans necessary to qualify the fiber placement process for application on the E/F program was accomplished.

Similar to the structural equivalence work that was done in Task 1, Task 2 also generated the information necessary to demonstrate equivalence of fiber placement to the current manual layup method. This was necessary since this laminate is fabricated using a different fiber than Task 1. Based on the acceptable results, fiber placement has been selected as the baseline process for F/A-18E/F engine inlet ducts.

Two full scale inlet duct mandrels were fabricated for Task 2. They have graphite/bismaleimide skins with the backup structure attached to a steel shaft.

Full-scale inlet ducts have been fabricated by both Northrop and Hercules. The first set of ducts had panels removed from them for post buckling tests. The second set of ducts were used for Fuel/Hammershock tests. The third set of ducts are being used as demonstration articles. Each of these ducts have been subjected to dimensional and other nondestructive examinations as well as gathering data during processing.

A baseline design has been developed in Task 3 for a low observable inlet duct that consisted of a duct wall fabricated of honeycomb sandwich construction. This design is capable of attenuating electromagnetic energy. Since this program is a feasibility study and not tied to a particular aircraft or weapon platform, materials selections were based on previous experience.

Baseline hand layed up control panels and fiber placed comparison panels were fabricated and tested. The results of the mechanical and electrical testing indicates equivalency between hand layed up and fiber placed low observable laminates.

A Phase II addition to this effort is in work. This activity will identify fiber placement process limits for advanced inlet designs as well as establish a methodology to link design properties to process variation and process control requirements.

The F/A-18E/F fuselage skins were reviewed and evaluated in Task 4 to determine the best candidate for this program. The criteria used in the selection process was the skin should contain features which could cause potential problems in implementing fiber placement processing for fuselage skins. Also, the skin should have sufficient complexity to ensure that successful demonstration of fiber placement processing for this skin will result in confidence for applying this technology to any other F/A-18E/F skin. Using this criteria an F/A-18E/F Upper Side Skin was selected for use on this program.

Development test panels have been designed to incorporate those features which are considered most critical by both manufacturing and engineering personnel. Test plans have been developed to generate the necessary information to demonstrate the equivalence of the fiber placed skin to the current manual layup method.

CONCLUSION

In summary, this program has combined the resources of several member companies of the GreatLakes Composites Consortium to develop, fabricate, and evaluate this technology for implementation on Navy/DOD applications. Specifically this technology is being transitioned into the F/A-18E/F program where it has been baselined as the manufacturing process for fabrication of the Horizontal Stabilator Skins, Engine Inlet Ducts, and Aft Fuselage Panels on the F/A-18E/F program.

The program has generated fiber placed component designs, created a common material specification to address the impact of incoming material quality, created a common process specification, developed tooling for designs that incorporate the fiber placement process, performed structural equivalency testing to verify that fiber placed laminates are equivalent to hand layup, and fabricated and tested subcomponent and full scale hardware. These activities have produced the required data to incorporate fiber placement technology with low risk into the F/A-18E/F program. This technology lends itself as a state of the art, low cost solution, and as a substitute for existing and future composite hand layup designs.

Throughout the program technology transfer has taken place through regular (quarterly) progress reviews and through written reports. Program plans and results have been reviewed and discussed with McDonnell Douglas Corporation, Northrop Corporation, Hercules Composite Products, Grumman Corporation, Rockwell International, Cincinnati Milacron, GLCC and NAVAIR.

ADVANCED FIBER PLACEMENT PROGRAM

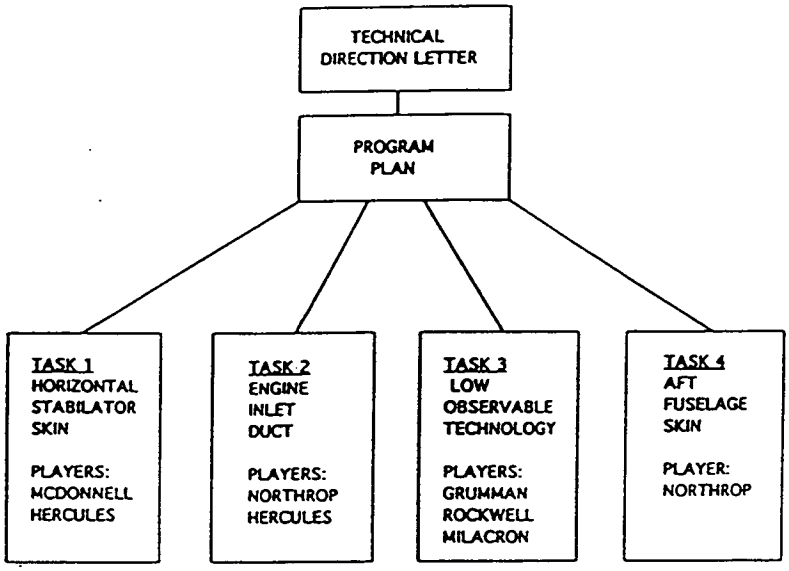


Figure 1

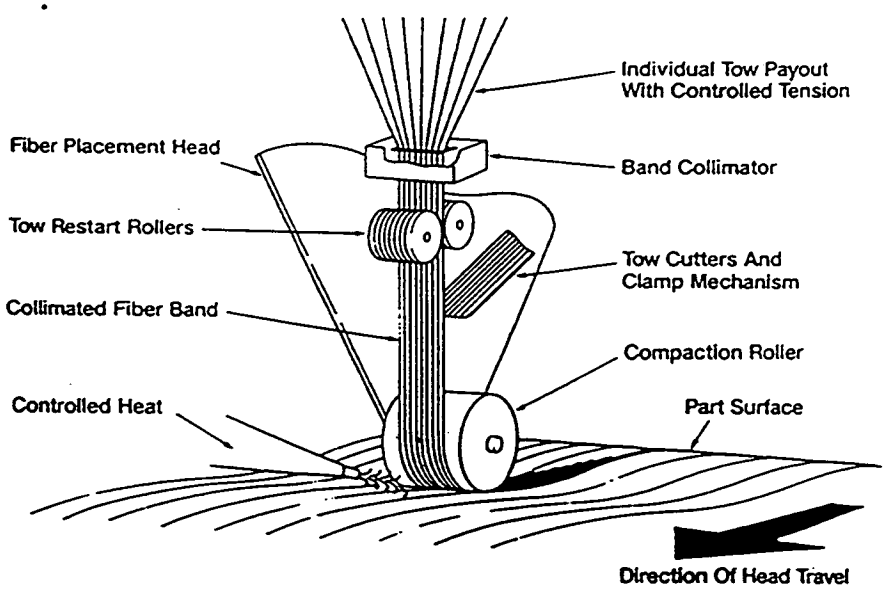


Figure 2

THE COMMERCIAL PAYOFFS OF THE BALLISTIC MISSILE DEFENSE ORGANIZATION'S DIAMOND TECHNOLOGY INITIATIVE: PRESENT AND FUTURE

Ted Lynch
National Technology Transfer Center
Alexandria, VA 22314

ABSTRACT

In 1986, the Ballistic Missile Defense Organization (BMDO) Innovative Science and Technology (IS&T) program began one of its most ambitious projects: the Diamond Technology Initiative. This program's goal is to bring the promise of diamond coating technology to missile defense uses, particularly those made possible by wide bandgap diamond microelectronics. In the process, BMDO has also encouraged commercial use of diamond coating technology. This paper provides an introduction to some of the key issues related to commercialization of diamond coating technology, thereby serving as a starting point for companies who might want to take advantage of BMDO-sponsored diamond research.

INTRODUCTION

Name a material property—almost any will do—and chances are good that diamond is at one extreme or the other. The hardest material known, the best electrical insulator, the best heat conductor, the slickest material, the best semiconductor when doped . . . diamond's exceptional properties, which is repeated in more detail in figure 1, has been recited many times over, each time with the implication that anyone who can harness these properties will become very rich.

For most applications, though, diamond's potential has gone untapped. Until very recently, only diamond's hardness and, of course, its physical beauty as a gemstone have made it valuable. However, these two properties alone accounted for a $5 billion worldwide market for natural diamond (1). Chemical vapor deposition (CVD) of diamond (and related diamond thin-film deposition techniques) has received so much attention in recent years because it can deposit diamond on substrates of varying composition and shape. This ability promises to turn diamond into an engineered material, one whose exceptional properties can be tailored to meet the demands of many different applications.

Though diamond CVD has only recently become the focus of popular attention, the technique was first developed in 1952 by William Eversole of Union Carbide. In diamond CVD, researchers mix hydrocarbon gas in a low-pressure reaction chamber and energize it with a heat source, plasma, or microwave. The resulting vapor deposits on a suitable substrate in the form of diamond crystals. In Eversole's work, too much graphitic carbon was mixed in with the diamond coating, which meant that anyone who wanted to use the coating had to go through an expensive and time-consuming process to remove the graphite. This limitation left diamond CVD on the shelf for years.

Figure 1: Properties and Applications of Diamond Coatings

With an extraordinary range of properties, diamond coatings could be used in many areas, from optical coatings to electrical heat spreaders. The following list highlights some of the properties and the resulting uses:[1]

Hardness: Diamond has a hardness of 10,000 kg/mm^2, two times harder than its nearest natural competitor, boron carbide. This hardness makes diamond an excellent coating for machine cutting tools or other areas where wear protection is important.

Coefficient of Friction: With a coefficient of friction of 0.05 to 0.1, diamond equals or betters Teflon®'s "slipperiness." Diamond's low coefficient of friction combines with its hardness to provide an excellent set of tribiological properties (properties related to the interaction between a material's surface and that of other materials), making it an ideal coating for bearings, rotating seals, extruders, dies, and cookware.

Tensile and Compressive Strength: Diamond can withstand stretching forces of 290 kg/mm^2, nearly ten times better than alumina. Diamond also has the highest compressive strength at 14 million pounds per square inch, making it an excellent candidate for fiber-reinforced composites. This strength also permits thin membranes for optics and acoustic diaphragms.

Chemical Inertness and Radiation Resistance: Diamond is inert to most corrosive chemicals and is very stable in high-radiation environments. These properties combine with diamond's electrical and wear-resistant characteristics to make it attractive in biomedical implants, chemical and petroleum processing and exploration, and electrical devices and sensors for satellites and medical instruments.

Electrical Resistivity: Diamond, with a resistivity of $1x10^{16}$, insulates against electricity ten times better than alumina. This resistivity, combined with its high thermal conductivity, makes diamond an ideal heat spreader for microelectronic devices and also for dielectric applications such as high-power, thin-film diamond capacitors.

Wide Bandgap Semiconductor: When doped, n-type diamond has the second highest electron mobility (2,200 cm^2/volt-sec), while p-type diamond has the highest hole mobility (1,600 cm^2/volt-sec). Diamond's wide bandgap (5.5 electron volts) permits many applications in high-temperature, high-frequency, and high-power electronics. Because it is stable in high-radiation environments, diamond is also attractive for radiation-hardened electronics.

Thermal Conductivity: For every degree Kelvin in temperature difference, diamond can conduct 2000 Watts of heat per meter, nearly five times more than its next closest competitor, silver. This makes diamond substrates ideal for a wide range of thermal management applications, such as the electronic heat spreader.

Thermal Shock: Diamond can withstand thermal stresses of 10 million Watts per meter, 1,000 times better than Zerodur®. This property is also important for many thermal management applications.

Optical Transmission: Diamond can transmit light with wavelengths from 0.22 to over 100 micrometers, compared to 0.2 to 4 micrometers for silica. This transmission range, combined with a hardness that will prevent damage to the optic, makes diamond an ideal optical coating. Applications abound in x-ray instruments, x-ray lithography membranes, infrared sensors and windows, and microwave tube windows. Diamond's high refractive index (2.41 at 590 nm) also makes it attractive for optics and electro-optics.

[1] This list is adapted from several sources, primarily the Office of Naval Research. For more information, see references (8) and (9).

At about the same time, researchers at General Electric produced small diamond stones by subjecting bulk graphite to extremely high temperatures and pressures. Since developing this high-pressure, high-temperature (HPHT) technique, GE has sold industrial-grade synthetic diamond used in cutting tools, grinding abrasives, watch bearings, and similar applications. A few other firms have joined GE as producers of synthetic diamond and in 1991 HPHT diamond synthesis accounted for a worldwide market of $647 million (1).

While huge multinational firms built businesses around HPHT diamond, CVD diamond synthesis went largely ignored until the late 1970s. Around this time, a group of Russian researchers added hydrogen to the CVD precursor mix of hydrocarbon gases and began producing high-quality diamond thin-film coatings for the first time. While researchers disagree on the role of hydrogen, one prevailing model states that when the hydrogen gas disassociates, hydrogen atoms bind with surface termination sites of the diamond lattice. This binding prevents the carbon molecules' diamond bonds from decomposing into their graphitic form.

Unfortunately, because the Russians that produced these findings had poor standing in the scientific community (a previous paper of theirs on polywater, a supposedly polymeric form of water, had been thoroughly discredited), their results were not taken seriously in the United States until the 1980s. At this time, several groups of Japanese researchers confirmed the findings of the Russians. When the news eventually spread to the United States, several research teams began CVD research, but at funding levels far lower than those in Japan.

The Strategic Defense Initiative Organization (now BMDO) Innovative Science and Technology (IS&T) office, along with the Office of Naval Research, was one of the first Federal agencies to provide significant levels of funding to these U.S. efforts. The result has been the development of new and more efficient growth techniques, a more basic understanding of how growth occurs, and some solid applications developments (some of which have found their way to the marketplace). However, the big payoff—the one that the BMDO Diamond Technology Initiative set its sights on from the beginning—will not come until diamond semiconductors can be produced with cost/performance benefits that exceed those of devices made from other materials. To do so will require some fundamental breakthroughs in how diamond is produced. As a result, the BMDO Diamond Technology Initiative continues a broad-based research program to produce large-area, single-crystal diamond coatings at a cost competitive with other electronics-grade materials.

GROWTH TECHNIQUES

Today, more than ten different methods for growing high-quality diamond and diamond-like coatings exist, a list that will certainly grow in the future. Most of these methods are some variant of the original CVD technique. What differentiates them is the form of energy used to turn hydrocarbon precursors into diamond. For instance, filament-assisted CVD places a wire heat element at the substrate, while plasma-enhanced CVD energizes a plasma local to the substrate surface. Other CVD processes include microwave-assisted CVD, laser-assisted CVD, and various combinations of these, such as microwave plasma-enhanced CVD.

The oxy-acetylene torch method, or combustion growth, has also received considerable attention. Although technically a CVD process, it is quite different from other CVD techniques because it occurs at room conditions (atmospheric pressure, no special clean-room requirements). In the technique, oxygen and acetylene burn to produce a mixture of carbon monoxide, carbon dioxide, hydrocarbon radicals and hydrogen in the hottest part of the flame; if this part of the flame impinges

on a water-cooled substrate, these ingredients mix to produce a diamond film. Combustion growth's biggest advantage is extremely low capital and production costs. However, the method has yet to supplant traditional CVD processes because it is difficult to control the quality of the film. Impurities in the oxy-acetylene tank are common, and the growth conditions are more difficult to control than in a traditional CVD reaction chamber.

In laser ablation, another growth technique, a pulsed laser heats a carbon source, thereby producing an energized vapor of carbon atoms that settles on a substrate in a diamond form. Different laser ablation processes have produced several forms of diamond at substrate temperatures significantly lower than in CVD (35°C versus 1000°C). The lower substrate temperature allows researchers to deposit diamond on a wider variety of surfaces since the substrate is less likely to crack during deposition or cooling. Like combustion growth, this process does not provide as precise control of the deposition process as CVD does. Also, laser ablation rarely produces pure diamond, but rather some form of diamond-like carbon (diamond crystals mixed with impurities such as hydrogen or graphite).

In addition, some physical vapor deposition techniques—such as atomic layer epitaxy (ALE), molecular beam epitaxy (MBE), and ion implantation—are under development that have much slower growth rates than CVD, but provide extremely fine control over the surface reactions. These methods have such precise control because growth does not occur through a large-scale chemical reaction as in CVD, but by localized reactions that limit growth to as little as one atomic layer at a time.

So far, no process has emerged as clearly superior to the others, technically or economically, and it isn't likely that any will in the near future. According to the technical cost model of Busch and Dismukes, four CVD processes—combustion flame, DC arc jet, hot filament, and microwave—vary in cost from $92 per carat for combustion flame to $22 per carat for DC arc jet. However, by the year 2020 all four techniques could produce diamond coatings for less than $5 per carat (2). In addition, non-CVD processes should fill several niches. For example, laser ablation and other processes should fill a low-cost niche even though they produce lower-quality diamond-like carbons. After all, diamond's material properties exceed other materials' by so much that lower-cost diamond-like carbons can meet many engineering needs (3). On the other end of the spectrum, interest in developing a way to deposit the large-area, single-crystal coatings needed for microelectronic applications will guarantee that new growth techniques continue to receive considerable attention.

MARKETS FOR DIAMOND COATINGS

Diamond coating technology today is at a crossroads. It has been several years since *Science* named it the "Molecule of the Year," and the early excitement of discovering a modern-day alchemy has subsided. Solid market developments are now needed to sustain the early enthusiasm. So far, products made with diamond coatings have made their first tentative steps into the marketplace, and several companies are positioning themselves for spiraling growth in the next decade.

For its part, the BMDO Diamond Technology Initiative has helped bring three products to the marketplace, nurture three start-up companies that recently completed initial public offerings of common stock, foster 22 collaborative ventures designed to introduce future products, and generate at least 51 patent applications (31 of which have been granted to date). Most of the commercial progress has occurred in small-volume markets like x-ray windows, custom coatings, and certain machine-tool applications. Other larger markets, such as flat-panel displays that use diamond cold cathodes, are just now emerging.

Crystallume (Menlo Park, CA) provided the first notable commercial success, both for the BMDO Diamond Technology Initiative and the diamond CVD world in general, when in 1989 it became the first company to introduce a CVD diamond-coated product (an ultra-thin diamond coated x-ray window). Since then, Crystallume has introduced two other products (a microelectronic heat spreader and a diamond-coated tungsten carbide cutting tool) and completed an initial public offering of stock in March 1994. This stock offering has helped Crystallume raise money needed to introduce higher-volume products, such as diamond-coated rotating tools, diamond-coated ceramic bearings, and radiation-resistant microelectronics.

SI Diamond Technology, Inc. (Houston, TX), another participant in the BMDO Diamond Technology Initiative, became the first publicly owned diamond start-up when the company completed an initial public offering in early 1993. In one of SI Diamond's many ventures, the company is jointly developing (with the Microelectronics and Computer Technology Corporation) a diamond cold cathode electron source used in flat-panel display screens; in June 1994, SI Diamond announced a proof of concept for such a display (the world's first) at the Society for Information Display International Symposium. In addition, the company opened a commercial diamond coating facility in 1994.

Other notable commercial stories include the following:

- Advanced Technology Materials, Inc. also recently became a publicly owned company and is developing its own diamond-based flat-panel display in a joint venture with Silicon Video Corporation and the Massachusetts Institute of Technology's Lincoln Laboratory.

- Research Triangle Institute recently entered a joint venture with the 3M Company to refine its diamond-coating process and scale it to production requirements.

- Penn State University's Diamond and Related Materials Consortium (recently renamed the Diamond and Related Materials Information Consortium) has helped several U.S. companies launch diamond-coated products.

Despite this progress, the total sales value of products based on CVD diamond was no more than $10 million in 1992 (1), nothing more than a blip on the screen compared to sales of natural and HPHT synthetic diamond. So far, CVD diamond has found nothing more than a few niche markets coating small, high-value items where performance needs greatly outweigh cost considerations.

However, several near-term, higher-volume applications are on the horizon, including tool coatings, wear parts, and passive electronic devices that provide thermal management and electrical isolation. Further down the road, the largest market for diamond coatings should be for active electronic devices such as microprocessors, memory circuits and power semiconductors. Most of these devices must be made of large-area, single-crystal diamond and doped with n- and p-type materials, capabilities still in their infancy.

One active electronic device that can be made using current diamond coating technology is diamond cold cathodes (4). As a result, flat-panel displays provide one of the most promising markets for diamond thin films. Due to diamond's unique electronic properties, diamond cold cathodes will emit electrons at much lower power levels than other materials. In addition, unlike virtually any other cathode, it is not easily poisoned by exposure to air, water, and other environments (3). Because diamond cold cathodes can provide these advantages without any advances in coating technology, they have rapidly proceeded into device integration. As mentioned earlier, two participants in the BMDO Diamond Technology Initiative have flat-panel display development projects underway.

The market for flat-panel displays is currently $3.7 billion and should reach $9.4 billion by 2000 (5). While diamond-based displays won't account for the full share of that market, diamond-based field-emission displays (FEDs) will be easier to manufacture than today's leading contender, active matrix liquid crystal displays, and will be much brighter than conventional FEDs. As a result, diamond-based FEDs should capture a large share of that $9.4 billion market. If so, diamond-based FEDs could provide the revenue stream that will allow firms involved in diamond coating research to continue developing other active semiconductor devices.

Diamond coatings are clearly making commercial headway. Still, diamond research has a long way to go before it lives up to early expectations. With so many technical variables, it's impossible to tell just how big the diamond market will be. To overcome this problem, Russell has developed a scenario-based forecast that estimates how fast markets for diamond coating technology will grow in the next 30 years. Each scenario—one pessimistic, one most probable, and one optimistic—depends on how well diamond CVD researchers address cost barriers and technical hurdles (1).

In the most probable scenario, the total market for diamond thin-film coatings would reach $15 billion by 2020, tripling today's market for diamond grits, powders, and stones. Of the $15 billion, $12 billion would be for microelectronic devices. Most of that $12 billion would result from diamond cold cathodes, with some military and small niche markets also playing a role. In Russell's most optimistic scenario, single-crystal diamond films prove to be a viable competitor to silicon semiconductors, and the market exceeds a trillion dollars. However, in his least optimistic scenario, technical progress proceeds slowly and cost issues limit applications development. In this case, the market would be about $5 billion, no larger than today's market for natural diamond (6).

RESEARCH AND DEVELOPMENT ACTIVITIES

While firms involved in diamond CVD research have matured to a new, more competitive position in the last few years, many basic research questions remain unanswered. For instance, researchers still cannot grow single-crystal diamond on a non-diamond substrate, a key requirement if anyone is going to make diamond semiconductors that are competitive with silicon and other materials. As a result, the BMDO IS&T office continues to fund several major research efforts, primarily focused on diamond microelectronics. These efforts can be broken into three categories: new production techniques, growth theory refinements and new device developments.

PRODUCTION

While researchers today can grow polycrystalline coatings on any number of substrates, most of the high-payoff applications in diamond electronics require single-crystal coatings. Several researchers have demonstrated single-crystal growth, but they have done so only on natural diamond substrates (a process known as homoepitaxial growth). Unfortunately, homoepitaxial growth is too expensive and, because large-area diamond substrates do not exist in nature, cannot occur over a large enough area to make diamond microelectronics cost-competitive with today's devices, even in highly specialized applications.

The best way to make large-area, single-crystal diamond coatings cost-competitive is to grow them on a substrate other than diamond (a process known as heteroepitaxial growth). Unfortunately, lattice mismatches between diamond and non-diamond substrates create epitaxial stresses that cause the diamond coating to crack or decompose into graphite. Lattice mismatches are also a major problem in many other diamond coating applications, such as machine tool coatings and optical coatings, where these stresses create poor adhesion between the diamond coating and, say, a tungsten carbide tool bit.

One way to encourage heteroepitaxial growth is to make a non-diamond substrate look more like diamond. By doing so, researchers can increase the number of sites where nucleation occurs. A non-diamond substrate can be made to look more like diamond by scratching the substrate with a diamond powder or etching diamond-shaped pits into the substrate. One participant in the BMDO Diamond Technology Initiative, Epion, Inc. developed expertise in this etching process during a project to produce thin diamond lenses. Epion holds a patent for this technique, which is known as selective nucleation of diamond.

Another possibility avoids the problem of heteroepitaxial growth altogether. Instead, researchers are working to manufacture large-area, diamond substrates, something nature doesn't provide. Two BMDO-funded research groups—Research Triangle Institute (RTI) and the Massachusetts Institute of Technology's Lincoln Laboratory—are making large-area substrates by piecing together less expensive, smaller diamond crystals (either natural or synthetically grown). In addition, a company that received funding from the BMDO Small Business Innovation Research program—Linares Management Associates, Inc.—has developed a proprietary technique to produce free-standing, homoepitaxial diamond substrates.

An entirely different approach seeks to grow diamond on a non-diamond substrate such as silicon or copper by minimizing the lattice stresses between the materials. To do so, researchers grow an interlayer of another material between the substrate and the diamond. Because this interlayer more closely matches the lattice structure of diamond, epitaxial stresses are reduced and diamond is more likely to grow. Three BMDO-funded groups are working in this area. Wayne State researchers are working on cubic-boron nitride interlayers grown between silicon and diamond, while RTI researchers are working on molybdenum interlayers grown between copper and diamond. In addition, researchers at North Carolina State University discovered, with the help of a scanning tunneling microscope, that wherever single-crystal diamond starts to grow on silicon (which so far only occurs over microscopically small areas), a silicon carbide interlayer naturally forms between the two. Using this knowledge, the North Carolina State researchers have developed a new growth technique known as bias-enhanced nucleation that shows promise for single-crystal growth on silicon.

More fundamental approaches seek to alter the growth chemistry of diamond to prevent the uneven growth that characterizes polycrystalline diamond. To prevent this uneven growth, researchers have begun to employ self-limiting growth, which is provided through two routes. The first route uses new growth chemistries, such as halogen replacements for hydrogen, while the second route uses new growth processes. A good example of such a process is atomic layer epitaxy, which limits growth to an atom-thick layer at a time.

Researchers are investigating new growth chemistries in the hopes of finding a chemical that better prevents diamond seed crystals from decomposing into graphite. One BMDO-funded group, Aerodyne Research, Inc., is investigating self-limiting reactions between diamond seed crystals and halogens. Halogens such as fluorine and chlorine are more difficult to remove from the seed crystal than hydrogen, which means they better maintain the diamond structure. While Aerodyne is primarily focusing on fluorine, the company has also investigated the role of chlorine, water, oxygen and alkyl halides. In addition, RTI and Penn State University are investigating the role of oxygen in CVD; one possibility involves a role similar to hydrogen and halogens in preserving diamond's crystalline structure. Rice University researchers have also developed several new growth chemistries, including one that uses chlorine to more efficiently produce atomic hydrogen. The technique, called chlorine-activated CVD, provides higher growth rates at lower substrate and gas temperatures. Another technique developed at Rice, fluorocarbon pyrolysis CVD, takes

advantage of elemental fluorine's ability to react with graphite but not with diamond. The technique uses a cyclic growth/etch process to remove graphitic carbon from the seed crystal.

BMDO has also funded several promising new growth techniques, including one at West Virginia University and another at SI Diamond Technology, Inc. The West Virginia technique is known as catalyzed diamond epitaxy, or CDE, and occurs in an ultrahigh vacuum CVD chamber. CDE offers well-controlled growth, much like atomic layer epitaxy. As a result, West Virginia researchers can use silicon instead of hydrogen to stabilize the diamond surface structure and generate vacant surface sites. Silicon bonds to surface termination sites more weakly than hydrogen, which results in faster growth rates. SI Diamond's technique, in turn, is known as seeded supersonic beam deposition. This intriguing process uses the collision between gas feedstock molecules and the substrate to provide the activation energy needed to produce diamond coatings. The supersonic expansion of a light carrier gas provides the necessary acceleration of the gas feedstock molecules. Because this expansion can precisely control kinetic energy, the technique offers well-controlled, epitaxial growth.

GROWTH THEORY

Any effort to develop new production techniques must be complemented with an improved theoretical understanding of how diamond growth occurs. In hopes that new theoretical insights will lead to improved growth techniques, BMDO has funded several such theoretical investigations. The two leading theoretical investigations are underway at Penn State University and the Naval Research Laboratory: both are involved in molecular-level studies of diamond growth.

In addition, methods to better characterize diamond films have complemented this theoretical work. In particular, both Ohio University and North Carolina State University have produced important insights into diamond growth through their work in new characterization techniques. Ohio University's work centers around two forms of electron emission microscopy, which allows researchers to directly observe changes in the surface of diamond at the relatively high pressures and temperatures required for CVD diamond growth. As a result of this ability, Ohio University was an important contributor to the discovery that diamond emits electrons in a low-field, room-temperature environment, leading to today's work in diamond-based flat-panel displays. North Carolina State University researchers, in turn, have employed a wide variety of characterization techniques, from scanning tunneling microscopy to photoemission spectroscopy. One of the most important results of this work is the discovery that small-scale heteroepitaxy on silicon is accompanied by a silicon carbide interlayer, a discovery discussed earlier in relation to interlayers.

DEVICE DEVELOPMENTS

Another thrust of the BMDO Diamond Technology Initiative involves efforts to make working microelectronic devices using diamond. The two major issues here involve doping and forming contacts. Because it is extremely difficult to add impurities to diamond without destroying its crystalline structure, researchers have successfully doped diamond only with boron and carbon to make a p-type semiconductor. So far, no one has been able to reliably add n-type dopants. In addition, researchers have had some trouble forming electrical contacts that meet key technical parameters for diamond electronics, such as high-temperature survivability (greater than 500°C), strong adhesion, low contact resistance, and precise dimensional control. These problems, though, do not seem to be as troublesome as those associated with doping.

Researchers at the Naval Command, Control and Ocean Surveillance Center (NCCOSC) developed some of the world's first working diamond transistors, metal-insulator-semiconductor field-effect transistors or MISFETs. Using two of these MISFET transistors, they have also formed a simple

circuit with a voltage gain of 2x. To produce these transistors, the NCCOSC researchers developed a patented technique for forming reliable ohmic contacts to diamond and a doping technique that uses ion implantation of carbon and boron.

MIT's Lincoln Laboratory has also formed working diamond microelectronics: research there currently revolves around improving the rectifying characteristics of mercury-diamond Schottky diodes. They have found that plasma treatments with some gases (N_2 or CF_4 with 8.5 percent O_2) reduce the leakage current in Schottky p-type diamond diodes (other gases, such as N_2O, H_2 or O_2 can increase the leakage current). In addition, the Lincoln Laboratory researchers have developed an annealing process that can substantially increase the forward conductance of the diode.

Finally, ThermoTrex Corporation is developing three-terminal, diamond transistors that circumvent the need for doping. Instead, the company's transistor uses an electron-beam that, when directed on the diamond transistor, creates carrier pairs in the diamond and temporarily turns the diamond into a conductor. With the electron-beam serving as the transistor's switching terminal, ThermoTrex has built diamond switches and millimeter wave amplifiers.

CONCLUSION

Now that diamond can be applied as a coating to substrates of various compositions and shapes, many observers see a revolution in the world of engineered materials, some even declaring that we are about to enter a "New Age of Diamond." To those who hope to take advantage of diamond's extraordinary range of properties, Russell offers some important insights into the structure that revolution might take (1).

Historically, the HPHT diamond market has had a high barrier to entry. HPHT systems cost several million dollars and as a result, the industry was dominated by a few large firms. In contrast, CVD systems cost between $30,000 and $60,000. One recent article in *Life* even describes how "high school student Lea Potts can make diamonds with her dad's welding torch." (7) With lower barriers to entry, the emphasis in diamond coatings will be on innovation, not access to the tools of production. As a result, even though the diamond coating research has significantly matured in the past decade, there is still room for new entrants to the field.

To help new entrants learn from the experiences of those who have gone before them, the BMDO Office of Technology Applications has produced a booklet describing the research efforts of 25 contractors who have participated in the BMDO Diamond Technology Initiative. To obtain a copy of this booklet, call (703) 518-8800 ext. 500.

REFERENCES

1. Russell, C. J. "CVD Diamond Markets in the 21st Century," *Synthetic Diamond*, ed. Spear and Dismukes, pp. 625-647, (John Wiley & Sons, Inc.: 1994).

2. Busch, John V. and Dismukes, John P. "A Comparative Assessment of CVD Diamond Manufacturing Technology and Economics," *Synthetic Diamond*, ed. Spear and Dismukes, pp. 581-624, (John Wiley & Sons, Inc.: 1994).

3. Yoder, Max N. "The Vision of Diamond as an Engineered Material," *Synthetic Diamond*, ed. Spear and Dismukes, pp. 3-17, (John Wiley & Sons, Inc.: 1994).

4. Kania, Don. "The Promise and Challenge for CVD Diamond Electronics," Presentation to the Gorham Advanced Materials Conference, March 8, 1994.

5. Schmidt, Howard K. "Commercializing the Diamond FED," Presentation to the Gorham Advanced Materials Conference, March 8, 1994.

6. Russell, C. Jim. "Forecasting CVD Diamond Growth," Presentation to the Gorham Advanced Materials Conference, March 8, 1994.

7. "Diamonds in the Family Garage," *Life*, March 1993.

8. Yoder, Max. "The Diamond Technology Initiative," Presentation to the Gorham Advanced Materials Conference, October 15-17, 1989.

9. "CVD Diamond Application Areas," Promotional slide describing properties of diamond, produced by Crystallume, 1989.

ACOUSTIC ATTENUATION AND VIBRATION DAMPING MATERIALS

Gerald B. Thomas
Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola, FL 32508-1046

William B. Cushman
Poiesis Research, Inc.
1315 Finley Drive
Pensacola, FL 32514

ABSTRACT

In the course of creating an enhanced hearing protector for use in very high noise environments, a composite technology capable of effectively blocking acoustical energy, particularly low frequency acoustical energy, was developed. Existing sound-absorptive materials are generally very good as barriers for acoustical noise above 500 Hz but are uniformly poor at lower frequencies. In fact, few manufacturers even provide data for frequencies below 125 Hz and, for those that do, attenuation values of about 15 dB at 31.5 Hz have represented practical maximums for relatively thin (i.e., <.25") configurations. We have developed composite materials that are true bulk attenuators and routinely offer 30+ dB of attenuation at 31.5 Hz while retaining excellent attenuation at the higher frequencies and good weight characteristics. This proprietary composite technology has significantly improved the acoustical energy absorbing properties of a wide range of materials including silicone and organic rubbers, polyurethanes, and epoxy resins. Potential applications of the technology range from hearing protection to the automotive and building construction industries to electronic packaging to sporting goods. Comparison data and current applications research are described.

INTRODUCTION

While developing a new-concept earcup for protection from very high levels of environmental noise, a class of materials particularly effective at blocking low-frequency acoustical energy was created.

Because the design of the new hearing protector required acoustical isolation of its critical parts, research was begun into the isolation characteristics of commercially available sound absorbing materials. With minor variations, all of the available materials exhibited similar sound blocking characteristics: excellent blocking of frequencies above 500 Hz and generally poor attenuation at frequencies below 500 Hz. Furthermore, few manufacturers even report attenuation values below 125 Hz, a region of particular interest to us because an initial application of the new hearing protector was to be in military helicopters, which have large amounts of low-frequency acoustical energy.

Figure 1 illustrates the acoustical energy-blocking characteristics of various types of commercially available materials as well as of in-house fabricated materials. These materials ranged from proprietary polymers to foil-backed petroleum laminates to laboratory-created traditional composites. Extant energy-absorbing foams were not included because of their generally very poor low-frequency attenuation characteristics. As can be seen, the energy- blocking capabilities of these conventional materials are excellent above about 500 Hz but are wanting at the lower frequencies. Their weights, in the tested configuration, ranged from about 7g to about 12g.

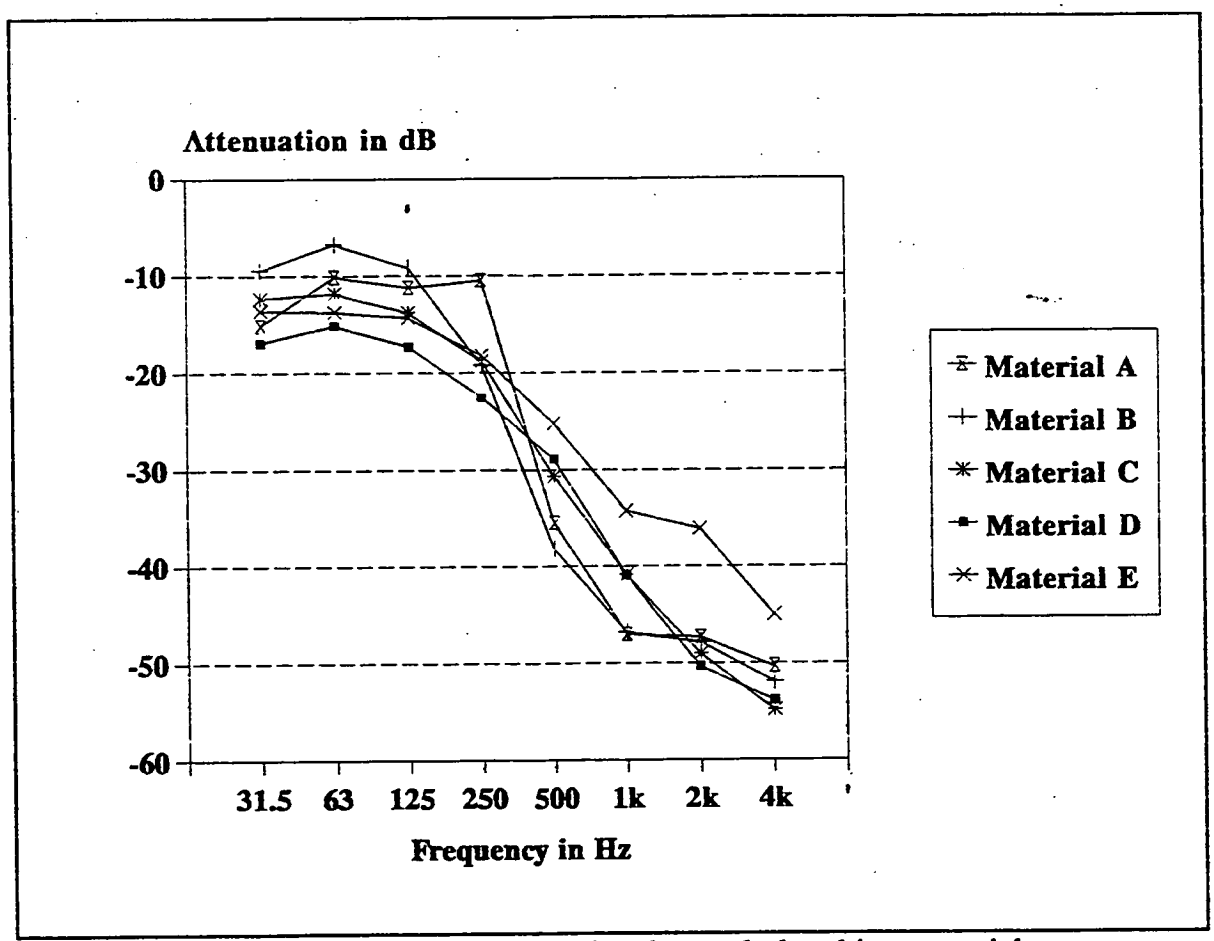


Figure 1. Sound attenuation of conventional sound absorbing materials.

These materials, as well as all others reported in this paper, were tested in a high-noise semi-reverberant test chamber at overall pink noise levels of 120 dB (SPL). Materials were placed over a sensing microphone housed in a flat plate coupler, and the resulting spectrum was submitted to a Fast Fourier Transform, the results of which were subtracted from a similar transform generated with the microphone uncovered. The derived attenuation difference reflected the amount of acoustical energy blocked by the material.

EXPERIMENTAL MATERIALS

Confronted with the problem of inadquate low-frequency acoustical energy isolation, we began fabricating our own composite materials, largely following traditional "Mass Law" dictums. Increasing the mass of an experimental polyurethane isolation gasket by loading the polymer with 75% (by weight) lead powder did increase its low-frequency energy absorption characteristics significantly. Unfortunately, it also tripled the weight of the gasket (to 29.8g) and because weight was a serious consideration in our hearing protector design, this mass loaded gasket was deemed unsuitable for this application. We, therefore, began exploring alternative approaches to the problem and developed a working hypothesis around a conceptually promising new approach.

This new approach involved the careful selection of filler materials based on certain inherent characteristics and on the characteristics of the chosen base material. The exact formulation of material proportions is extremely critical and is empirically determined. (For legal reasons, the exact nature of the underlying mechanisms cannot be disclosed pursuant to nondisclosure agreements with potential licensees; full disclosure will occur with patent issuance.)

After 40 attempts at implementing the concept, Gasket Material-41 (GM-41) was the result. GM-41 is a relatively high-durometer polyurethane to which the composite technology was first successfully applied. Figure 2 contains the mean of the attenuation values for the commerically available and traditional materials displayed in Figure 1, the curve for the unacceptably heavy lead powder polyurethane, and the curve for GM-41. It will be noted that GM-41 is more effective than the lead composite material at attenuating the low frequencies but,in fact, weighs approximately the same (11.2g) as the commercially available materials.

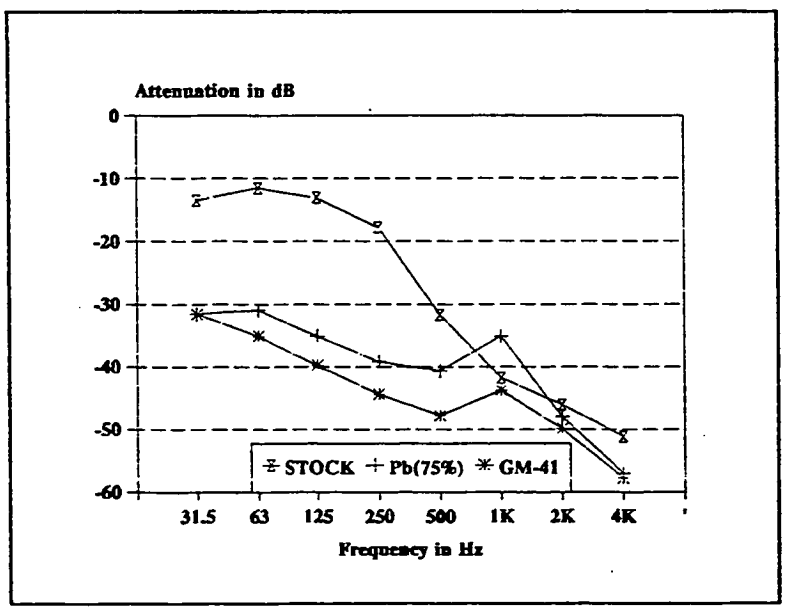


Figure 2. Mean attenuation of conventional, Pb-loaded, and GM-41 materials.

Subsequent research indicated that the total attenuation of the composite material was greater than the sum of the attenuation of its constituent parts. This indicated that an interaction was occurring, which produced a synergistic attenuation effect.

Additional research confirmed that the resulting composite was a true bulk attenuator (i.e., the amount of attenuation was dependent on the thickness of the material). Figure 3 illustrates the relative attenuation of different thicknesses of a silicone rubber material to which the technology had been applied. It will be noted that there is an approximate 6 dB change in attenuation as the material's thickness is halved or doubled.

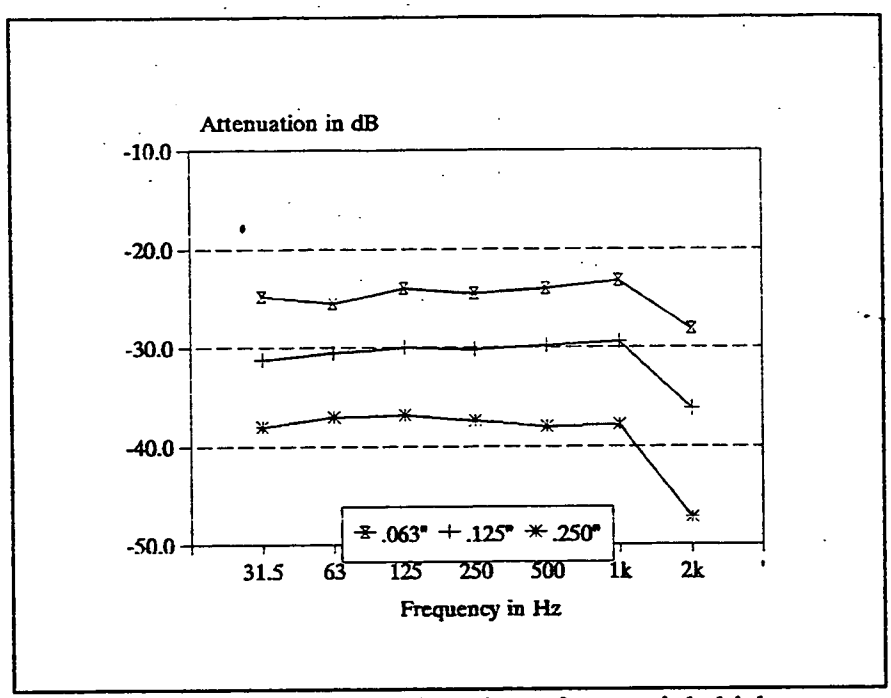


Figure 3. Attenuation as a function of material thickness.

Physical Characteristics

Because this remains an emerging technology, not all physical parameters associated with the technique have been measured. That is, we have not yet conducted tests of the effects of the technology on tensile strength, temperature range, shear resistance, etc. However, it is our subjective opinion that the base material to which the technology is applied will largely determine the composite's basic physical properties.

Manufacturability

To date, this technology has been applied to such materials as virgin and reprocessed silicone rubbers, polyurethanes, carbon-based rubbers, and epoxy resins. Further, it appears not to alter the basic manufacturability of the base material. That is, so far it has been extruded, cast, and blown with no apparent problems.

APPLICATIONS

Hearing Protection

Because our initial interest was in creating a more effective hearing protector (in the form of an advanced earcup), much of our research effort has centered around this goal.

Conventional earcups are typically constructed of ABS or similar plastics which, while inexpensive and easy to manufacture, do not act as bulk attenuators and provide little protection from the lower acoustical frequencies. With sponsorship by the Naval Medical Research and Development Command and by the Office of Naval Research, we have been able to complete research on a new earcup material currently being evaluated for possible use in future hearing protectors.

Figure 4 contains the attenuation curves for this epoxy resin based material along with a representative curve for conventional earcup materials. These curves are only illustrative because earcup wall thicknesses differed slightly among the samples, and the curve for the experimental earcup is extrapolated to some degree. This extrapolation was necessary at frequencies above 125 Hz because with the experimental earcup the limits of our measuring equipment were being exceeded.

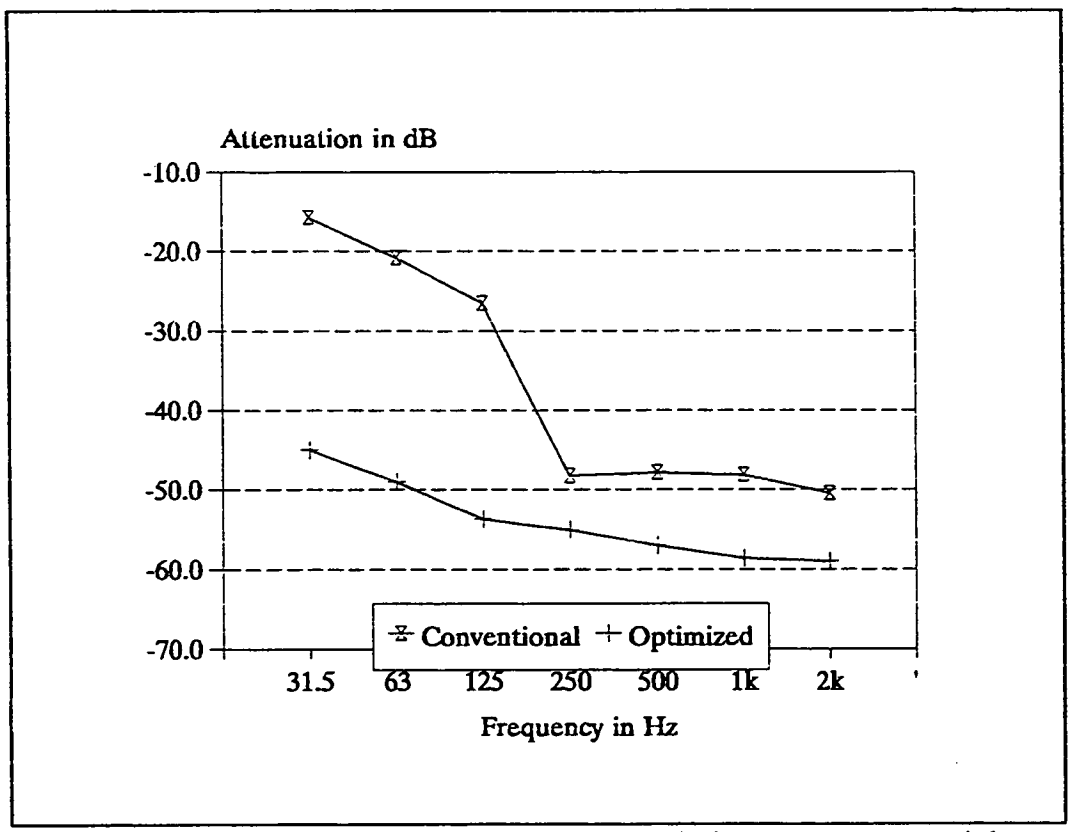


Figure 4. Attenuation of conventional and optimized earcup materials.

Current research in this area is centering around applying the technology to ear seal materials. Future research will investigate the application of the technology to other thermosetting plastics and thermoplastics for possible use in ear plugs and helmets.

Transportation Industry

As we were nearing completion of the earlier phases of the research into this technology, members of the Gulf Coast Alliance for Technology Transfer (of which our laboratory is a member) were informed of our endeavor. One company, an automotive supplier, immediately expressed a keen interest. As a result, we began applying the technology to the family of rubber materials routinely used in the automotive industry.

Figure 5 illustrates the improvement in noise attenuation gained by applying the technology to a virgin silicone rubber (thickness: .20").

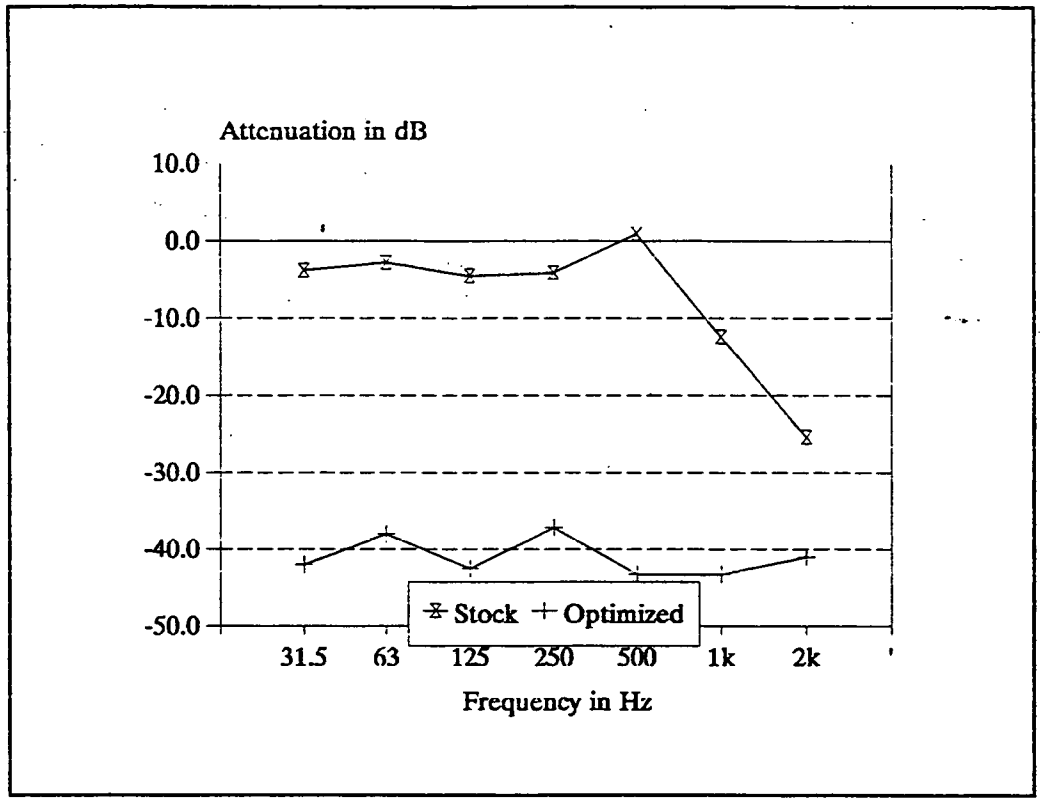


Figure 5. Attenuation of stock and optimized silicone rubber.

Similar results were obtained when the technology was applied to reprocessed silicone rubber and to carbon based rubbers.

Figure 6 presents a "by request" test run comparing a material currently used by the Toyota Motor Company in their steering column boots with a similar material to which the subject technology had been applied. It will be noted that the Toyota material is virtually acoustically transparent at the lower frequencies while an optimized silicone rubber of identical thickness (.040") is approximately 20 dB better. Similar results were obtained when other car manufacturers submitted samples for comparison testing. It is anticipated that this technology could be applied to a number of automobile components including bushings, weather stripping, panels, hoses, etc.

In addition to the automotive industry, interest has been expressed by truck, marine, and aerospace companies. For example, Boeing Helicopter is investigating the material for use in their Chinook helicopters, and the U.S. Navy has inquired about its use in its Landing Craft Air Cushion vehicles.

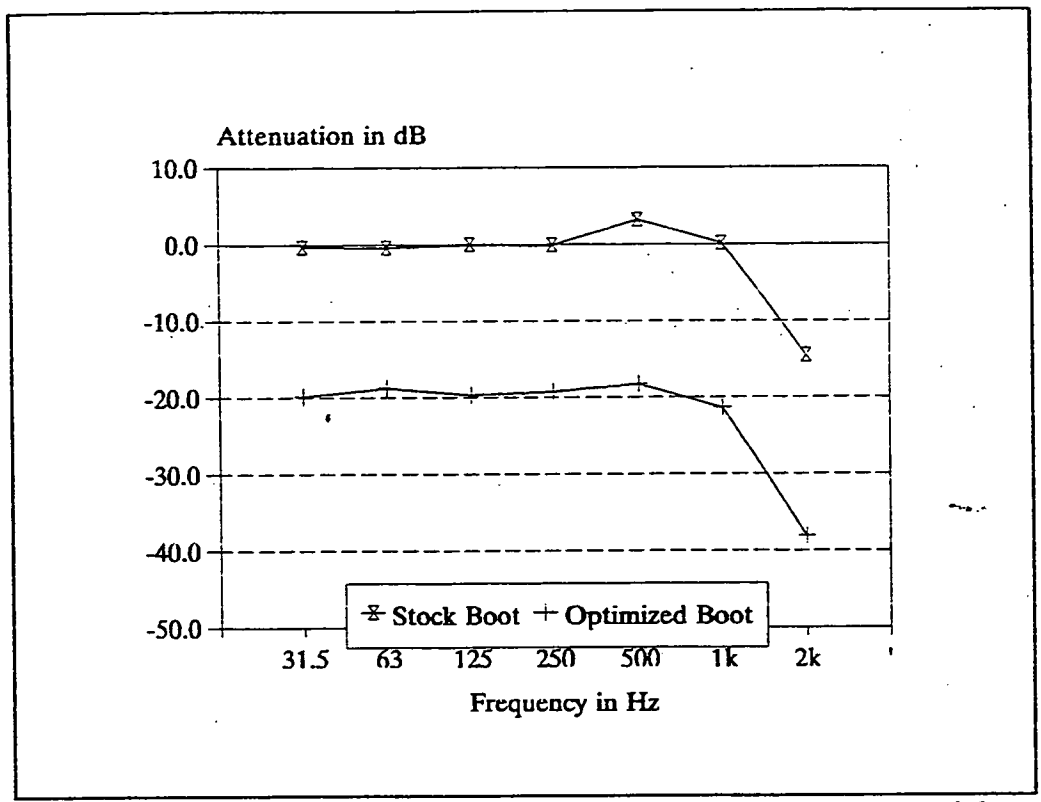


Figure 6. Attenuation of stock and optimized steering boot material.

Building Construction

Current state-of-the art techniques for blocking the vertical transmission of sound in multi-story buildings include the "floating floor" technique. This presently is the most effective system for reducing upper floor to lower floor noise transmission but is complex and expensive to build. Approximately .25" of our improved silicone rubber yields the same or better noise attenuation at what would presumably be a significant cost reduction.

Although it has not been empircally verified, the authors believe that the technology could be applied to a gypsum substrate so that sheet rock walling could more effectively block the horizontal transmission of sound and be less expensive than the traditional "double wall with air space" alternative. Similarly, a single layer of improved silicone rubber in a multi-ply plywood should result in improved sound isolation.

Electronic Packaging

Shock isolation for fragile electronic circuitry is another area where the present technology holds promise. Because initial shock waves in dropped electronic devices tend to have periods equivalent to frequencies below 500 Hz, materials employing the present technology could provide energy absorption superior to existing shock-absorbing materials.

The exact formulas for candidate materials would have to be empirically derived, however, because all of our work to date has concentrated on air-borne (as opposed to structure-borne) pressure waves.

Within the area of air-borne acoustical energy isolation of electronic devices, we were asked to fashion a transducer housing for a new electronic stethoscope. This stethoscope was developed by a researcher at the Harvard Medical College and the Veterans Administration for use in medical evacuation helicopters and ambulances, venues where large amounts of low frequency noise are present. After having tried several commercially available sound-absorbing materials, the researcher selected a silicone rubber housing which had been optimized with the subject technology.

Sports Equipment

Energy absorption in sports equipment is an important safety and performance consideration. In some instances, maximal energy absorption is desirable (e.g., helmet linings) while in others an optimal balance of absorption and reflection is sought (e.g., running shoes). This technology appears to lend itself to both situations because the degree of energy absorption can be controlled but, again, this supposition has not been empirically verified.

Other

As with any emerging technology, much additional research remains to be undertaken. For example, the authors believe that a sprayable material incorporating the technology can be developed. Further, to date, absolutely no research has been conducted into the area of laminates, where even improved performance can be expected. Successful implementation of these and other techniques should result in additional areas of application.

SUMMARY

An emerging technology for improving the low-frequency acoustical energy-absorbing qualities of a number of materials was discussed. This composite technology has been successfully applied to a number of materials and manufacturing techniques, and its potential applications range across a broad spectrum of products.

DEVELOPMENT OF ENVIRONMENTALLY COMPATIBLE SOLID FILM LUBRICANTS

Howard L. Novak
USBI Co.
P.O. Box 21212; USB-IRD
Kennedy Space Center, Florida 32815

Phillip B. Hall
National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Department EH-14
Marshall Space Flight Center, Alabama 35812

ABSTRACT

Multi-body launch vehicles require the use of Solid Film Lubricants (SFLs) to allow for unrestricted relative motion between structural assemblies and components during lift-off and ascent into orbit. The Space Shuttle Solid Rocket Booster (SRB), uses a dual coat, ceramic-bonded high temperature SFL in several locations such as restraint hardware between the SRB aft skirt and the Mobile Launch Platform (MLP), the aft SRB/External Tank (ET) attach struts, and the forward skirt SRB/ET attach ball assembly. The proposed National Launch System (NLS) may require similar applications of SFLs for attachment and restraint hardware. A family of environmentally compatible non-lead/antimony bearing alternative SFLs have been developed including a compatible repair material. In addition, commercial applications for SFLs on transportation equipment, all types of lubricated fasteners, and energy related equipment allow for wide usage's of these new lubricants. The new SFLs trade named BOOSTERLUBE is a family of single layer thin film (0.001 inch maximum) coatings that are a unique mixture of non-hazardous pigments in a compatible resin system that allows for low temperature curing (450°F). Significant savings in energy and processing time as well as elimination of hazardous material usage and disposal would result from the non-toxic one-step SFL application. Compatible air-dry field repair lubricants will help eliminate disassembly of launch vehicle restraint hardware during critical time sensitive assembly operations.

INTRODUCTION

USBI Co. is responsible for the assembly and refurbishment of the non-motor components of the SRB as part of the Space Shuttle system shown in Figures 1 and 2, and which is developed and managed by Marshall Space Flight Center (MSFC) in Huntsville, Alabama. Programs are underway to develop and evaluate environmentally acceptable SFLs for use on aerospace flight hardware in order to eliminate lead, mitigate corrosion by substitution of graphite pigments, and ultimately extend the useful service life of these unique and expensive hardware items. Figure 3 shows the typical areas of lubricant application on SRB flight hardware. All SRB bearing material is made from Inconel 718 which interfaces with either Inconel 718 or 15-5 P.H. materials. The initial study focused on the replacement of a lead bearing dual coat SFL with that of a unique environmentally compatible single coat system. The existing dual coat SFL system required a complex masking and basecoat application and then curing at 1000°F, followed by additional masking and the application of a topcoat followed by a second curing operation at 500°F. The basecoat contains lead and other hazardous air pollutants as does the topcoat and extensive precautions are taken during the application and disposal of such materials. The new family of single coat SFLs trade named BOOSTERLUBE do not contain lead or hazardous air pollutants and can be easily applied in a single step with curing at 450°F. A compatible air dry SFL was also developed as a field repair material. Both the MSFC, and the Pratt & Whitney Government Engine - Space Propulsion (P&W, GESP) Tribology Laboratories assisted in the testing of the new lubricants. E/M Corporation located in West Lafayette, Indiana helped manufacture the unique SFL system. A significant joint effort was required between USBI Co. and E/M Corp. in order to establish a production test article in a relatively short time frame. Initial screening of all lubricants was performed at E/M Corp. The new SFL also had to be compatible with the existing production equipment at Kennedy Space Center in order to make the process economical. Subsequent application of SFLs to all qualification test articles were performed by production technicians at USBI Co. production facilities.

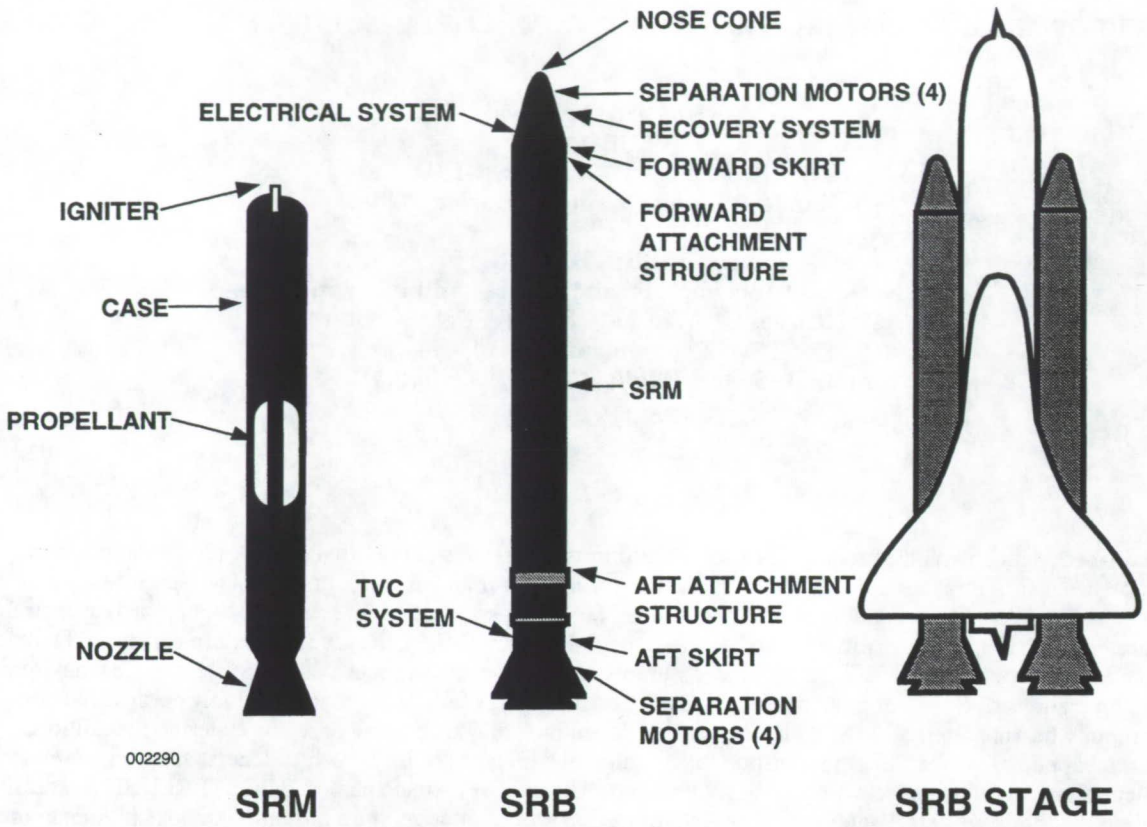


Figure 1. Space Shuttle's SRB

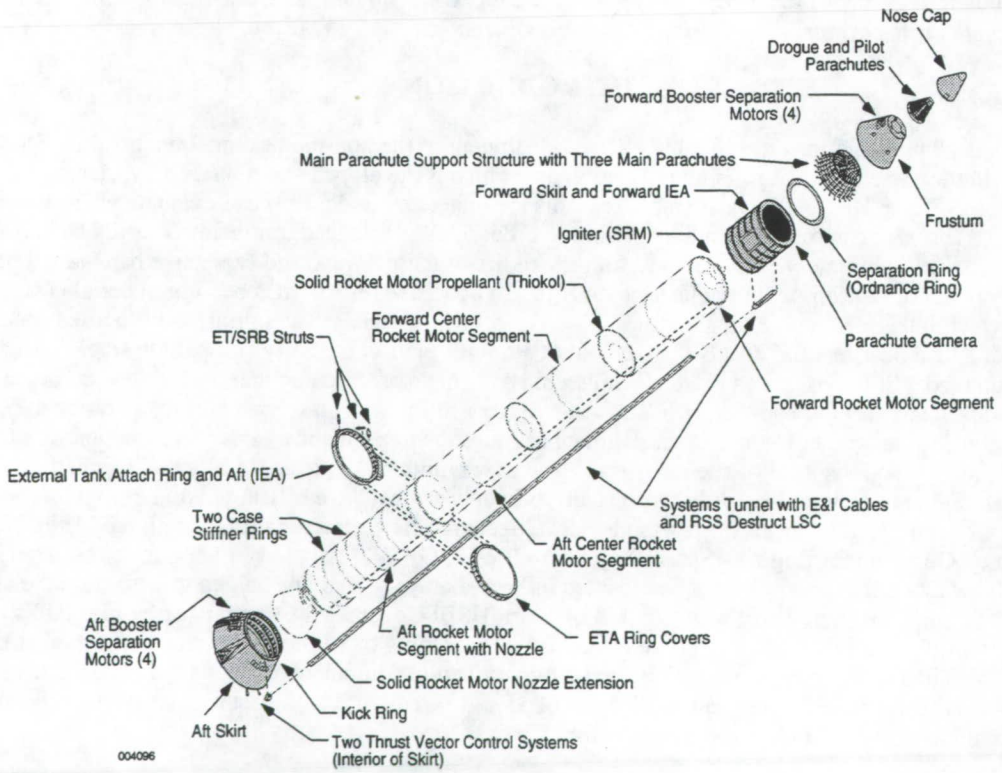


Figure 2. Solid Rocket Booster

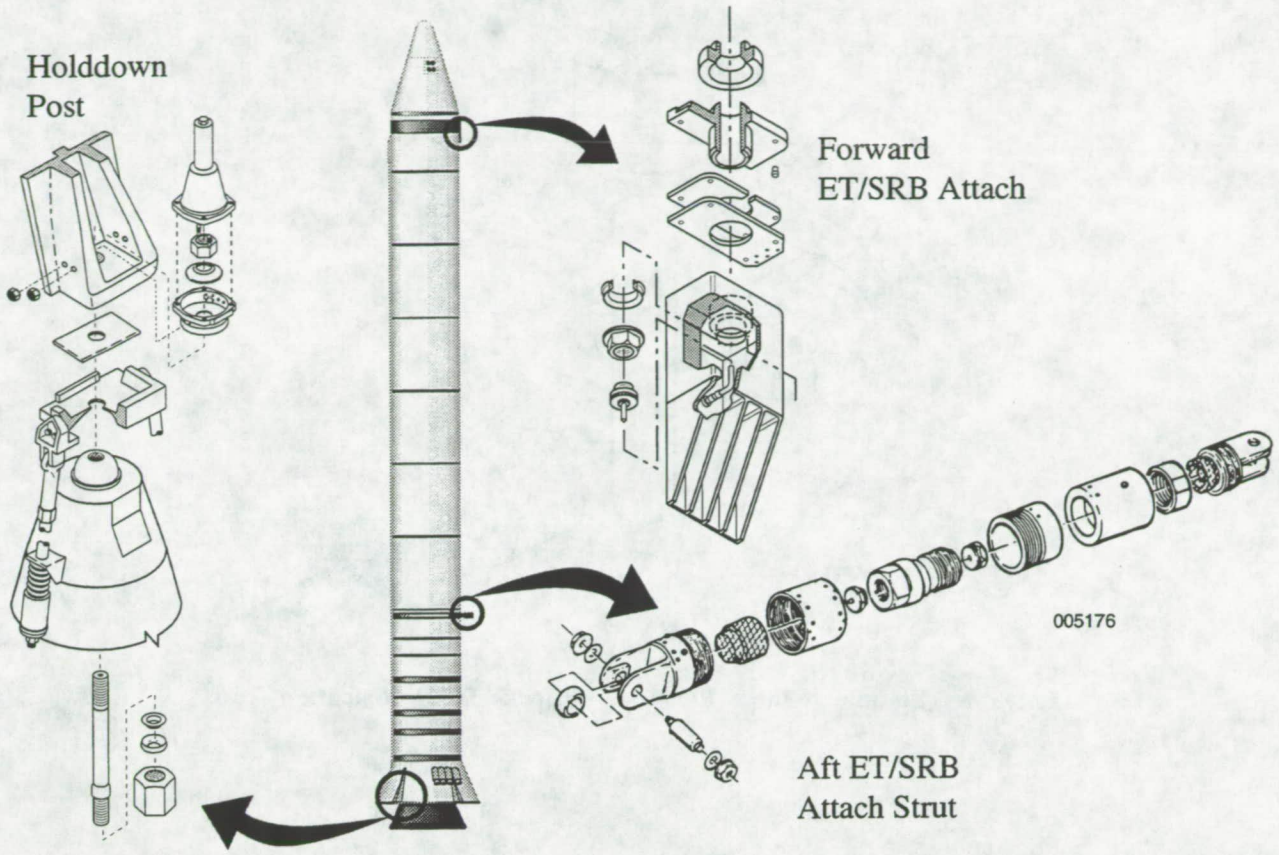


Figure 3. Typical Areas of Lubricant Application on SRB Flight Hardware

DISCUSSION

Pigment and Resin Evaluation and Selection

The first phase of this program involved the selection and screening of five resin systems consisting of epoxy, silicone, phenolic, polyimide and epoxy/silicone blends. Pigments consisting of molybdenum disulfide, graphite, antimony trioxide, lead oxide and boron oxide were all evaluated. The existing dual coat ceramic bonded lead oxide bearing system was used as a baseline for coefficient of friction, load bearing capacity, and endurance life. Testing of the various SFL systems were performed in accordance with ASTM D 2625 Parts A & B (1) (Falex Pin and V-Blocks, testing endurance life and load carrying capacity) as shown in Figure 4; ASTM D 2714 (1) (Falex Block-on-Ring, testing coefficient of friction and wear life) shown in Figure 5, and a mono-ball test developed by MSFC and an alternate device developed by USBI Co. and P&W, GESP shown in Figure 6. The MSFC mono-ball test simulates the very high load conditions at the holddown post and upper external tank attach point, and is the most critical of all tests for verifying SFL performance. Screening the above lubricant systems resulted in the selection of an epoxy/silicone resin blend with pigments consisting of molybdenum disulfide, graphite, and boron oxide. The substitution of graphite with boron oxide, where required, provided a means of eliminating galvanic corrosion when in contact with dissimilar metals in a sea water or salt fog environment. Development of a new device for extreme pressure testing of SFLs allowed USBI Co. and MSFC to run their test programs in parallel, thus saving many months of serial time. Another future part of this program involved the potential application of Ion Vapor Deposited (IVD) aluminum to CRES fasteners that experience dissimilar metal interfaces and potential galvanic corrosion. This is the case on the SRB between 2219 aluminum alloy structures and A286 fasteners. Future testing and evaluation is contemplated for BOOSTERLUBE in such environments.

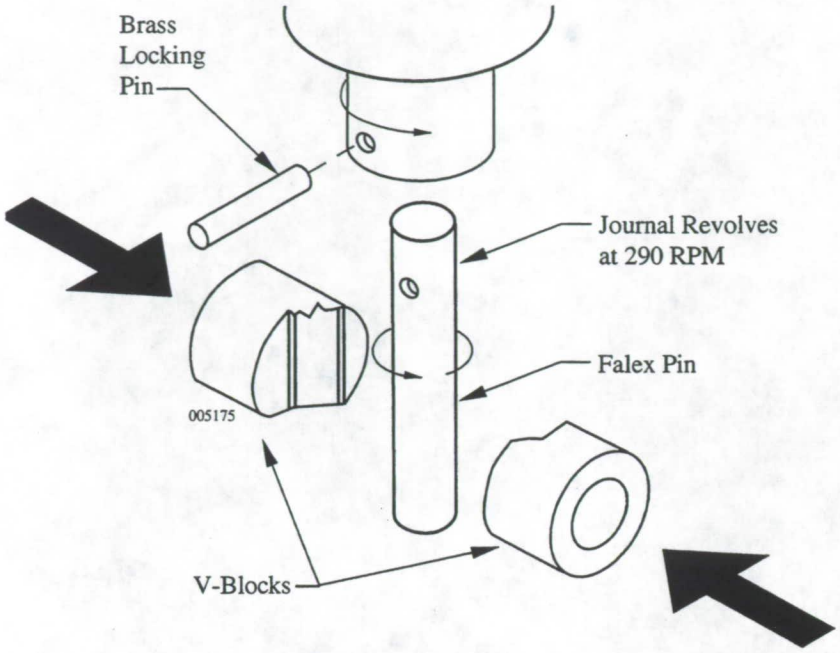


Figure 4. Standard Falex Pin and V-Blocks for Lubrication Tests

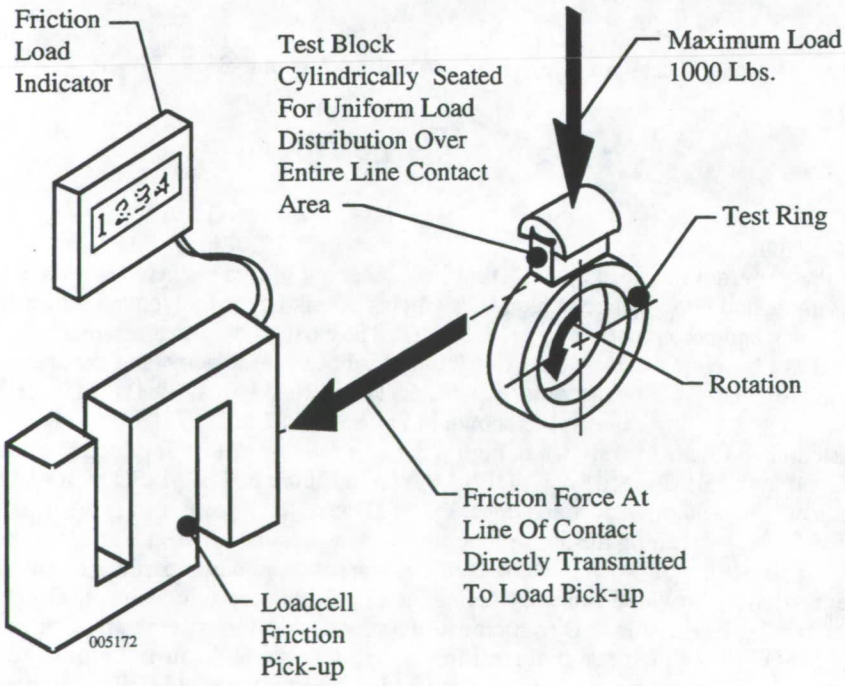


Figure 5. Function Diagram of the Falex Block-On-Ring Testing Machine

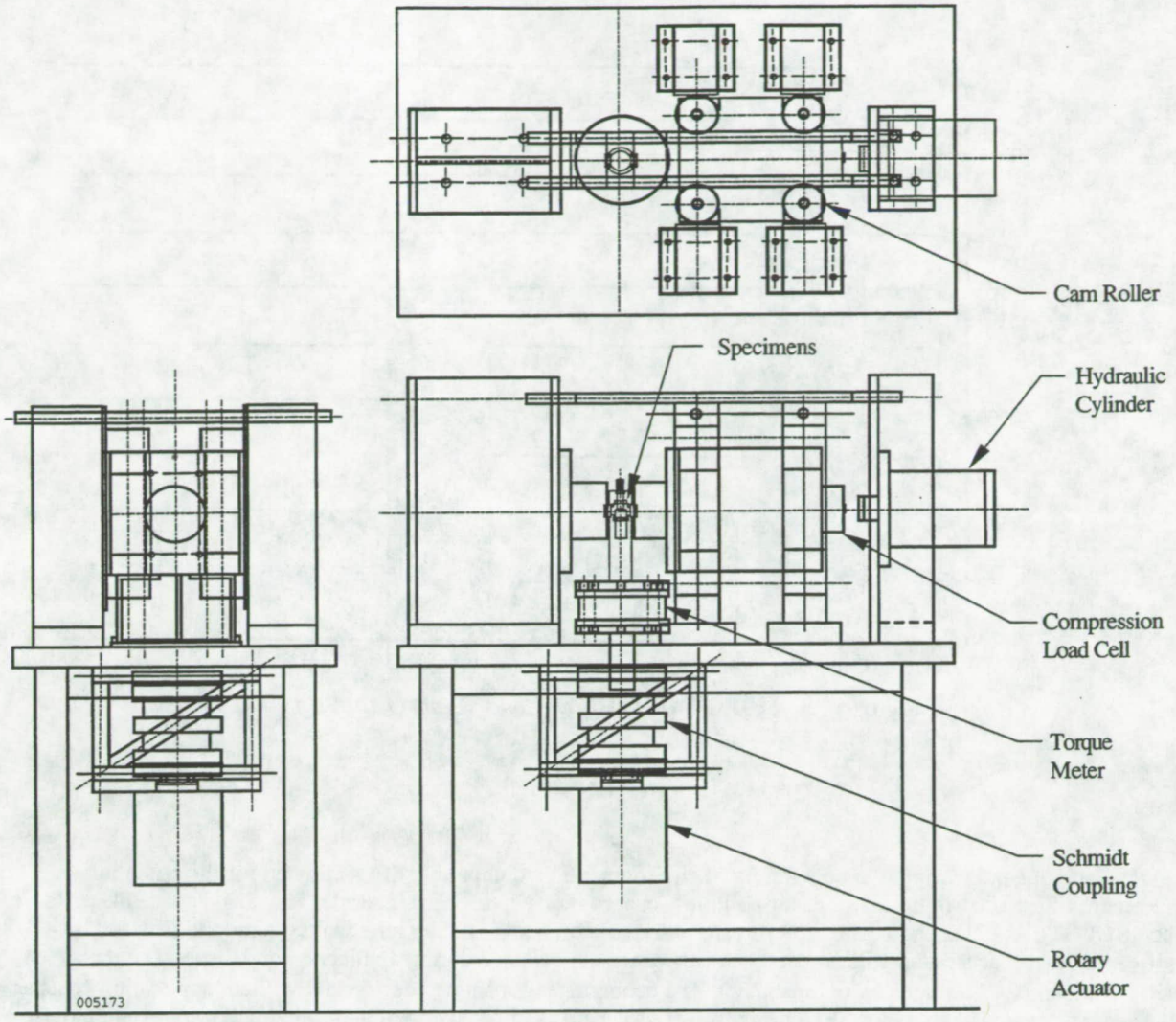


Figure 6. USBI Co./P&W Mono-ball Tester

Test Results

Initial testing of BOOSTERLUBE properties at the manufacturer's site, using the Falex Pin and V-Block, ASTM D 2625 Parts A & B and the Falex Block-on-Ring, ASTM D 2714 confirmed the acceptable properties of the Silicone/Epoxy resin system. The Block-on-Ring test showed a significant improvement in wear life of the BOOSTERLUBE materials as compared to the Ceramic Bonded dual coat SFL. Coefficients of friction (μ) were 0.8 - 0.12 and durability of the new SFLs were excellent as confirmed by burnishing with a lint-free cloth. One of the problems associated with the original dual coat system was the relatively soft topcoat that was prone to damage. Significant inspection time was required to determine whether surface marks were caused by actual damage or by burnishing during normal handling. BOOSTERLUBE is a much tougher resin system and is less susceptible to marking and/or damage. A compatible air dry repair lubricant was also developed to allow for minor area repairs to lubricated parts that are inadvertently damaged in the field. Testing of its properties using ASTM procedures and the MSFC mono-ball tester are underway. Final tests of BOOSTERLUBE were performed on optimized pigment volume concentrations using MoS2/graphite and MoS2/boron oxide. This, in combination with optimum blends of silicone and epoxy resin systems, allowed for tailoring properties to their required applications. All test specimens were precision cleaned, grit blasted, lubricated, and oven cured by production technicians and on production line facilities at USBI Co. operations. Specimens were then sent to MSFC, Tribology Labs. for qualification testing. Results of testing can be seen in Figure 7. The most relevant of all tests, MSFC/EH14 High Load Test (mono-ball test) showed superior properties of BOOSTERLUBE (WL-1158-20) over the existing ceramic bonded dual coat SFL (Ceramic B) and second only to Ceramic A, a material that also contains lead and hazardous air pollutants.

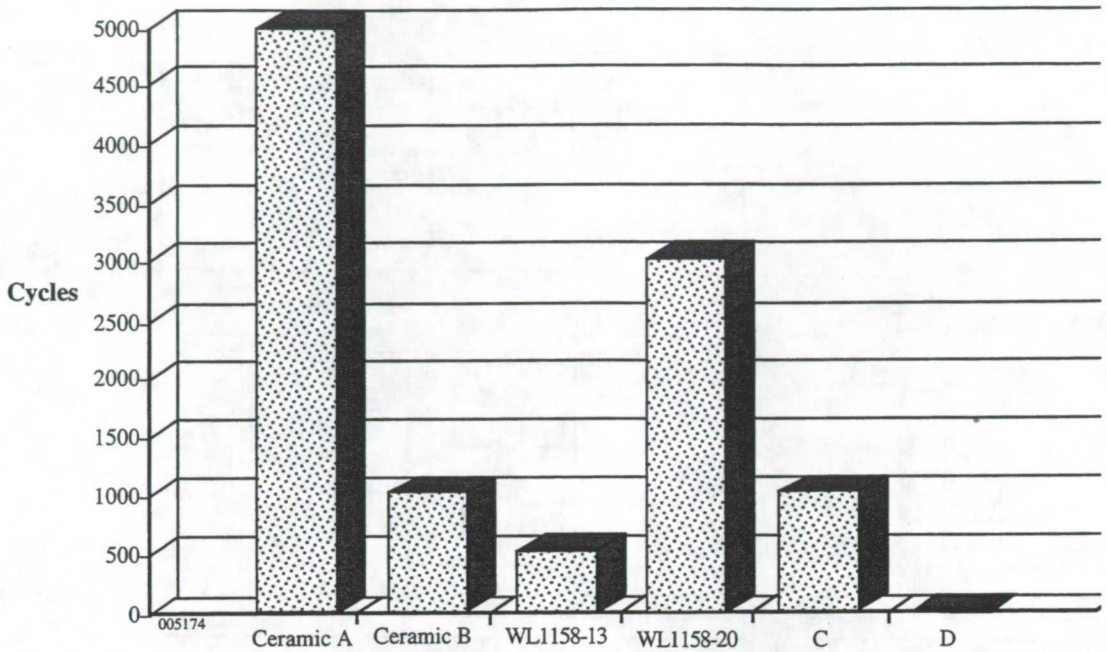


Figure 7. MSFC/EH14 High Load Test (110 KSI)

CONCLUSION

Development and testing of a new family of environmentally compatible SFLs allowed for the potential replacement of lead bearing, hazardous air polluting, energy inefficient, SFL materials. The single coat SFL trade named BOOSTERLUBE, showed excellent performance as compared to other lubricant systems and met with significant operator acceptance. Eliminating lead exposure to personnel during application and disposal of those materials was in itself a major accomplishment. The economics of a single coat system will cut application costs as well as energy related charges for SFL curing, by 50%. The costs of disposing of and utilizing hazardous materials continue to escalate and in time will be prohibitive. Applications for many other fields than aerospace exist. Energy and power generation, chemicals, transportation, machinery of all types, exo-atmosphere and low earth orbit vehicles can benefit from implementation.

ACKNOWLEDGMENTS

The authors would like to thank the various USBI departments both in Huntsville, Alabama and Kennedy Space Center, Florida for supporting and assisting in program management, production operations, logistics and testing. Many thanks to NASA Marshall Space Flight Center Materials and Processes Laboratory, Tribology Branch, and the SRB Program Office personnel who were supportive of this program and to Pratt & Whitney Government Engines Space Propulsion Tribology Branch for help in developing the new lubricant test machine. Additional thanks are given to E/M Co. management and laboratory personnel who helped make this program a success.

REFERENCES

(1). Annual Book of ASTM Standards, Vol. 05.02, "Petroleum Products and Lubricants," ASTM, Philadelphia, PA 1992.

Unicoat—Durable, Environmentally Compliant
Polyurethane Paint

NOT AVAILABLE AT THIS TIME

MAGNETIC LEVITATION TRANSPORT SYSTEM FOR USE IN UNDERGROUND MINING AND MATERIALS HANDLING

J. J. Geraghty
Electronics Engineer
U. S. Bureau of Mines
Denver Research Center
Building 20, DFC
Denver, CO 80225

William E. Wright
Mechanical Engineer
U. S. Bureau of Mines
Denver Research Center
Building 20, DFC
Denver, CO 80225

J. A. Lombardi
Mining Engineer
U. S. Bureau of Mines
Denver Research Center
Building 20, DFC
Denver, CO 80225

ABSTRACT

The U.S. Bureau of Mines (USBM) is developing a magnetic levitation (maglev) materials transport system to improve the safety and reduce the cost of underground mining and materials handling. The transport system combines permanent magnets, for levitation of the materials containers, with an electronic control system to maintain noncontact positioning of the containers in a dedicated transit corridor.

The Maglev Transport System uses two types of permanent magnets: neodymium-iron-boron (NdFeB) and ceramic-5, arranged in arrays. An array of NdFeB magnets is contained in the base of each levitated materials container; an array of ceramic magnets lines the bottom of the transit corridor; the orientation of the magnets is such that the two arrays repel each other. The electronic position control system, located on the levitated materials containers, overcomes the inherent lateral instability of the repelling magnet arrays. Cost estimates, based on current design data, would make the Maglev Transport System cost competitive with a conveyor belt system of equivalent capacity.

INTRODUCTION

For many years, the standard materials transport system in underground coal mining has been

the conveyor belt. Conveyor belts are the most cost-effective of current haulage methods, but they introduce numerous safety hazards. The primary hazard is that of inadvertent worker contact with an operating belt. Conveyor belt systems also contribute to the amount of respirable dust to which miners are exposed and have been found to cause mine fires from overheated rollers or belt friction points [1]. The costs associated with the resultant injuries and fatalities from belt system safety hazards have an adverse effect on the overall cost of coal production.

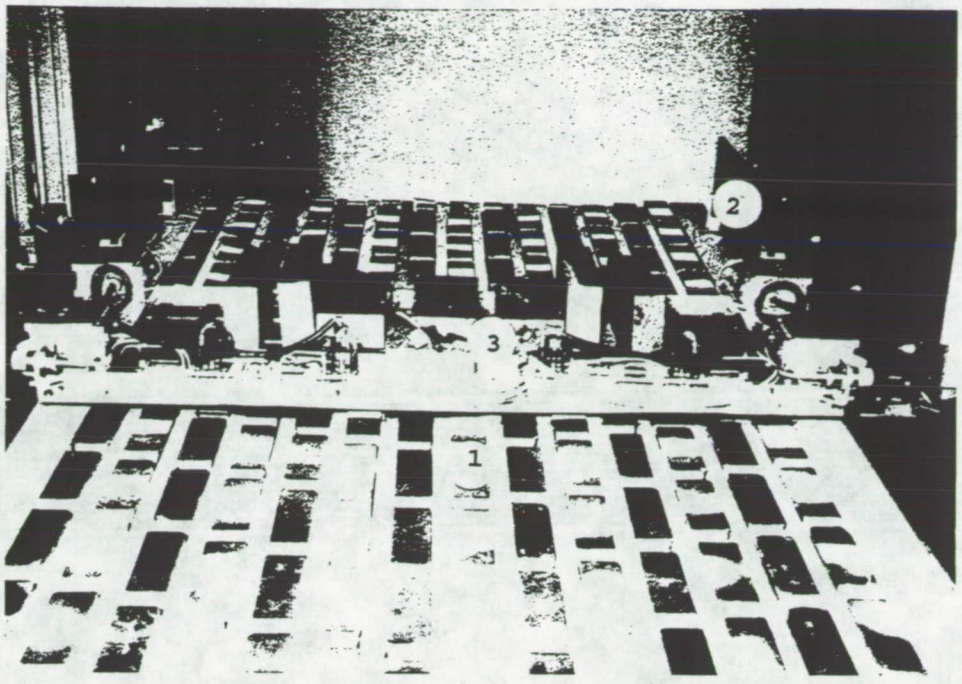
The U.S. Bureau of Mines (USBM) recognized the need for a system that would improve the safety and reduce the cost of underground coal haulage. A coal haulage system integration and design study identified pipeline transport systems as the safest, most promising alternative to conveyor belt systems. Coal transport in pipelines was reviewed and analyses were made as to why past pipeline transport attempts were unsuccessful. From these analyses, the containerized transport of coal in a pipeline was determined to be viable and to meet current coal transport requirements. It was decided that the optimum containerized coal, pipeline transport system design would incorporate magnetically levitated, air-driven containers.

It was determined that maglev transport system research would involve radically new haulage system engineering, constituting a high risk effort. By virtue of the inherent high risks and relatively high costs, the mining industry could not be expected to pursue this research. The high percentage of mining accidents attributable to underground coal haulage by conveyor belts, coupled with the USBM's commitment to improving the safety and efficiency of the Nation's mines, prompted the USBM to undertake this research.

SYSTEM DESCRIPTION

The Maglev Transport System is being designed specifically for the transport of coal from an underground coal mine to a surface facility. The system design calls for individual or linked, magnetically levitated materials containers to be pneumatically propelled through a concrete pipeline. The bottom of the concrete pipeline would be lined with permanent magnets, for container levitation, and have steel siderails embedded in the interior to enable container steerage.

The Maglev Transport System is being developed as a half-commercial-scale prototype model (see figure 1). The model is composed of three components: a transit corridor, a levitated materials container, and a container-mounted electronic position control system. The transit corridor and levitated container house arrays of permanent magnets. The magnet arrays are oriented to repel each other, thereby effecting suspension of the materials container above the transit corridor. The electronic position control system maintains noncontact positioning of the materials container within the transit corridor.



Key: (1) transit corridor magnets
(2) transit corridor steel siderails
(3) levitated materials container

Figure 1. Prototype Maglev Transport System. The materials container, loaded with approximately 85 kg, is magnetically suspended 4 cm above the transit corridor.

Transit Corridor

The prototype transit corridor, which is approximately 122 cm wide, consists of a wooden base with attached steel siderails. The base houses a chevron-shaped array of oriented ceramic-5 magnets, arranged in 14 alternating polarity columns (see figure 2). The spacing between the columns and rows of magnets is dictated by two factors. First, the spacing must be sufficient to avoid significant field interference between adjacent magnets. Such interference acts to reduce the repelling suspension force of the corridor array [2]. Second, the spacing must be set such that there is a nearly continuous area of magnet interface between the corridor and the materials container during transit. This second factor is interdependent with the spacing of the magnet array in the levitated materials container.

A steel siderail is mounted on both sides of the corridor extending the entire corridor length. The siderails, which are 15.24 cm tall by 1.27 cm thick, are a necessary element for the operation of the electronic position control system. The electronic position control system

utilizes proximity sensors and electromagnets to determine and maintain a centered position of the levitated container within the transit corridor. The sensors and electromagnets require the presence of a soft magnetic material, such as steel, on both sides of the transit corridor.

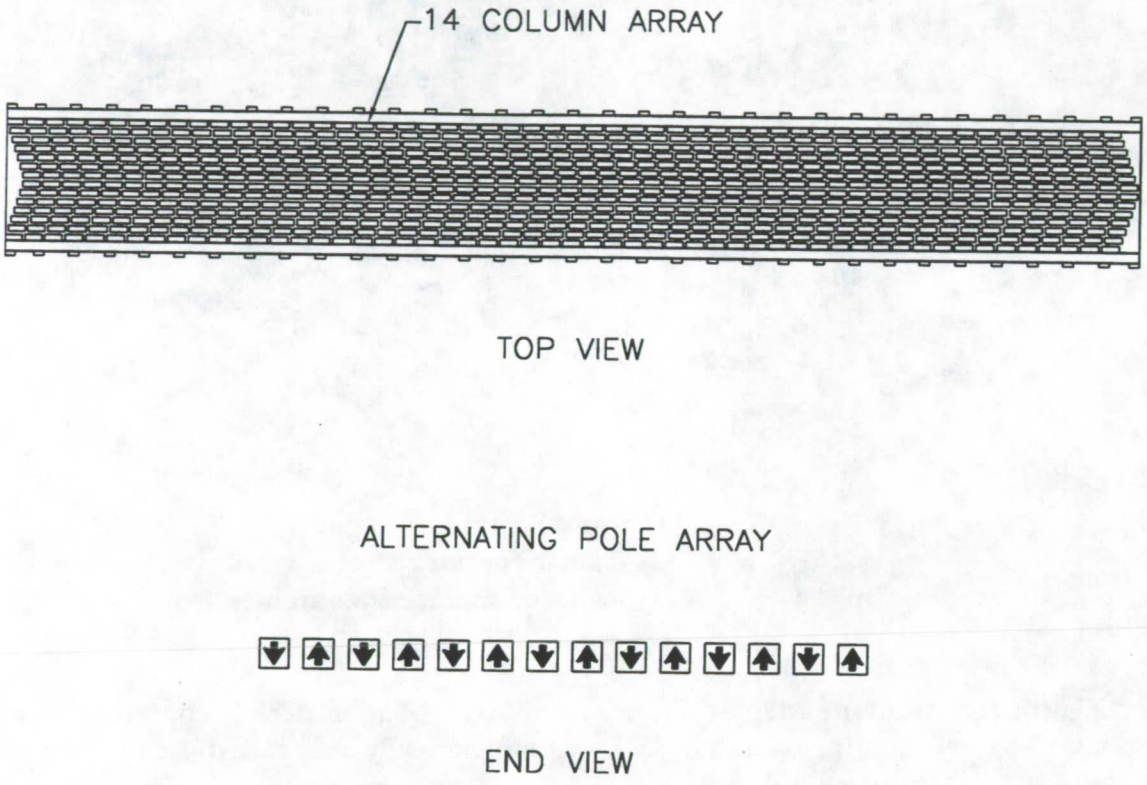


Figure 2. Transit corridor magnet array.

Materials Container

The prototype materials container consists of a nonmagnetic, rectangular body that houses a rectangular array of oriented neodymium-iron-boron magnets. The prototype container, which is approximately 119 cm wide, 60 cm long, and 5 cm tall, represents the base of an actual materials container unit, without the sidewalls. The weight of the prototype materials container, including the magnets and onboard electronic position control system, is approximately 58 kg. The container magnet array is arranged in alternating polarity columns, opposite to the corridor magnet polarities, such that the interface is strictly repulsive (see figure 3). The container magnet array spacing is governed by the same two factors as the

transit corridor array—field interference and continuity of magnet interface area.

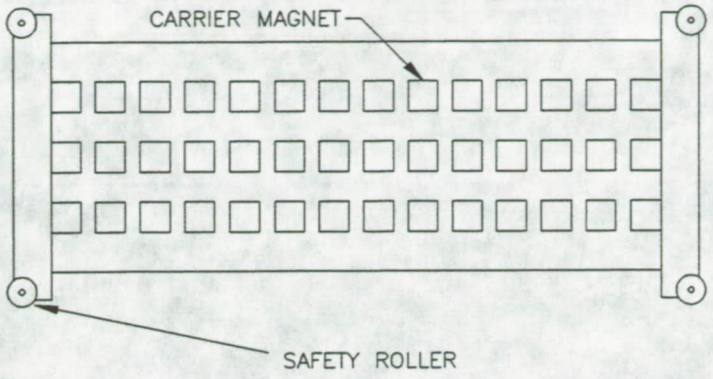
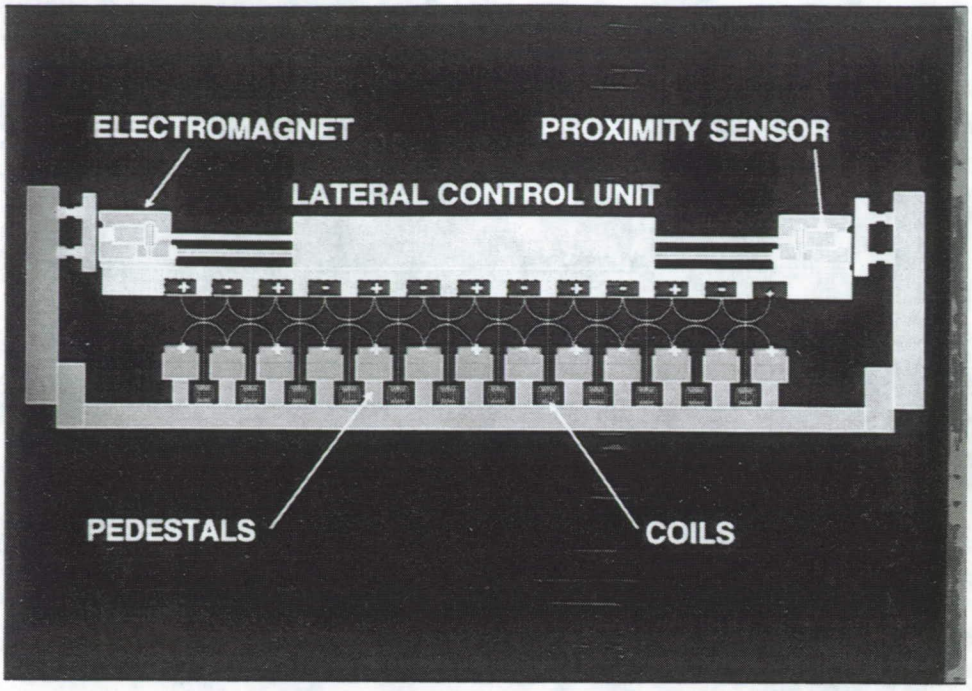


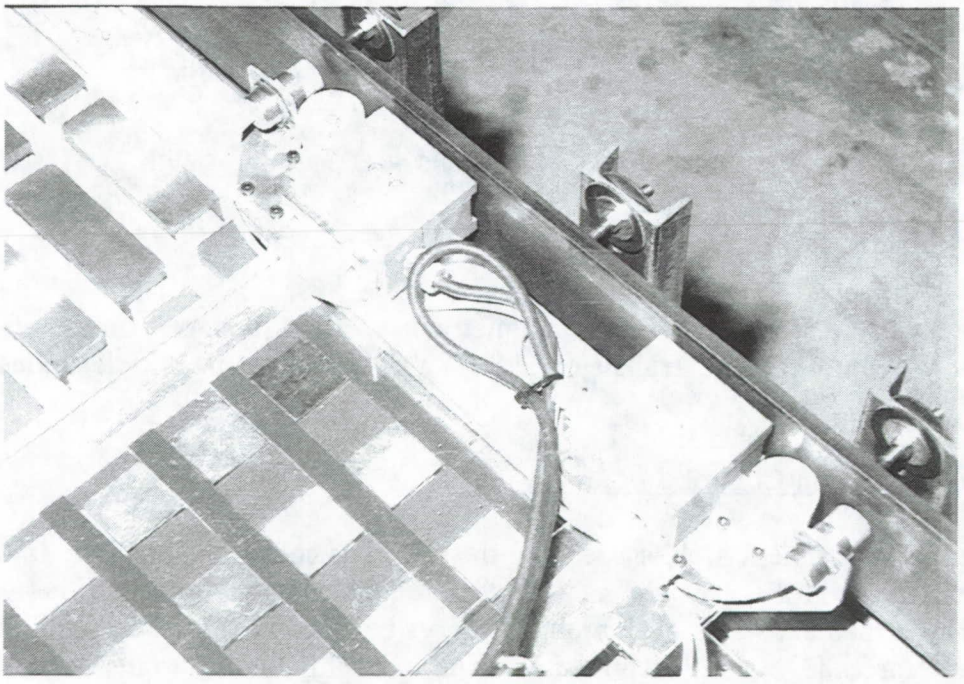
Figure 3. (a) Materials container magnet array. (b) Interface of magnets in materials container and transit corridor - arrows represent magnet poles.

Electronic Position Control System

The repulsive magnetic field suspending the materials container above the transit corridor acts, in accordance with Earnshaw's theorem, to force the container toward either corridor siderail. Earnshaw's theorem states that an object cannot be suspended in stable equilibrium by purely magnetic forces [3]. To overcome this lateral instability, an electronic position control system was installed on the levitated container (see figures 4,5). The control system uses four proximity sensors, one fore and aft on the two container sides facing the siderails, to gather information on the position of the container relative to the siderails. The proximity sensor outputs are processed by a proprietary lateral control circuit. Circuit outputs control four electromagnetic actuators, one located next to each proximity sensor. The electromagnets attract to the steel siderails when energized, and accordingly, maintain a centerline position of the materials container between the corridor siderails.



(a)



(b)

Figure 4. Electronic position control system. (a) Drawing of system installed on the materials container. (b) Proximity sensors and electromagnets on one side of the materials container.

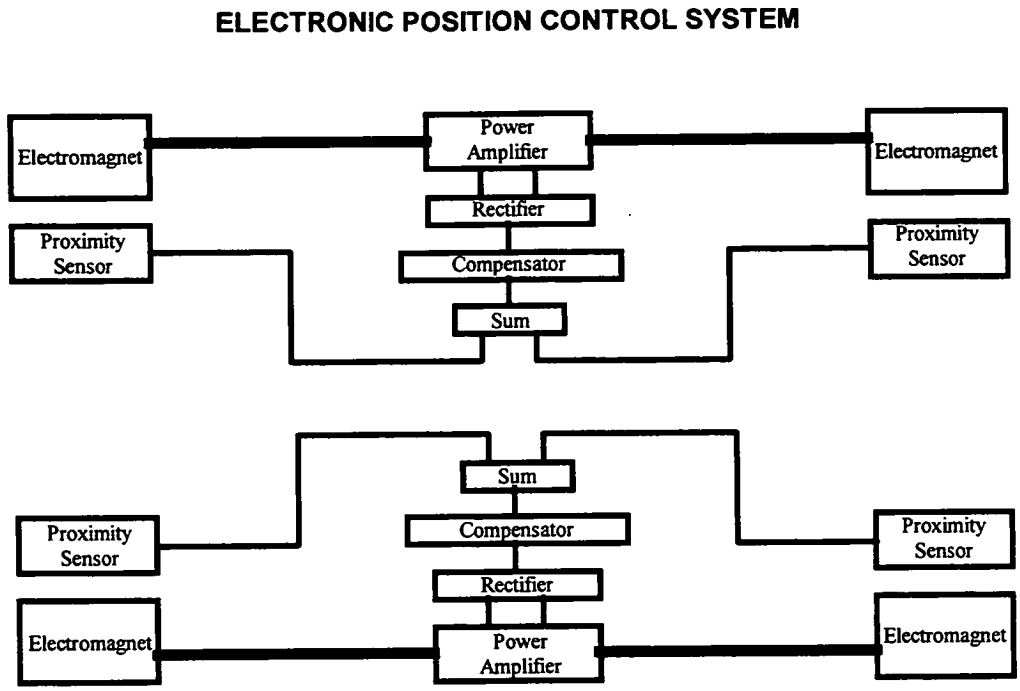


Figure 5. Block diagram of the electronic position control system.

The electronic position control system requires a direct current power source for operation. An onboard power generation system has been conceived, but not yet developed. Electrical conductors, perhaps coils, attached to the underside of the levitated materials container, could possibly be used to provide some or all of the power required by the electronic position control system. During container transit, the hanging conductors would intersect the magnetic field present between and/or above the columns of corridor magnets. The current induced in the coils by the coil-field interaction is converted to the necessary voltages and then distributed to the sensors, electromagnets, and lateral control circuit. Data was collected to provide a geometric profile of the magnetic field intensities. By incorporating this magnetic field profile into a computer model, the optimum power generation coil configuration can be determined.

LATERAL CONTROL

A major milestone in the research efforts was the development of an electronic positioning system to overcome the inherent lateral instability of the levitated materials container. The magnetically suspended container exerts lateral forces on the transit corridor siderails due to the orientation of the repulsive magnetic fields. An electronic position control system was developed to provide centered, contactless levitation of the materials container within the transit corridor.

A study was conducted to quantify the lateral force of the suspended materials container toward the corridor siderails. The study provided a profile of the change in lateral force

relative to the degree of lateral position error, ranging from no position error (container centered) to maximum position error (container contacting siderail). Figure 6 presents typical lateral load curves. Data were collected for various materials container load weights. The data were used to estimate the forces necessary to manage the lateral instability of the levitated materials container for varying load weights and lateral gaps (container to siderail).

An electronic position control system was developed specifically for the prototype Maglev Transport System. Electronic control systems can be constructed using analog or digital components, or a combination of the two. The analog control system format was deemed most appropriate for the development of a prototype electronic control system largely because of the difficulties associated with synchronizing delay times in digital system design.

In developing the electronic position control system, a computer model of the prototype transport system and proposed position control system was constructed. Transport system physical parameters, control system implementation methods, and component data were incorporated in the computer model. The computer model was used for analysis and simulation studies to evaluate the performance of the proposed position control system on the prototype transport system. The results of the analysis and simulation studies were used to assist in the selection of control structures and parameter values to provide stable operation.

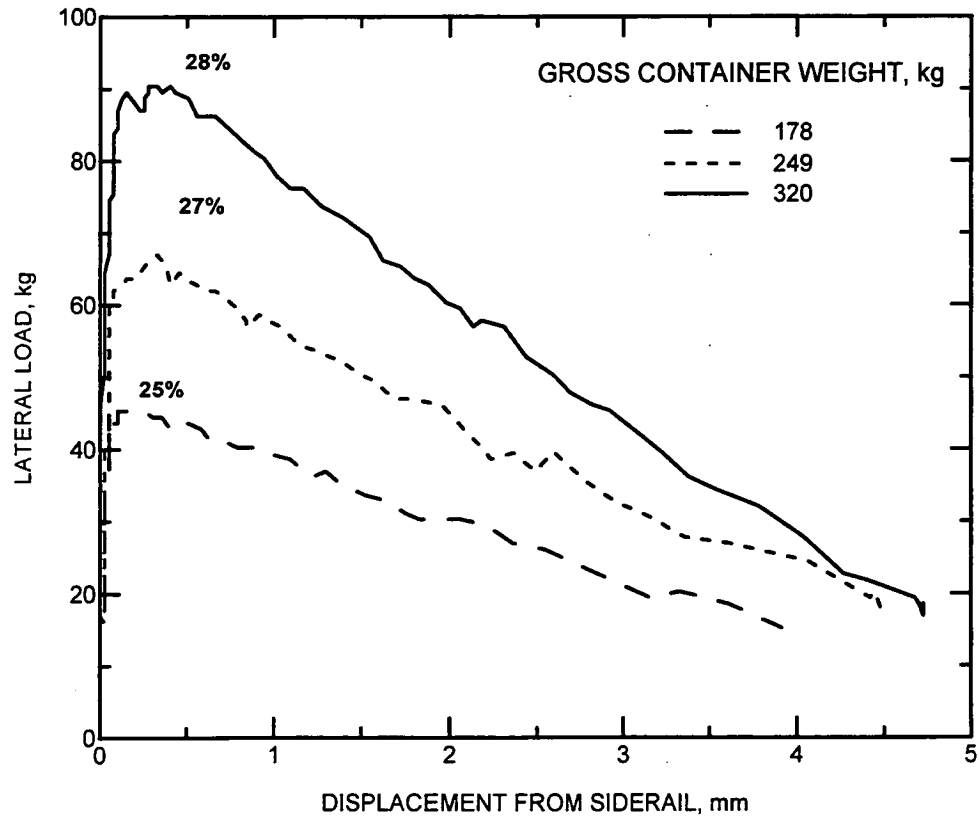


Figure 6. Lateral load versus materials container lateral position error, for three different container payloads.

Technical Description

The electronic position control system uses four analog, inductive proximity sensors, with two installed on both sides of the container that face the steel corridor siderails (see figure 4). Each sensor provides a linear 0-20 mA output signal directly proportional to the distance between the sensor face and a metal target, the siderail. The proximity sensors are interfaced to the lateral control circuit such that the outputs of the two front sensors are continuously compared to each other. The two rear sensors are interfaced to the lateral control circuit in the same manner. When the levitated materials container is centered between the corridor siderails, all four proximity sensors produce the same level of output current. When the levitated container is off-center, sensor output current levels differ. Comparing proximity sensor output current levels provides a position error signal that directly represents the magnitude and direction of container position error.

The position error data are used to energize electromagnets on the levitated materials container in proportion to the difference between container position and transit corridor centerline. The greater the position error, the greater the level at which the appropriate electromagnets are energized. The strength of the attraction between the electromagnets and the corridor steel siderails varies in accordance with the level at which the electromagnets are energized. By continuously energizing the appropriate electromagnets in response to materials container position error, the levitated container is kept centered between the corridor siderails.

The electronic position control system uses four electromagnets, one located next to each proximity sensor. The electromagnets are 12V DC bi-polar models, each rated at 499 kg holding power.

The proximity sensors and electromagnets are the inputs and outputs, respectively, of the lateral control circuit. The lateral control circuit conditions the sensor outputs, performs first-order, lead-lag dynamic phase compensation, and distributes DC power to the electromagnets. The DC power is supplied by a regulated DC power supply connected to the lateral control circuit.

System Performance

The electronic position control system was installed on the magnetically suspended materials container to evaluate its performance. The position control system demonstrated centered, noncontact positioning of an empty, levitated container traversing a 2.4 m test section of the transit corridor (at approximately 0.25 m/s). The electronic position control system maintained a uniform gap of approximately 3 mm between the corridor siderails and the materials container while traversing the corridor, thus providing totally noncontact, frictionless movement of the levitated container.

The electronic position control system performance is currently being evaluated with load added to the levitated materials container. Noncontact levitation of the materials container has been demonstrated while traversing 2.4 m of the transit corridor with a payload of approximately 153 kg in the container (211 kg gross container weight).

RECOMMENDATIONS

Further research is necessary to complete the development and testing of the Maglev Transport System. The remaining research can be divided into two main categories: operation and optimization.

The remaining operational research must be performed to allow the completion of the prototype construction and system testing. Research included in this category is: lateral control of a fully loaded (320 kg gross wt.) materials container, analysis of the dynamic behavior of multiple, mechanically linked, loaded materials container trains traveling in the pipeline at the intended operating speed, onboard power generation system studies, and propulsion system studies.

Much of the Maglev Transport System design appears to allow for optimization which would reduce the complexity, cost , and quantity of materials in the original System design. Optimization research should include: analysis of an electronic position control system operating off only one siderail instead of both, transposition of the permanent magnet array shapes, lamination of the steel siderails, and use of a hybrid, pneumatic/linear induction propulsion system. Computer simulation studies have indicated that the positioning of the materials containers can be accomplished by sensing only one siderail. Analysis of the System performance in an environment containing curves and change of grade has indicated that the performance can be improved if the chevron shaped permanent magnet array is located in the materials container and the orthogonal array in the transit corridor. The efficiency of the positioning and propulsion subsystems may be improved by using laminated steel siderails. As the materials containers travel through the pipeline, eddy currents induced in the siderails lead to a loss of attractive force and the development of a drag force [4]. By laminating the siderails, the induced eddy currents should be minimized. Finally, the propulsion system efficiency may be improved by utilizing a linear induction motor, for acceleration of the loaded materials containers, linked to a pneumatic pipeline to maintain the velocity of the containers. Recent advancements in the area of linear induction motors and the large air volume requirements for acceleration by pneumatic pressure warrant further investigation of such a hybrid propulsion system.

SUMMARY

The USBM has developed magnetic levitation technology utilizng permanent magnet suspension coupled with an electronic position control system. Research indicates that this enabling technology may be suitable for development of a maglev pipeline transport system. This innovative materials transport system design appears as a promising means to improve the safety and reduce the cost of underground mining and materials handling. Underground coal transport via a maglev pipeline transport system offers many safety benefits over conveyor belt transport, including (1) a reduction in fire hazards due to the elimination of mechanical friction in rotating idler bearings and use of a nonflammable system enclosure (concrete pipeline), and (2) improved worker safety through reduced exposure to respirable dust and moving parts (rotating belt and belt-idler assemblies).

USBM research into a permanent magnet maglev haulage system has developed technology to provide totally noncontact, frictionless transport of haulage containers. Cost estimates, incorporating current design data, indicate that the operating and capital costs of the Maglev Transport System will be comparable to those of an equivalent capacity conveyor belt system (the costing test case assumed a 4 km straight-line haul up a 5% grade). In addition, the minimized component wear provides for a substantially longer operating life than a belt system.

REFERENCES

1. Poland, A. P., and J. A. Lombardi. Coal Transport Using Permanent Magnet Levitation. Paper in Proceedings of the International Symposium on Mine Mechanization and Automation. Vol. II, 1991, pp. 11-13 - 11-24.

2. Poland, A. P., and J. A. Lombardi. Containerized Haulage Using Permanent Magnet Levitation. Paper in Proceedings of the ISDT Annual Technical Conference. 1993.

3. McCaig, M. Permanent Magnets in Theory and Practice. Halsted Press, 1977, p. 249.

4. Rodger, D., N. Allen, P. C. Coles, S. Street, P. J. Leonard, and J. F. Eastham. Finite Element Calculation of Forces on a DC Magnet Moving Over an Iron Rail. To be published in IEEE Trans. Mag., Nov., 1994.

NONLINEAR OPTICAL POLYMERS FOR PHOTONIC APPLICATIONS

Geoffrey A. Lindsay
U. S. Navy, NAWCWPNS
Research & Technology Division 474220D
One Administration Circle
China Lake, CA 93555-6001

ABSTRACT

Photonics is data processing using light waves (photons) as the carrier signal, in contrast to *electronics*, in which electrons are the carrier signal. The trend in data transmission and processing is away from purely electronic systems and towards photonic and hybrid electronic-photonic systems. High speed, low power, and massively parallel processing possibilities are driving the growth in photonics. The materials described here are primarily targeted for electro-optic applications in which light is processed in channel waveguides on integrated microcircuits (optical chips); and, they are also being considered for frequency doubling applications (blue light). These materials, called second-order nonlinear optical polymers (NLOP), have been demonstrated in devices for very high speed optical switching and modulating (> 40 GHz). Examples of promising applications are optical interconnects for computer boards and chips, integrated optical circuits for photonic control in phased-array radar and fiber-optic gyroscopes, and beam-steering devices for reading optical memory.

INTRODUCTION

Switching an optical signal from one track to another in a waveguide, or modulating the phase or amplitude of an optical signal, requires an electro-optic material with macroscopic polar order that undergoes a change in index of refraction when an electric field is applied across the waveguide. For frequency doubling, the same material can be used, but an applied electric field is not required. Dye molecules, called chromophores, are the active components of organic nonlinear optical materials. The molecular origin of optical nonlinearity can be pictured as the electrical polarization of the chromophore as it interacts with coherent electromagnetic radiation and with other applied electric fields [reference 1]. For electro-optic switching and frequency doubling applications, we are interested in second-order nonlinearity which requires that the material not have a center of symmetry. In other words, for second-order behavior, the chromophore must be asymmetric, and the chromophores must be aligned in the same direction (i.e., have polar alignment). Thus, it is the chromophore and its orientation that dictate the macroscopic nonlinear optical properties of the material.

The mathematical framework and description of the macroscopic nonlinear optical properties of NLOP films are well described in books by Prasad and Williams in 1991 [reference 2], and Chemla and Zyss in 1987 [reference 3]. The linear and nonlinear optical behavior of a macroscopic material can be described by a Taylor's series expansion of the dipole moment per unit volume, P, as a function of the applied electric field(s), E:

$$P = e_0\chi^{(1)}E + e_0\chi^{(2)}EE + e_0\chi^{(3)}EEE + \ldots \quad (1)$$

where e_0 is the vacuum permittivity, and $\chi^{(\xi)}$ are the susceptibilities. The first term describes ordinary linear behavior (reflection and absorption). The second term (the first nonlinear term) describes the interaction of two

electric fields (or the square of one field), be they from a laser, r.f., or d.c. fields. Interactions described by this second term are sometimes called "three-wave mixing," i.e., $(\omega_3;\omega_1,\omega_2)$ where ω_1 and ω_2 are the applied frequencies (fields), and ω_3 is the resulting optical field. The third term in equation 1 describes four-wave mixing, and is important here only in electric-field induced second-order harmonic (EFISH) effects.

In general, $\chi^{(2)}$, the second-order susceptibility (a third-rank tensor), is related to a chromophore's molecular nonlinear optical coefficient, β, as follows [reference 4a]:

$$\chi^{(2)} = N\beta f^{\omega} f^{2\omega} \langle \cos^3(\theta) \rangle \qquad (2)$$

where N is the number of chromophores per unit volume, f^{ω} and $f^{2\omega}$ are the "local field factors" approximated by Lorentz-Lorentz expressions ($f = (n^2 + 2)/2$, where n is the refractive index), and $\langle \cos^3(\theta) \rangle$ is the average orientation of the transition dipole moments of the chromophores with respect to the direction normal to the plane of the polymer film. Chromophores usually have their transition dipole moment nearly parallel to the long axis of their molecular structure.

The molecular nonlinear optical coefficient, β, is also called the molecular second-order susceptibility, or first molecular hyperpolarizability. A very successful model (the two-state model) for predicting the magnitude of a chromophore's β can be derived by considering the interaction of the ground state of the chomophore (the highest occupied orbital, S_0), and the first excited state (lowest unoccupied molecular orbital, S_1) [references 4]. The most important factors are that the chromophore have a low energy of transition between ground and first excited state, a large change in dipole moment in going between the ground state to the first excited state, and a large transition dipole moment (a large oscillator strength).

Competing Materials are single crystals of inorganic compounds, such as lithium niobate ($LiNbO_3$) and GaAs, with inherent polar order. Although $LiNbO_3$ has been investigated for use in hybrid photonic-electronic devices for over 20 years, it has a very high dielectric constant (ten times higher than that of polymers), hence, power loss becomes significant when the speed of modulation is increased above 10 GHz. Fabrication of components is also difficult and expensive because components must be cut from $LiNbO_3$ single crystals. The III-V materials have low electro-optic nonlinear coefficients, hence, large voltages are required to effect the switching and modulation. These materials are also difficult to integrate with silicon microelectronic devices, where as polymer resists and planarizing layers are used routinely in silicon chips.

POLYMER TOPOLOGICAL DESIGN AND SYNTHESIS

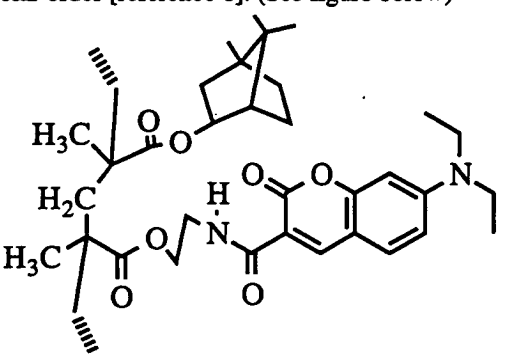
The polymer structure dictates the processability and temporal stability of the final product. Polymers of interest are amorphous glasses, because they are very transparent and scatter very little light. Chromophores in the polymers can be aligned by electric field poling near or above the glass transition temperature, then cooled with the field on [references 5 and 6].

Guest-host (G-H) Systems. These are solid solutions of small chromophoric molecules and high molecular weight polymers. Generally a G-H system contains about 10 to 20% by weight of the chromophore (higher levels tend to phase separate causing light scattering, and the glass transition temperature of the polymer decreases due to plasticization). G-H systems are generally undesirable because the chromophores are labile: they evaporate or sublime at elevated temperatures, they diffuse to the surface, and the small chromophores, which are easily absorbed

through the skin, can be toxic, mutagenic, teratogenic, and carcinogenic. High molecular weight polymers cannot be absorbed through the skin. Therefore, attaching the chomophore to the polymer, makes it much safer.

Sidechain (SC) Polymers. Our group and others have made significant advances in this area by attaching the chromophore to a polymerizable group, such as an acrylate, and polymerizing the resulting substituted monomer, creating a sidechain NLOP. Most of the early reported SC polymers were derived from the free radical copolymerization of functional chromophores and methyl methacrylate [references 7 -14]. The anilino-azo-nitrobenzene type of chromophores, such as the commercially available Disperse Red 1 and Disperse Red 19, have been attached as sidechains through their hydroxy ethyl groups and resulted in quite successful materials [references 9 - 12]. The first generation SC methacrylic polymers are easy to process, have excellent clarity, and have some of the best NLO figures of merit yet reported, for example, see reports from Hoechst Celanese [reference 13] and Akzo-Nobel [reference 14] on polymers containing the 4-dimethylamino-4'-nitrostilbene (DANS) chromophore. The glass transition temperature of these methacrylic polymers is generally about 140° to 170°C. Several SC acrylate NLO polymers are for sale to the public [reference 16].

Our group has developed a series of coumarin-based NLOP which have hydrogen bonding groups that suppress sub-Tg relaxation of the polar order [reference 8]. (See figure below)

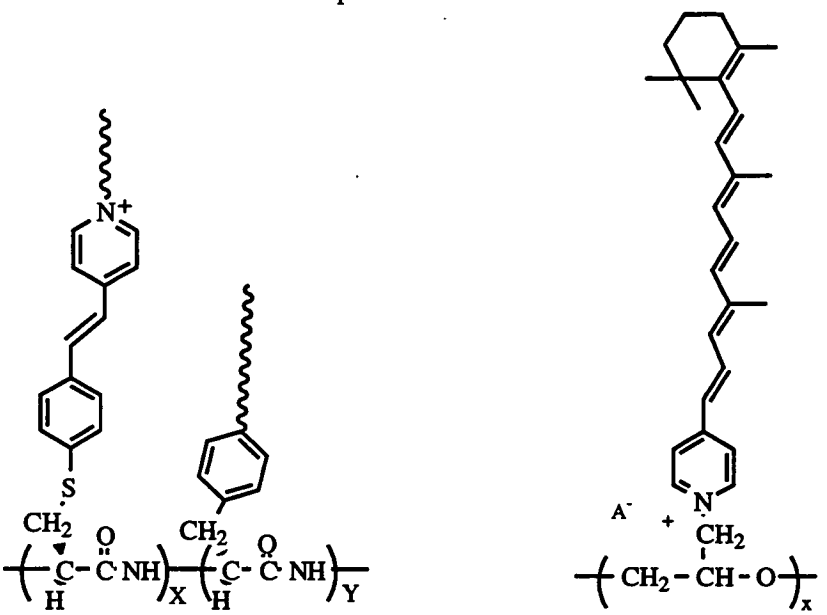


These NLOP are described in U.S. Patent 5,286,803 which issued in March, 1994, and is titled "FUNCTIONAL AND POLYMERIZABLE COUMARIN DYES FOR COUMARIN DYE-SUBSTITUTED POLYMERS WHICH EXHIBIT NONLINEAR OPTICAL PROPERTIES." The polymers are light yellow, have a very sharp absorption peak centered near 418 nm, and are very transparent above 500 nm. Their electro-optic coefficient, r_{33}, ranges from 12 pm/V at 477 nm to about 3 pm/V at 780 nm [reference 17]. They have outstanding resistance to laser damage (damage threshold > 120 GW/cm2), and should be considered for frequency doubling applications. Using the isobutylmethacrylate (IBMA) comonomer in place of MMA, increased the glass transition temperature of the copolymer to 170°C. The increased glass transition temperature of the IBMA copolymer translated into greater long term temporal stability. For the IBMA-coumarin NLOP, the second-order nonlinear optical coefficient, d_{33}, estimated from frequency doubling experiments using a Nd:YAG at 1.06 mm, was 11 pm/V. The index of refraction of these polymers ranges from 1.56 at a wavelength of 1.06 microns, to 1.59 at 0.633 microns.

Our group has also developed red-colored azomethine sidechain NLOP's [reference 9], a tricyanovinyl sidechain NLOP, and a host of other chromophores with larger nonlinear optical coefficients amenable for sidechain polymers (reference 18).

Our group has developed sidechain NLOP designed for Langmuir-Blodgett deposition which are described in U.S. Patent 5,162,453 entitled "DYE SUBSTITUTED POLYMERS CONTAINING HYDROPHOBICALLY TERMINATED STILBAZOLIUM RADICALS" and issued on Nov. 10, 1992. Most of these compositions are

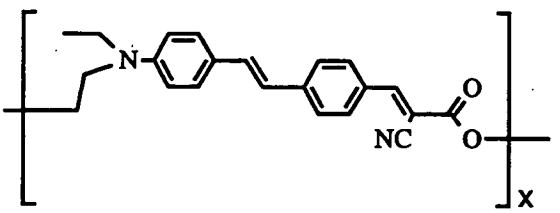
based on the stilbazolium chromophore which contains an ionic charge, although one of the sidechains was derived from t-retinal. This patent describes a sidechain NLOP having backbones derived from oxiranes, oxetanes, aziridines, azetidines, and amino acids. Several of these compositions are shown below:



The chiral polypeptide backbone may offer a template for improving the polar order in sidechain NLOP if a β-sheet conformation can be achieved (reference 19). The t-retinyl polymer has second-order NLO properties as well as being active at radio frequencies.

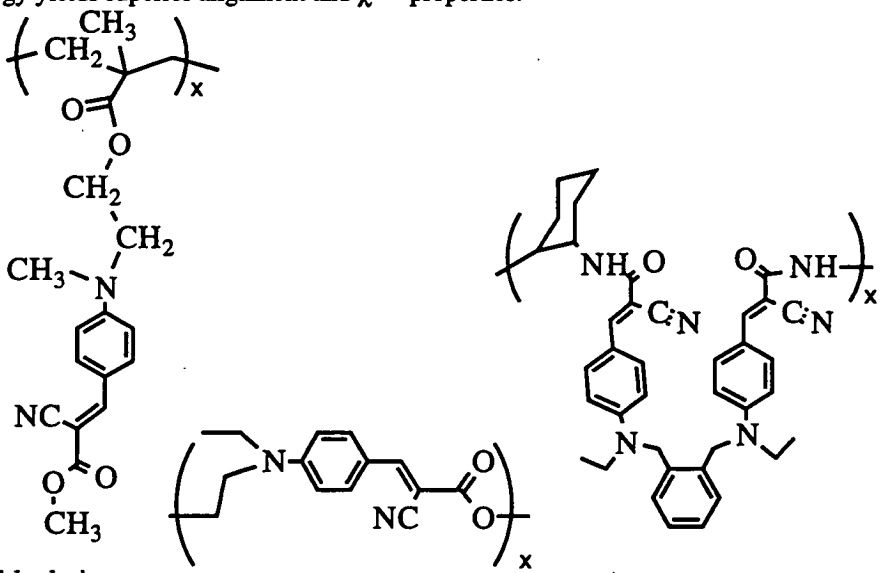
Mainchain (MC) Polymers. These polymers are comprised of chromophores which are linked in the backbone at both ends of the chromophore, and the majority of the chromophore forms part of the backbone. The rest of the backbone of mainchain polymers are connecting groups, bridging one chromophore to the next. The chromophores can be linked in a head-to-tail configuration (isoregic) [reference 20], in a head-to-head configuration (syndioregic) [reference 21], or randomly head-to-tail and head-to-head configurations (aregic) [reference 22]. The mainchain chromophoric topology is inherently more stable than the sidechain topology because the chromophore has one less degree of freedom of motion.

Isoregic NLOP. Our group has made several notable milestones using the isoregic architecture, including development of the first high temperature film [reference 23], which shows essentially no activity loss when ramping the temperature to within 10°C of the glass transition temperature [reference 24]. These NLOP have achieved the highest chromophore density, essentially 100%. The patent covering these compositions, "MAIN CHAIN CHROMOPHORIC POLYMERS WITH SECOND ORDER NONLINEAR OPTICAL PROPERTIES" is U.S. Patent 5,212,269 issuing on May 18, 1993. One of the compositions is shown below:



Syndioregic NLOP. Continuing research on this series of polymers, our group invented another unique NLOP structure. This structure evolved from questioning the need for the isoregic topology: is it really necessary to have the dipoles pointing in the same direction along the polymer backbone, or could the polymer be designed to fold in a way that would give polar order without requiring complete end-to-end alignment of the bulk polymer chain? The folded accordion (syndioregic) structure was proposed as a way to allow chains to orient more easily in an electric field or on a Langmuir-Blodgett (LB) trough. The accordion structure also aided solubility of the polymer for ease of film formation and clarity of the film. The first NLO accordion polymers were prepared using transesterification with flexible diols and had rather low Tg's (e.g., 50° to 100°C). In the course of the research it was found that 1,2-diaminocyclohexane, which is commercially available from DuPont, used for the bridging group was amazingly potent in increasing Tg and solubility [references 25a and 26]. U.S. Patent 5,247,055 describing these NLOP, entitled "ACCORDION-LIKE POLYMERS FOR NONLINEAR APPLICATIONS" was issued on Sept. 21, 1993.

A comparison of 3 polymer topologies, all containing essentially the same chromophore, argues that the syndioregic topology yields superior alignment and $\chi^{(2)}$ properties:



Sidechain Homopolymer
T_g = 114°C
d_{33} = 22 pm/V
[reference 15b]

Isoregic Mainchain
T_g = 105°C
d_{33} = 7 pm/V
[reference 23]

Syndioregic Mainchain
T_g = 205°C
d_{33} = 30 pm/V
[reference 25b]

If one modifies the syndioregic polymer (shown above on the right) by substituting the ethyl group on the amine donor with a methyl group, the Tg is increased to 212°C, and it is estimated that the Tg will reach 220°C at higher molecular weights (see table below).

SAMPLE NUMBER	M_n (Daltons)	T_g (°C) (±2°C)
ETHYL-SUBSTITUTED:		
trans-1400-16	7,300	193
trans-1400-58	15,600	201
trans-1400-19	19,800	206
trans-1400-15	30,000	208

METHYL-SUBSTITUTED:		
trans-1400-23	8,800	204
trans-1400-91	8,200	206
trans-1460-29	12,000	206
trans-1460-7	9,900	207
trans-1400-26	13,400	212

Soon to be published measurements on this high temperature accordion polymer show it to have an electro-optic coefficient between 8 and 10 pm/V using 1.3 micron light in a sandwiched waveguide configuration [ref. 27].

Crosslinked Polymers. Crosslinked (XL) polymers offer the promise of greater temporal stability (at least below the Tg). XL polymers by their very nature must undergo a chemical reaction (the crosslinking reaction) during or after the alignment process. Therefore, this class of waveguide material is more difficult to manufacture, and for this reason, could lead to greater variation in the quality of the waveguide. Early reports indicated greater optical losses due to density fluctuations (inhomogeneity in the crosslink density). However, more recent reports indicate that losses can approach those obtained with linear polymers (thermoplastics). The inventions described in the U. S. Navy patents reported here include crosslinkabe versions of the NLOP.

NLOP FILM FABRICATION PROCESSES

Electric-Field Poling imparts noncentrosymmetry to the film which is necessary for its second-order nonlinear optical properties. Thin films of the polymer are prepared for poling by spin-coating a liquid solution containing about 10% of the polymer onto a solid substrate. The solvent is removed by baking the film just above the glass transition temperature (Tg). An electric field is applied across the film by corona poling the film as it sits on a grounded conductor plane, or by charging two electrodes contacting the film (there may also be thin cladding or buffer layers between the electrodes and the nonlinear optical polymer). The films are typically poled at a field strength of about 50 to several hundred volts per micron. The electric field is applied at or above Tg for 10 to 30 minutes, then the film is cooled with the field on. After the external field is removed, a net alignment of dipole moments can remain essentially locked in the film for years as long as the temperature of the film remains well below the Tg.

Langmuir-Blodgett (LB) Processing. In LB processing, the polymer molecules are designed to have hydrophilic and hydrophobic groups which cause the polymer to float in a preferred conformation on the water. These hydrophilic/hydrophobic forces are useful in orienting the chromophores normal to the plane of the film, removing the centrosymmetry. To make films by LB processing, an organic compound is floated on water in a computer-controlled trough; and, a solid substrate is dipped through the air-water interface, depositing a single molecular layer on the substrate. Thicker films comprised of multilayers of polymer are built up by dipping the substrate repeatedly into and/or out of the LB trough, depositing one molecular layer per stroke [reference 28]. Turn-key multi-compartment troughs are available from several commercial suppliers, for example, NIMA (Coventry CV4 7EZ, England), NLE (Nagoya, 468, Japan), and KSV (SF-00380 Helsinki, Findland).

Multilayer LB films can be formed in three different configurations. Historically these are called "X"-, "Y"- and "Z"-type films, where X is made by depositing always on the down-stroke, Z is made by depositing always on the up-stroke, and Y is made by alternating up- and down-strokes. The Y configuration is thermodynamically more stable (e.g., sometimes X and Z configurations spontaneously rearrange in the solid state to the Y configuration)

because the surface energies are minimized by having hydrophobic-hydrophobic and hydrophilic-hydrophilic contacts. Hence, one may interleave a chromophoric polymer with an optically inert polymer layer to form a stable, polarized film in the Y configuration. Alternatively, for higher chromophore density in the Y configuration, one may use a two-compartment LB trough, building up -(AB)n- bilayers of two different chromophoric polymers. In doing so, in order to have their respective dipole moments point in the same direction through the thickness of the film, their chromophores must point in the opposite sense to each other with respect to the hydrophilic and hydrophobic parts of the polymers to which they are attached [reference 29].

LB processing has the advantage over electric-field poling in that it can be done at room temperature (or lower), hence the kT Brownian motion is much less. Furthermore, formal ionic charges on the polymer will not hinder the ordering process, in fact they can be taken advantage of in designing hydrophilicity into the polymer. This film processing procedure has been reduced to practice for sidechain polymers [references 29 and 30] and syndioregic mainchain chromophoric polymers [reference 31].

Cladding, Channels and Ribs. In waveguide switches and modulators, the light generally travels along the path of highest index of refraction. The path must go between metal electrodes in the active switching or modulating section. For most applications, a low index of refraction cladding or buffer material must separate the optical waveguide from the metal electrodes in order to reduce optical loss (the evanescent wave at the surface of the polymer waveguide is absorbed where it touches the metal electrode). The cladding layer must be very transparent so as not to cause absorption or scattering losses from the evanescent wave. If electric field poling is used, the cladding layers should be slightly higher in conductivity insuring that the maximum electric field is applied over the nonlinear optical polymer layer. This can usually be accomplished by selecting a cladding material that has a slightly lower Tg than the NLO polymer.

The optical paths on an integrated device can be created before or after aligning the chromophores in a nonlinear optical polymer film. NLOP can be used for passive optical waveguides (and also for cladding material) simply by bleaching the chromophores with ultraviolet light which lowers the index of refraction [reference 32], or using the polymer in its randomly ordered state. A channel (or a 2D-confined path) for the light is easily patterned in a NLOP film by photobleaching the polymer while masking the pattern of a channel. Other types of 2D-confined waveguides are ribs made by etching the NLOP layer [reference 33], or are grooves made by etching the lower cladding and back filling it with the NLOP . Etching is done by conventional patterning techniques used in microlithography, such as reactive ion etching, UV crosslinking and liquid developing, and UV laser ablation. Many of the polymer waveguide processing techniques have been fine tuned for passive polymer waveguides [reference 34].

PHYSICAL AND CHEMICAL STABILITY

Amorphous polymer glasses containing polarity-aligned chromophores are in a nonequilibrium state. Brownian motion (thermal energy) works to randomize the order, although this occurs very, very slowly (over many years) in a properly designed glass. The rate of randomization depends mainly on the use-temperature relative to the glass transition temperature (Tg), and on local free volume, which in turn is proportional to the rate of cooling below the Tg, and the time-temperature history of the sample [reference 35]. There are two types of stability. Physical stability refers to stability of alignment of the chromophore (physical aging [reference 36]). Chemical stability refers to the integrity of the original chemical structure of the chromophore. Chemical changes can be brought about by heat, light, reactions with oxygen, or any combination of the above. Recent reports have focused attention on the role of singlet oxygen on unsaturated chromophores, such as DANS [reference 37]. Chemical

changes can range from isomerization, cyclization, oxidation, bond cleavage, free radical additions, all of which change the refractive index and nonlinear optical properties. These reactions can be avoided by proper selection of chromophore, processing conditions, and packaging for the device.

The temporal stability of a NLOP film refers to how long the physical alignment and chemical integrity of the chromophores are maintained at a given temperature. One is concerned with the temporal stability under end-use and storage conditions (typically 80 to 120°C). The processing stability refers to how well the polymer holds up under film processing and packaging conditions. These conditions are often reach high temperatures (typically 250 to 350°C) for tens of minutes during which the polymer films are baked but the chromophores are not yet aligned. NLOP samples and films can be screened by differential scanning calorimetry and monitoring the optical properties as a function of time and temperature, but the NLOP must be tested in an actual waveguide device before one can state with confidence its degree of stability.

DEVICES AND APPLICATIONS

The first commercial applications for NLOP will clearly take advantage of their low dielectric constant at high speeds (>20 GHz), the ease of integration of these films on low cost silicon microelectronics devices, and the ease of aligning optical fibers with NLOP channel waveguides using Si V-groove technology. The state of the art NLOP optical modulator, a traveling wave modulator with a 3-dB electrical bandwidth of more than 40 GHz using a $V\pi$ of 6 Volts was reported by Teng in 1992 [reference 38]. There are several groups close to reproducing this work, and also attempting to make 100 GHz amplitude modulators.

Frequency doubling or second harmonic generation (SHG) is another application under intense investigation. Frequency doubling through the thickness of an NLOP film has achieved the shortest pulse of UV ($\approx$ 10 fs @ 310 nm) yet reported [reference 39]. Although not optimized, higher efficiencies have been achieved with optical waveguides in which the phase matching is accomplished by tailoring the thickness of several layers to achieve a high overlap integral of the fundamental and second-harmonic modal field [references 40, 41 and 42].

ACKNOWLEDGEMENTS

The support of our sponsors, the Advanced Research Projects Agency and the Office of Naval Research, is gratefully acknowledged. These results were made possible by the invaluable contributions of collaborators at China Lake, CA: Andrew P. Chafin, John W. Fischer, Robert C. Hall, Richard A. Hollins, Ronald A. Henry (dec.), James M. Hoover, Melvin P. Nadler, Robin A. Nissan, Larry H. Merwin, John D. Stenger-Smith, Rena Y. Yee; and at NAWCAD (U.S. Navy), Warminster, PA: Warren N. Herman; and at the Univ. of Maryland, Baltimore Campus: L. Michael Hayden; and at MICOM (U.S. Army), Huntsville, AL: Paul Ashley.

REFERENCES

1. Garito, A.; Wong, K. Y.; Zamani-Khamiri, O.; in *Nonlinear Optical and Electroactive Polymers*, P. N. Prasad and D. R. Ulrich, Eds., Plennum Press, NY, 1988.
2. Prasad, P.N.; Williams, D.J. *Introduction to Nonlinear Optical Effects in Molecules and Polymers* ; John Wiley & Sons: New York, 1991.

3. Chemla, D.S.; Zyss, J., Eds.; Academic: New York, 1987; Vol. 1 and 2.
4. (a) Singer, K.D.; Kuzyk, M.G.; Sohn, J.E. in *Nonlinear Optical and Electroactive Polymers*, P. N. Prasad and D. R. Ulrich, Eds., Plennum Press, NY, 1988; and (b) Oudar, J. L.; Chemla, D. S. J. Chem. Phys. **1977**, *66*, 2664.
5. Singer, K.D.; Kuzyk, M.G.; Holland, W.R.; Sohn, J.E.; Lalama, S.L.; Comizzoli, R.B.; Katz, H.E.; Schilling, M.L. *Appl. Phys. Lett.* **1988**, *53*, 1800.
6. Mortazavi, M.A.; Knoesen, A.; Kowel, S.T.; Higgins, B.G.; Dienes, A. *J. Opt. Soc. Am. B* **1989**, *6(4)*, 733.
7. (a) Leslie, T. M.; et al. *Mol.Cryst. Liq. Cryst.* **1987**, *153*, 451; (b) Esselin, S.; et al. *Proc. SPIE* **1988**, *971*, 120.
8. Hayden, L. M.; et al. *J. Appl. Phys.***1990**, *68(2)*, 456-465.
9. (a) Amano, M.; Kaino, T. *J. Appl. Phys.* **1990**, *68(12)*, 6024.
10. Ye, C.; Marks, T.; Yang, J.; Wong, G. K. *Macromolecules* **1987**, *20*, 2324.
11. Singer, K. D.; Sohn, J. E.; King, L. A.; Gordon, H. M.; Katz, H. E.; Dirk, C. W. *J. Opt. Soc. Am. B* **1989**, *6(7)*, 1339.
12. Feringa, B. L.; deLange, B.; Jager, W. F.; Schudde, E. P. *J. Chem. Soc., Commun.* **1990**, 804.
13. DeMartino, R. N.; et al. in *Nonlinear Optical and Electroactive Polymers*, P. N. Prasad and D. R. Ulrich, Eds., Plennum Press: New York, 1988.
14. Mohlmann, G. R.; et al. *Proceedings of SPIE* **1990**, *1337*, 215.
15. (a) S'heeren, G.; Persoons, A.; Rondou, P.; Wiersma, J.; van Beylen, M.; Samyn, C. *Makromol. Chem.* **1993**, *194*, 1733; (b) Rondou, P.; Van Beylen, M.; Samyn, C.; S'heeren, G. Persoons, A. *Makromol. Chem.* **1992**, *193*, 3045.
16. (a) AdTech Systems Research, Inc, contact Dr. Sam Sinha, 1342 N. Fairfield Rd., Beavercreek, OH, 45432 (phone 513-426-3329); and (b) IBM Almaden Research Center, contact Dr. Dan Dawson, 650 Harry Road, San Jose, CA 95120-1501 (408-927-1617).
17. Mortazavi, M.A.; et al. *Applied Physics* B, **1991**, *53*, 287.
18. Lindsay, G. A.; et al. *Proc. SPIE* **1994**, *2143*, 19.
19. Lindsay, G.A.; Schram, A. C.; Hoover, J.M.; Nissan, R.A.; Stenger-Smith, J. *Proc. of SPIE* **1993**, *1853*, 77.
20. Fuso, F.; Padia, A. B.; Hall, H. K., Jr. *Macromolecules* **1991**, *24*, 1710.
21. Lindsay, G. A.; Henry, R. A.; Hoover, J. M.; Knoesen, A.; Mortazavi, M. A. *Macromolecules* **1992**, *25*, 4888-4894.
22. Xu, C.; Wu, B; Dalton, L. R.; Ranon, P. M.; Shi, Y.; Steier, W.H., Becker, M. W. *Chemical Materials* **1993**, *5*, 1439-1444.
23. Stenger-Smith, J. D.; Fischer, J. W.; Henry, R. A.; Hoover, J. M.; Lindsay, G. A.; Hayden, L. M. *Makromol. Chem., Rapid Commun.* **1990**, *11*, 141.
24. Barry, S.; Soane, D. *Mat. Res. Soc. Symp. Proc.* **1992** *277*, 223.
25. (a) Stenger-Smith, J. D.; Henry, R. A.; Hoover, J. M.; Lindsay, G. A.; Nadler, M. P.; Nissan, R. A. *J. Poly. Sci.: Part A: Polym. Chem.***1993**, *31*, 2899; and (b) private communication from W. N. Herman, U.S. Navy, NAWCAD, Warminster, PA.
26. Lindsay, G. A.; et al. in *Organic Thin FIlms for Photonic Applications* Technical Digest 1993, Volume 17 (Optical Society of America, Washington, D.C. 1993), pp. 14 - 17.
27. Private communication from Dr. Paul Ashley, U. S. Army, Redstone Arsenal, Huntsville, AL.
28. Gaines, G.L. *Insoluble Monolayers at Liquid-Gas Interfaces*, Interscience: New York, 1966.
29. Hall, R. C.; Lindsay, G. A.; Anderson, B.; Kowel, S. T.; Higgins, B.G.; Stroeve, P. *Mat. Res. Soc. Symp. Proc.***1988**, *109*, 351-356.
30. Motschmann, H. R.; Penner, T. L.; Armstron, N. J.; Ezenyilimba *J. Phys. Chem.* **1993**, *97 (216)*, 3933.

31. Hoover, J. M.; R. A. Henry, G. A. Lindsay, S. F. Nee, J. D. Stenger-Smith, in *Organic Materials for Nonlinear Optics III*, G. J. Ashwell and D. Bloor, Eds.; the Royal Society of Chemistry, 1993; Special Publication 137.
32. Zyung,T.; Hwang, W.-Y.; Kim, J.-J. *Appl. Phys. Lett.* 1994, *64 (26)*, 3527.
33. Wang, W.; et al. *Appl. Phys. Lett* **1994**, *65 (8)*, 929.
34. Booth, B. L. in Polymers in Lightwave and Integrated Optics, Technology and Applications, L. A. Hornak, Ed.; Marcel Dekker, New York, 1992.
35. Lindsay, G. A. *Trends in Polymer Science* **1993**, *1(5)*, 138.
36. Struik, L.C.E. *Physical Aging in Amorphous Polymers and other Materials*, Elsevier: Amsterdam, 1978.
37. Mortazavi, M.; Yoon, H.; Teng, C. *J. Appl. Phys.* **1993**, *74*, 4871.
38. Teng, C. C. *Appl. Phys. Lett.* **1992**, *60*, 1538.
39. Yankelevich, D. R.; Knoesen, A.; Dienes, A.; Schoenlein, R. W.; Shank, C. V. *IEEE J. Quantum Elect.***1992**, *vol. QE-28*, 2398.
40. Florsheimer, M.; Küpfer, M.; Bosshard, H.; Looser, H.; Günter, P. *Adv. Mater.* **1992**, *4*, 795.
41. Otomo, A.; Mittler-Neher, S.; Bosshard, C.; Stegeman, G. I.; Horsthuis, W. H. G.; Mohlmann, G. R. *Appl. Phys. Lett.* **1993**, *63(25)*, 3405.
42. Clays, K.; Schildkraut, J. S.; Williams, D. J. *J. Opt Soc. Am. B* **1994**, *11(4)*, 655.

ADVANCED MATERIALS FOR STRENGTHENING AND REINFORCING CONCRETE

Piyush K. Dutta
U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, NH 03755-1290

Richard G. Lampo
US. Army Construction Engineering Research Laboratory
Champaign, IL 61826-9005

ABSTRACT

Concrete engineering history may have reached a new stage with the advent and availability of the new, low-cost, high performance structural composites. Strengthening of concrete beams with internally or externally bonded fiber -reinforced plastics (FRP) has been proved feasible to increase the load carrying capacity and stiffness of existing structures. Composites made with glass, aramid, or carbon fibers are being increasingly considered for prestressing concrete. To replace the corroding steel rebars, FRP reinforced system may eventually be used in concrete bridge decks, or other outdoor concrete flooring structures. However, at present, quantitative guidelines for composite reinforcement of concrete are not available. This paper reviews the current state of FRP composite reinforcements of concrete and focus on the R&D thrusts, the lack of which appears to be hindering the development of the quantitative guidelines.

INTRODUCTION

Availability of the new, low cost, high performance fiber reinforced plastic (FRP) structural composites for strengthening concrete columns, piles, piers, beams, decks, and other concrete structures has given an impetus to the development of a new technology in construction industry. The applications of the forms of FRP in concrete range from using high strength cables or tendons for prestressing or post-tensioning of concrete columns or beams to using the steel rebar-profiled reinforcing FRP rods, gratings of different profiles and sections, continuous-fiber isogrids for imbedded concrete reinforcement, external wrappings of columns, or simple bonding of FRP plates to the concrete structure. Recently the US Army Corps of Engineers (COE) through its Construction Productivity Advanced Research (CPAR) program has introduced a large number of demonstration projects for these concepts. The driving force for these developments in the concrete technology is the decaying infrastructure of the country. It has been estimated by the Federal Highway Administration (FHWA) that the USA currently faces a $167 billion cost to repair deficient bridges and roads. The deterioration has been caused by corrosion due to marine environments, high chloride content in the air, use of de-icing salts, alkali-solute reactions

(ASR), poor initial design, and poor construction and/or maintenance. There is also an increased awareness of seismic- induced damages which have happened as recently as last year (1994) in southern California causing extensive damages to highway bridges and other structures.

Inherent lightness, placement flexibility, corrosion resistance, high strength to weight ratio, low relaxation/creep, high fatigue resistance, and magnetic transparency of FRP composites make them attractive and logical substitutes for steel structural elements which have been historically used for strengthening and reinforcing concrete. For the purpose of this paper, we will focus on both the internal and external reinforcement technique of concrete structures and review the experiences from a number of case histories.

FRP REINFORCING BARS

In recent times, FRP reinforcing bars (Figure 1) are receiving increasing attention as the tension element in reinforced concrete [Roll, 1991]. As available in market, these rebars, as long rods, are made of very fine continuous glass fiber strands which are bound together with a thermosetting polymer. Wu et al [1990] have reported that E-glass reinforced composite rods from which these rebars are made may have tensile strength in excess of 100 ksi and longitudinal elastic modulus about 7,500 ksi. In tensile tests the bars fail without any significant yield (brittle failure). The rods are produced by a process known as "pultrusion" which involves pulling the continuous glass strands through a resin impregnating bath and a shape-forming heated die in which it cures. Since glass is commonly used as the reinforcing fibers in these rebars, these rebars are commonly designated as GFRP. Table -1 gives a comparison of mechanical properties of the steel rebars and the FRP rebars.

Table -1 Comparison of mechanical properties of steel and FRP rebars [Faza, 1995]

	Steel Rebar	FRP Rebar
Specific Gravity	7.9	1.5 - 2.0
Tensile Strength MPa (ksi)	483-690 (70-100)	517-1207 (75-175)
Yield Strength MPa (ksi)	276-414 (40-60)	---
Compressive Strength MPa (ksi)	276-414 (40-60)	310-482 (45-70)
Tensile Modulus, GPa (Msi)	200 (29)	41-55 (5.9-8.0)
Coeff. thermal expansion 10^{-6}/°C (°F)	11.7 (6.5)	9.9 (5.5)

Currently there are quite a few FRP rebar companies actively marketing their products in the USA. Most FRP rebars contain by volume about 55% E-glass fiber and about 45% thermoset resin. The sizes (diameter) of the rebars follow the size designations of the steel rebars (e.g. #3, #4, or #7 rebars). Faza (1995) has reported a number of successful applications of rebars in the USA including applications in sea walls, hospital

MRI, Reactor pad, compass calibration pad, mill roofs, laser test facilities, highway barriers, residential foundations, and bridge decks.

One of the most critical problems to be overcome in large-scale applications of the FRP rebars is developing improved bond strength with concrete. Some available designs provide a helically convex surface (Figure 1) made with a strand spirally wound and cured on the surface. Other designs suggest use of sand or grit coating on the rebars. A recent design includes a pultruded ribbed surface. A comparative survey of the bond quality of these surface modifications is still not available.

There are several other major barriers to FRP rebar applications. These include lack of sufficient data on durability or performance under extreme environments. Creep, fatigue, and corrosion from alkaline environment of concrete need to be investigated. Recent research by the author, on a market brand of FRP rebar subjected to a constant longitudinal load of 50 percent of ultimate load continuously over periods in excess of six months, at low (-20°C,-4°F), room temperature (20°C,68°F), and high (49°C, 120°F) temperatures, have shown practically no creep [Dutta, 1995; Kumar et al., 1995] have studied the fatigue performance of concrete slabs with FRP rebars and concluded that the loss of stiffness was significant. The impediments to the use of FRP rebars also include high material costs, lack of design guide lines, and lack of standard quality of products, or industry standardization. Recently, standardizing committees have been formed by the ASTM, ACI, and the ASCE to address the issues of developing materials standards, test methods, design specifications, and design guide lines. A rapid progress is expected from these standards-setting bodies.. Some researchers [Faza, 1995; Kumar et al., 1995] have recently presented equations for evaluating the FRP rebars and designing with it. Availability of these design equations and ACI endorsement will certainly facilitate the acceptance of FRP in concrete reinforcement.

Confidence in using FRP rebars in load critical structures will grow through demonstration projects. The COE CPAR demonstration project of the FRP reinforcement of concrete will include construction of a 33.5m (110 ft) Mckinleyville Bridge in the Brooke county , WV. The design and construction of this bridge will be performed by the WVDOT and the Constructed Facilities Center of the West Virginia University [GangaRao, 1995]. Other construction efforts are also well underway in the various parts of the country.

ALTERNATIVE METHODS OF FRP REINFORCEMENT

FRP Grating Reinforcement

Several studies have concerned alternatives to rebar type reinforcements to reinforcing grids of FRP. A novel approach has been discussed by Bank et al (1991) which includes commercially available molded and pultruded FRP gratings (Figure 2). These gratings are typically used for walkways and platforms in plants and factories. They have a number of advantages over the FRP rebars, the major being that the grating can be produced in variable longitudinal and transverse member spacings according to the reinforcement design requirement, and the overall dimensions of the grate can be configured at the plant or field to fit the concrete slab forms, thus requiring no time-

consuming "rebar" placement in the field. The lighter weight of the grate relative to steel rebar configuration may reduce the overall deadweight of the slab.

Bank and his coworkers' work included a number of commercial molded or pultruded FRP grates. They investigated the workability, of the concrete with the FRP gratings. Special attention was paid to avoidance of voids and ensuring that the concrete infiltrate the FRP grating. The general conclusion of this study is that although the FRP grating reinforced slabs can attain the strength of the steel reinforced slab, they can not attain the flexural stiffness of steel reinforced slabs. However, attaining the higher stiffness is not an insurmountable problem as the solutions could be found in redesigning the grating, geometrical shape.

FRP Isogrid reinforcement

Isogrids differ from the gratings in that these are manufactured in bi-or tri-directional orientation of continuous fibers cured with resin, and thus can provide higher strength and stiffness than the FRP gratings. In FRP gratings the transverse bars (cross-rods) generally pass through the holes in longitudinal bars, whereas in isogrids the continuous fibers weave through the junctions.(Figure 3). The concept of using FRP composite isogrids in civil engineering structures, especially as a bridge deck was proposed by Koury and Dutta in 1994. A COE-CPAR project has recently been initiated to demonstrate the concept. Load deflection characteristics of such concrete composite structures with grid reinforcement have also been studied by Larralde and Zerva (1991), and Goodspeed et al. (1991). Both studies have shown feasibility of such reinforcement, but no comparison with steel reinforcement is available.

PRETENSIONING AND POST-TENSIONING

The FRP tendon applications in pretensioned concrete beams have also received increased attention in recent past [Gerritse and Schurhoff 1986, Dolan 1990, Iyer and Sen 1991]. The interest was again primarily dictated by the corrosion problems of steel prestressed elements exposed to marine or deicing salt environment. Prestressing tendons made of steels to day possess a tensile strength of 1.87 GPa (270,000 psi) and Young's modulus approximately 200 GPa (29,000,000 psi). Any material replacing steel must meet its high strength, ductility, modulus, low cost, and ready availability. The competing fiberglass and graphite composite pultruded rods have the tensile strengths in excess of 2.0 GPa (300,000 psi) . The Young's modulus of the glass fiber FRP tendon is around 62 GPa (9,000,000 psi), for graphite fiber FRP 138 GPa (20,000,000 psi). The FRP tendons have almost no ductility and cost is considerably higher than steel. The major advantage of FRP tendons is their corrosion resistance and therefore longer life. However, when used as the prestressing element in pretensioned beams the tendons are in direct contact with the concrete. This raises a serious concern about the glass fiber FRP tendon's durability in the concrete's alkaline environment. Therefore a series of tests has recently been undertaken to study this problem (Poster, 1994). In a study by Sen et al., (1993), it has been observed

that when epoxy resin is used as the matrix of the S2 glass fiber prestressing tendon, severe degradation of glass has occurred when the specimens were exposed to dry/wet cycles.

Currently two experimental studies to investigate the feasibility of using aramid and carbon fiber prestressing tendons are being conducted by the Florida Department of Transportation [Shahawy, 1995]. Preliminary results from both carbon and the aramid FRP pretensioned beams showed bond degradation. Carbon fiber FRP pretensioned concrete pile and pile caps experiments have also recently been conducted by Iyer [1995], and Iyer and Khubchandani [1995] at Port Hueneme, California under a COE/CPAR demonstration program. Design details which can eventually provide design guidance for the FRP cable pretensioning methods have been developed in this process.

Post-tensioning of concrete structures involves a process for the transfer of prestress to the structure. A suitable tendon-anchorage assembly is the key to the system. For high-strength prestressing steel a variety of tendon-anchorage assembly are available but those are not suitable for the FRP tendons, because FRP elements do not tolerate transverse pressure. Considerable effort has recently been applied in developing suitable tendon-anchorage assemblies (Rostasy and Budelmann, 1991; Iyer et al., 1991; Porter and Barnes, 1991; Kakihara et al., 1991; Kim and Meier, 1991).

STRENGTHENING, REPAIRING, AND RETROFITTING

FRP Composite Column Wraps

A widely used technique for retrofit of concrete columns is the use of cylindrical steel jackets placed around the existing columns [Priestley and Seible, 1991]. To create a constraining effect epoxy or concrete is pumped into the gap between the existing column and the jacket. There are several disadvantages of this method. First, this method is particularly time consuming and difficult, and then the steel `jackets themselves corrode over time. However, the confining action is effective in increasing both compressive strength and ductility of concrete structure. The use of FRP composite wraps as confining layers for concrete columns has thus become an increasingly popular idea in recent years [Harmon and Slattery 1992, Priestly et al. 1992, Hasegawa et al. 1992, Katsumata et al. 1988, Karbhari et al. 1993, Ma 1993].

Various FRP materials for column wrappings have been tried in recent past. These include glass, aramid and carbon fiber composites. The efficiency of the process depends on the final compaction of the fabric on the column. Most commonly used methods of wrapping include a variation of hand lay-up and wet lay-up/winding. Automated lay-up/winding system has also been developed. To improve the compaction of the FRP on the concrete column Karbhari et al [1993] has recently developed the resin-infusion technique. In this process the resin is injected under a vacuum film using a resin distribution medium. Since the vacuum is maintained until the resin is cured, the compaction level is generally high.

Similar to the bonding of steel plates for repairing concrete structures FRP composites have also been considered for use in such repairs [Uemura et al. 1994]. The associated advantages as described before are obvious in terms of lighter weight and corrosion resistance. In Switzerland very promising results have been obtained by using

carbon fiber FRP for concrete repair [Meier and Kaiser, 1991]. Since such repairs involve only small amount of FRP composites, despite the high costs carbon fiber composites are the most preferred system. Prestressing of the bonded strengthening elements improves the performance. In USA similar work has been reported by Triantafillou and Plevris [1991]. They have shown that strengthening or repair by FRP bonding to the concrete structure can improve the load carrying capacity as well as the stiffness. Concerns in this technique revolve around the behavior under sustained loading, fatigue, thermal cycling, and humidity cycling. Recently Shahawy [1995] reported results of load cycling of a few FRP repaired beams which clearly show a substantial increase in both the stiffness and the ultimate strength capacity of the repaired beams. Figure 4 illustrates this result.

CONCLUSIONS

There are several focus areas in concrete industries where FRP composites are best suited for their corrosion resistance, light weight, high strength, design flexibility, low maintenance, and parts consolidation. While these advantages are tremendous, FRP composites may not be required in many applications where traditional materials alone are best suited. The key to successful application lies in a synergistic approach. In combination with other traditional materials like steel and concrete, composites can be the super material of the future, the hybrids. Many studies concern translating the benefits of FRP composites into new designs, and new combination of materials build civil engineering structures that will out-perform competitive products on the basis of cost, performance, and productivity. Engineering research is needed to optimize and standardize materials and processes so that each material is used where its particular cost is minimal and physical, mechanical, and durability performance is maximized.

REFERENCES

Roll, R.D., (1991),Use of GFRP Rebar in Concrete Structures, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 93-98.

Wu, W.P., GangaRao, H., and Prucz, J.C. (1990). Mechanical Properties of Fiber Reinforced Plastic Bars, Internal Report, Constructed Facilities Center, College of Engineering, West Virginia University.

Faza, S.S. (1995), Properties of FRP Reinforcing Bars, Fiber Reinforced Plastics Workshop, Office of Technology Applications, FHWA, January 1995

GangaRao, H., (1995) Mckinleyville Jointless Bridge with FRP Bars in Concrete Deck, Fiber Reinforced Plastics Workshop, Office of Technology Applications, FHWA, January 1995

Dutta, P.K., (1995), Results of FRP Rebar Creep tests, CRREL Technical Note.

Kumar, S.V., Faza, S.S., GangaRao, H.V.S., Meodad, M.A., (1995) Fatigue Performance of Concrete Slabs with Glass Fiber Reinforced Plastic (RP) Rebars, Proceedings Composite Institute's 50th Annual Conference & Expo 95, Cincinnati, Jan 30- Feb. 1. Session 21C.

Bank, L.C., Xi, Z., and Mossalam, S.M.(1991), Experimental study of FRP grating reinforced concrete slabs, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 111-122.

Koury, J.L. and Dutta, P.K., (1993). Design and Analysis of Composite Isogrid for Bridge Construction, Proceedings Technology 2003, NASA Conference Publication 3249, pp. 377-384.

Larralde, J., and Zerva, A., (1991), Load Deflection Performance of FRP Grating Concrete Composites, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 271-277.

Goodspeed, C., Schmeckpeper, E., Gross, T., Henry, R., Yost, J., Zhang, M., (1991), Cyclical Testing of Concrete Beams, Reinforced with Fiber Reinforced Plastic (FRP) Grids, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 278-287.

Gerritse, A., and Schurhoff, H.J.,(1986), Prestressing with Aramid Tendons, Proceedings, 10th Congress, Federation Internationale de la Precontrainte, New Delhi, India, 7 pp.

C.W. Dolan, (1990), Developments in Non-metallic Prestressing Tendons, PCI Journal, V35, No.5, Sept-Oct, pp. 80-88.

Sen, R., Mariscal, D., Shahawy, M., (1993), Durability of Fiberglass Pretensioned Beams, ACI Structural Journal, Sept-Oct 1993, pp. 525-533

Porter, M. (1994). Personal Communication.

Shahawy, M., (1995) Current Areas of FRP Research in Florida, Fiber Reinforced Plastics Workshop, Office of Technology Applications, FHWA, January 1995

Iyer , S.L., (1995), FRP Reinforced Concrete, Fiber Reinforced Plastics Workshop, Office of Technology Applications, FHWA, January 1995

Iyer, S.L., and Khubchandani, A. (1995), Design and Construction of U.S. Navy Pier at Port Hueneme using Composite Cable. Proceedings Composite Institute's 50th Annual Conference & Expo 95, Cincinnati, Jan 30- Feb. 1. Session 21F.

Priestley, M.J.N. and Seible, F., (1991), Design of Retrofit Measures for Concrete Bridges, in Seismic Assessment and Retrofit of Bridges, University of California, San Diego, Structural Systems Research Project Report. No. 55RP91/03, pp. 197-250.

Harmon, T.G. and Slattery K.T. (1992), Advanced Composite Confinement of Concrete, in Proceedings of the 1st International Conference on Advanced Composite Materials in Bridges and Structures, Canadian Society for Civil Engineering, pp. 299-306.

Priestly, M.J.N., Seible, F. and Fyfe, E. ,(1992), Column Seismic Retrofit using Fiberglass/epoxy Jackets, in Proceedings of the 1st International Conference on Advanced Composite Materials in Bridges and Structures, Canadian Society for Civil Engineering, pp. 287-298

Hasegawa, A., Matsuda, T., Higashida, N., (1992), Carbon Fiber Reinforced Earthquake-Resistant Retrofitting for RC Piers, Doboku Gijutsu, 97 [3]. pp. 43-50

Katsumata H., Kobatake, Y. and Takeda, T., (1988), A Study on Strengthening with Carbon Fiber for Earthquake- resistant Capacity of Existing Reinforced Concrete Columns, Proceedings of the 9th World Conference on Earthquake Engineering, Tokyo, Vol. VII, pp. 517-522

Karbhari, V. M. , Eckeel, D.A., and Tunis, G.C., (1993), Strengthening of Concrete Column Stubs Through Resin Infused Composite wraps, University of Delaware Center for Composite Materials Technical Report CCM Report 93-23.

Ma, G. (1993), Bridge Column Retrofit with CF-Reinforced Materials, in Proceedings High Performance Composites for Civil Engineering Applications, SAMPE Regional Seminar, September 27-28,

Meier,U., and Kaiser, H.,(1991), Strengthening of Structures with CFRP Laminates, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 224-232.

Uemura, M., Aoyama,K., and Kliger, H.S., (1994), A New Reinforcing Fiber Material and Its Applications in the Repair of Concrete Structures, Proceedings of the 39th International SAMPE Symposium and Exhibition April 11-14, Vol 39, pp 347-354.

Triantafillou, T.C. and Plevris, N.,(1991), Post-strengthening of R/C Beams with Epoxy-bonded Fiber Composite Materials, in Advanced Composites Materials in Civil Engineering Structures, Ed. S. L. Iyer, and R. Sen, ASCE, pp. 245-256..

List of Figures:

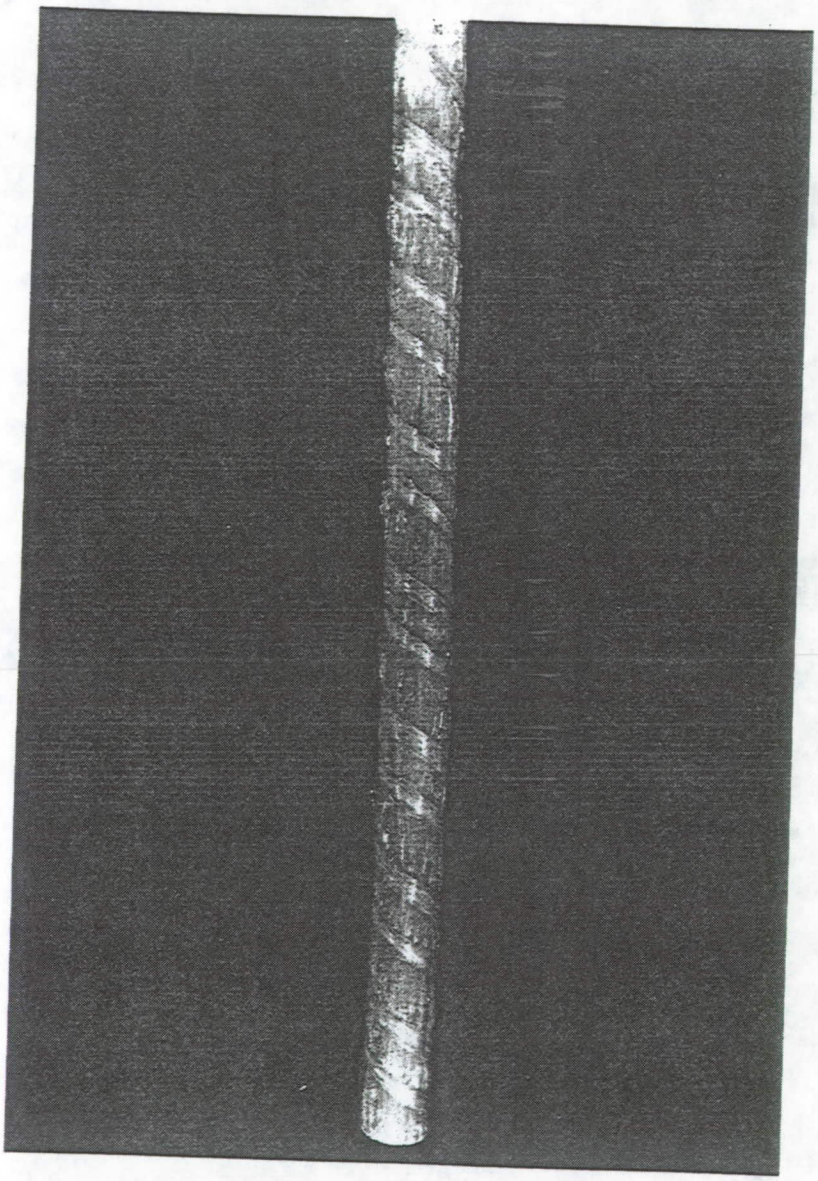


Figure 1. Commercially available concrete reinforcing FRP bar.

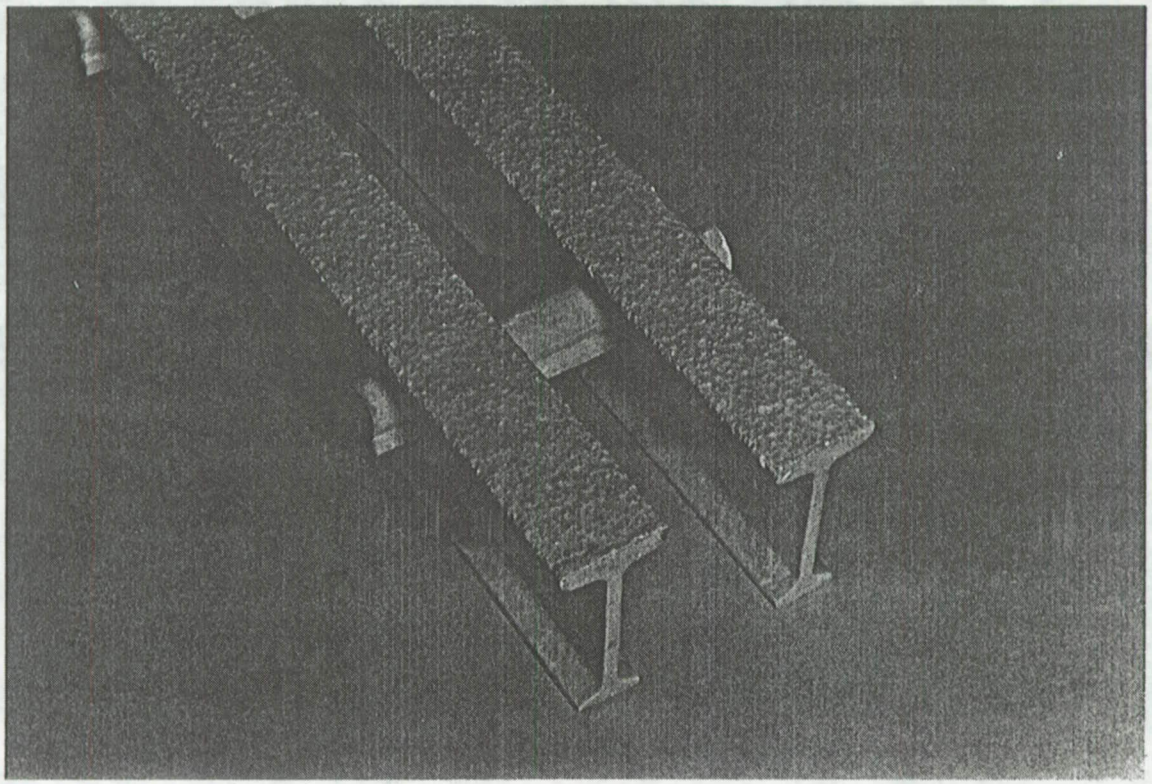


Figure 2. FRP gratings used for experimental studies of concrete reinforcement.

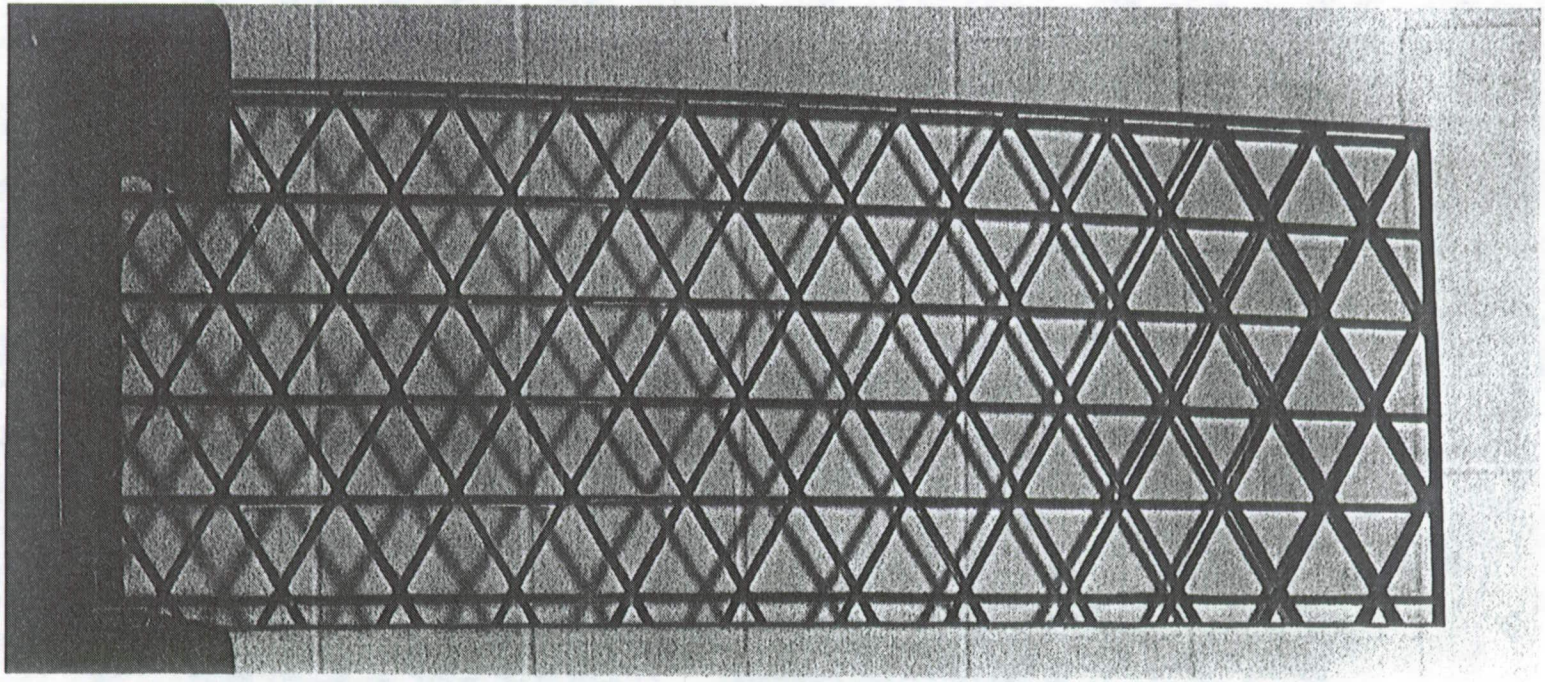


Figure 3. FRP isogrid as a conceptual material for reinforcing concrete.

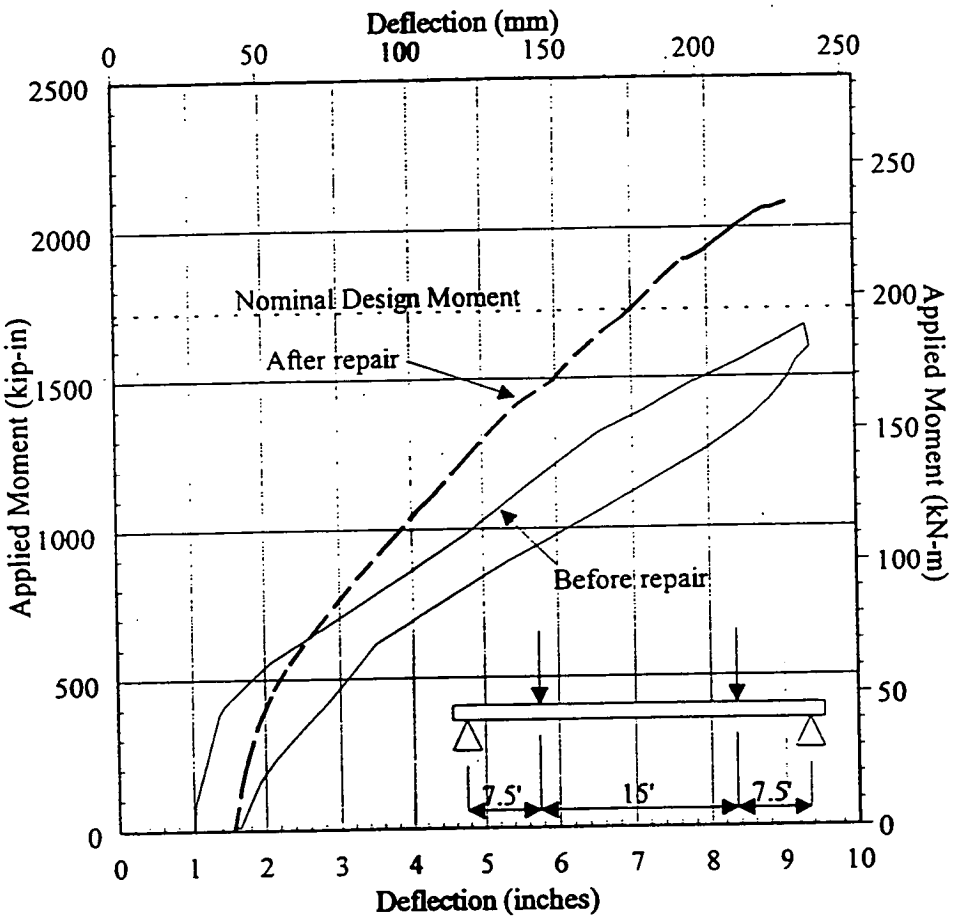


Figure 4. Comparison of original and carbon FRP repaired concrete beam (after Shahawy, M., 1995) .

HIGHER DURABILITY, HIGHER MODULUS ELASTOMERS THROUGH A DOUBLE NETWORK ARCHITECTURE

Patrick G. Santangelo
Chemistry Department, Code 6120
U.S. Naval Research Laboratory
Washington, DC 20375

C. Michael Roland
Chemistry Department, Code 6120
U.S. Naval Research Laboratory
Washington, DC 20375

ABSTRACT

The formation of a double network is a processing technique designed to obtain improved mechanical and failure properties in elastomers. With this technique, an elastomer comprised of two networks is produced which confers "rigid rod"-like mechanical properties while retaining the superior processing of flexible chain polymers. The process involves subjecting a pre-cured elastomeric material to an orientation technique and further curing the material thereby obtaining a superior combination of stiffness and durability. An elastomeric material with enhanced modulus and improved failure properties would have a significant impact in many areas. Recent advances that demonstrate the broad range of capabilities of the technique are discussed using natural rubber with various curatives.

INTRODUCTION

Among the methods for increasing the modulus and strength of elastomers, crosslinking and orientation are the most common approaches used in industry. The standard practice of increasing the crosslink density in order to obtain higher modulus has a well-known drawback. The elastomer becomes brittle, with a reduction in the ultimate properties such as tensile strength, crack growth resistance, fatigue life, and tear strength. Thus, it is generally accepted that with conventional rubber technology, there is a trade-off between stiffness and durability. This traditional compromise can be circumvented using the double network technique.

Orientation is an example of another standard technique for increasing the modulus and strength of materials comprised of flexible chain polymers. The benefit of using orientation to enhance crystallization and mechanical properties is tempered in elastomers by its nonequilibrium nature. However, the double network architecture circumvents this problem, producing orientation that is truly stable in the thermodynamic sense. In a conventional elastomer, the crosslinking produces a set of chain configurations whose equilibrium condition corresponds to a macroscopic state of zero stress, and the mechanical and failure properties are limited by this condition. However, when the network is crosslinked a second time while deformed, the subsequent state of elastic equilibrium no longer corresponds to zero strain. The final configuration is one of net orientation, with consequent mechanical property benefits. Most

intriguing is the fact that the net orientation is not associated with a decrease in processibility, as would be the case for conventional processes.

DOUBLE NETWORK FORMATION

The fabrication technique involves the orientation of pre-cured elastomer, such as natural rubber, polybutadiene, and styrene-butadiene rubber. While maintaining the orientation, the elastomer is furthered cured (i.e., more cross-links are introduced) forming a second network. The final dimensions, or residual strain, of the article reflect the relative apportionment of crosslink densities from the two acts of curing, along with the orientation during the second cure. The modulus of the double network elastomer produced by this procedure is higher than a conventional rubber having the same total crosslink density.

The orientation techniques that can used include uniaxial extension, biaxial extension, simple shear, planar shear, and inflation. The crosslinking is accomplished using suitable methods including essentially all conventional curing procedures (i.e., sulfur vulcanization, "EV" curing, peroxide curing, electron beam irradiation, and gamma ray irradiation).

RESULTS

Modulus

By using the double network technique, the modulus of an elastomer can be enhanced while maintaining a constant crosslink density. In this example, crosslinking was accomplished by a peroxide cure of natural rubber where the number of crosslinks was weighted towards the second network. The total crosslinks as well as their ratio between the two steps were maintained constant. An isotropic cure constituted the first network, while the second network was introduced at various strains under uniaxial strain. Samples from the first crosslinking were cured in the unstrained state simultaneously with the stretched samples, thus yielding conventional networks of the same crosslink density.

Displayed in Figure 1 is the residual strain dependence on the stress-strain data (measured parallel to the direction of orientation) for natural rubber double networks formed at various strains. The residual strains are as follows: 1.0(✱); 1.5(●); 2.3(♦); 3.6(×); and 4.3(▲). Also included in Figure 1 is the curve for a (conventional) single network with the same total crosslink density. In Figure 2, the ratio of the double to single network moduli are shown as a function of residual strain. The moduli are taken from Figure 1 at a strain of 1.1. An enhancement in the double network moduli are seen for residual strains greater than 170%. For residual strains less than 170%, the double networks exhibit smaller moduli. This has been previously observed experimentally, but attributed to scission of bonds in the original network. In the present experiments, the use of carbon-carbon network junctions is expected to minimize changes in the first network during the second crosslinking. These results are also consistent with our earlier calculations.

In the present example the strain during the second cure stage was varied, while the crosslink distribution between networks remained fixed. The *total* crosslink density is factored out by ratioing all moduli by the modulus of the single network of equivalent crosslink density.

Hence, by choosing a fixed distribution of crosslinks between the two networks, the strain at which the second crosslinking occurs is the independent variable for the double networks in Figure 2.

If the same second stage curing strain is used, but the distribution between networks is altered, the modulus enhancement can change. Moreover, double networks having the same total crosslink density and residual strain can be obtained by varying both the second stage curing strain and the crosslink apportionment. Such elastomers exhibit different behavior; that is, the residual strain does not uniquely define the double network material.

Tensile Fatigue and Strength Measurements

In investigating the tensile properties of double networks, it is advantageous to compare the results with conventional rubbers that have either (I) the same total crosslink density or (II) the same modulus as the double network. Displayed in Figure 3 is the stress-strain isotherms for the natural rubber double networks and single networks (I) with the same total crosslink density. As expected from the previous example, the moduli of double networks with a residual strain equal to 200% are higher than that of single networks. The stress-strain curve for single networks (II) which have moduli equal to the modulus of a double network at 10 % elongation is also shown in Figure 3. In order to facilitate the survivability of the double network rubbers at large strains during the second crosslinking stage, an ultrafast sulfur curing method was incorporated for all elastomers used in this example which allowed the use of lower curing temperatures, and minimizes the contribution of polysulfidic linkages.

While higher modulus is conventionally obtained by simply increasing the number of crosslinks in an elastomer, higher crosslink densities have a deleterious effect upon failure properties (e.g. fatigue life and tensile strength). Figure 4 displays the fatigue life of single networks I and II as a function of strain energy. As the strain energy increases (i.e. larger dynamic elongation) the fatigue life of higher crosslinked single network II decreases below that of the single network I. With double network, in which the total crosslinks are the same as single network II but apportioned between two composite networks, a higher modulus is attained without the loss of durability. In fact, as shown in Figure 4, double networks exhibit a markedly longer fatigue life than both conventional networks I and II throughout the experimentally accessible strain energies.

To the extent that the fatigue life is greater than that observed for conventional elastomers, the double network technology circumvents the usual compromise between stiffness and durability. This is an important result because the fatigue life and other failure properties (e.g. tensile strength) of conventional elastomers exhibit a maximum as a function of crosslink density.

At the usual crosslink densities of commercial rubbers, higher modulus entails reduced strength. The corresponding tensile strength values for single network I, single network II, and the double network are 9.9 ± 1.6, 8.4 ± 1.8, and 6.5 ± 1.4 MPa; respectively. The low tensile strength values reported are a consequence of the ring geometry used in the tensile strength measurements. The tensile strength of the higher crosslinked single network II is smaller as expected. Notwithstanding the enhanced crystallizability of the natural rubber double network,

the lower tensile strength measured for the double network possible is possibly due to scission as chain segments of the first composite network approach the limit of their extensibility thereby decreasing the observed tensile strength.

Transverse Properties

While the stress-strain data in Figures 1-4 refer to the direction parallel to the strain during curing, Figure 5 displays the stress-strain data obtained in both the perpendicular and parallel directions for double networks with a constant crosslink density but different curing strains. As in the previous example, an ultrafast sulfur curing method was used. Also included in Figure 5 is the curve for a single network of the same crosslink density. At low strains, no significant differences were observed. For each double network at high strains, the perpendicular modulus was less than the modulus in the direction of the original crosslinking strain, but comparable to the modulus of the single network. Whether the upturn in the parallel stress-strain curve is a result of finite chain extensibility or a manifestation of enhanced crystallizability is a matter currently under investigation.

Tensile strength values (ASTM D4482) were measured for the double networks in both the parallel and perpendicular direction, and these results are shown in Table I. The striking feature is the tensile strength measured for the double networks in the transverse direction are equal to the tensile strength of the single network. The smaller tensile strength values for the double networks in the parallel direction are consistent with the tensile strength data obtained with the ring geometry used in the previous example.

Single Network	Double Network I		Double Network II	
Tensile Strength (MPa)	Parallel	Perpendicular	Parallel	Perpendicular
24.2 ± 4.3	9.0 ± 3.4	20.2 ± 2.9	14.7 ± 3.4	21.2 ± 2.9

Table I. Tensile strength of the single and double networks with the same total crosslink density. Double network I has a larger residual strain than double network II.

SUMMARY

The technology described herein focusses on the modulus obtained when an elastomer is crosslinked a second time while oriented forming a double network. The modulus of the double network can be made greater than the modulus of a conventional, isotropic elastomer of equal crosslink density. The modulus enhancement of the double network is governed by the apportionment of crosslinks between the first and second cure steps, along with the magnitude of the orientation used during the second crosslinking stage.

In addition, for elastomers which strain crystallize, an enhancement of strain crystallizability can be obtained by virtue of the double network. This leads to significantly better durability, since the strain crystallization is the primary mechanism for failure resistance in such elastomers. Thus, while conventional processing technology forces a trade-off between durability

and modulus, this technology enables higher modulus to be realized while simultaneously improving fatigue failure.

Besides the double network technique, the only other method for i) increasing the modulus, ii) maintaining a constant crosslink density and (iii) improving the durability of elastomers is carbon black reinforcement. However, carbon black fillers yield significant hysteresis, which is detrimental in many applications. Of course, double networks can be used in conjunction with carbon black reinforcement.

Wherein the double network modulus is significantly amplified while maintaining durability, elastomers of better mechanical properties are obtained by methods readily incorporated into existing manufacturing processes. The following are some ideas concerning specific implementations of the double network architecture.

Regenerative braking systems use a deformed elastomer to store energy. Although not yet commercial viable, these devices offer the potential for significant increases in fuel economy. During braking in regenerative braking systems, a portion of the vehicle's kinetic energy is temporarily stored via elastic deformation of an elastomer, rather than being entirely dissipated as friction. If the elastomer is designed via double network technology, larger amounts of energy could be stored for a given deformation.

Aerosol cans using an inflated elastomer as the propulsion mechanism rather than compressed gas have enormous advantages. The main advantages of these self-pressurized containers is the environmental appeal of eliminating gaseous propellants and a reduced likelihood of explosions during incineration. Another role for double networks is in the replacement of the conventionally crosslinked elastomers currently used in such devices. Retraction of a cylindrical rubber sleeve upon opening of a discharge valve evacuates the contents. To the extent that the material properties of the rubber, along with its shape, determine the delivery pressure, the double network architecture offers a route to improving on the technology of nongaseous pressurizing devices.

During radial tire production, a tire carcass is calendared, expanded, then cured. Introduction of crosslinks (for example, on-the-fly electron irradiation) of the calendared stock will enable utilization of double network technology in the tire building process virtually without alteration of current production methods. Significant improvements in the mechanical properties of tire plies can result.

REFERENCES

J.P Berry, J. Scanlan, and W.F. Watson, *Trans. Far. Soc.* **52**, 1137 (1956).
A. Greene, K.J. Smith, and A. Ciferri, *Trans. Far. Soc.* **61**, 2772 (1965).
T.J. Pond and A.G. Thomas, Proceedings Int. Rubber Conf. (Venice 1979), p. 810.
C.M. Roland and M.L. Warzel, *Rubber Chem. Technol.* **63**, 285 (1990)
C.M. Roland and K.L. Peng, *Rubber Chem. Technol.* **64**, 790 (1991).
P.G. Santangelo and C.M. Roland, *Rubber Chem. Technol.* **67**, 359 (1994).
K.J. Smith and R.J. Gaylord, *J. Poly. Sci. A-2* **10**, 283 (1972).

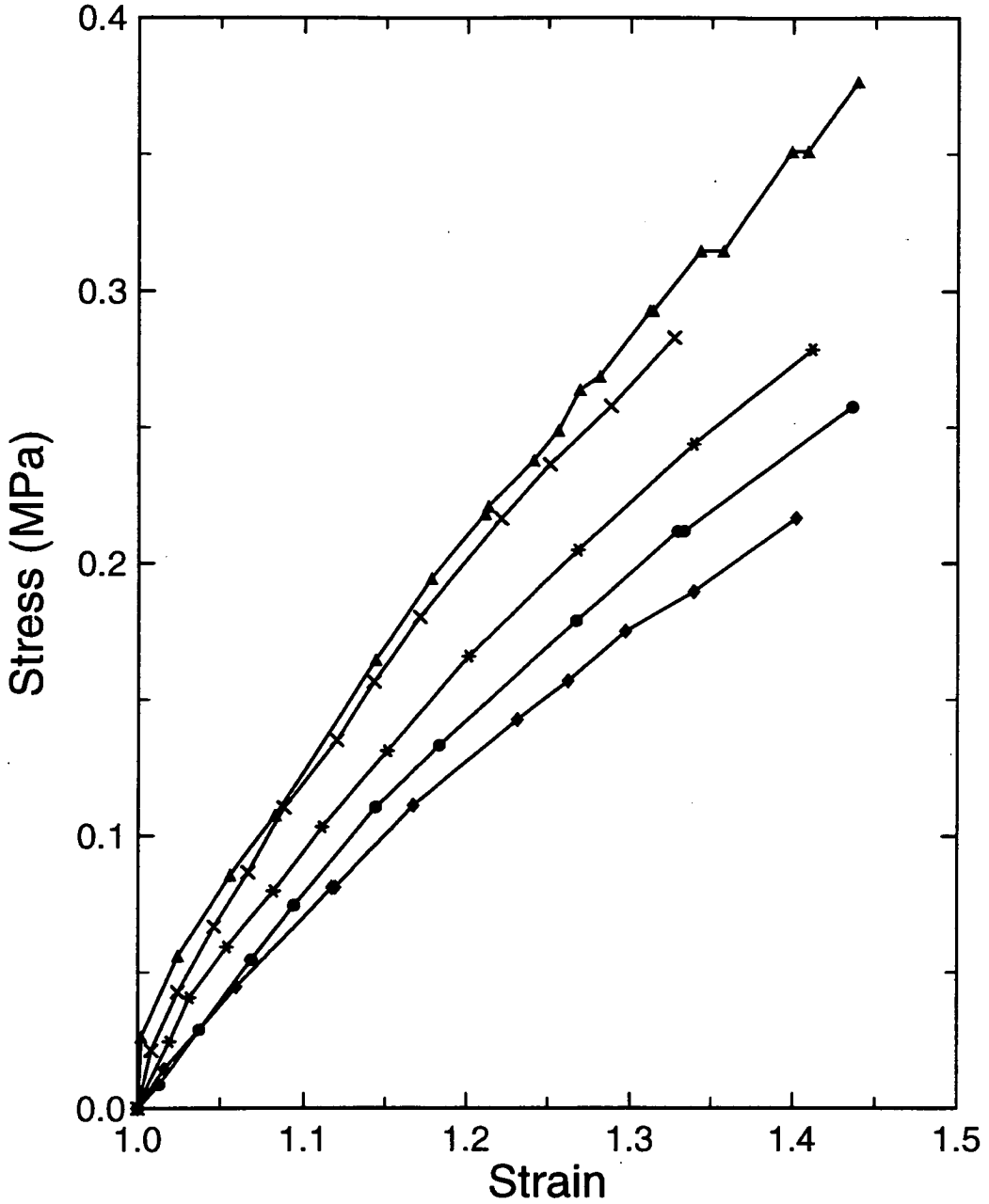


Figure 1. The residual strain dependence on the stress-strain data for double networks formed at various strains (residual strain: 1.0 (✱), 1.5 (●), 2.3 (♦), 3.6 (×), and 4.3 (▲)). The double networks exhibit higher stress than the single network (✱) of equal crosslink density for residual strains greater than 170%.

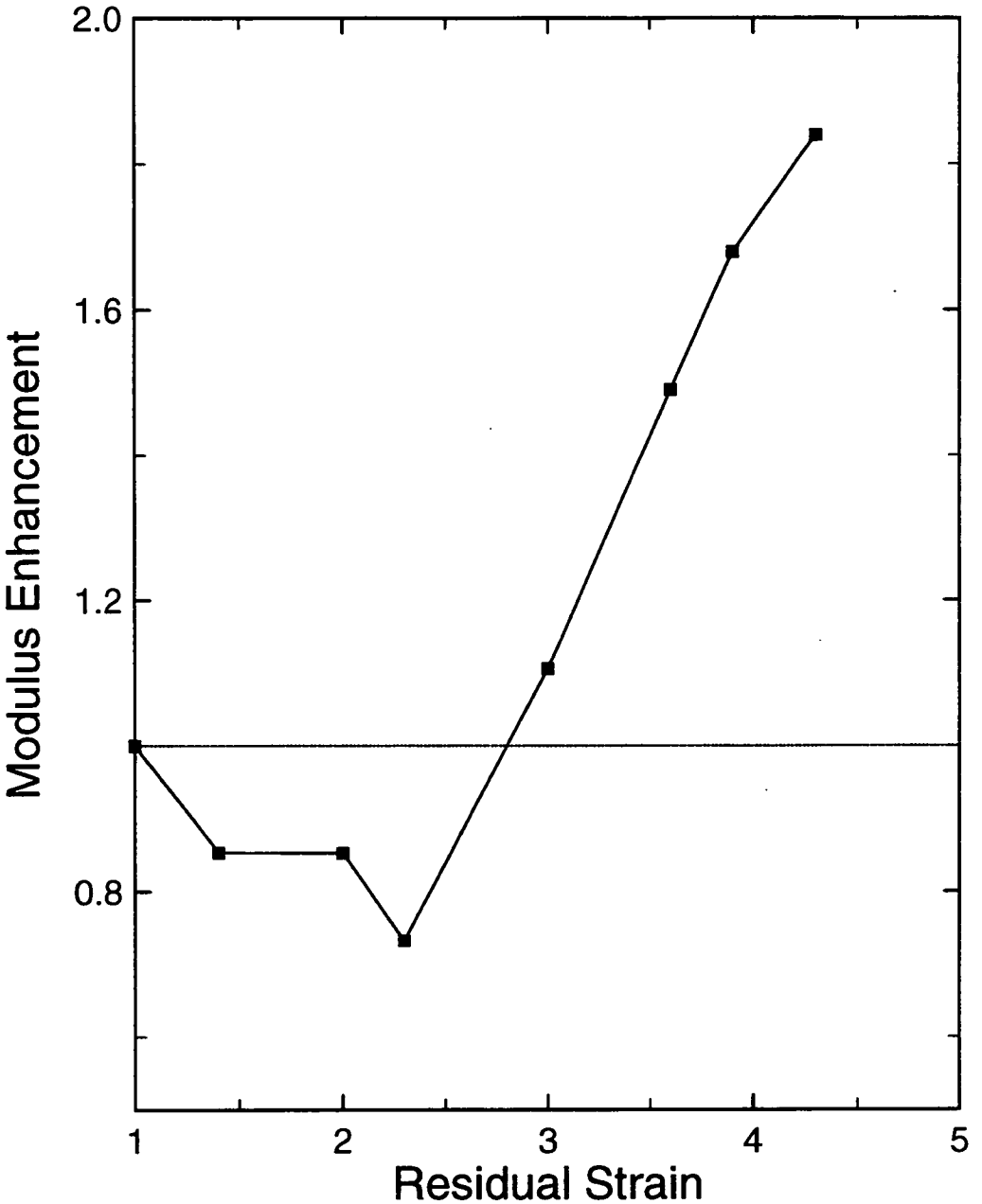


Figure 2. The ratio of the double network modulus to that of the corresponding single network for various residual strains. The data is taken from Figure 1 at a strain of 10%. An enhancement in the double network moduli are seen for residual strains greater than 170%.

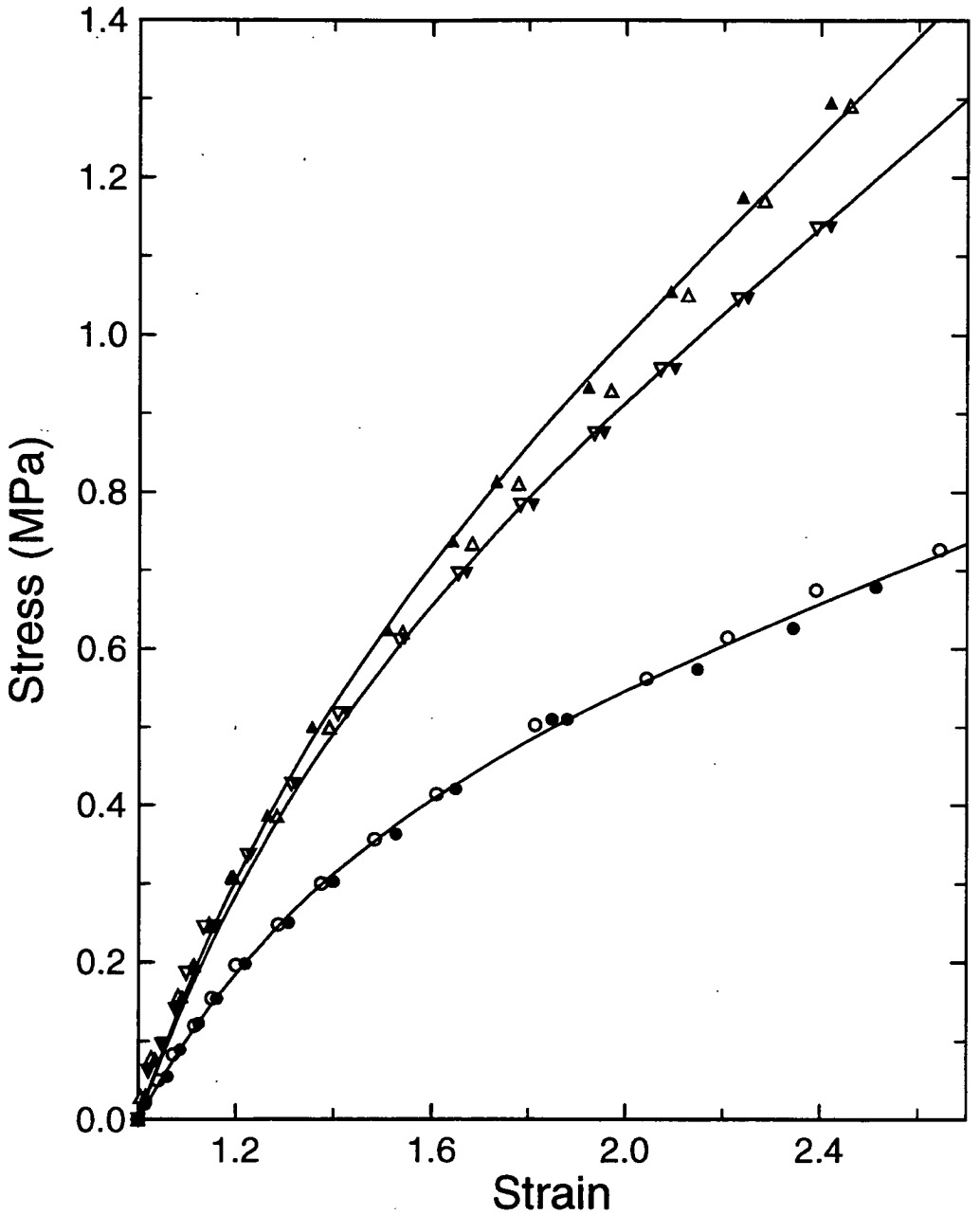


Figure 3. The stress-strain data for natural rubber double networks (▲, Δ) with a residual strain of 200%. Also shown are the stress-strain data for single networks that have either (●, ○) the same total crosslink density or (▼, ▽) the same modulus (i.e., at 10% strain) as the double networks. The open and closed symbols represent individual samples for a given network. The double networks exhibit enhanced moduli when compared to either single network at strains above 20 %.

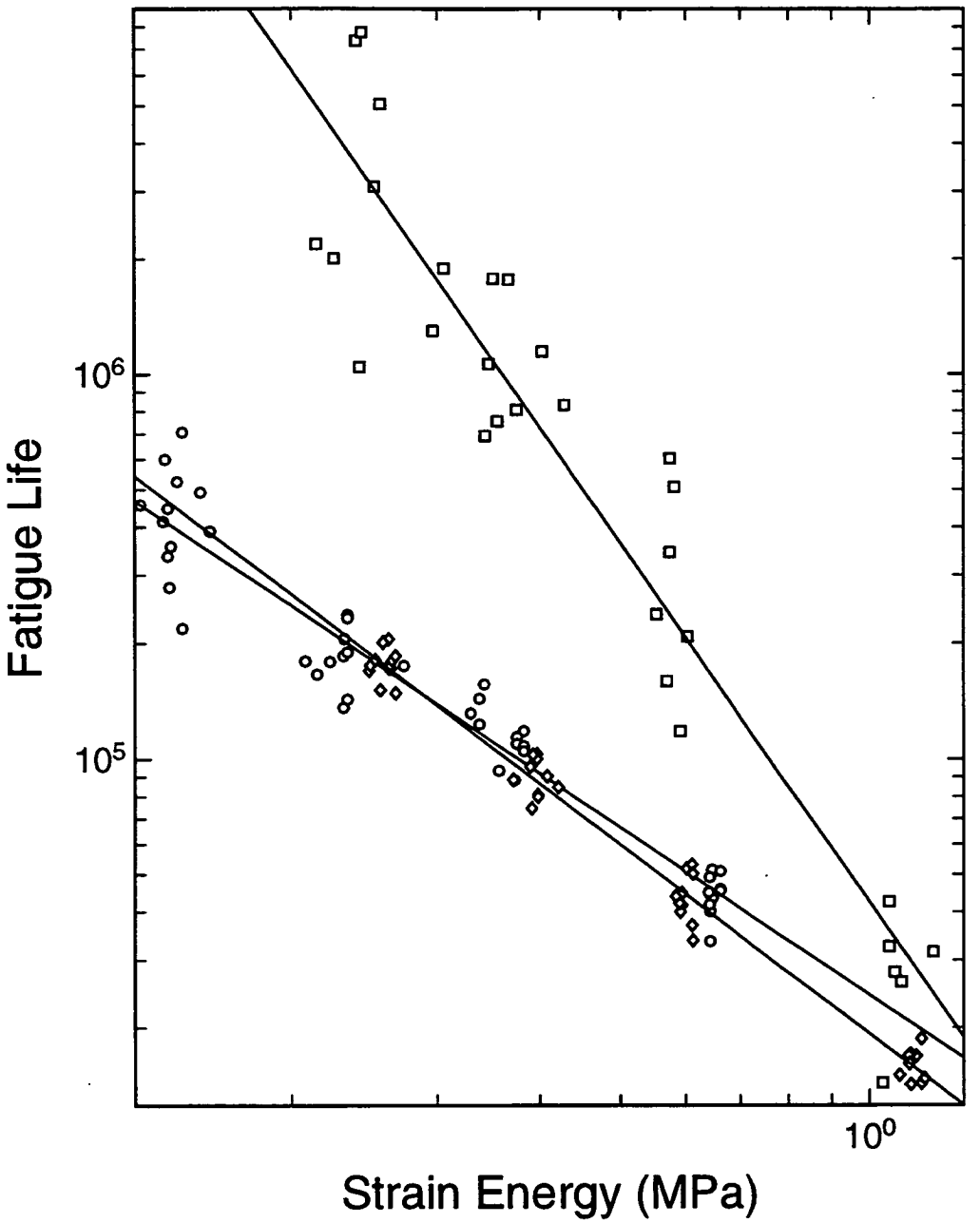


Figure 4. The fatigue life as a function of strain energy is shown for a double network (□), along with the data for single networks that have either (○) the same total crosslink density or (◊) the same modulus (i.e., at 10% strain) as the double network. The corresponding stress-strain data is shown in Figure 3. The double network has a noticeably longer fatigue life over the entire range of strain energies.

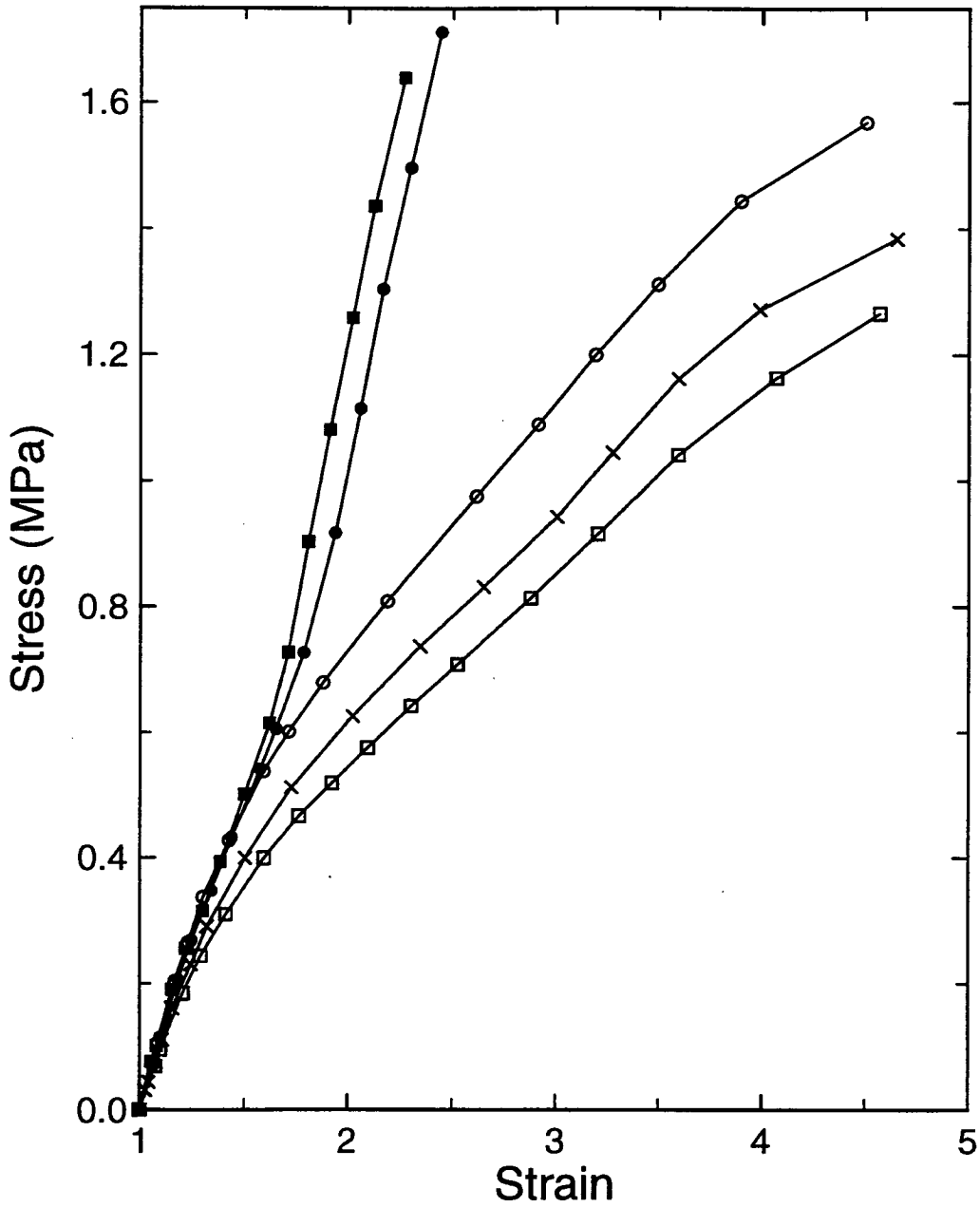


Figure 5. The stress-strain data for two double networks that were formed with a constant crosslink density but different residual strains (I: square, II: circle; see Table I) are displayed. The data was obtained in the directions perpendicular (open) and parallel (solid) to the orientation strain used during the second stage crosslinking. Double network I has a larger residual strain than double network II. Also shown is the curve for a single network (X) of the same crosslink density. An anisotropy in the modulus of a double network is clearly demonstrated.

Power and Energy

COMMERCIAL TECHNOLOGIES FROM THE SP-100 PROGRAM

Vincent C. Truscello
SP-100 Project Manager
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109
(818) 354-1820

Jack F. Mondt
SP-100 Project Deputy Manager
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109
(818) 354-1900

Toshio Fujita
Member of Technical Staff
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109
(818) 354-8700

ABSTRACT

For more than a decade, the Jet Propulsion Laboratory (JPL) and Los Alamos National Laboratory (LANL) have managed a multi-agency funded effort to develop a space reactor power system. This SP-100 Program has developed technologies required for space power systems that can be implemented in the industrial and commercial sectors to improve our competitiveness in the global economy. Initial steps taken to transfer this technology from the laboratories to industrial and commercial entities within the United States include: (1) identifying specific technologies having commercial potential; (2) distributing information describing the identified technologies and interacting with interested commercial and industrial entities to develop application-specific details and requirements; and, (3) providing a technological data base that leads to transfer of technology or the forming of teaming arrangements to accomplish the transfer by tailoring the technology to meet application-specific requirements.

SP-100 technologies having commercial potential encompass fabrication processes, devices, and components. Examples are a process for bonding refractory metals to graphite, a device to sense the position of an actuator and a component to enable rotating machines to operate without supplying lubrication (a self-lubricating ball bearing). Shortly after the National Aeronautics and Space Administration (NASA) Regional Technology Transfer Centers widely disseminated information covering SP-100 technologies, over one hundred expressions of interest were received. These early responses indicate that there is a large potential benefit in transferring SP-100 technology. Interactions with industrial and commercial entities have identified a substantial need for creating teaming arrangements involving the interested entity and personnel from laboratories and their contractors, who have the knowledge and ability to tailor the technology to meet application-specific requirements.

INTRODUCTION

The SP-100 Program recently completed the development of technologies for space reactor power systems (see References 1-6). The effort, which was carried out from 1983 to 1994, was jointly supported by the Department of Energy (DOE), the National Aeronautics and Space Administration (NASA), and the Ballistic Missile Defense

Organization (BMDO), that was formerly named the Strategic Defense Initiative Office (SDIO). The technology development activities were directed and managed by a Jet Propulsion Laboratory (JPL) and Los Alamos National Laboratory (LANL) Project Office, and developed by a team of contractors and national laboratories led by Martin Marietta Corporation (MMC).

The technologies developed for the SP-100 Space Reactor Power System (SRPS) are available for implementation when required for future commercial, NASA and military space missions. Very importantly, many of the individual technologies are applicable to multiple uses in the industrial and commercial sectors. In concert with the national effort to improve our competitive posture by transferring government developed technologies to our industrial and commercial sectors, JPL, together with other laboratory and contractor personnel familiar with SP-100 technologies, is undertaking an effort focused specifically on transferring these technologies. This effort is supported by the Commercial Program Office at JPL that includes the Technology Affiliates Program, Intellectual Property Management, and Technical Information Dissemination.

This paper describes the SP-100 technology transfer effort in terms of its overall strategic plan, steps initiated as of July, 1994, and the resulting progress. Conclusions are drawn regarding the potential benefits that can be achieved and the scope of the activity needed to successfully capture the full benefits of transferring SP-100 technologies.

SP-100 TECHNOLOGY TRANSFER EFFORT

As depicted in Figure 1, the overall SP-100 technology transfer effort consists of four basic phases. The actions required for the accomplishment of these phases are to: (1) identify and describe the SP-100 technologies; (2) distribute SP-100 technology information to industrial U.S. companies; (3) establish how the needs of the user can be met by this technology; and, (4) implement technology transfer.

The first phase requires a systematic review of the technology development activities undertaken within the SP-100 Program. Each of the individual technologies are assessed in terms of their potential for being used to improve commercial products. Those SP-100 technologies judged to have the highest commercial potential are selected. Information that describes these technologies and their identified potential commercial application is then prepared.

In the second phase, this information is distributed to a network of NASA Technology Transfer Centers and is provided also directly to U.S. industrial and commercial companies involved with the identified potential applications. During the third phase, specific needs of the commercial user are identified, and ways in which the SP-100 technology can meet these needs are established.

The fourth phase uses the specific knowledge obtained in the third phase as the basis for implementing the transfer of technology. In some cases, simply providing the SP-100 technology data base to a potential user is sufficient to accomplish the transfer of technology. For these cases, the technology is either directly applicable or the interested company has the knowledge, ability, and resources to tailor the technology to meet his needs. In other cases, teaming arrangements must be implemented and resources provided to support the interested industrial or commercial company in modifying the technology to meet the requirements of their application. Technical support can be provided by SP-100 component suppliers or laboratory personnel who are familiar with the technology and possess the expertise required for the tailoring effort. Resources are provided by the government to match industry funds, as required, to modify the technology.

Description of SP-100 Technologies

Following a systematic review and assessment of the technologies developed in the SP-100 Program, a set of technologies judged to have high potential for commercialization was selected. The selected SP-100 technologies are categorized as processes, devices, and components.

The process technologies, along with their identified commercial possibilities, are listed in Table 1. Bonding techniques which join different materials and have low electrical and thermal resistance have a wide range of applications. Explosive forming is used to mechanically join dissimilar materials. Hard surface coatings that remain bonded to the substrate material have many applications.

Table 2 covers device technologies and their potential commercial application. The device category contains a collection of very different devices. The devices include heat exchangers and heat pipes for heat transfer, thermoelectrics for power conversion and cooling, electrical actuators for moving control segments, sensors for measuring position and temperature, and gas/liquid separators for both microgravity and gravity applications.

The third category, SP-100 component technology, is presented in Table 3. This category consists of a wide range of components having broad commercial possibilities. The new technology components are carbon-carbon tubes for heat exchangers, compliant pads for joining materials having different coefficients of thermal expansion, high temperature coils for motors, pressure transducer adaptations for use with high-temperature and/or corrosive fluids, self-lubricating ball bearings for rotary equipment at high temperatures, and very high temperature metal-ceramic electrical insulators to isolate high temperature, high-voltage environments.

Information Dissemination

Working closely with NASA Regional Technology Transfer Centers (RTTCs), the SP-100 technology transfer project has successfully disseminated information to industry covering the selected technologies available from the SP-100 program. Over 2000 companies have been sent data sheets describing the technologies and suggesting potential commercial applications. The RTTCs have also placed notices within their regularly published newsletters describing the technologies available and directing interested companies to JPL. Other sources of public notice have included BMDO's publication, "*The Update*", the NASA publication, "*Innovation*", and a private sector newsletter called, "*Advanced Manufacturing Technology*".

Interactions with Potential Users

As of July, 1994, more than 135 companies have indicated an interest in one or more of the selected SP-100 technologies. Each of these companies was contacted. In general, this contact led to providing them with detailed reports covering the technology and a series of follow-on discussions that clarified the application and defined specific steps that would be required to implement the technology. In some cases, SP-100 contractors involved in the development were brought into the discussion to clarify specific detailed issues.

Implementation of Technology Transfer

After application-specific details and requirements are determined through the interaction process, the implementation effort is divided into two general categories. In the first category, the provision of the SP-100 technology data base together with contacts to hardware contractors is sufficient to accomplish the technology transfer. As of July, 1994, a number of companies have begun using technologies developed under the SP-100 Program. The self-lubricating bearing is an example of a component in this category. Uses for these bearings have been found for equipment designed for a space shuttle experiment, vacuum pumps operating in a vacuum chamber, and fail-safe support system for a kinematic flywheel storage device being tested under hard vacuum conditions. The high-temperature coil is being baselined on generators for a More Electric Airplane concept. The process to bond rhenium to niobium metal has found a home in chambers for small thrusters while our hard coating processes are being considered for wear resistant surfaces to protect machinery, and as a coating to reduce erosion and increase the lifetime in water jet orifices.

In the second category, the SP-100 technology must be tailored, or in some cases, further developed to meet application-specific requirements. Early experience indicates that this constitutes the largest category with commensurate potential benefits. In general, the most effective way of developing the technology is to form a teaming arrangement between the industrial user and personnel from laboratories and their contractors who have

knowledge about the technology. The industrial user must take the lead and define his requirements for this technology. Then, the laboratories and/or their contractors can modify the technology based on their experience.

As a direct result of early interactions, a number of cooperative teaming arrangements are being formed to transfer technologies, including the gas separator, compliant pad, the Johnson noise thermometer, bonding of ceramics to metals, carbon-carbon tubes, and self-lubricating bearings. We presently are working with a number of companies, both large and small, in developing programs to modify these components for other applications. A small company servicing the soft drink industry has found a novel way to process the liquid syrup through the use of the gas separator. Several companies have proposed the use of the compliant pad for transporting either heat or electricity through a conductor connection in ceramic parts. The pad prevents the connected ceramic parts from being structurally damaged. The Johnson Noise thermometer is being considered by at least one industrial company for measuring the temperature of jet engines operating at temperatures in excess of 3000° F, while our processes to bond ceramics or graphite to metals are being considered for optical assemblies, turbine wheels, feed-through for an x-ray accelerator, and fabricating the internal structure for large energy storage capacitors.

A number of companies want to modify our basic self-lubricating ball bearing design to allow operation at high temperature in a gas environment, or to enable the bearings to either conduct electricity or to be cleaned at high temperatures. Finally, our carbon-carbon structure is being considered by industry for fabricating various forms of heat exchangers for a variety of commercial applications.

CONCLUSIONS

The conclusions drawn from early activities are:

1) Broad opportunities exist for transferring SP-100 technologies to a wide range of industrial and commercial entities;

2) SP-100 technologies must be tailored to meet the specific requirements of each commercial or industrial entity's application;

3) Teaming arrangements are needed to tailor SP-100 technologies, where the arrangement links an industrial or commercial entity which knows the market and its requirements with a national laboratory and the component developer who can quickly and cost effectively modify the SP-100 technology to meet the commercial requirements; and,

4) Exploration of technology transfer opportunities requires a proactive effort encompassing: (1) identification of technologies having commercial potential; (2) distribution of information; (3) delineation of specific applications and their associated requirements; and, (4) transfer of technology either by providing a data base or more generally forming teaming arrangements to tailor the technology.

ACKNOWLEDGMENTS

The work reported herein was performed by the SP-100 Project Office, Jet Propulsion Laboratory, under contract with the U. S. Department of Energy, Oakland Operations Office. The authors wish to acknowledge the support and assistance provided by the NASA Regional Technology Transfer Centers, the Commercial Program Office at the Jet Propulsion Laboratory, and a team of SP-100 contractors.

REFERENCES

1. Mondt, J. F., Truscello, V. C., and Marriott, A. T., *"SP-100 Power Program"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

2. Marriott, A. T. and Fujita, T., *"Evolution of SP-100 System Designs"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

3. DeMuth, S. F., *"SP-100 Reactor Subsystem Development"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

4. DeMuth, S. F., *"SP-100 Control Drive Assembly Development Plan"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

5. Buksa, J. J. (1994), *"SP-100 Heat Transport Technology Development"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

6. England, C. and Ewell, R. C. (1994), *"Progress on the SP-100 Power Conversion Subsystem"*, Eleventh Symposium, Space Nuclear Power Systems, Albuquerque, NM, 1-94.

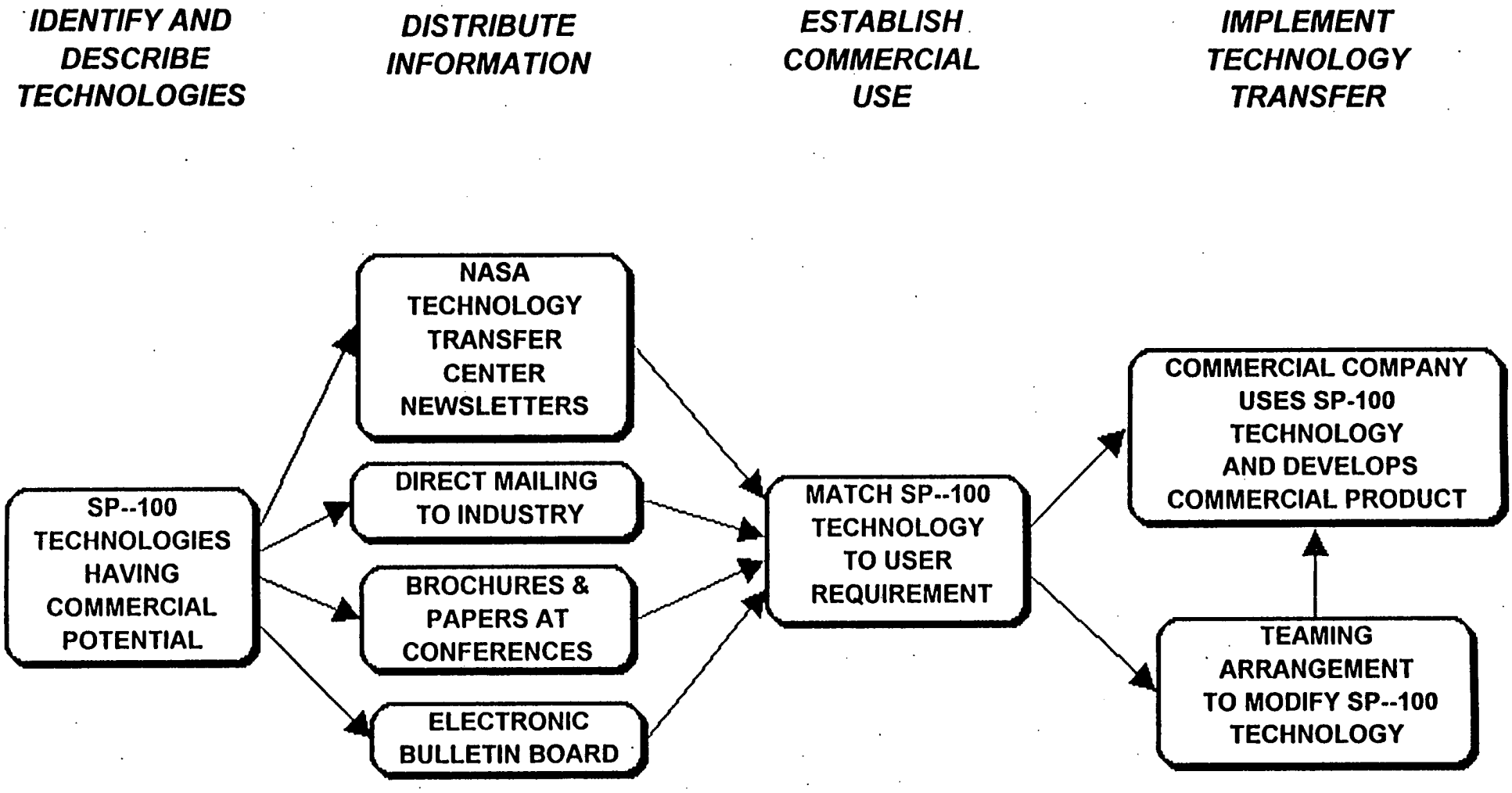


Figure 1. SP-100 Technology Transfer Effort

TABLE 1. SP-100 PROCESS-ORIENTED TECHNOLOGIES HAVING COMMERCIAL POTENTIAL

PROCESS TECHNOLOGY	COMMERCIAL POSSIBILITIES
Bonding of Rhenium to Niobium	Application of corrosion resistant metal liners to a substrate such as a rocket motor. Techniques applicable to bonding other materials with different coefficients of thermal expansion.
Bonding Metals to Sapphire (Alumina)	Bonding clear insulators to metals in lamps, lasers, special-purpose windows, and high-voltage isolation devices. Techniques for bonding are applicable to amorphous glasses as well as single-crystal oxides.
Bonding Refractory Metals to Graphite	Applications include prosthetic devices, rocket motors, lightweight aircraft components, and brakes. Basic techniques are applicable to bonding of commercial alloys.
Explosively Formed Copper Plate with Niobium Clad	Uses encompass multi-layered structures requiring mechanical bonding of dissimilar metals. Techniques are applicable to metals having different coefficients of expansion.
Hard Surface Coatings on Substrates	Applying surface coatings to substrates to provide wear resistance to sliding surfaces in machinery as well as erosion protection for surfaces exposed to gas and liquid flows.

TABLE 2. SP-100 DEVICE TECHNOLOGIES HAVING COMMERCIAL POTENTIAL

DEVICE TECHNOLOGY	COMMERCIAL POSSIBILITIES
Compact Heat Exchanger	Very small flat plate liquid-to-gas heat exchangers or any flat plate heat exchanger for transferring heat from a very corrosive fluid to a gas. Fabrication processes accommodate materials such as stainless steel, copper, aluminum, as well as refractory materials.
Conductively Coupled Thermoelectric Converter	Generation of electrical energy by using waste heat from industrial and home furnaces. The thermoelectric converter array would be incorporated into the thermal insulation of the furnaces.
Gas Separator	Replacement of current liquid-gas separators with moving parts that are used in micro-gravity, e.g., water-air separators used for environmental control on the space station. Compact liquid-gas separators for ground applications, e.g., removal of gases from aerated water.
High Temperature Linear Position Sensors	Use of Linear Variable Differential Transformers (LVDTs) and positions switches in high-temperature locations for improved control of industrial processes.
High Temperature Motor-Clutch-Brake Assemblies	Industrial applications having radiation or high-temperature environments that require motor-clutch-brake units for actuation of process equipment.
Johnson Noise Thermometer	Measurement of temperatures in furnaces and in radiation environments without encountering the decalibrating effects of temperature or radiation-induced changes in the temperature-sensing element.
Liquid Metal Heat Pipe for High-Temperature Operation	Spread heat uniformly over a heat transfer surface via an array of pipes to enhance waste rejection or transfer of heat to another medium. Also, the pipes can transport concentrated heat from one location to another with only a small temperature drop.
Self-Powered TEM Pump	The thermoelectric-electromagnetic (TEM) pump is able to pump electrically conducting fluids (e.g., liquid metals) at high temperatures. The avoidance of moving parts improves reliability when incorporated in industrial processes.

TABLE 3. SP-100 COMPONENT TECHNOLOGIES HAVING COMMERCIAL POTENTIAL

COMPONENT TECHNOLOGY	COMMERCIAL POSSIBILITIES
Carbon-Carbon Tubes with Integral Fins and Liners	Tubular heat exchangers for applications such as gas or oil-fired furnaces. Low weight and high strength favor heat exchanger applications on mobile systems.
Complaint Pad	Joining surfaces of materials having different coefficients of thermal expansion, while providing good thermal conduction and electrical transmission. For example, the pad could be used to join a ceramic superconducting cable with a conventional cable.
High Temperature Coils	High temperature devices that use electromagnetic coils and conductors. These devices include motors, position sensors, solenoid valves, and switches.
Pressure Transducer Adaptations for Hostile Environment	Industrial and scientific research applications requiring measurement of pressure drops in either fluids at high temperatures exceeding transducer limits or fluids requiring special containment materials.
Self-Lubricating Ball Bearings	Equipment involving rotary motion that operate at high temperatures in either a vacuum or gas environment. The need for a separate lubrication system is eliminated.
Very High Temperature Electrical Insulator	Industrial applications requiring insulator technology to isolate high-voltage environments at high temperatures while maintaining excellent thermal contact.

Lithium Rechargeable Batteries for Space Applications

NOT AVAILABLE AT THIS TIME

PIEZOELECTRIC MOTOR DEVELOPMENT AT ALLIEDSIGNAL INC., KANSAS CITY DIVISION*

Robert B. Pressly
Staff Engineer
AlliedSignal Inc., Kansas City Division
Kansas City, Missouri 64141-6159

Charles P. Mentesana
Staff Engineer
AlliedSignal Inc., Kansas City Division
Kansas City, Missouri 64141-6159

ABSTRACT

The Kansas City Division of AlliedSignal Inc. has been investigating the fabrication and use of piezoelectric motors in mechanisms for United States Department of Energy (DOE) weapons applications for about four years. These motors exhibit advantages over solenoids and other electromagnetic actuators. Prototype processes have been developed for complete fabrication of motors from stock materials, including abrasive machining of piezoelectric ceramics and more traditional machining of other motor components, electrode plating and sputtering, electric poling, cleaning, bonding and assembly. Drive circuits have been fabricated and motor controls are being developed. Laboratory facilities have been established for electrical/mechanical testing and evaluation of piezo materials and completed motors. Recent project efforts have focused on the potential of piezoelectric devices for commercial and industrial use. A broad range of various motor types and application areas has been identified, primarily in Japan. The Japanese have been developing piezo motors for many years and have more recently begun commercialization. Piezoelectric motor and actuator technology is emerging in the United States and quickly gaining in commercial interest. The Kansas City Division is continuing development of piezoelectric motors and actuators for defense applications while supporting and participating in the commercialization of piezoelectric devices with private industry through various technology transfer and cooperative development initiatives.

INTRODUCTION

The fabrication and use of piezoelectric motors to replace solenoids and other electromagnetic actuators in mechanisms for DOE weapons applications has been an ongoing investigation of the Kansas City Division of AlliedSignal Inc. for approximately four years. The project has been successful as evidenced by the development of prototype processes for the complete fabrication of motors from stock materials and the establishment of electrical/mechanical testing capabilities to evaluate piezo materials and completed motors. The project is also supported by work being conducted at the University of Missouri at Rolla for mechanical analysis and advanced development of piezoelectric ceramic materials and fabrication processes.

Appreciation of the immense possibilities for piezoelectric devices, coupled with the present down turn in defense related production, has allowed project efforts to focus on evaluating the potential of piezoelectric devices for commercial and industrial use. A wide range of diverse motor types and application areas has been identified, primarily in Japan, where the Japanese have been developing piezoelectric motor technology for many years and have more recently begun commercialization.

* Operated for the United States Department of Energy under Contract Number DE-AC04-76-DP00613

Commercial interest has spread to Europe and now appears to be gaining in the United States. The greatest potential in U.S. markets is for fractional horsepower intermittent duty devices. Application areas for this emerging technology include motors and actuators for 1) powering automobile windows, seats, sun roofs, windshield wipers, antennas, etc., 2) auto-focusing camera lenses, film winders and shutters, 3) optical lens and mirror positioning and adjustment, 4) computer disc drives, printer drivers and print heads, 5) watches and clocks, 6) space applications: adaptive space structures and gyroscopes, 7) fluid pumps, valves and fluid controls, 8) robotics: high torque/low weight finger and joint manipulation, 9) defense safing mechanisms, 10) powering devices in high magnetic fields, 11) medical devices and instruments and 12) many others.

In support of its primary mission as a prime contractor for the DOE, the Kansas City Division is continuing development of piezoelectric motors and actuators for defense-related applications while supporting and participating in the commercialization of piezoelectric devices with private industry through various technology transfer and cooperative development efforts.

TRAVELING WAVE PIEZOELECTRIC MOTORS

General

The first piezoelectric effect (conversion of mechanical or strain energy to electrical energy) was discovered by the Curie brothers (Pierre and Jacques) in 1880. It has been used extensively for a variety of transducers to sense pressure, impact, acceleration, etc., as well as other applications such as sonar, electric circuit oscillators, phonograph pick-ups, ignition systems and many others. The reverse effect, conversion of electrical energy to mechanical, was predicted by Lippmann in 1881 and verified by the Curies later that same year. This effect has also been exploited for application areas such as sonar, ultrasonic detection devices, microphones, speakers, smoke alarms, audio buzzers, and TV remote controls. More recently, intense interest and efforts have been focused on producing piezoelectric actuators and motors. These range from simple low frequency bending elements, stack (extension) actuators and inch-worm type linear and rotary motors to a number of unique concepts that exploit structural, resonant-frequency vibrations. The traveling wave motor is of the resonant frequency type and is one of a number of new concept piezoelectric actuators. The traveling wave motor is also frequently referred to in literature as an ultrasonic motor (USM).

Benefits

Traveling wave ultrasonic piezoelectric motors offer new design options and exhibit a number of advantages over conventional electromagnetic motors. Advantages have been demonstrated in output torque, response time, operation mode, motor noise and size. Benefits of traveling wave motors include the following:

- High Torque to Size Ratio
- High Torque @ Low Speed -- allows direct drive
- Reduced Inertia and Bounce
- Fast Response
- Precision Movement / Controllable
- Friction Locking with Power Off
- Flat Profile Enhances Packaging
- Non-Sparking
- Quiet
- Efficient
- Insensitive to Magnetic Fields

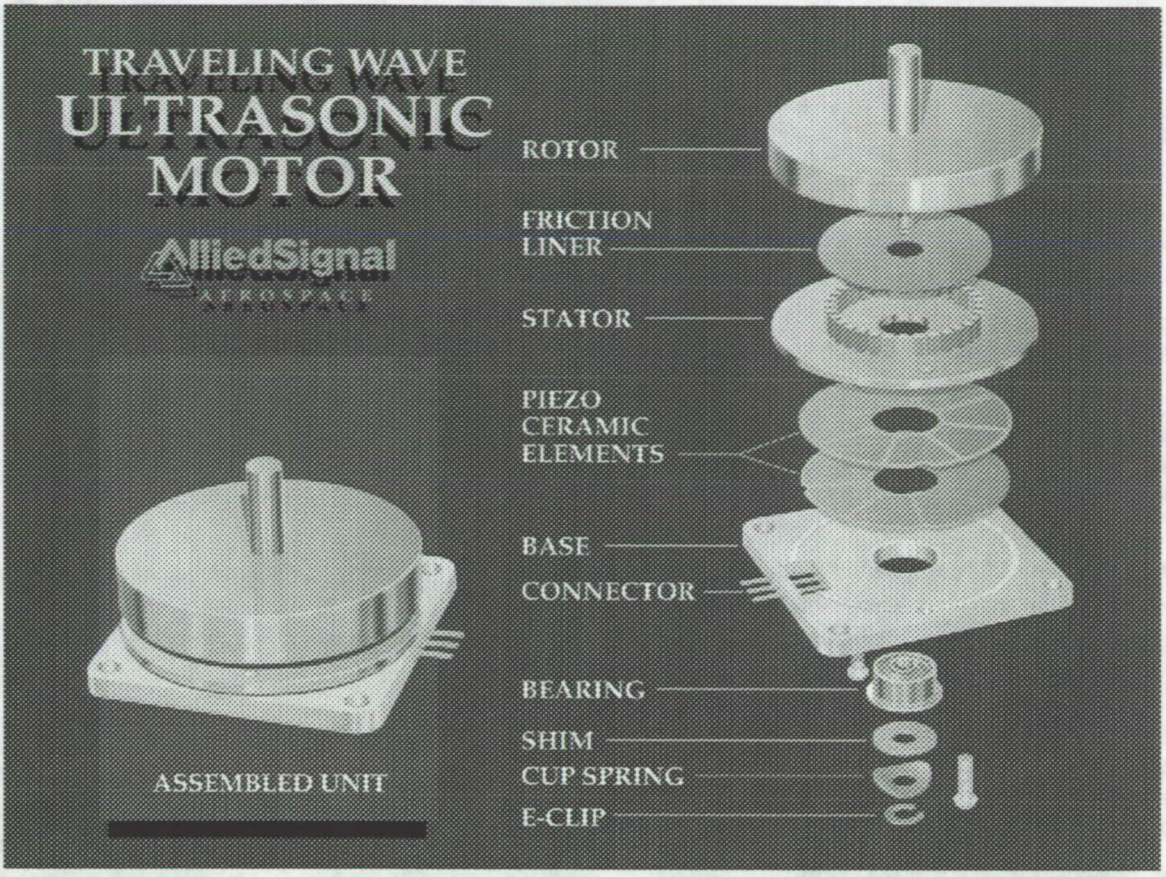


Figure 1. Assembled and exploded view of a traveling wave ultrasonic motor

Kansas City Division's Development

To date several prototype ultrasonic (traveling wave type) motors have been fabricated and tested. These include one linear motor and numerous rotary motors. The linear motor was utilized as an initial learning tool. The primary focus has been on the development and fabrication of rotor motors. Two sizes of rotary disc motors have been produced: 1.5-inch and 0.7-inch diameters. Torque and motion in each of these motors are produced by the generation of traveling waves with two piezoelectric ceramic disc elements. The most recent undertaking is the design and fabrication of a 0.315-inch diameter rotary ring motor. Torque and motion of this motor are produced by the generation of traveling waves with only one piezoelectric ceramic ring element.

Figure 1 shows an exploded view of an ultrasonic motor. Traveling rotational waves are created by the two thin piezoelectric ceramic elements and transmitted to the stator. These traveling waves result from the manner in which the two crystals are physically oriented and electrically driven. This phenomenon will be described in greater detail later. The two ceramic elements and the stator are rigidly bonded together and attached to the base to form a stationary unit. An electronic drive circuit (not shown) is designed to drive the stator at or near the desired mechanically resonant frequency. Operating near resonance is necessary to cause a relatively large movement of the tooth-like projections on the vibrating stator. A friction material liner is rigidly bonded to the rotor. The rotor is spring-loaded against the stator by cup washers on the rotor shaft and secured with an E-clip on the shaft end. As the resonant waves travel around the stator, the toothed projections execute an elliptic motion. As they move, these teeth push against the friction liner, resulting in rotary motion. Figure 2 illustrates the driving principle, showing a right traveling wave at two different instances in time.

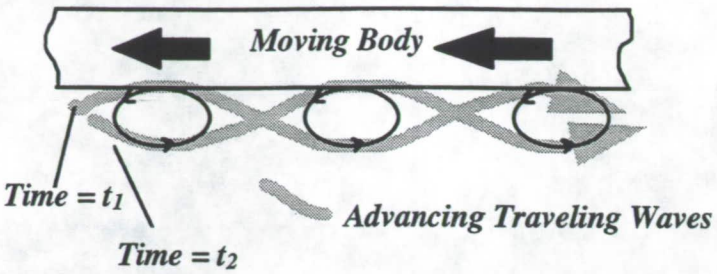


Figure 2. Traveling wave driving principle

As the wave moves through the vibrating body (stator), every point on the surface will execute an elliptic path as shown. Since the moving body (rotor) is pressed firmly against the stator, this elliptic motion transfers torque to the rotor, propelling it in a direction opposite to that of the traveling wave. Figure 3 is an (enormously exaggerated) illustration of the elliptic motion of one of the stator teeth.

The actual movement of the teeth is extremely small. The elliptic path for the 1.5-inch diameter rotary motor is estimated to be approximately circular, with a diameter less than 0.0005 inch. The top surface of each tooth traverses this small circular path at about 70 kilohertz or about once every 14 microseconds. Many people who have touched this vibrating ring-shaped surface report feeling a very slight vibration or a reduced friction as their finger moves across the surface. However, almost any small flat object placed on the circular ring of teeth will rotate.

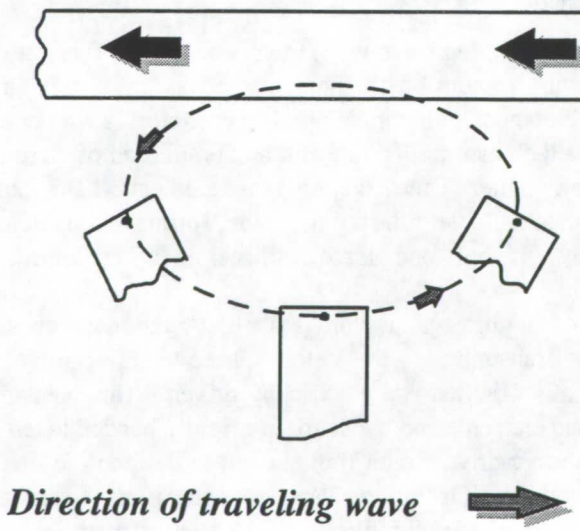


Figure 3. Illustration of elliptic motion of the stator teeth

Generating Waves and Elliptic Motion

Elliptic motion of the stator teeth is produced by resonant frequency traveling waves in the stator. The traveling waves are generated by a specific temporal and spatial combination of two standing waves. Each standing wave is generated by a segmented piezoelectric ceramic element. Although the actual generation of standing and traveling waves by circular piezoelectric ceramic disc elements is fairly complex, the principle in achieving them is best understood by the simplified example of a straight elastic linear body attached to piezoelectric ceramic bars.

Standing Waves

A high-frequency standing wave can be generated in a single piezoelectric bar. The bar is electrically segmented such that adjacent segments are polarized in opposite directions. The segmented piezo bar is attached to an elastic body. The elastic body represents the motor's stator. Once a DC voltage is applied, opposing sides of adjacent segments will contract while expansion occurs at each of the remaining segmented sides. The result is a flexing of the elastic body in the shape of a *(+)SINE* wave. When the polarity of the DC voltage is reversed, flexing of the elastic body in the opposite direction produces a *(-)SINE* wave. A standing wave is generated by the continuous flexing of the elastic body back and forth by the application of an AC signal where polarities are reversed at the signal frequency. A full wavelength is defined by two segments. So if there are four segments, there will be two standing waves that are in phase with each other at every instant of time. (Eight segments producing four wavelengths have been used in the Kansas City Division rotary disc motors.)

Traveling waves.

Standing waves can be combined to produce traveling waves. Two segmented piezoelectric bars, each driven by an AC signal of the same frequency, are arranged such that the segments of one are offset by one-half segment with respect to the other. Since two segments represent a full wavelength, this one-half segment offset causes the standing waves of one element to be translated one-quarter wavelength out of phase with the standing waves of the other. The AC signal driving one bar is a *SINE* wave while that driving the other bar is a *COSINE* wave. These driver signals therefore cause the waves to also be one-quarter wavelength out of phase in time. The combination of these two standing waves of the same frequency which are both spatially and temporally one-quarter wavelength out of phase results in a traveling wave.

This summation of two standing waves, one-quarter wavelength out of phase in both time and space, producing a traveling wave, is easily demonstrated mathematically. The standing wave is represented by Equation (1).

$$y_1 = a \sin(2\pi\ x/\lambda)\cos(\omega\ t) \quad (1)$$

where

$y_1 =$ the amplitude at any time, t,
$a =$ the peak amplitude,
$x =$ distance along the wave,
$\lambda =$ the wavelength and
$\omega =$ the circular frequency.

The first term (*sin*) is the spatial variation and the second term (*cos*) is the variation with time. A standing wave one-quarter wavelength out of phase spatially from Equation (1) is obtained by replacing the *sin* term with a *-cos*

term and a similar shift in time can be effected by replacing the second *cos* term with a *sin* term. Therefore, the second standing wave can be represented by Equation (2).

$$y_2 = -a\cos(2\pi\ x/\lambda)\sin(\omega t) \tag{2}$$

Adding Equations (1) and (2)

$$\begin{aligned} y = y_1 + y_2 &= a\sin(2\pi\ x/\lambda)\cos(\omega t) \\ &\quad -a\cos(2\pi\ x/\lambda)\sin(\omega t) \end{aligned} \tag{3}$$

The right side of (3) is recognized as a trigonometric identity, which reduces (3) to the following equation,

$$y = a\sin[(2\pi\ x/\lambda) - \omega t] \tag{4}$$

which represents a right traveling wave.

Elliptic Motion

The generation of elliptic motion from a traveling wave can be shown mathematically by analyzing the path of a point on a stator tooth as a wave passes through. For a small amplitude wave traveling through the stator, it is reasonably assumed that the tooth remains perpendicular to the wave at all times. This example is illustrated in Figure 4. Vector l represents a stator tooth. A point near the bottom of the tooth (P_0) is fixed at $x = 0$ as the wave moves to the right. Figure 4 depicts the wave and vector at one instant in time. At some other instant in time, point P_0 would remain on the y-axis but be at a different height and l would change slope. While constraining l to remain perpendicular to the wave, a description of the path of the opposite end of l, point P_1, is as follows.

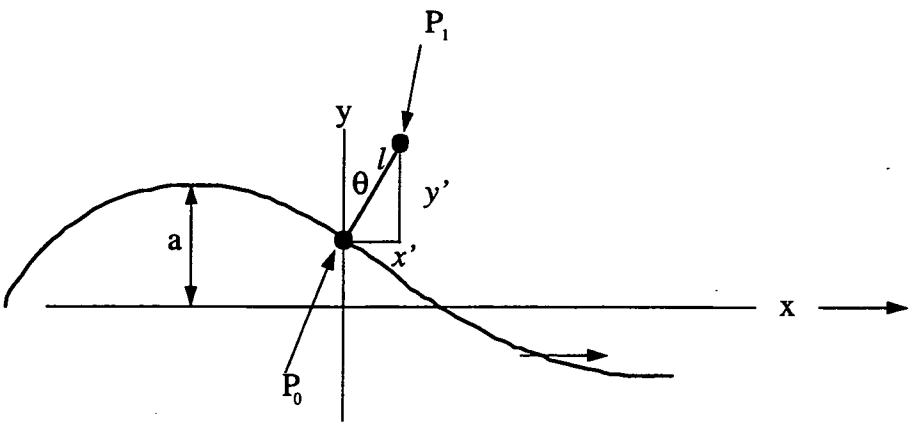


Figure 4. Elliptic Motion Analysis

Since Equation (4) represents the wave at any point in time, the slope can be obtained by differentiation.

$$\frac{dy}{dx} = \frac{2\pi a}{\lambda} \mathrm{Cos}\left(\frac{2\pi x}{\lambda} - \omega t\right) \tag{5}$$

The slope, *m*, normal to the wave at any position *x* is given by the negative reciprocal of Equation (5).

$$m = -\frac{\lambda}{2\pi a} \frac{1}{\mathrm{Cos}\left(\frac{2\pi x}{\lambda} - \omega t\right)} \tag{6}$$

Since we have placed the bottom of vector, *l*, at $x = 0$,

$$m = \frac{-\lambda}{2\pi a} \frac{1}{cos\,\omega t} \tag{7}$$

and

$$y = -a\,sin\,\omega t \tag{8}$$

Since *x* and *y* define a point on the wave corresponding to P_0, the coordinates of P_1 are given by

$$x' = x + l\,sin\,\theta \tag{9}$$

and since $x = 0$,

$$x' = l\,sin\,\theta \tag{10}$$

and

$$y' = y + l\,cos\,\theta \tag{11}$$

Substituting Equation (8) into (11) gives Equation (12).

$$y' = a\sin\omega t + l\cos\theta \tag{12}$$

Whereas the wave shown in Figure 4 is highly exaggerated for analysis, the actual amplitude is very small. The corresponding angle, θ, is estimated to be about 0.36 degrees. For small θ, $sin\theta \cong \theta$, $cos\theta \cong 1$ and $tan\theta \cong \theta$. Then Equations 10 and 12 reduce to

$$x' = l\theta \tag{13}$$

and

$$y' = l - a\,sin\omega t \tag{14}$$

From Figure 4, slope m is also given by

$$m = \frac{1}{tan\theta}$$

Equating this to Equation 7,

$$m = \frac{1}{tan\theta} = -\frac{\lambda}{2\pi a}\frac{1}{\cos\omega t}$$

$$tan\theta = -\frac{2\pi a}{\lambda} cos\omega t$$

and since $tan\theta = \theta$ for small angles,

$$\theta = -\frac{2\pi a}{\lambda} cos\omega t \tag{15}$$

Substituting Equation 15 into Equation 13 gives

$$x' = \frac{2\pi l a}{\lambda} cos\,\omega t \tag{16}$$

If we translate our origin to a new coordinate system (to center the excursions of P_1) where

$$x'' = x' \text{ and } y'' = y' - 1$$

then

$$x'' = x' = \frac{-2\pi l a}{\lambda} \cos\omega t \tag{17}$$

and

$$y'' = y' - l = -a\,sin\,\omega t \tag{18}$$

Substituting Equations 17 and 18 into the trigonometric identity,

$$\sin^2\theta + \cos^2\theta = 1$$

gives

$$\left(\frac{y''}{a}\right)^2 + \left(\frac{x''}{2\pi\, l a/\lambda}\right)^2 = 1 \tag{19}$$

which is clearly the equation of an ellipse centered at $x = 0$ and $y = 1$ with major diameter = $2a$ and minor diameter = $4\pi l a/\lambda$ as illustrated in Figure 5.

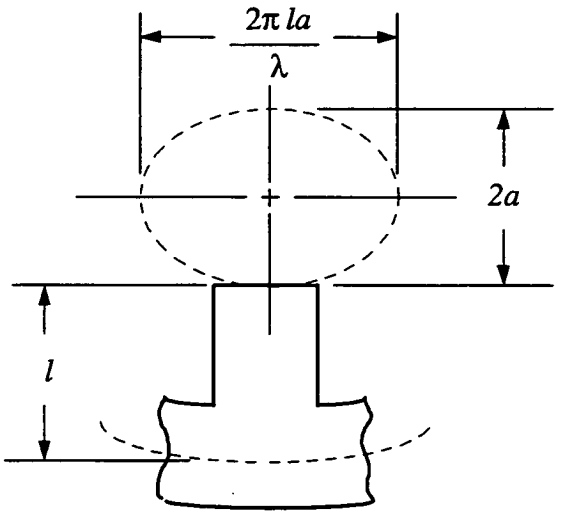


Figure 5. Stator tooth at the lowest point of its elliptic path

Kansas City Division's Continued Development and Commercialization

The Kansas City Division has made great strides in understanding the principles and applications of piezoelectric motors. A full compliment of prototype processes has been developed for complete fabrication of motors from stock materials, including abrasive machining of piezoelectric ceramics and more traditional machining of other motor components, electrode plating and sputtering, electric poling, cleaning, bonding and assembly. Drive circuits have been manufactured and motor controls are being developed. Finite element analysis (FEA) capabilities have been utilized to optimize motor designs. Laboratory facilities have been established for electrical and mechanical testing and evaluation of piezoelectric materials and completed motors.

The Kansas City Division is continuing development of piezoelectric motors for traditional defense-related applications while supporting the commercialization of piezoelectric motors with private industry through various technology transfer and cooperative development efforts. The most recent effort is a technology commercialization initiative with AMMPEC, Inc. (Advance Materials & Manufacturing Processes for Economic Competitiveness). This multiphase initiative is intended to stimulate growth of small, regional businesses by enhancing technologies and defining customers for the ultimate creation of new jobs. The experience gained with piezoelectric motors, coupled with the vast manufacturing knowledge and resources of the Kansas City Division, establishes the solid foundation necessary to step forward with the production of piezoelectric motors. A commitment to piezoelectric material advancement, motor designs optimization, along with manufacturing process improvements are all necessary to ensure piezoelectric motors realize their potential within the United States.

BIBLIOGRAPHY

Ikeda, Takuro. *Fundamentals of Piezoelectricity.* Oxford: Oxford University Press, 1990

Jaffe , Cook, and Jaffe. *Piezoelectric Ceramics.* London: Academic Press Inc. Ltd., 1971

Panasonic Ultrasonic Motor Technical Reference Manual. Electric Motor Division, Matsushita Electrical Industrial Co., Ltd.

Program Manual for Piezo Design Aid. Piezoelectric Products, Inc., 1985

Sashida, Toshiika and Kenjo, Takashi. *An Introduction to Ultrasonic Motors.* Oxford: Oxford University Press, 1993

Uchino, Kenji. "Electrostrictive Actuators: Materials and Applications", *Ceramic Bulletin,* Vol. 65, No. 4 (1986), pp. 647-652

Ueha, S. and Tomikawa, Y. *Ultrasonic Motors: Theory and Applications.* Oxford: Oxford University Press, 1993

IMHEX® MOLTEN CARBONATE FUEL CELLS

T.G. Benjamin
R.R. Woods
R.M. Laurens
V.J. Petraglia

M-C Power Corporation
8040 South Madison Street
Burr Ridge, IL 60521

ABSTRACT

M-C Power Corporation (MCP) was formed in 1987 to commercialize molten carbonate fuel cell (MCFC) stacks based on the electrochemical technology and *I*nternally *M*anifolded *H*eat *EX*changer (IMHEX®) stack concept developed by the Institute of Gas Technology. Since M-C Power's formation, the primary activities have been to scale-up this technology to commercial-size, 1 m^2 active components and to develop and demonstrate commercially viable manufacturing processes. Today, M-C Power's facilities consist of over 8,000 m^2 of floor space dedicated to the manufacture and testing of commercial-scale MCFC components. Our developmental manufacturing capabilities could produce approximately 3 MW of MCFC components annually if operated continuously on a three shift basis.

M-C Power and its commercialization team, Bechtel Corporation, Stewart & Stevenson Services, Inc. and the Institute of Gas Technology (IGT), are conducting a coordinated program to develop and deliver market responsive power plants by the end of the decade. A major component of the IMHEX® Commercialization Program includes design, fabrication, and operation of demonstration power plants. The first of these power plants is a 250-kW demonstration in late 1994 at the Unocal Corporation's Fred L. Hartley Research Center in Brea, California. M-C Power's second demonstration power plant is a 250-kW integrated unit to be operated by San Diego Gas & Electric Company in mid 1995.

This paper will present results of short, commercial-area stack tests, status of the 250-kW demonstrations, and the implementation of continuous manufacturing for cell components. Over the past two years, significant technical progress has been made. Three commercial-area stacks were successfully tested; a 250-kW stack acceptance test facility was designed, fabricated, installed, and commissioned; the world's largest single MCFC stack (250-kW class) has been assembled and is ready for conditioning and testing; continuous manufacturing has been achieved; active component manufacturing capacity has been doubled and component yields are near the targets set to meet commercialization goals.

Fuel Cells

A fuel cell is an electrochemical power generation device which directly (electrochemically) converts the chemical energy of a fuel into electrical energy without an intermediate combustion step. Therefore, fuel cells are not subject to the Carnot limitations and fuel cells can achieve electrical efficiencies as high as 60% based on higher heating value. Like a battery, a fuel cell consists of two electrodes (anode and cathode) separated by an electrolyte. Electrochemical oxidation takes place at the anode and electrochemical reduction takes place at the cathode. The electricity is carried ionically through the electrolyte from one electrode to the other. The circuit is completed when the electrons are directed to the external circuit as shown in Figure 1 for the molten carbonate fuel cell.

The electrolyte determines the fuel cell type and its operating temperature as shown in Table 1. The overall reaction for all types of fuel cells is the oxidation of hydrogen: $H_2 + 1/2\ O_2 \rightarrow H_2O$ + electricity + heat.

Electrolyte	Temp (°C)	Cell Reactions	
		Anode	Cathode
Aq. KOH	60-120	$H_2+2OH^- \rightarrow H_2O+2e^-$	$H_2O+\frac{1}{2}O_2+2e^- \rightarrow 2OH^-$
Sulfonic Acid	30-50	$H_2 \rightarrow 2H^++2e^-$	$\frac{1}{2}O_2+2H^++2e^- \rightarrow H_2O$
Phosphoric Acid	200-400	$H_2 \rightarrow 2H^++2e^-$	$\frac{1}{2}O_2+2H^++2e^- \rightarrow H_2O$
(Li,K,Na) CO_3	600-700	$H_2+CO_3^= \rightarrow H_2O+CO_2+2e^-$	$CO_2+\frac{1}{2}O_2+2e^- \rightarrow CO_3^=$
Stabilized Zirconia	850-1,000	$H_2+O^= \rightarrow H_2O+2e^-$	$\frac{1}{2}O_2+2e^- \rightarrow O^=$

Table 1
Types of Fuel Cells

M-C Power's molten carbonate fuel cell technology uses a mixture of carbonate salts that are molten at the 650°C operating temperature. The molten salt provides the ionic conductivity required for the process as shown in Figure 1. The electrolyte is distributed between the anode, cathode, and matrix by capillary forces. In the MCFC, the anode reaction takes place at the 3-phase interface between the catalytic Ni-Cr anode, the gas phase, and the liquid electrolyte. The $CO_3^=$ ions consumed at the anode are conducted through the carbonate electrolyte from the NiO cathode where they were generated. The electrolyte is held in place by a porous matrix of lithium aluminate. The electrons generated at the anode are collected by a metal current collector and conducted through the metal separator plate to the cathode of the adjacent cell where they react with oxygen and carbon dioxide to form carbonate ions. Connections to the external load are made at the ends of the stack. Figure 2 is a schematic representation of an IMHEX® cell.

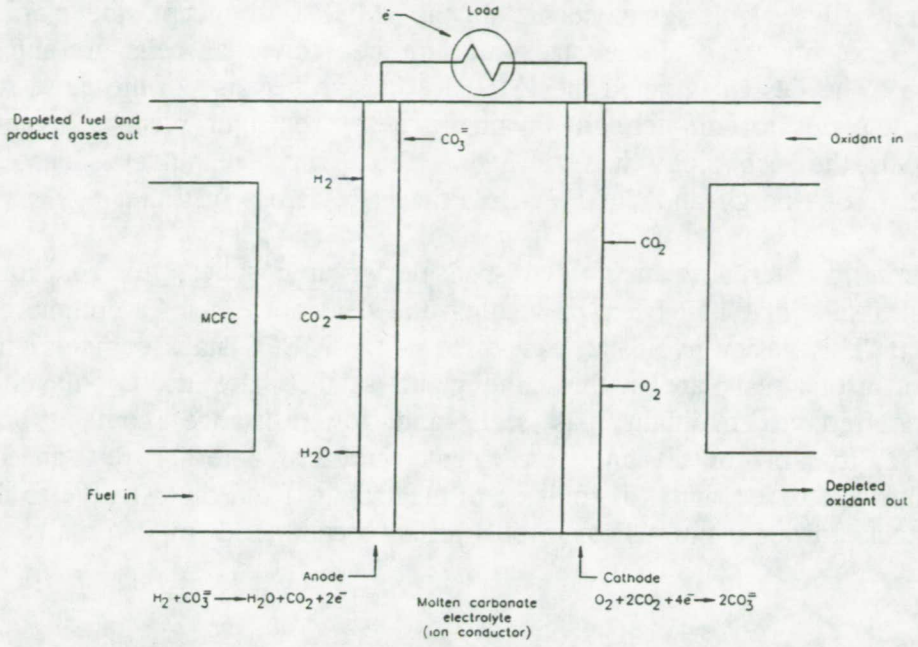


Figure 1
Schematic Representation of a Molten Carbonate Fuel Cell

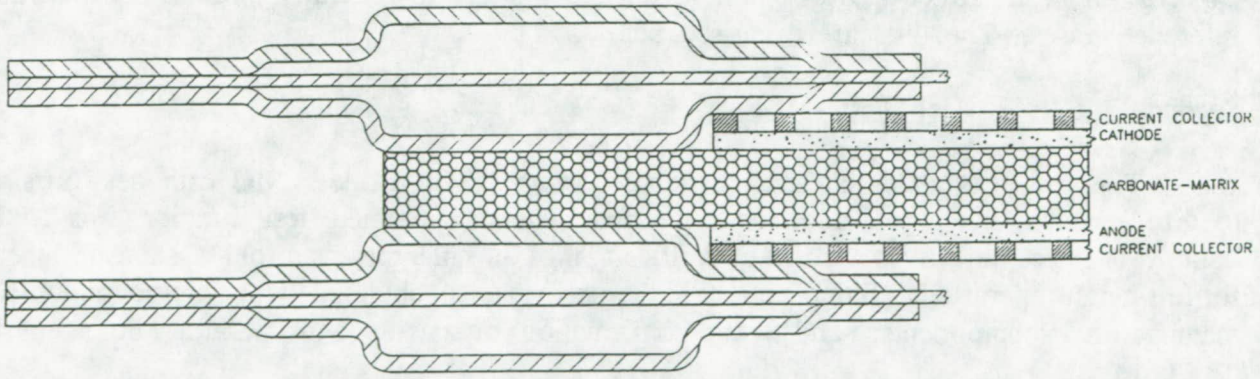


Figure 2
Schematic Cross-Section of IMHEX® Cell Package

M-C Power's MCFC technology is based on the internally manifolded heat exchanger (IMHEX®) fuel cell stack design concept. In the IMHEX® concept, the manifolds which distribute the reactant gases to the stack and to the individual cells are integral to the separator plate. The design is called the IMHEX® design because of the close resemblance of the separator plates to commercially available heat exchanger plates. Invented by IGT during the 1980s, the technology was verified at the $0.1 m^2$ level under sponsorship of the U.S. Department of Energy and the Electric Power Research Institute.

Compared to externally manifolded stack design approaches, this concept eliminates many technical issues that had been preventing the development of a commercially-viable MCFC product. This design avoids at least three of the MCFC stack engineering issues: (1) because the manifold seals are in the same plane as the plate itself, only unidirectional forces are required to maintain gas seals and low-resistance contact between cell components; (2) loss of contact and consequent increased internal resistance caused by shrinkage of the electrodes and; (3) transfer of electrolyte from the positive to the negative end of the stack through porous manifold gasket seals used in externally manifolded configurations.

Stack Development

Three commercial-area, short stacks (nominal 20-kW) were tested at the M-C Power 20-kW Test Facility during the past two years. The primary purpose of these tests was the resolution of commercialization design issues. As such, these tests were of short duration (1,000 to 2,000 hours) and used state-of-the-art active components and hardware. Development stacks at M-C Power are usually initially operated on reference gases to establish baseline performance comparisons. After the baseline performance has been established, systems gases are introduced to determine the stack performance on the gases expected in an integrated power plant. The system gases are generally more dilute than the reference gases. The tests are discussed below.

Second Full Area Stack Test (MCP-4)

This test was a 20-cell stack assembled with Ni-Cr anodes, NiO cathodes (in-situ oxidized), and conventional electrolyte matrix. The programmed test duration was 1,000 hours. The test objectives were: 1) verification of an enlarged manifold system designed for full-height operation; 2) duplication of the performance obtained with semi-continuously manufactured components; and 3) the confirmation of start-up and operating procedures.

The test operated for 1,560 hours (777 hours of operation on reference gases and 643 hours on dry system gases). Maximum power output was 23 kW on reference gases and 21 kW on system gases. Start-up procedures developed during operation of the first full-area stack (MCP-3) were successfully duplicated. The electrochemical performance duplicated that of the first commercial-area stack and bench-scale cells. The stack achieved more than 1 kW per cell during operation with system gases at atmospheric pressure and confirmed model predictions. Pressure drop measurements confirmed the validity of the empirical

pressure drop model. The gas analysis and stack balances, however, indicated the presence of an outboard leak which was confirmed during post-test teardown and inspection.

Third Full Area Stack Test (MCP-5)

This test was also a 20-cell stack, using components similar to Stack MCP-4. The test was scheduled for a duration of 1,000 hours. Test objectives were: 1) the verification of electrochemical performance on wet system gases; 2) data acquisition on alternative fuel flow configurations; and 3) the validation of cell performance after hot hold conditions.

The test operated for 500 hours (400 hours on reference gases and 100 hours on wet system gases). After electrolyte distribution and during pre-test stack sealing evaluation, anomalous pressure drop measurements were observed. Diagnostic testing indicated that a feed valve was closed rather than open preventing adequate flow on the anode side of the cell. Further, gas measurements indicated the existence of a small outboard leak. Because of the leak, an immediate analysis of the stack axial clamping force was initiated, using a stack structural model. Operation on reference and wet system gases was started to establish stack baseline performance. As expected, performance was substandard because of the valving error. After 200 hours of operation, the stack was shut down for repairs.

Two stack modifications were made. The first was to change the stack clamping system for better control of the axial load. The stack structural analysis revealed that the seal force, at areas where outboard seals were compromised, was substantially lower with the original configuration. Implementation of the new system provided uniform seal pressure. The second change was the development of a method to repair outboard seals using a matrix/carbonate cement. The stack was re-started and the stack performance duplicated the operation prior to the thermal cycle on reference and wet system gases. This result is important because it confirmed the resilience of an IMHEX® MCFC stack after thermocycle and mechanical repair. Gas analysis and stack mass balance also confirmed the integrity of the seal repair. Upon confirmation of the changes, the test was terminated. Post-test inspection showed the seal repair to be sound.

Fourth Full Area Stack Test (MCP-6)

This test was also a 20-cell stack. The components were similar to those in the earlier stacks, except that they were manufactured by continuous processes. The objectives of the test were: 1) validation of a new cell assembly approach; 2) confirmation of cell performance using wet system gases; 3) validation of the hot hold conditions; and 4) confirmation of the axial sealing ability of the new clamping system.

Prior to stack assembly, extensive testing was conducted, using pressure sensitive paper, to qualitatively evaluate the new cell assembly approach. In the new approach, the cell is assembled with the active components set a small, predetermined distance below the seal plane. Therefore, unless good force distribution is obtained, poor electrical contact may result. Good sealing was obtained and the active area force distribution was adequate.

A total of 1071 hours of operation was logged (533 hours on reference gases and 538 hours on wet system gases). Maximum power output was 23 kW on reference gases and 21 kW on wet system gases. Baseline crossover and stack mass balances indicated that good outboard seal efficiency was achieved with the new assembly approach. An issue that required validation during stack operation was that, in fact, cell internal resistances (IRs) were not compromised with the new assembly philosophy. Shown in Figure 3 is a comparison of the IRs for Stacks MCP-4 and MCP-6. Improvement rather than deterioration was achieved.

As mentioned above, all active components (anodes, cathodes, and electrolyte/matrix) were made by continuous manufacturing processes. Figure 4 shows polarization data for the three full-area tests; modest gains in stack voltages were obtained as manufacturing experience improved. Thus, M-C Power has been able to achieve continuous manufacturing of MCFC components without loss in stack performance. In fact, the voltages are identical to the electrochemical model predictions and well-controlled bench-scale tests.

Acceptance Test Facility

To ensure that the quality of the full-height, full-area stacks (250 kW) meets specifications, M-C Power designed, fabricated, installed, and commissioned an Acceptance Test Facility (ATF) at its facility in Burr Ridge, Illinois. The ATF is a skid-mounted unit for stack conditioning which includes the following steps: 1) binder removal from tape cast components; 2) electrolyte melting and distribution; 3) cathode oxidation; and 4) open circuit and part load (400 Amps) testing. The ATF was designed to meet power plant code criteria and is fully instrumented. A PLC-based system provides data acquisition and control. The unit is ready for operation.

Commercial Stack Assembly and Testing

M-C Power's immediate plans include two 250-kW demonstrations and a 1-MW demonstration. M-C Power will operate a partially integrated, grid-connected 250-kW powerplant in late 1994 at the Unocal Fred L. Hartley Research Center in Brea, California. An integrated and packaged 250-kW power plant demonstration is scheduled for mid-1995 at the Miramar Naval Air Station in San Diego, California. The 1-MW demonstration will be a market-entry prototype at Southern California Edison Company.

M-C Power has completed the assembly of the Unocal 250-kW stack. Figure 5 shows the assembled stack. In general, all of the hardware used was validated during the short-height, full-area tests described above. In addition, extensive finite element modeling was used to ensure that each component meets strict transportation, seismic, and operating criteria. The stack has been installed in the ATF and startup is imminent.

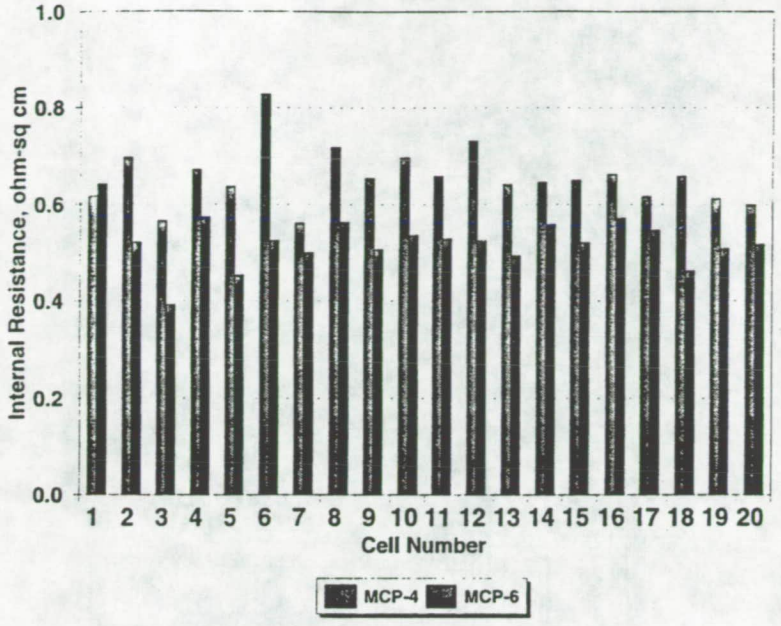


Figure 3
Comparison of Cell Internal Resistances (Stack MCP-4 vs Stack MCP-6)

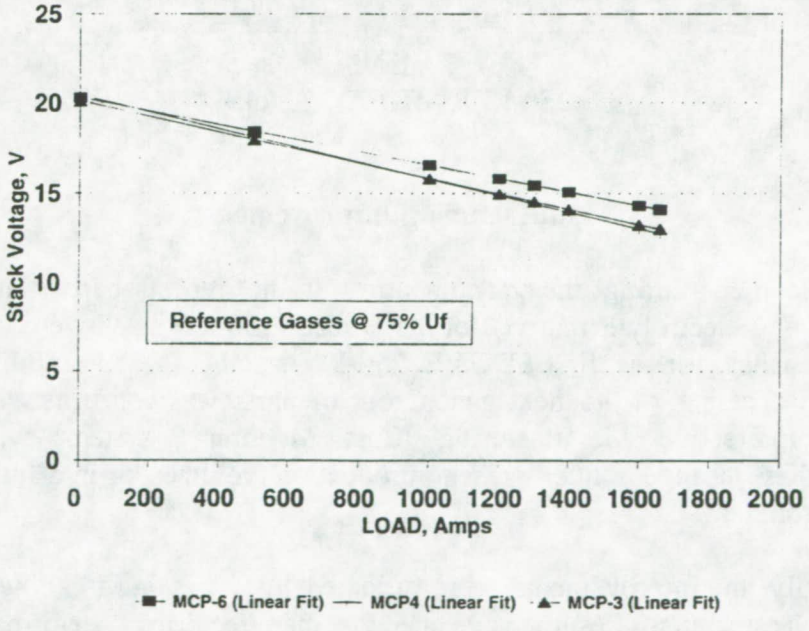


Figure 4
Stack Performance vs Manufacturing Approach

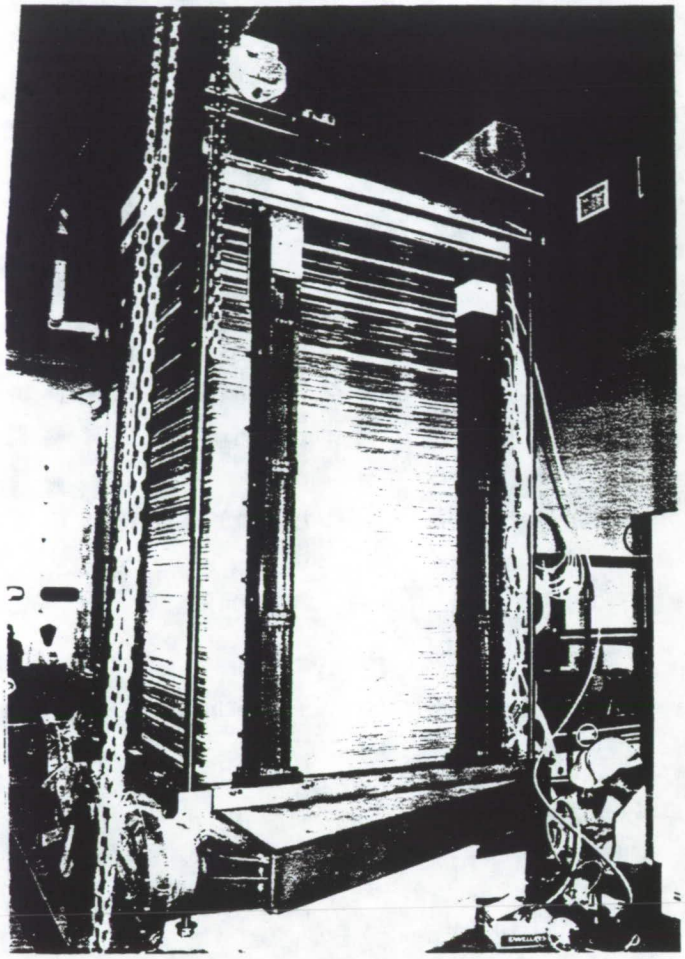


Figure 5
250-kW IMHEX® Stack

Manufacturing Improvements

Lessons learned during the manufacture of the repeat parts (anodes, cathodes, separator plates, and electrolyte matrix) for the Unocal 250-kW stack were incorporated into fabrication approaches for the first SDG&E 250-kW stack. The advisability of changes to the design and processing of the next generation of parts was confirmed by analysis using M-C Power's "Interactive Manufacturing Cost Modeling" system, which allows the examination of these changes under present production volumes or in a future high-volume production environment.

Additionally, the improvements were validated by testing in M-C Power's Bench-Scale Quality Control Test Units. Results have met the manufacturing performance projections for increased through-put. Overall electrode manufacturing through-put has improved 188% (see Table 2). The separator plate design was altered to decrease the number of pieces required to produce a plate assembly. Total plate details were reduced from 13 to three. Separator plate assembly costs were reduced by 25% (see Table 3).

Electrode Fabrication	Unocal	San Diego #1
Relative Casting Speed	1	1.4
Relative Casting Yield	1	1.24
Relative Sintering Speed	1	1.3
Relative Sintering Yield	1	1.28

Table 2
IMHEX® Electrode Manufacturing Improvement

Separator Fabrication	Unocal	San Diego #1
Number of Pieces	13	3
Relative Cost	1	0.75
Relative Pressing & Welding Time	1	0.72

Table 3
IMHEX® Plate Manufacturing Improvement

MAGNETIC BEARINGS AT DRAPER LABORATORY

Anthony S. Kondoleon
William P. Kelleher
Peter D. Possel

C.S. Draper Laboratory
Cambrige, Ma. 02139

ABSTRACT

Magnetic bearings, unlike traditional mechanical bearings, consist of a series of components when mated together form a stabilized system. The correct design of the actuator and sensor will provide a cost effective device with low power requirements. The proper choice of a control system utilizes the variables necessary to control the system in an efficient manner. The specific application will determine the optimum design of the magnetic bearing system including the touch down bearing. Draper for the past 30 years has been a leader in all these fields. This paper summaries the results carried out at Draper in the field of magnetic bearing development. A 3-D radial magnetic bearing is detailed in this paper. Data obtained from recently completed projects using this design are included in the paper. One project was for high radial load (1000 pound) application. A second was for a high speed (35,000 RPM), low loss flywheel application. The development of a low loss axial magnetic bearing is also included in this paper.

INTRODUCTION

The present state-of-the-art magnetic bearings are a result of continuing development for industrial applications. Operational units are emerging for large pipeline and refinery gas compressors where safety, reliability and low maintenance are a high priority. Other commercial applications include high speed spindles for precision machining, sealed industrial pumps for safety and environmental concerns and turbomolecular vacuum pumps for elimination of oil systems. Projects in the electric utility industry show they may soon be used in feed pumps, fans, generators and turbines. A large number of prototype systems have been developed to meet these specialized needs.

Many side benefits have been recognized after placing a magnetic bearing system in operation. One such benefit is in equipment monitoring and diagnostics from both on-site and remote locations. Failure prediction and analysis are the by-products from these systems. The feedback data from field equipment has resulted in improved new equipment design. **A close relationship between equipment maker, magnetic bearing supplier and user is required to develop the full potential of this technology.**

The present commercially available magnetic bearing systems are large compared to the bearings they replace. This has led to their use in equipment where size and weight are not of primary importance. Momentum is building for commercial use, but the high initial cost has limited magnetic bearings to niches where there is significant cost/benefit improvement. The keys to greatly expanded use of this technology is a reduction in size, weight, power and cost of the total system.

Draper has been developing magnetic bearings for over 30 years, initially for inertial instruments suspensions and then for other military and space applications. The emphasis in recent years at Draper has been the investigation and proto-typing of magnetic bearing systems for main shaft support in gas turbine engines. Commercial magnetic bearings were unsuitable for this application because of their excessive size and weight, the limitations on operating temperature and speed range and the high power requirements and inefficient and expensive controllers.

Draper has addressed all of these issues. **We have made significant advances in digital controllers, reduced the size, weight and power requirements of the actuator and have reduced the eddy current losses in the actuator which limits the rotational speed of the system.** Bearing actuator designs are simple and control methodologies have been developed and simplified using low cost mass produced electronics. Patent applications have been filed for actuators designs and control techniques.

BACKGROUND

Initial work at Draper on magnetic suspensions was applied to floated inertial instruments. The first of these suspensions was an active suspension developed in the early fifties. In 1953 it was discovered that passive suspensions using tuned AC fields were possible and the technology was applied to floated instrument designs. These passive suspensions have been a mainstay in all precision inertial instruments for the last 30 years [1]. In parallel, work has continued at Draper on active magnetic suspensions, emphasizing instrument applications. In 1975, Draper began to apply this technology to the flywheel energy storage problem (low loss bearings) under a grant from the National Science Foundation (NSF). Since 1975, both analytical and hardware development activities have been continued toward developing electromagnetic actuators and associated controls with emphasis on a total magnetic bearing system.

Five-Axis Active Flywheel Bearing

This project in the mid 1970's resulted in a system which magnetically suspended a twelve pound flywheel at operating speeds to several thousand rpm's. The magnetic bearing system actively controlled five degrees of freedom (three translational and two lateral rotational axes) during operation. The actuators chosen for this system were permanent magnet biased electromagnets. Analog circuitry provided the control function.

Electromagnetic Isolator

This program in the late 1970's produced two major accomplishments. The first was to understand the performance limits of a system based on electromagnetic isolation (e.g., sensors, modeling, actuators, power). The second was to examine actuator performance in a single degree of freedom system. The single degree of freedom system allowed experiments to be run more directly than with the five-axis system. This system suspended a ten pound mass using pure electromagnetics. Both analog and digital controls were evaluated with this unit.

Flywheel Energy Systems

The project called CARES was organized in the early 1980's for the development of a system that simultaneously stored energy in a flywheel and was capable of transferring the associated stored momentum for attitude control. In addition to developing a large angle magnetic suspension (LAMS), special motor-generator, power electronics, flywheel, and controls were developed. The LAMS was based on a Lorentz force actuator. A hybrid analog/digital controller actively controlled five degrees of freedom. Draper also conducted extensive work on rotor dynamics on this project. Emphasis on the ability of magnetic bearings to change their dynamic performance in

providing forces to the rotor [2] was examined. Benefits from this approach include: suppression of whirl-mode instabilities, reduction of synchronous (mass imbalance) vibrations and eliminating gyroscopic torque couplings.

Slew Actuator-Isolator Development

This 1987 program was to determine the feasibility of using a large angle magnetic suspension system to meet the high torque, low noise requirements for a retarget and slew application. Hardware modification to the CARES system was done to achieve a demonstration of the torquing and the vibration isolation capabilities of a magnetic bearing system. A result from this program was the development of a permanent magnet axial bearing actuator.

Jet Engine Magnetic Bearing Effort

In 1987 a joint effort with MIT's Gas Turbine Laboratory and Draper was formed to investigate the feasibility of using magnetic bearings for main shaft support in gas turbine engines. Included here was an assessment of then state-of-the-art magnetic bearing technology to quantify load carrying capability and maximum operating temperatures. A bearing compartment design was developed using a new Draper actuator configuration for a proposed gas turbine.

Magnetic Bearing Test Bed

This project was started in 1988 with the objective to develop magnetic bearing technology to larger rotating machinery, as a means to control and minimize transmitted vibration. The initial design of the test bed was to demonstrate and quantify vibration reduction for a 150 lb., 5 foot rotor system. A novel approach to centralized control design methodology, involving de-coupling algorithms were developed. This Adapted Force Balancing (AFB) algorithm demonstrated significant vibration reduction as well as showing vibration reduction potential. A low power radial actuator was developed and constructed with an integrated sensor assembly.

MAGNETIC BEARING SYSTEM

A magnetic bearing is comprised of a number of components when mated together form a suspension system. The major components are the Magnetic Bearing Actuator, the Magnetic Bearing Control and Sensor System and the Touchdown Bearing. The difficulty in the design of a total system is the determination of where and when to use magnetic suspensions. **Figure 1** shows a graph of rotor speed vs. rotor support weight. The figure shows low speeds and small shaft diameters are an area where mechanical bearings are better suited for support than a magnetic bearing system. As the speeds increase, the load the bearing can support decreases because of the internal loads generated in the bearings. To increase the load carrying capacity, larger bearings are required which in turn lowers the speed limit the of the shaft. At a point, mechanical bearings become unsuitable and magnetic bearings become more attractive. Table I list some area comparisons and the support system suitable for these areas.

DRAPER RADIAL MAGNETIC BEARING ACTUATOR (FOUR POLE-3-D)

Present commercial radial magnetic bearing actuators use motor-like configurations having two dimensional flux paths with 8 or 12 poles. These designs have two major drawbacks. The first is their high reluctance flux paths which require high input power to the control coils during operation. The second is the large number of poles create high frequency flux cycling in the rotor at rotational speed producing high eddy current loss and heating of the rotor. Both of these conditions limit the rotational speed and efficiency of the bearing system. **Draper has developed and built a radial**

actuator which produces high force at low power and with low rotational losses at high operating speed.

Figures 2 shows a magnetic bearing we developed for a 1000 pound radial load. It is a four pole device with endplates providing a continuous 360° return path for three dimensional flux flow between poles, rotor and end plates. The end plate air gap area is large compared to the main pole area to provide a low reluctance flux path. Excited control coils produce attractive forces at the pole gaps in two mutually perpendicular axes. The power is low because of the low reluctance magnetic circuit in the actuator. The reduced pole count, reduces the frequency of rotor currents and operating all four poles at the same magnetic polarity (Homo-Polar) reduces the cyclical flux density peak-to-peak amplitude in the rotating rotor. Both conditions reduce eddy current heating. The reduction of eddy currents in the rotor is an increase in efficiency of the bearing.

Magnetic bearings are operated in the biased mode to linearize the control function. This operation is the condition where the pole gaps contain one half the flux density that is required at the maximum operating condition. At no load, with all poles attracting the rotor with equal force, a state of equilibrium exists. Biasing may be achieved by either the use of permanent magnets (PM) in the magnetic circuit or by current in coils (EM). Each of these methods have advantages. With PM biasing, the no load power requirements will be limited to the power loss due to eddy currents in the magnetic circuit. This is a desired condition where input power is at a premium, like applications such as space based flywheels or range extender systems. EM biased designs are attractive where the static forces are high or are desired to be adjustable during operation of the actuator.

An inside-out radial magnetic bearing actuator was developed at Draper in 1993 based on a PM biased design. This actuator was to be a part of a range extender flywheel system which was to operate up to 35,000 rpms. **Figure 3** shows the rotor and stator components of the actuator. **Figure 4** shows a plot of the control current vs. pole force for this design.

AXIAL MAGNETIC BEARING

A double acting axial magnetic bearing actuator was developed in 1993 as part of a range extender flywheel system. The performance goal of the axial bearing was to support a static flywheel weight of 53 Lbs. under dynamic loads of ± 120 Lbs. The design was satisfied with the unique actuator shown in **Figure 5**. Static load capacity was achieved using a P.M. biased design provided by high energy neodymium iron boron magnets supported by a structure made from titanium. The low loss actuator design contributed towards an overall high system bandwidth. **Figure 6** shows the plot of control current vs. axial force for one of the delivered units.

MAGNETIC BEARING CONTROL SYSTEM

Advances in the design and implementation of the magnetic bearing control systems are required if magnetic bearings are to be accepted for wide spread use in industry. Magnetic bearing control system development at Draper have focused on a variety of topics. One area is in the field of systems control is optimal control system design. There are many specific design techniques which can be categorized as optimal, We limit the technology here to optimal multivariable continuous time control systems. The design premise of these controllers is access to the system state variables which are applied in the feedback loop through a gain matrix to optimize a pre-defined performance

parameter. There are enhancements of the classical optimal controller which include optimal observers to estimate state variables when they are not directly available for measurement. One technique is Linear Quadratic Regulator with Loop Transfer Recovery (LQG/LTR). Both LQG/LTR and classical optimal control design were applied to magnetic bearing systems at Draper in 1988 [3, 4]. This work provided the background for more advanced multivariable design techniques, such as H^a, which combine time domain optimization with frequency domain loop shaping to produce a robust control system.

For a quality magnetic bearing system to maintain integrity throughout the operating envelope, a high performance control system is required. For a control system, high performance implies a high operating bandwidth. **Analysis carried out in 1990, [5], showed a magnetic bearing control system operated in the "decentralized" configuration is subject to bandwidth limitations which would be unacceptable for a moving platform system. The analysis further showed that a "centralized" control produced a system with significant bandwidth and hence performance improvements over "decentralized" controllers.** LQG/LTR controllers described above are of the centralized type. An alternative "centralized" control technique which de-coupled the translational and rotational dynamics of the Jeffcott model of a flexible rotor has also been developed at Draper [6].

A major sub-topic of control system design deals with synchronous vibration compensation, referred to as auto balancing. There are traditional techniques such as notch filters which can be applied under certain conditions which effectively eliminate synchronous vibration due to mass unbalance. However, with a wide bandwidth control system, the conditions under which notch filtering techniques can be applied are limited. **A Draper patented technique called "Adaptive Forced Balancing" (AFB) effectively compensates for synchronous vibrations over the entire control system bandwidth without producing any de stabilizing effects.** The adaptive algorithm simultaneously tracks and compensated mass unbalance amplitude, phase and frequency forcing a rotating shaft to spin about its center of mass [7].

In applications where high reliability is essential, such as an aircraft engines, fault tolerant controllers are necessary. Draper has been a leader in this field for over twenty years. In 1990 a prototype single board fault tolerant controller, based on Transputer Technology was designed and built. Table II shows the comparison of this Advanced Uniprocessor to the Triplex BiProcessor Technology used in today's Full Authority Digital Engine Controllers (FADEC).

The Draper designed magnetic bearing actuator has been successfully operated on a rotor dynamics test rig. The controller designed to stabilize the system was a digital control system based on a Texas Instruments TMS320C30 floating point digital signal processor (DSP). A commercial DSP board was used to implement the real-time control laws, at sampling rates up to 10 kHz. **The adaptive forced balancing algorithm was implemented in the real-time software and successfully operated on the Draper test bed. Control system bandwidths up to 100 Hz were achieved.**

TOUCHDOWN BEARINGS

Draper has been a leader in the design and development of precision bearing assemblies since the 1950's. This includes ball bearings, gas bearings and journal bearings. At this time, over one million space based hours have been accumulated on Draper built ball bearing assemblies without a failure. Hemispherical gas bearings

assemblies designed and built by Draper into precision inertial instruments have an estimated MBTF of over 250,000 hours. A Draper designed journal bearing is contained in a pump 2 inches long by two inches in diameter which runs at 20,000 rpm's and produces 90 watts of hydraulic power. Draper is currently developing machining techniques and technologies which will enable wide spread use of ceramics to be utilized in precision roller bearing designs. The development of a touchdown bearing comprised of a ceramic material with a dry lubricant base is currently being researched at Draper.

The technology used for a touchdown bearing is dependent on the application. If fluid is present, a journal bearing may be the most cost effective approach. If vapors are present, then a gas bearing may be the design of choice. If these options are not practical then a roller element bearing would be necessary. Draper has and continues to maintain the expertise in all these fields to properly design and manufacture an optimum touchdown bearing for most applications.

CONCLUSIONS

To produce a cost effective magnetic bearing requires a full systems approach to the problem. Having just the actuator on hand will not provide a magnetic bearing. The mating of the magnetic actuator technology with control techniques and touchdown bearing designs will produce the desired effect of having a cost effective long life, low power magnetic bearing. Other important issues for the successful application of magnetic bearings include power drive systems, digital control computers and sensor which are light weight, fault tolerant and compact. Draper has been and continues to be a leader in the pioneering fault tolerant systems. Draper has the unique ability to apply these critical technologies to the design and development of magnetic bearings.

References

1] Magnetic and Electric Suspensions, G. Oberbeck, MIT Press 1974.

2] PhD thesis, MIT, September 1986, B. Johnson, Active Control of a Flexible Two Mass Rotor

3] SM thesis, MIT, May 1988, D. McCallum 'Dynamic Modeling and Control of a Magnetic Bearing Suspended Rotor System'

4] SM thesis, MIT, May 1988, P. LaRocca 'A Multivariable Controller for an Electromagnetic Bearing-shaft System'

5] SM thesis, MIT, May 1989, B. Silverman 'Validation of Finite Element Solution for Timoshenko Rotors with Disk'

6] MIT-EIP, mechanical engineering, 1988 - 1990, T. Sacoman SM thesis, MIT, (May 1991), 'Characterization of High Speed Magnetic Bearings'

7] Ph.D thesis, Northeastern, (May 1991), S. Beal 'Future Directions on the Application of Modern Control Theory for Flexible Rotor-shaft Dynamics via Electromagnetic Bearings'

Table I

Rotor Support System Trade-Offs

	Ground Base	Vehicle Base	Space Base
Support Mass	Large	Small to Medium	Small to Medium
Enviromental Input	Low	Large	Large at Launch Low During Mission
Size & Power	Secondary	Prime	Prime
Support System	Magnetic	Mechanical	Trade-Off of Mechnical vs Magnetic

Table II

COMPARISON OF FAULT TOLERANT PROCESSOR TECHNOLOGIES

	TRIPLEX BIPROCESSOR	ADVANCED UNIPROCESSOR
VOLUME	1.5 FT^3	.04 FT^3
WEIGHT	61 LBS	2.5 LBS
POWER	70 WATTS	20 WATTS
MIPS	2	30
MFLOPS	.4	4.3
MEM.	4 M BYTES	8 M BYTES
APPROX. IC'S PER CHANNEL	530	45

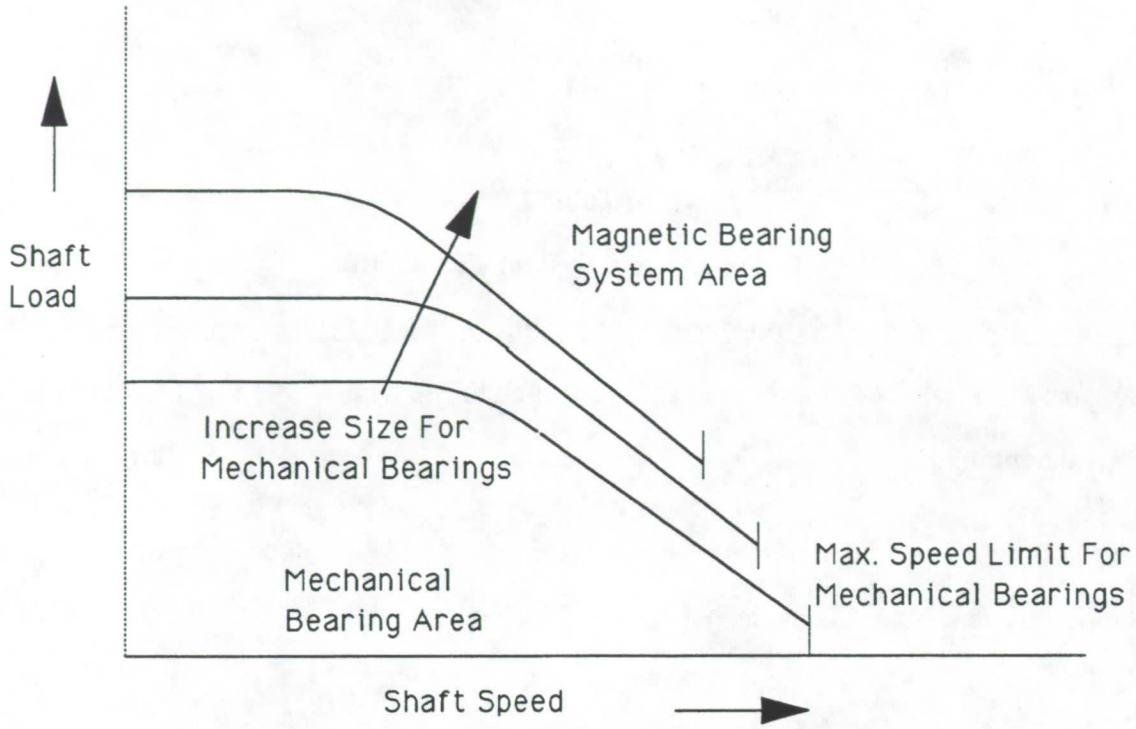


FIGURE 1

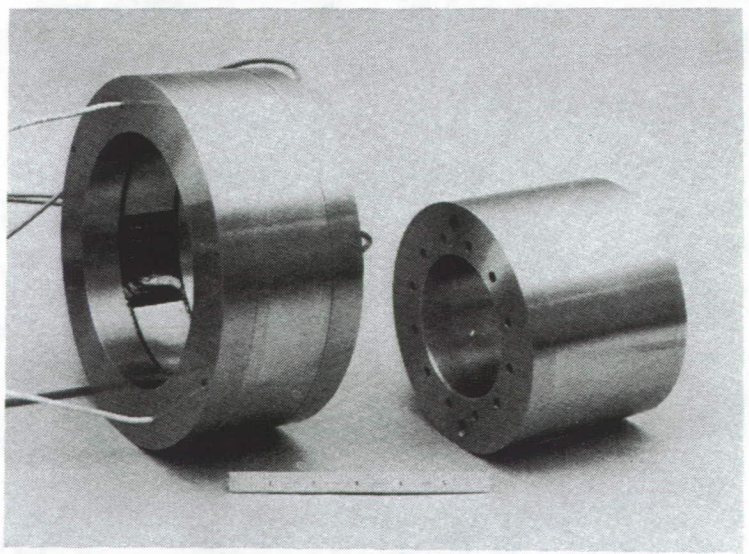


FIGURE 2

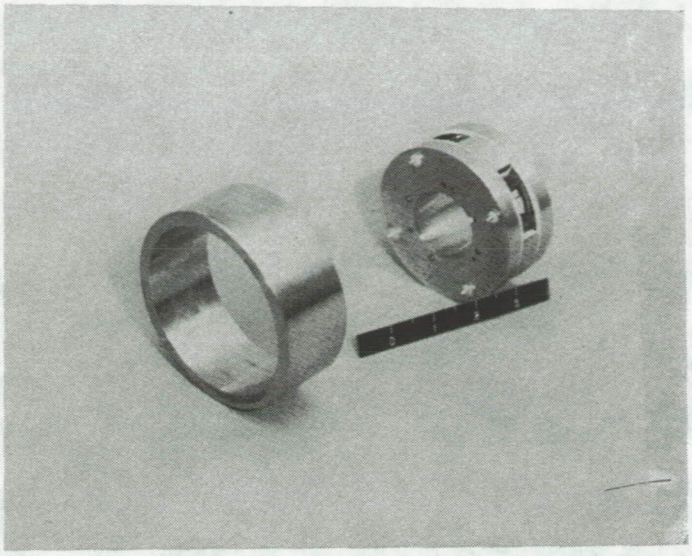


FIGURE 3

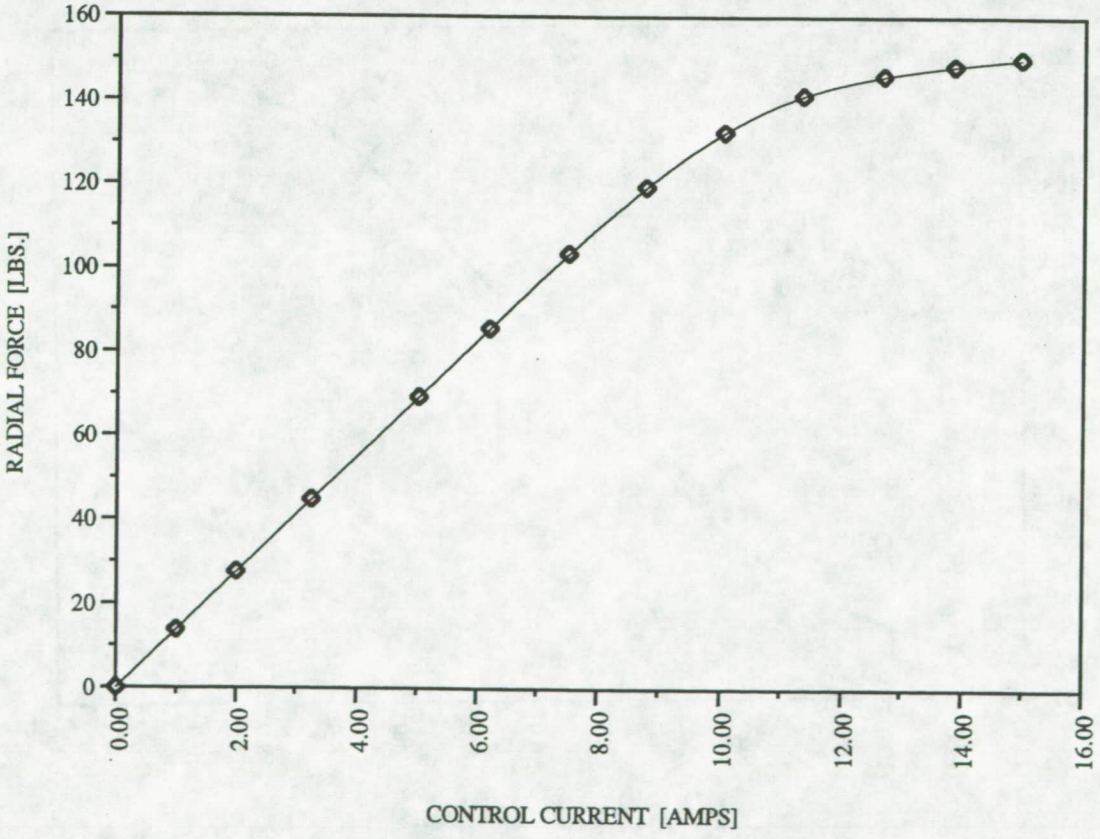


FIGURE 4

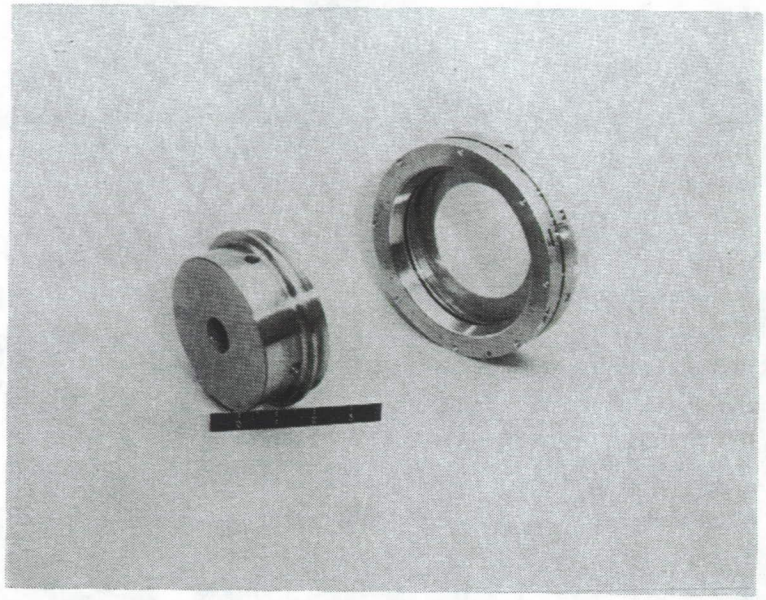


FIGURE 5

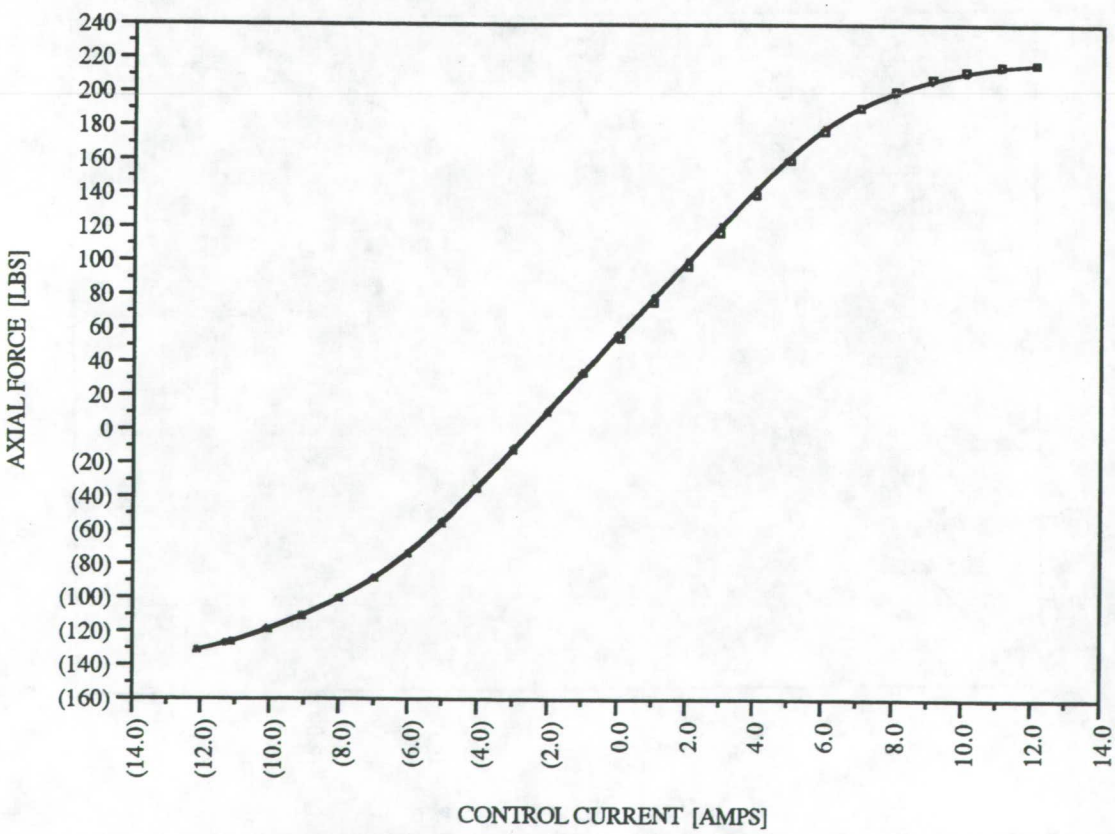


FIGURE 6

AIR-LIFT HYDROHOIST: AN INNOVATIVE UNDERGROUND-TO-SURFACE COAL HANDLING SYSTEM

Philip D. Lindahl, Mining Engineer
U. S. Bureau of Mines, Denver Research Center
Denver, CO 80225-0086

William W. Lutzens, Supervisory Mining Engineer
U.S. Bureau of Mines, Denver Research Center
Denver, CO 80225-0086

ABSTRACT

The U. S. Bureau of Mines (USBM) initiated research to improve the safety and efficiency of transporting coal from underground coal mines using a balanced air-lift hydrohoist (ALH). The ALH includes a U-tube pipeline that can be installed in a mine using an existing shaft or two surface-drilled holes connecting the mine level to the surface. Bench- and pilot-scale models were constructed and tested to evaluate the feasibility of hydraulically hoisting a coal-water slurry in a riser pipe by injecting compressed air microbubbles. A prototype injector was developed and tested that continuously compacts 5-cm minus, run-of-mine coal in a screw feeder and injects a thick coal slurry into the ALH. Initial production rates of 40 metric tph through a 20-cm-diam riser pipeline were achieved with a 50-m-high pilot-scale model, with potential to increase the model productivity to 100 tph through modification of the coal injector. Results indicate the ALH is a safe, feasible, and cost-effective alternative to conventional coal haulage methods, with many applications in the mining and construction industries for hoisting materials other than coal. The USBM is currently seeking mines or underground construction projects that are interested in a commercial-scale test of this alternative material hoisting system.

INTRODUCTION

The balanced ALH is based on the concept that vertical pipeline flow can be generated by injecting compressed air into one column of a two-column U-shaped hydraulic pipeline. The fluid density is reduced in the air-water column, creating a pressure differential between the two columns. This pressure differential results in an upward flow of the air-water column within the "riser pipe," and a corresponding downward flow of the water-only column within the "downcomer pipe." When coal is introduced into the upward flowing column, a coal hoist is created. Coal is ideally suited to an ALH, as coal is relatively light and normally hydrophobic, i.e., it tends to repel water and attract air. The coal is hoisted primarily by water flow, assisted by air flow.

Previous applications of airlifts to hoist slurry mixtures at mines include removal of cuttings from large-diameter rotary-drilled shafts (1) and hoisting coal from hydraulic mines in the Donbass Coalfield, Ukraine (2). These applications do not use a downcomer pipe but utilize a large reservoir of water that fills a shaft or large sump, continuously supplying water

to the riser pipeline. Airlifts were also used during the period 1960 to 1980 to hoist manganese nodules from the ocean floor. The ocean served as a reservoir and an airlift provided the required water displacement, with resulting upward movement in the riser pipe, to hoist an ocean bed slurry.

Hydraulic hoisting of 8-cm or less run-of-mine coal with centrifugal pumps was proven a feasible technology in the United States with the Consolidation Coal Company "Loveridge Coarse Coal Slurry System" (3). Over two million tons of coal during the period 1980 to 1983 were vertically hoisted 270 m in a steel pipe mounted in a borehole at the Loveridge Mine and transported 3.9 km to a preparation plant. The system proved to be safe and required relatively little maintenance. However, additional quantities of very fine-sized coal were generated as the slurry passed through a series of 13 centrifugal pumps that lifted the slurry and transported it to the preparation plant. Fine-sized coal is more difficult than coarse-sized coal to recover and separate out unwanted moisture, which resulted in increased coal preparation costs.

SYSTEM DESCRIPTION

The system envisioned for coal mines is to locate the ALH in a main or submain entry near several continuous miner or longwall sections. Figure 1 shows a conceptual diagram of the coal ALH system, where two holes are drilled from the surface to the mine level and steel pipe is installed in each hole. A horizontal pipe connecting the two vertical pipes underground completes the U-tube design. Relatively dry, broken, run-of-mine coal is injected in the riser pipe and transported to the surface in the air-water medium. Alternately, similar vertical piping can be installed in an existing shaft, and coal can be injected into an interconnected horizontal underground pipeline some distance away from the shaft. Since coal does not pass through any pumps while it is lifted to the surface, size degradation in the ALH is relatively minor. The three-phase mixture is discharged into a storage tank on the surface where the air separates from the coal slurry. At this point, the coal slurry may be transported overland in a low pressure pipeline to a cleaning plant or dried on-site with a mobile preparation plant. If disposal of waste rock on the surface is environmentally unacceptable, it is conceivable that waste materials could be returned periodically to the mine in slurry form. When coal is not being hoisted, waste rock is sent back to the mine level through ALH piping and hydraulically stowed underground.

Primary advantages of the ALH include its safety, flexibility, and compact design, allowing an ALH to be incorporated into an existing mine without requiring a separate, dedicated haulage entry. Safety hazards eliminated with an ALH include coal float dust, methane generation, traffic congestion, and the open, moving parts associated with belt and rail haulage. The ALH has relatively low capital and operating costs, low maintenance, and the flexibility required to integrate with existing mine layouts and coal haulage systems. In addition, it can be operated as a primary haulage system or as a supplemental system transporting a product from a specific area of the mine. As a supplemental system, the ALH can be activated with little preparation and used during peak production demands when existing haulage capacity is inadequate.

The primary disadvantage of the ALH is that additional moisture is added to the coal that can lower its BTU value. The increase in moisture content is proportional to the addition of fines generated through mining and handling coal. However, underground breakage of coal is required to achieve a top size of one-fourth the diameter of the transport pipeline. Since grinding and pulverizing actions to coal are minimized, the additional moisture added to the final product should be relatively low. Water is normally removed from coal in a cleaning plant through screening, gravity separation, filtration, and thermal drying. The ALH system will probably not find application in mines that currently do not have a plant to wash coal. Pipeline blockages may occur due to the presence of tramp materials, loss of water or air flow, or excessive solids concentration in the pipeline. However, proper engineering design can minimize these occurrences and provide for rapid cleanout of the pipeline if a blockage occurs.

Potential safety hazards of the ALH system have been identified, which appear to be controllable through proper design and regular inspection. Failure of the coal seal in the injector is a potential safety hazard, and the injector must be designed to protect mining personnel from a possible slurry discharge. Pipe wear due to abrasive materials travelling through a pipeline is an area for concern, since breakage of the pipeline is a safety hazard. However, wear rates of vertical slurry pipe are significantly lower than wear rates of a horizontal slurry pipe, estimated at one-third or less. Recent improvements in wear-resistant epoxy and ceramic coatings for steel pipe should result in significantly reduced pipe wear rates for coal slurry applications (4). The ALH could be designed with automatic valves that immediately close with any sudden drop in hydraulic pressure to protect mining personnel from a large discharge of water or coal slurry.

The efficiency of an ALH can vary considerably depending on the air-water ratio, specific gravity and shape of solids, pipe diameter(s), initial air bubble size, method of air injection, and level of air-water-solids mixing. When operating parameters are optimized, the ALH can provide an efficient, cost-effective hoisting system. Sizeable hydraulic flow rates can be created in a balanced ALH with compressed air generated from a standard commercial-size air compressor, and the hydraulic flow rate created is infinitely variable depending on the quantity of air injected. The U-tube airlift can be supplemented with a pump on the upstream side of the coal injector to increase the hydraulic flow rate in the pipeline; however, this will also increase the hydraulic pressure at the point of coal injection, thus increasing the force required to inject coal into the moving water stream.

BENCH- AND PILOT-SCALE MODEL TEST PROGRAM

The USBM initiated research on the coal air-lift hydrohoist system by testing a 7-m-high bench-scale model in the Bureau of Reclamation Hydraulic Laboratory on the Denver Federal Center (5). The model was constructed to characterize and visually observe the interaction between air, water, and simulated coal for a wide range of flow conditions. A 28-cm-diam clear, plastic vertical pipe was used to observe hoisting of 5-cm-diam or less simulated coal particles. Simulated coal particles were pieces of rubber having the same specific gravity as coal. Test results with the 7-m-high model proved to be beneficial in developing a basic understanding of two- and three-phase flow phenomena in a large diameter

pipe. The model demonstrated that an air-lift hydrohoist operates satisfactorily when air bubbles are dispersed within a water-simulated coal mixture. The addition of air in a water-filled riser pipe resulted in reduced fluid density and increased particle rising velocity, and prevented particle hang-up along the pipe wall. Although the hoisting efficiency of three-phase mixtures could not be accurately assessed in this small model, positive effects were observed that would encourage large-scale testing of a coal ALH.

A 50-m-high pilot-scale ALH model was then constructed using an existing 60-m-high tower located 32 km northwest of Denver. Pictures of the completed model are shown in figures 2-a and 2-b. The model is currently providing data to define the basic performance and operational characteristics of a coal ALH before it is tested in an active underground mine. Construction of the pilot-scale model was completed in July 1992, with the first successful model test performed in July 1993.

The pilot-scale model is designed to hoist 25 metric tons of 5-cm or less run-of-mine coal in a U-tube pipeline during a 15-minute test run, with an equivalent output of 100 tph. A schematic diagram of the model is shown in figure 3. Coal is first loaded into a 30-ton-capacity bin using a 3-m^3 skip that is lifted by a 5-ton hoist mounted at the top of the tower. A 40-cm-diam prototype screw feeder installed at the bottom of the bin continuously injects coal into the ALH, as shown in figures 4a and 4b. The feeder, called the "labyrinth-seal coal injector," receives gravity-fed coal from the bin, continuously compacts coal in a reducer at the end of the screw to form a hydraulic seal, and discharges it into a moving pressurized stream of water. A cylindrical air-sparger having the same internal diameter as the 20-cm riser pipe injects 480 kPa (70 psi) air through a porous insert into the water stream. Coal is hoisted 50 m high in the air-water mixture to the ninth level of the tower, returns to ground level in a slurry drop pipe, and is discharged into a 113 m^3 (30,000 gal) settling basin. Coal slurry discharge is shown in figure 5. The downcomer and slurry drop pipes are both open at the top of the tower; each pipe stack has a large, 60-cm-diam, 6-m-long pipe section at the top that provides atmospheric venting, accommodates hydraulic surges, and minimizes water level fluctuations. Air separates from the coal slurry at the top of the tower in an open 60-cm-diam pipe. A consistent level of water is maintained in the 30-cm-diam downcomer pipe to provide hydrohoist water at a continuous pressure to a coal-water mixing chamber at the bottom of the U-tube. The USBM has filed two patent applications, one on the balanced ALH system and one on the labyrinth seal coal injector.

PILOT-SCALE MODEL OBSERVATIONS AND SYSTEM MODIFICATIONS

The 50-m-high ALH model has proved to be relatively easy to operate and reliable for hoisting coal. However, several modifications were made to the original design to achieve acceptable performance. Maintaining the proper water level in the downcomer pipe during airlift operations was initially difficult due to water level fluctuations in the downcomer pipe that were measured with a sonic water level sensing device mounted at the top of an enlarged downcomer pipe section. Water level fluctuations resulted from the turbulence of water and entrained air that is dumped into the downcomer pipe and the inability of instrumentation to immediately respond to level changes by adjusting the flow at the feed water pump. This equipment was later replaced by a system that measures the hydraulic pressure at the bottom

of the downcomer pipe using a slurry pipeline pressure sensor and a digital controller. Relatively constant water levels were then maintained with the pressure-sensing system throughout a wide range of flow conditions created during model testing.

The model was initially operated with a 19 kW (25 hp) circulating pump installed upstream of the point of coal injection to hoist coal without injected air. However, the air-lift provided a more consistent U-tube flow than the circulating pump, reduced the hydraulic pressure at the point of coal injection, and was generally preferred to hydraulic pumping due to the ease of operation and infinitely variable flow rates that can be generated in the riser pipe.

The initial coal injection system proved to be ineffective in delivering a continuous flow of coal to the ALH. Coal could not be delivered due to improper discharge reducer selection, absence of a screw rotor tip within the reducer, excessive frictional forces between the coal and screw feeder housing, and the lack of continuous flow of coal from the bin to the injector intake. After a series of design modifications to the injector, a dependable system was developed to continuously inject dry, crushed, run-of-mine coal into a pressurized pipeline. One of the most significant discoveries was the need for additional water injected at several locations into the screw feeder housing to create a thick slurry of coal and water. The purpose of injection water is to lubricate and substantially reduce friction at coal-steel interfaces within the injector. Injection water is obtained from water in the downcomer pipe, pressure-boosted by 270 kPa (30 psi) using an auxiliary pump, and delivered through hoses to the screw housing. The flow of water in the screw is regulated up to a maximum of 0.379 m^3/min (100 gpm) in proportion to the screw feeder rotational speed.

INITIAL DATA ANALYSIS FROM THE PILOT-SCALE MODEL

Several hours of model performance data were obtained, and a preliminary data analysis was made. The model performed successfully in the ranges of 1.3 to 3.7 standard m^3/min (50 to 140 scfm) air flow, 0.93 to 3.2 m^3/min (260 to 900 gpm) water flow, and a coal screw feeder rotational rate of 12 to 45 rpm. A maximum production rate of 40 metric tph of coal having a specific gravity of 1.37 was achieved, and work is in progress to improve the production rate through an increase in the screw feeder rotational speed. Initial data indicate the coal injection rate is directly proportional to the screw feeder rotational speed, providing a proper flow of dry coal feed into the injector is maintained. The ALH system appears capable of hoisting coal at the 100 tph design production rate, although the current screw feeder is undersized and is not expected to achieve more than 50 tph.

The volumetric percentages of air, water, and coal in the riser pipe varied considerably depending on the flow rates selected. The model successfully hoisted coal-water slurries ranging, by volume, from 10% coal to 90% water, up to a maximum of 35% coal to 65% water, with air flows of 1.7 to 2.7 standard m^3/min (60 to 100 scfm). Air flows were increased in proportion to the increasing weight of the slurry mixture in the riser pipe to maintain a consistent flow. Slurry flow out of the riser pipe approached characteristics of "slug flow," or cyclical flow, with interrupted coal slurry discharge, as the total weight of the air-water-coal mixture in the riser pipe approached the weight of the water in the downcomer

pipe. If the weight of the two columns reaches equilibrium, no additional transport water is delivered to the coal-water mixing chamber at the bottom of the ALH until the mixture in the riser pipe is lightened with additional air.

COMMERCIAL APPLICATIONS

Coal Mine Applications

Initial results from the pilot-scale model indicate the coal ALH is a cost-effective alternative to conventional hoisting systems. A preliminary cost analysis was performed by the Bureau of Mines Office of Process Evaluation, with the operating cost of a coal ALH estimated in 1992 dollars at $0.93 per ton compared to $1.21 per ton for a mechanical hoisting system. The capital cost for a coal ALH installed in a 182-m-deep mine was estimated at $1,453,000 compared to an estimated capital cost of $2,889,000 for the mechanical hoist. Hydrohoist capital costs include the following: essential hydrohoist components, borehole drilling and underground excavation, feeder breaker, and additional dewatering circuit for an existing preparation plant. The mechanical hoist includes a hoist and headframe, mine shaft, and hoist building. Based on pilot-scale model data, a 224 kW (300 hp) air compressor would be required to inject air at the bottom of the riser pipe to hoist a 25% by volume coal slurry in a 35-cm-diam riser pipeline. The ALH would produce a water velocity of 3.7 m/s entering the riser pipe and an estimated coal hoisting rate of 430 metric tph.

Applications for the coal ALH are numerous, but a recommended initial commercial application for this technology is to install a system in a mine that currently is experiencing insufficient hoisting capacity. Coal could initially be screened underground to separate out coarse-size fractions for transport in an existing mechanical hoist, while fine-sized raw coal would be transported through the ALH system. Another feasible application for a coal ALH would be to install the system in a mine that is currently producing multiple products created from different mining machines or sections. Coal-rock mixtures created by continuous miners developing entries with an increased vertical height could be transported through an ALH to the surface and overland to a coal cleaning plant. Coal production from a longwall section that is mining entirely within the coal seam could be transported by an existing belt conveyor directly to the coal tipple, thus bypassing the cleaning plant.

Other Mining Applications

The ALH appears to have a wide range of applications for transporting mined materials from underground excavations other than underground coal mines. Surface coal mines that transport coal from a deep pit may consider the ALH as an alternative method for transporting coal up a very steep incline. The ALH appears capable of hoisting virtually any mined material that is properly sized for the selected pipe diameter, providing the hydraulic velocity generated corresponds to the specific gravity of the material, and a suitable injector is incorporated to feed into the pressurized pipeline. The labyrinth seal coal injector may require modification to successfully inject materials that have physical properties largely different from coal, but the concept appears feasible if sufficient material fines are available

to form a hydraulic seal. Tunnel boring machines could transport muck directly to the surface through an ALH installed at intervals along the tunnel and relieve congestion underground due to a long conveyor system. The ALH may also be feasible for removing cuttings from large-diameter rotary-drilled shafts developed from the surface or up-reamed from the mine level. Nonmetallic minerals mined underground that may be applicable to the ALH system include trona, a sodium bicarbonate, and potash, a salt. Since raw trona ore is currently purified on the surface by dissolving it in water, it could be crushed underground, injected into a pressurized pipe with a screw feeder, and dissolved during transport to the surface in an ALH.

CONCLUSIONS

The U.S. Bureau of Mines has developed, tested, and applied for patents on a method of hoisting coal in a U-tube pipeline using an airlift system that appears to be commercially viable and safer than existing transport methods. Pilot-scale tests have demonstrated that coal can be continuously hoisted at production rates of 40 tph with the balanced ALH system using air flow rates ranging from 1.7 to 2.9 standard m^3/min (60 to 100 scfm) to induce gravity-based fluid flow. The existing pilot-scale ALH model should be capable of hoisting 100 tph coal, which is currently limited by the size of the injector. A preliminary cost analysis indicates the ALH system has approximately a 25% lower operating cost than a comparable mechanical hoist at a capital cost of about 50% the cost of a comparable mechanical hoist. In summary, the ALH is a safe, cost-effective hoisting system with numerous potential applications in coal and other mining industries.

REFERENCES

1. Friant, J.E., Blind Shaft Construction - New Equipment Update, Paper in Shaft Drilling Course and Symposium (Colorado School of Mines, Golden, CO, October 12-14, 1988). Session 8, pp. 4-10.

2. Mills, L.J., Hydraulic Mining in the USSR. The Mining Engineer. (London), Vol. 137, No. 203, 1978, pp. 655-663.

3. Alexander, D.W. and Shaw, R.L., Coarse Coal Slurry Transport at the Loveridge Mine. Journal of Pipelines, (Amsterdam), v. 4, 1984, pp. 235-240.

4. Micksell, A., Williams Technologies, Inc., Private communication during tour of Black Mesa Coal Slurry Pipeline, May 24, 1994.

5. Lindahl, P.D. and Faddick, R.R., Air-lift Hydrohoist - A Unique Coal Transportation System. Paper in Mine Mechanization and Automation (Proc. 1st Int. Symp., Golden, CO, June 10-13, 1991). Colorado School of Mines, 1991, pp. 11-1 to 11-11.

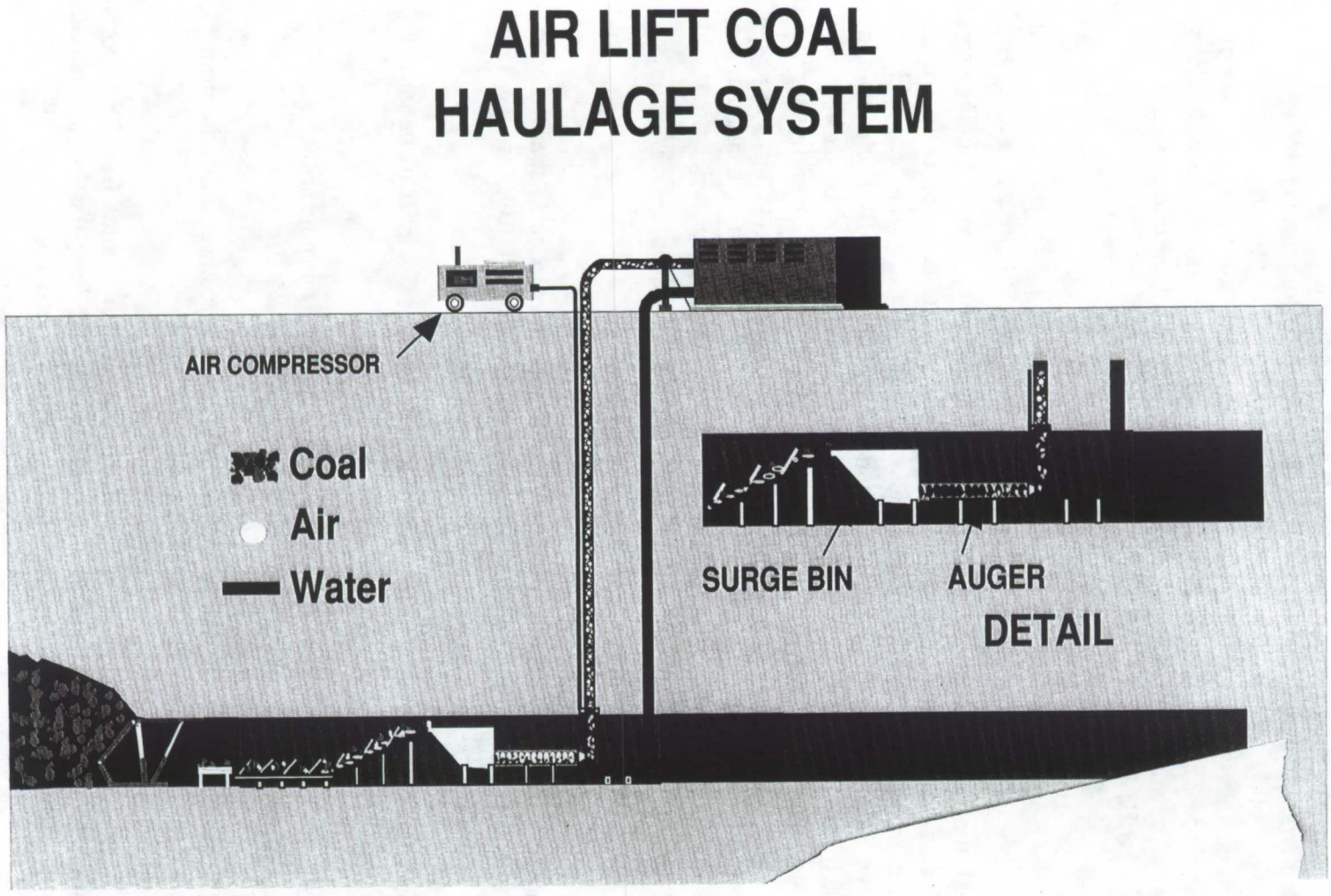


FIGURE 1. Conceptual diagram of the coal ALH system in an underground mine.

FIGURE 2a. Vertical hydrohoist piping attached to a 60-m-high tower.

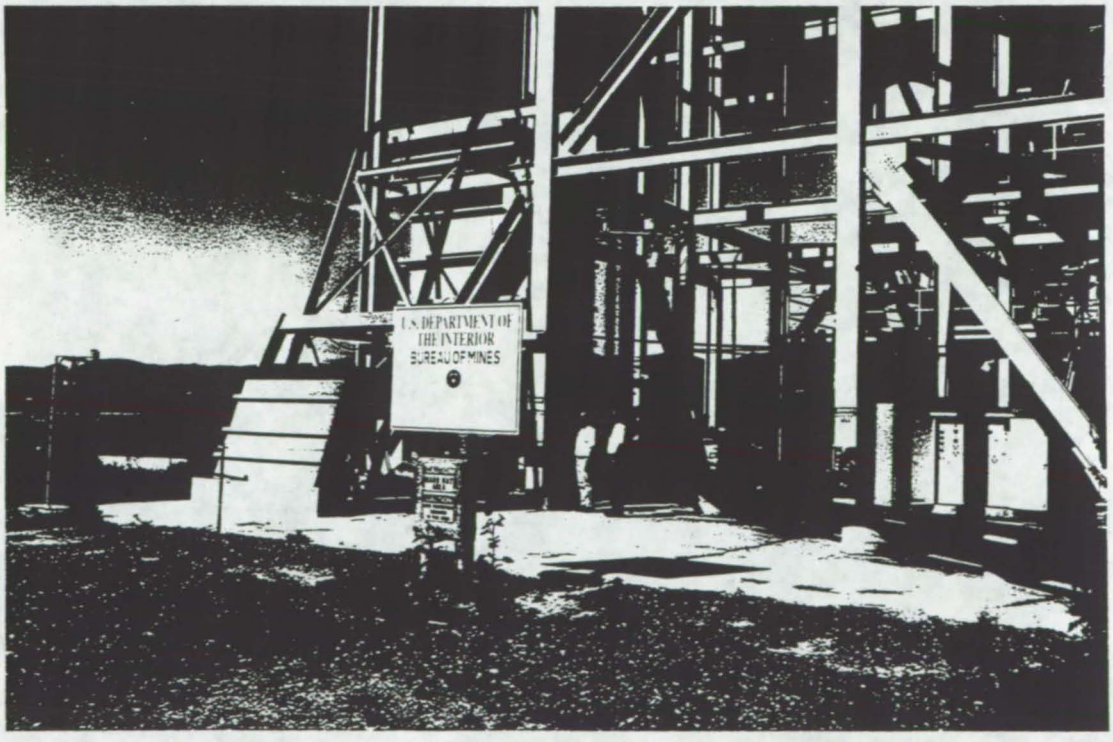


FIGURE 2b. Ground-level view of the pilot-scale ALH showing the coal skip, 30-ton-capacity coal bin, hydrohoist piping, and control panel.

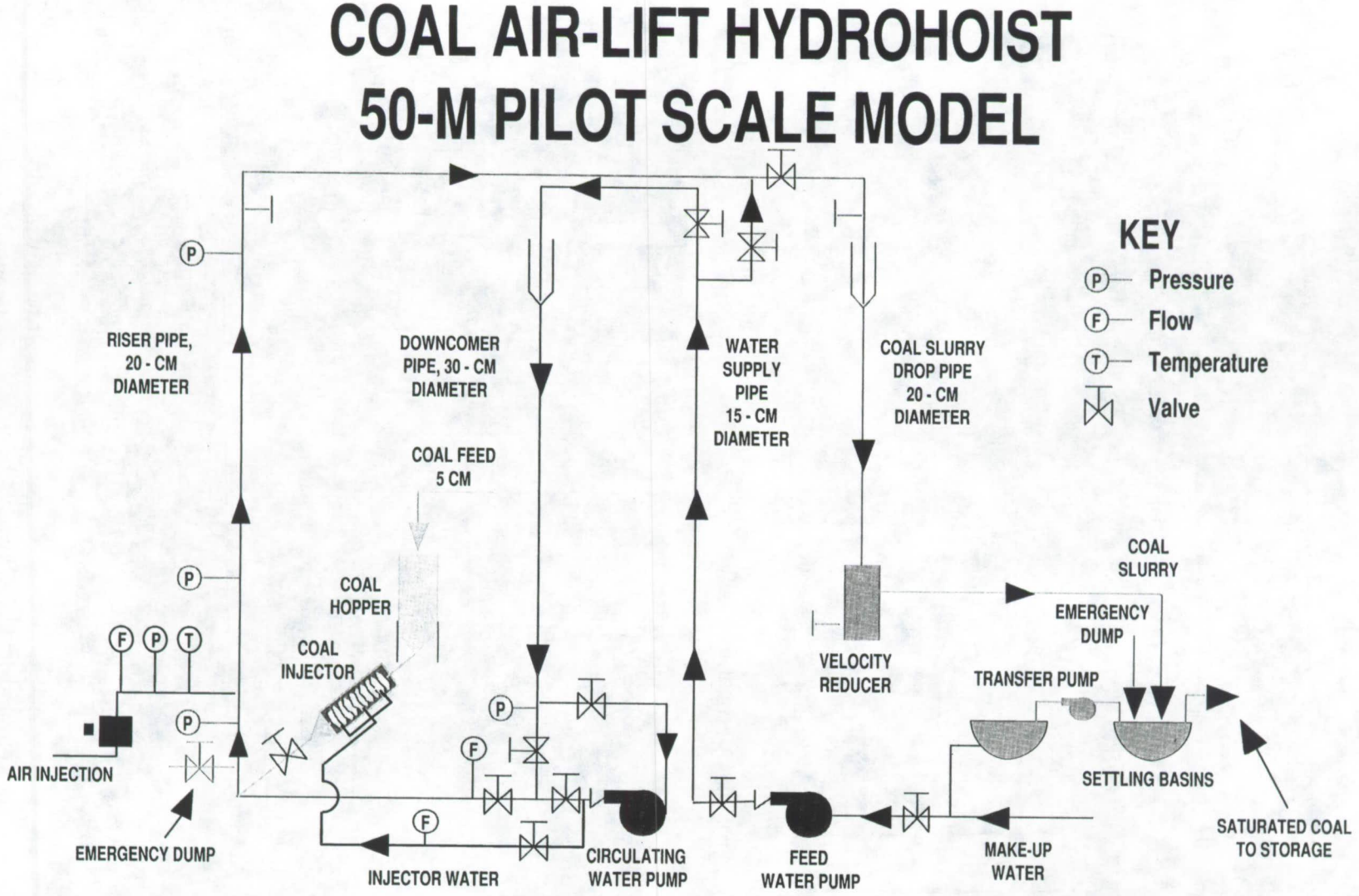


FIGURE 3. Schematic diagram of the 50-m-high coal ALH model.

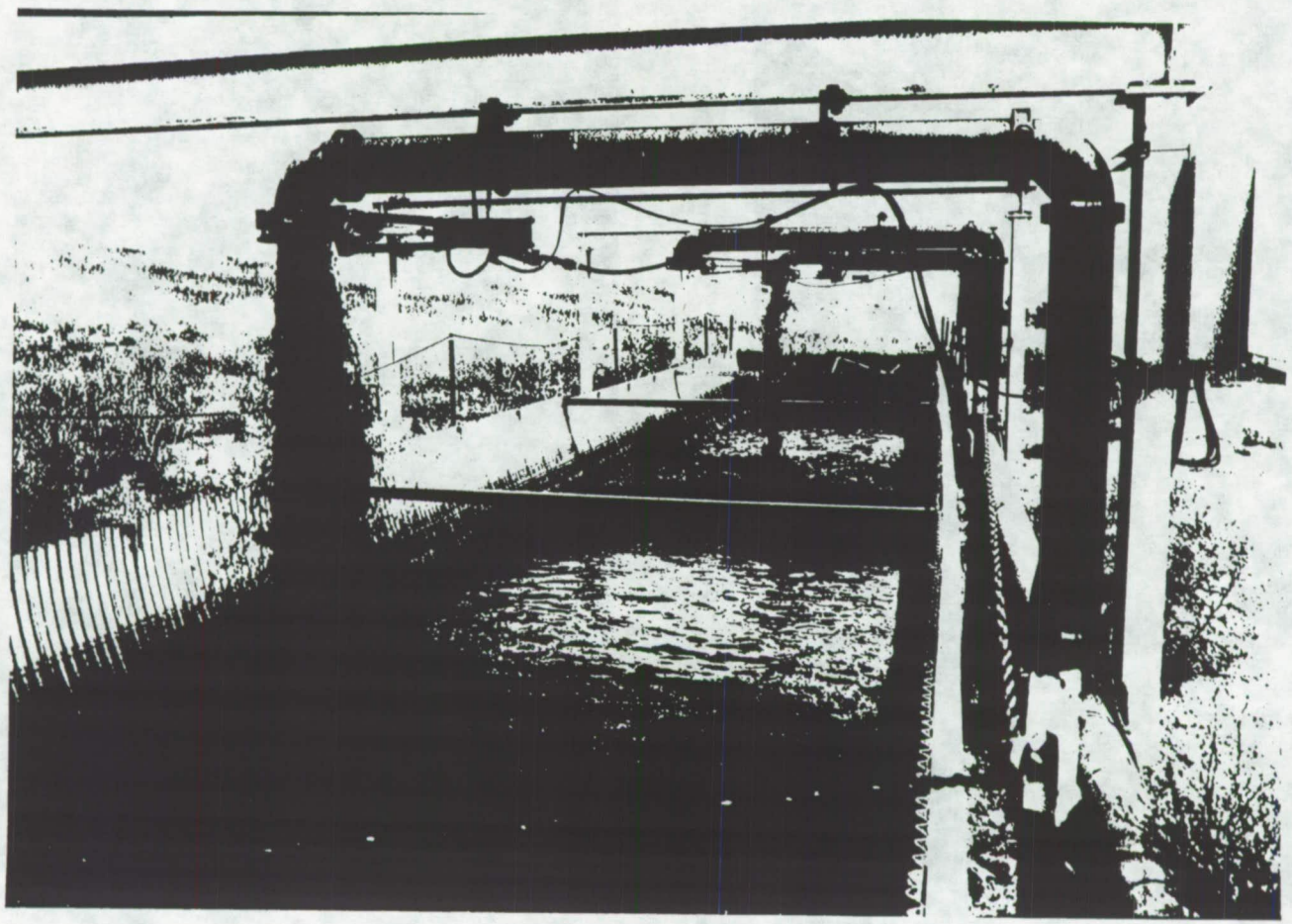


FIGURE 5. Coal slurry being discharged into a 114-m^3- (30,000-gal-) capacity settling basin after it is hoisted up the 50-m-high ALH and returned to ground level.

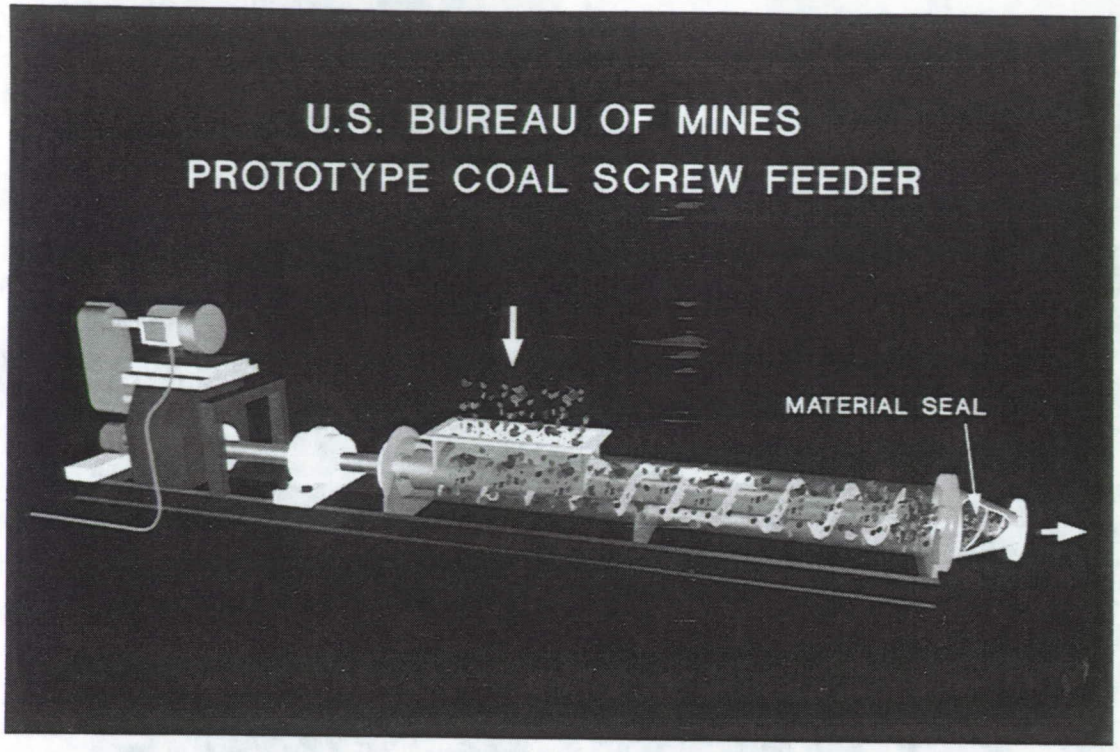


FIGURE 4a. Conceptual Drawing of the USBM "labyrinth seal coal injector."

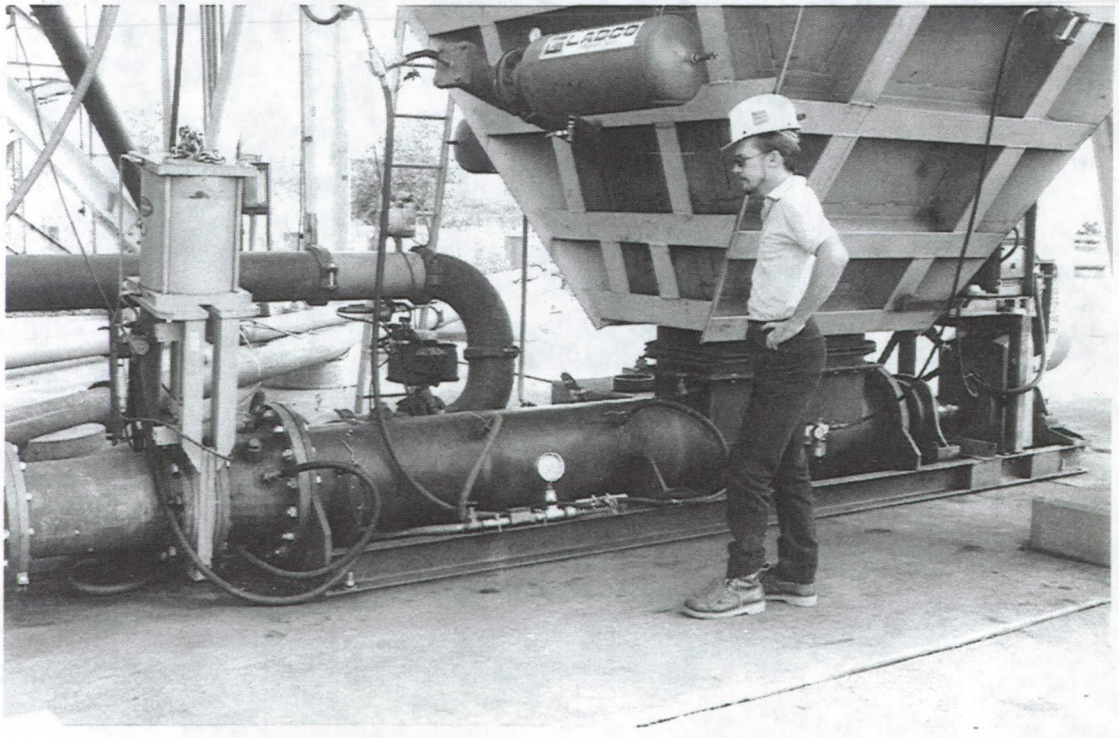


FIGURE 4b. Prototype coal injector installed below the 30-ton capacity bin.

HYBRID CHEMICAL/MECHANICAL HEAT PUMP

L. R. Grzyll, R. P. Scaringe, and F. Gui
Mainstream Engineering Corporation
Rockledge, FL 32955

ABSTRACT

A hybrid chemical/mechanical heat pump with a demonstrated improvement in COPc compared to traditional vapor-compression systems is discussed. Rather than employing a traditional refrigerant, this heat pump uses a mixture of liquid absorbent and any environmentally-safe refrigerant as the working fluid. This cycle has already demonstrated the ability to provide performance which is superior to the current performance leader, the vapor compression cycle, and will allow the use of ozone-friendly hydrofluorocarbon (HFC) refrigerants. An experimental prototype heat pump was built using an innovative lubrication-free two-phase rotary compressor. This compressor uses self-lubricating, low-friction vanes and effectively uses the absorbent liquid to provide sealing and improved compressor performance. The compressor also utilizes a bore geometry which is optimized for the thermodynamics of the two-phase compression process. The commercial potential of this heat pump cycle is outstanding. Because of the high-efficiency, low-cost, and environmental acceptability of the design, this cycle is ideal for heat pump applications in the 0-20 ton range. The cycle would find use in residential, commercial, and industrial heat pump applications, automotive air conditioning, and commercial and government aerospace applications. Another significant benefit of the cycle is that it is similar to existing vapor-compression designs so as to be maintainable by the existing HVAC community.

INTRODUCTION

The basic principle of the chemical/mechanical heat pump has been described by Mainstream and others in previous technical articles and reports [1-7]. This cycle has been referred to as a vapor-compression heat pump with solution circuit. The cycle uses an absorbent/refrigerant pair as the working fluid. The liquid absorbent/refrigerant mixture enters a generator/evaporator where heat is acquired from a heat source causing partial evaporation of the refrigerant from the liquid mixture. The pure refrigerant vapor passes to a compressor and remaining liquid mixture passes to a pump. The refrigerant vapor and liquid mixture are then recombined at a higher pressure and temperature in an absorber/condenser, where heat is rejected to a heat sink. The liquid solution leaving the absorber/condenser is then throttled to a lower pressure in a thermal expansion valve before entering the absorber/generator.

Previous analysis has shown that the vapor-compression heat pump with solution circuit heat pump has several advantages over the traditional mechanical vapor-compression system [1-7].

1. Because the system uses an absorbent/refrigerant mixture, the heat of vaporization of the refrigerant and the heat of solution of the absorbent/refrigerant pair are available to provide cooling, resulting in lower fluid flow rates, less mechanical work requirements, and higher COPs compared to a pure refrigerant vapor-compression system.

2. Because the constant-pressure vaporization of a refrigerant from a liquid mixture occurs over a gliding temperature range, the system approximates the Lorenz cycle which also results in higher COPs compared to a pure refrigerant system and more efficient heat exchangers.

3. Because there is both liquid solution and refrigerant vapor leaving the evaporator/generator, higher evaporator/generator heat transfer coefficients can be achieved compared to pure refrigerant evaporators since evaporator dry-out is not approached.

4. Because the system uses an absorbent/refrigerant pair, the system cooling capacity can be easily varied by changing the fluid composition.

The Vapor-Compression Heat Pump With Solution Circuit

A flow schematic of the vapor-compression heat pump with solution circuit is given in Figure 1. In the system, the refrigerant vapor is separated from the liquid absorbent/refrigerant solution in the generator/evaporator. A pump is used to raise the pressure of the liquid mixture while the refrigerant vapor is compressed in a compressor. The two streams are recombined at a higher temperature and pressure in the absorber/condenser.

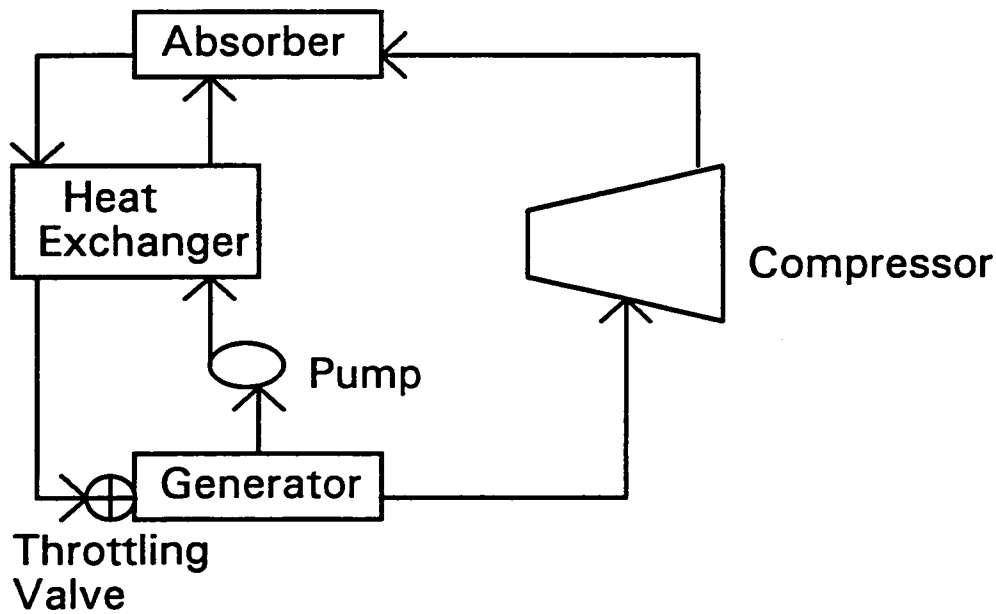


Figure 1. Flow schematic of the vapor-compression heat pump with solution circuit.

For aerospace applications, the separation of refrigerant vapor from the liquid mixture in the generator/evaporator is difficult, making this component complex and heavy. Another problem with the conventional system is how to balance the flow rates and pressure lifts in the liquid and vapor circuits. The individual liquid and vapor flow rates vary as the heat acquisition and heat rejection temperatures change, making this problem acute. Mainstream investigated these problems in a breadboard system. Attempts to balance the liquid and vapor flow rates and pressure lifts included the use of a complex control system and the selection of a pump and compressor with similar head-flow characteristics. Both of these approaches were found to be impractical for a reliable high-performance system.

The Chemical/Mechanical Heat Pump

In order to overcome these difficulties, Mainstream developed a single two-phase compressor to replace the pump and compressor of the vapor-compression heat pump with solution circuit. Use of this component eliminated the need to separate the vapor from the liquid mixture and also solved the flow and pressure balancing problems. This innovative two-phase compressor was designed to handle the presence of liquid in the vapor stream and still provide a high level of performance and reliability. The two-phase compressor evolved from the working principle of a positive displacement, sliding-vane rotary compressor, and therefore is probably limited to cooling loads below 20 tons (positive-displacement machines are typically practical for cooling loads up to about 20 tons and for applications over 20 tons the centrifugal machines dominate). Key features of this compressor include an innovative bore shape, self-lubricating vanes, and the use of absorbent solution liquid sealing instead of oil (lubricant) sealing. The two-phase compressor was designed and fabricated by Mainstream Engineering; specific details of the compressor design and fabrication are given elsewhere [7-8].

EXPERIMENTAL PERFORMANCE OF A BENCH-SCALE SYSTEM

A bench-scale prototype chemical/mechanical heat pump test loop was assembled using the innovative two-phase compressor described above. A schematic of this test loop is given in Figure 2, with pressure and temperature measurement locations given. The bench-scale prototype had four main components: an evaporator (generator), a condenser (absorber), the two-phase compressor, and an expansion valve. The evaporator consisted of two shell-and-tube heat exchangers joined in series, with a heated pumped water loop on the shell side providing the heat load. The condenser was also a shell-and-tube heat exchanger, with cooling water on the shell side. A manual thermal expansion valve was used. The torque and speed of rotation of the compressor were measured with a torque meter.

The heat pump was experimentally tested using chlorodifluoromethane (HCFC-22) as the refrigerant and N,N-dimethylformamide (DMF) as the absorbent. This fluid pair was selected because both the pure refrigerant and the absorbent/refrigerant mixture are well-characterized [4]. Tests were performed for different absorbent concentrations ranging from 0% (pure refrigerant) to 20%. The experimental cooling loads ranged from about 800 W to about 4.1 kW. Table 1 is a summary of the experimental data from the bench-scale prototype system.

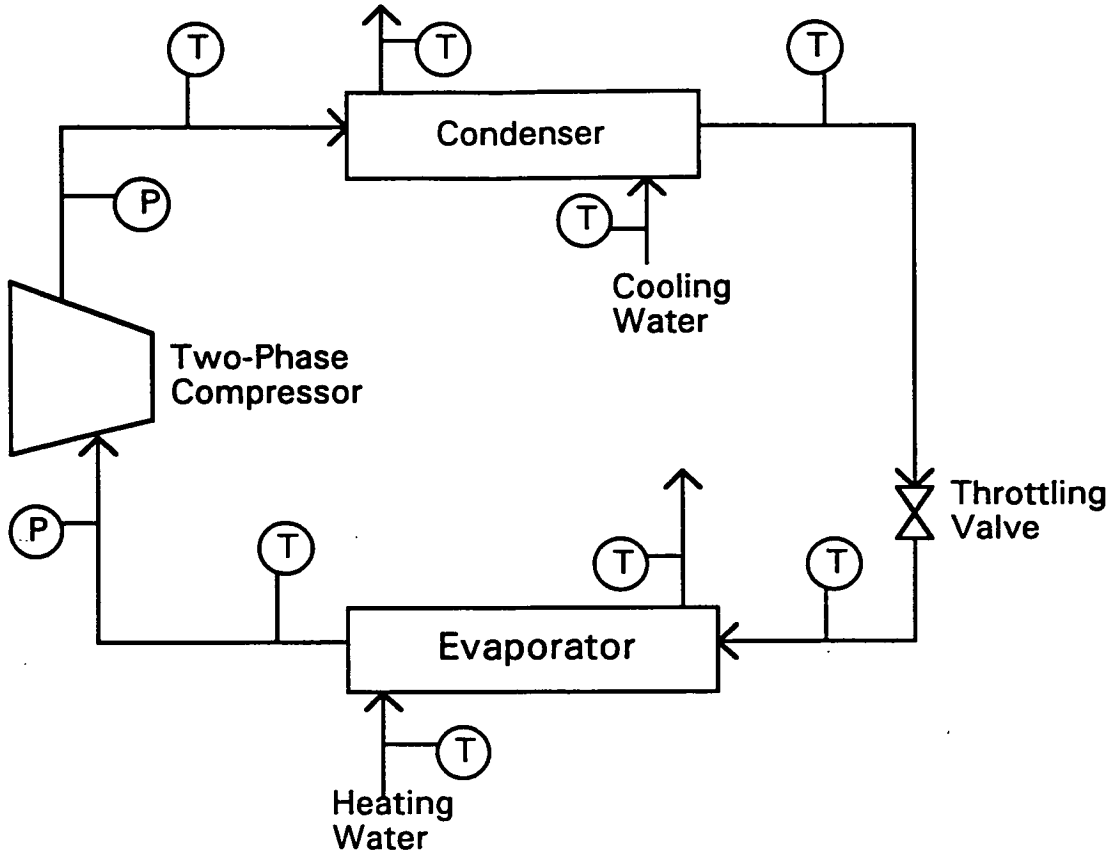


Figure 2. Schematic of the prototype chemical mechanical heat pump test loop.

Figure 3 is a plot of COPc of the system versus the heat pump temperature lift for 4 different absorbent concentrations (0%, 9%, 14%, and 20% DMF). The heat pump temperature lift is defined as the condenser outlet temperature minus the evaporator outlet temperature. Defining the temperature lift in this way assures that the working fluid to air temperature difference for the absorbent/refrigerant pair is equal to or greater than the temperature difference for the pure refrigerant. The experiments with 0% DMF correspond to a conventional mechanical vapor-compression system with a pure refrigerant. This figure shows the effect of absorbent concentration on the performance of the system. While only a limited amount of experimental data is available, Figure 3 shows that for all temperature lifts COPc increases with increasing absorbent concentration until a maximum is reached; above which the COPc decreases with a further increase in absorbent concentration. The optimum absorbent concentration was not determined from these test results, but the experimental results do show that the chemical/mechanical heat pump using the absorbent/refrigerant pair had higher performance than the standard mechanical vapor-compression system (i.e. 0% DMF). Note that this heat pump test loop was not designed to maximize COP but rather to allow for extensive examination of the cycle parameters. It is therefore not surprising that the COPc values for the 0% DMF concentration (pure HCFC-22) is lower than normal. However, the effect of the absorbent on the performance of the cycle is clearly visible since the DMF concentration was the only variable, all of the hardware being identical in all test runs.

Table 1 - Results of Bench-Top Experiments			
100% HCFC-22, 0% DMF			
Cooling Load (W)	**Pressure Ratio**	**Temp. Lift (K)**	**COPc**
1495	2.27	22.02	0.99
1501	2.29	22.08	1.00
1691	2.69	23.16	0.85
1663	2.70	23.15	0.84
91% HCFC-22, 9% DMF			
Cooling Load (W)	**Pressure Ratio**	**Temp. Lift (K)**	**COPc**
2079	2.46	18.75	1.36
2060	2.27	20.13	1.38
2017	2.71	25.54	1.15
1881	2.72	26.30	1.08
1799	2.78	26.26	1.06
1716	2.75	26.63	1.02
1650	2.74	27.14	1.00
1718	3.02	24.13	1.04
1542	3.12	27.77	1.06
1535	3.15	28.66	1.06
1417	3.22	29.36	0.99
1423	3.20	29.89	1.00
1266	3.22	30.80	0.92
86% HCFC-22, 14% DMF			
Cooling Load (W)	**Pressure Ratio**	**Temp. Lift (K)**	**COPc**
1190	2.59	23.69	0.89
1927	2.33	21.14	1.36
2034	2.34	20.89	1.45
4013	2.20	15.58	2.11
3728	2.19	16.32	1.99
3681	2.19	16.26	1.96
3223	2.16	17.67	1.88
1420	2.24	21.14	1.21
1263	2.35	22.28	1.06
80% HCFC-22, 20% DMF			
Cooling Load (W)	**Pressure Ratio**	**Temp. Lift (K)**	**COPc**
1677	2.15	16.23	1.37
1740	2.29	16.09	1.46
1688	2.31	16.09	1.41
1729	2.39	16.12	1.31
2735	2.27	13.23	1.88
2624	2.31	13.36	1.79
2511	2.36	13.61	1.70
2392	2.32	13.79	1.69

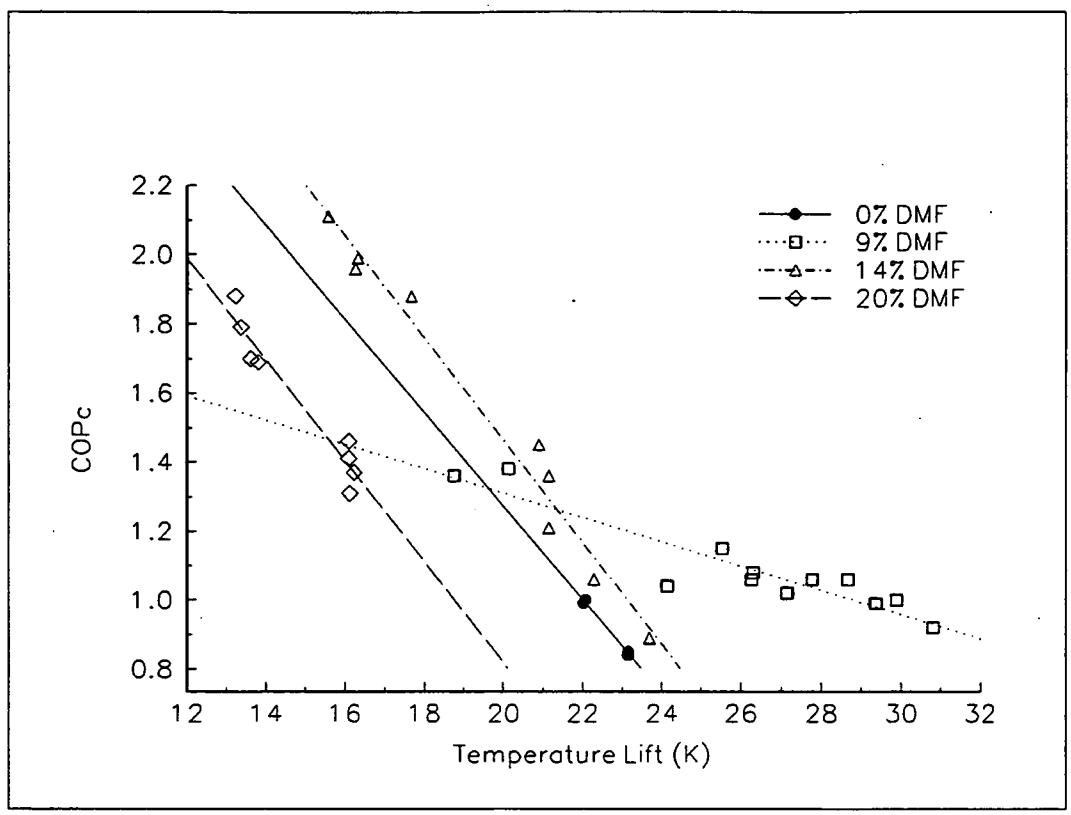


Figure 3. COPc versus temperature lift for various absorbent concentrations.

FURTHER DEVELOPMENT OF THE CHEMICAL/MECHANICAL HEAT PUMP

Mainstream is pursuing further development of the chemical/mechanical heat pump for both commercial and military applications. The goal is to develop a high-performance, high-reliability system that is lubrication-free (eliminating the problems of finding a suitable lubricant for the absorbent/refrigerant pair used in the system) and uses an environmentally-acceptable absorbent/refrigerant working fluid pair that.

In order to be environmentally acceptable, the absorbent/refrigerant pair should have the following properties:

- nontoxic
- nonflammable
- zero ozone depletion potential (ODP)
- thermally stable

The absorbent/refrigerant pair must also have the following thermodynamic characteristics for high performance and reliability

- suitable vapor pressures at system temperatures over the entire concentration range
- high heat of solution when mixed
- absorbent has high affinity for refrigerant vapor
- volatile refrigerant/non-volatile absorbent

An obvious choice for an environmentally acceptable refrigerant that has suitable thermodynamic properties for this application is HFC-134a, which is the replacement refrigerant of choice for CFC-12 and HCFC-22. HFC-134a is nontoxic, nonflammable, is thermally stable, and has zero ODP. What is needed is to identify suitable absorbents for HFC-134a. The absorbents must meet all the environmental and thermodynamic criteria listed above.

Mainstream has been active in many efforts over the past seven years aimed at working fluid identification, development, and characterization. This experience gives us critical insight into what is needed to screen, develop, and characterize working fluids for specific thermal control applications. A multi-step approach to find suitable absorbents for HFC-134a has been identified that will result in the screening of a large number of potential absorbents, the experimental characterization of a large number of absorbents, and the selection of the most suitable HFC-134a absorbent for the heat pump. This procedure is outlined below.

1. Screen a large list of potential compounds for use as absorbents of HFC-134a. Screening criteria would include absorbent volatility, affinity and solubility for HFC-134a, toxicity, flammability, environmental impact, materials compatibility, cost, and other physical or thermodynamic properties.

2. Select a list of promising absorbents from the screening procedure and characterize the key properties of the absorbent/HFC-134a pairs. The key properties of the absorbent/refrigerant pairs include the vapor pressure of the mixture as a function of temperature and concentration, the heat of vaporization of the mixture as a function of temperature and concentration, and liquid heat capacity of the mixture as a function of temperature and concentration.

3. Use the property characterizations to model the various absorbent/refrigerant pairs in a computerized chemical/mechanical heat pump simulator to screen their theoretical performance in the system. Mainstream has previously developed a chemical/mechanical heat pump simulator designed for fluid pair screening[4-5]. A comparison with competing thermal control technologies could also be made at this time.

4. Pick the most suitable pairs for experimental evaluation in a prototype chemical/mechanical heat pump.

COMMERCIAL APPLICATIONS OF THE CYCLE

In this day of concern over the depletion of the ozone layer and energy efficiency, this high-efficiency, environmentally-acceptable, hybrid heat pump would have applications in both the commercial and military environments. The two-phase compressor will also have many applications in both the commercial and military sectors as well.

Commercial systems such as home air conditioners, refrigerators, automotive air conditioners, air conditioners for electric vehicles, and other cooling load applications below about 100 tons will benefit from this technology. The fact that the COPc of this cycle is higher than conventional vapor-compression systems will result in energy savings for these appliances and systems. The evaporator and condenser heat exchangers for the chemical/mechanical heat pump are also more compact and lightweight than conventional vapor-compression heat exchangers because all of the heat transfer is to or from a two-phase liquid/vapor mixture. This is a more efficient heat transfer process than transferring heat to or from a superheated vapor, which occurs in a portion of the conventional vapor-compression heat pump heat exchangers. The fact that the system heat exchangers are more compact and lightweight than vapor-compression heat exchangers makes the system attractive where weight and space are a concern, such as automotive applications and window air conditioning applications. The use of a rotary rather than a reciprocating or scroll compressor means lower noise, which is a distinct benefit in residential applications. The fact that the cycle can also use non-ozone-depleting working fluids makes the cycle environmentally friendly while improving performance.

This technology will also be very attractive to both military and commercial applications since there is dramatic effort to eliminate ozone-depleting substances. The cycle is already receiving attention for aerospace applications, where small, modular heat pumps are being considered for orbiting habitats and lunar and Martian stations. The fact that the cycle has a high efficiency and is lightweight makes it especially attractive for aerospace applications. The cycle is also attractive for ground-based, aircraft, and shipboard military applications because of its high-performance, small weight and size, and environmental characteristics.

Because of the similarity of this cycle to conventional heat pump cycles, it will not present any service, maintenance, training, or education problems to the almost 500,000 refrigeration technicians in the United States. The absorbent liquid would simply be charged into a system in the same way that the refrigerant lubricant is currently handled. The basic heat pump components, namely the compressor (now a modified rotary), evaporator, condenser, and throttling valve remain essentially unchanged. In fact, service technician activities for the absorbent/refrigerant pairs would be closer to traditional service practices than the technician activities associated with the new refrigerant blends that are considered as "drop-in" replacements for CFCs.

The two-phase compressor developed as part of the chemical/mechanical heat pump also has significant commercial and military applications in its own right. For two-phase fluid pumping applications, the two-phase compressor has significant impact. This compressor will also find many commercial applications in the HVAC industry, where systems are designed to prevent two-phase flow.

ACKNOWLEDGMENT

Mainstream Engineering Corporation would like to acknowledge the support of technical monitor, Dr. Steve Benner, of NASA Goddard Space Flight Center. This work was performed under contract NAS5-31167.

REFERENCES

1. Mucic, V., "Two Media Resorption Compression Heat Pump With Solution Circuit," *Heat Recovery Systems*, 5(1), 27-31, 1985.

2. Pourreza, S. and Radermacher, R., "Calculation of the Performance of Vapor Compression Heat Pumps With Solution Circuits Using the Mixture R-22/DEGDME," *International Journal of Refrigeration*, 9, 245-250, 1986.

3. Stokar, M. and Trepp, Ch., "Compression Heat Pump With Solution Circuit - Part 1: Design and Experimental Results," *International Journal of Refrig*eration, 10, 87-96, 1987.

4. Silvestri, J. J., Scaringe, R. P., and Grzyll, L. R., "Performance of a Hybrid Chemical/Mechanical Heat Pump," *Proceedings of the 25th Intersociety Energy Conversion Engineering Conference*, 2, 239-43, 1990.

5. Grzyll, L. R., Silvestri, J. J., and Scaringe, R. P., "Development of a Hybrid Chemical/Mechanical Heat Pump," *Proceedings of the 26th Intersociety Energy Conversion Engineering Conference*, 2, 547-52, 1991.

6. Rahman, M. M., Gui, F., Grzyll, L. R., and Scaringe, R. P., "Modular Chemical/Mechanical Heat Pump for Spacecraft Thermal Bus Applications," Final Report, submitted to NASA-GSFC, June 1993.

7. Rahman, M. M., Gui, F., and Scaringe, R. P., "Design and Experimental Evaluation of a Hybrid Chemical/Mechanical Heat Pump," *Proceedings of the International Absorption Heat Pump Conference*, 487-92, 1994.

8. Gui, F., Rahman, M. M., and Scaringe, R. P., "Development of Compact, Lightweight, High-Performance Sliding-Vane Rotary Compressors for Heat Pump Applications," *Proceedings of the 28th Intersociety Energy Conversion Engineering Conference*, Vol 1, 885-90, 1993.

REAL-TIME UTILITY METERING and DSM via TELEPHONE: DOE-SPONSORED FIELD TRIAL RESULTS

Charles J. Cain, President
UTILEX, Inc.
P.O. Box 991
Greenville, NC 27834

ABSTRACT

During two years of DOE-sponsored field trials a new telephonic technology has reliably performed complex utility-metering functions without special or replacement meters; centralized real-time load research without load recorders; automated real-time outage detection and plotting; and real-time measurement of the results of load management actions; using simple utility meters, existing ordinary telephone lines, and a personal computer running only ProComm and Lotus 123. The prototype system communicated with one remote site per second; commercial units will achieve 100 sites/second and will also send control signals, data and text to remote locations in real time. Compatible with today's telephone networks and with their planned conversion to broadband systems, and readily adaptable to coaxial cable equipment, UTILEX's Bridge to Broadband[tm] not only promises to reduce electric and gas utility costs, improve quality of service and give economical real-time two-way communication with customers, but to develop data traffic to speed the expansion of broadband systems into non-urban areas.

BACKGROUND

The results below are quite different from what we set out to achieve, which was to develop a new telecommunications technology that would make automatic meter reading (AMR) truly practical. But in doing that, we also learned how to provide real-time telecommunication using ordinary telephone lines, and how it can make a lot of expensive special-purpose utility equipment unnecessary, letting simple, ordinary meters and non-proprietary software do some surprisingly complex things. Our two years of DOE-sponsored field trials in two North Carolina towns have made it clear that real-time telecommunication will cut utility costs, improve customer service and help customers conserve electricity, gas and water. Our system is now patented and FCC-approved, and we are undertaking larger pilot projects with selected utilities.

UTILEX, Inc., a North Carolina startup company, is the successor to an earlier company which developed, demonstrated, and patented in twenty countries an inexpensive sensor (then known as the Cain Encoder[tm]) that obtains direct dial readings from ordinary utility meters. The encoder has no moving parts, clips easily onto the register of the existing meter, and is transparent so the dials of the meter can still be read visually. It works by scanning the meter's dials with an electric field, measuring the disturbances the meter hands cause in the fields, and outputting the dials' readings; it was the first meter-reading device ever to use

standard ASCII code as an output. And because it doesn't count pulses, it doesn't suffer from cumulative errors, doesn't require constant monitoring, and needs no batteries or data storage at the meter site. UTILEX now owns all rights to the encoder and has added some improvements, with more to come.

This encoder was a great technical success, accumulating about 18 million consecutive accurate meter readings in field trials with telephone, TV cable and powerline carrier systems. It won high praise from electric and gas utilities, but had no market: there was no satisfactory telecommunications link to the remote meter site. We had, in effect, invented the carburetor before anyone had developed a car. We expected suitable telecommunications systems to appear, but that never happened, and it became clear that until it did AMR would never be viable. In 1989 we formed UTILEX to find a solution that would create a market for the encoder, and were fortunate enough to be fully sponsored by the Department of Energy under an SBIR grant for $550,000.

Most of what follows relates to electricity, but of course most of the same considerations apply to gas and water, and our technology is well suited to all three. Similarly, the references to telephone systems are equally applicable to coaxial systems.

The UTILEX solution

Of the potentially applicable communications technologies--telephone, radio, TV-cable, optical and powerline carrier--only telephone is ubiquitous, reliable, secure and bidirectional--properties that are ideal for carrying data. Telephone networks are not vulnerable to frequency congestion, atmospheric conditions or reflections from buildings; don't require the utility to broadcast its revenue data over the public airwaves; and are capable of high data speeds. But telephone systems have been unsuited for most important utility applications because of two properties of the telephone switch: (a) it is entirely dedicated to the voice subscriber, blocking any other signals whenever the subscriber's telephone is busy or ringing; and (b) it requires seven to thirty seconds to access each subscriber line, which is impractically slow for utility use.

We have overcome these two obstacles by using a novel (and patented) "distributed out-of-band" approach. Briefly, small "modem units" of our own design, dispersed throughout the utility's service area and accessible by a conventional telephone call from the utility's computer, are able to communicate bidirectionally with companion units (MIUs) at each subscriber location within a radius of several miles via carefully selected and inaudibly high (i.e., out-of-band) frequencies. Our signals have no effect on the telephone system or service, and can travel to and from the remote site at all times, even while the subscriber's line is in use, whether for voice, FAX, or modem--or even if the subscriber's telephone service has been disconnected. Because remote sites are not individually dialled the system is quite fast (in our field

trials it idled along at one remote site per second, but commercial equipment will reach 100 sites/second) and is site-specific, immune to the frequent changes in telephone numbers which plague many telephone-based AMR devices.

We have successfully tested our prototypes on four types of telephone switches; on copper loops, remote terminals and pair-gain systems; on loaded and non-loaded lines; on lines that were below voice-grade standards; and even on a line serving a coin-operated telephone. Five telcos have judged our approach to be compatible with POTS, using analog or digital lines. But equally important, it is easily adapted to coaxial-cable systems and is upward-compatible with the broadband systems now being planned and built. Indeed, it is the ideal Bridge To Broadband[tm], as we discuss below.

Teaching Old Meters New Tricks

For just one example of what this new signalling capability offers to the utility industry, let's look at what meters do and what they cost to buy and to read.

meter type	cost: to buy	to read	what you get
watthour	$30	0.60/mo	total energy used
Time-Of-Day	150	0.60/mo	TOD data
Demand	300-900	0.60-25/mo	Interval demand data
Load recorder	300-900	12-25/mo	Load profile data

Figure 1: Major Types of Electricity Metering Equipment

The simplest utility meters are those used for residential accounts. They indicate only how much electricity (or gas) has passed through them, must be read visually, and cost only about $30. They don't store any data electronically, but are read at their site, usually monthly. Each reading typically costs only fifty cents--so low that AMR systems are very hard to justify.

Monthly readings from ordinary watthour meters can be thought of as "demand readings" in which the demand interval is one month. These meters are simple and inexpensive because the utility can easily collect their readings at the end of each demand interval--that is, each month--by sending a person to the site, so the meter doesn't have to store any data.

To find out anything more than how much energy was used during the month, much shorter demand intervals are required: a few hours for Time-Of-Day metering, or a few minutes for demand metering and load research. Because a person can't visit the site that often, the meter has to store the data from each time interval until it can later be collected. And to prevent losing data during power outages the meter must have backup batteries. All this makes these meters considerably more expensive, even though they are really doing the same fundamental thing as their simpler cousins--

measuring how much of the utility's product has flowed through them during each demand interval. They are expensive not because of their function, but because humans can't collect their data often enough.

But our real-time technology can get the information from the meter whenever we need it--even while the telephone line is busy. We don't need to store data at the meter site, so a very simple meter suffices for even the shortest demand interval, hence for the most complex rate structures. And because our communication with the meter site is bidirectional we can also send load management signals to the site, obtain confirmation that they were acted upon, and measure the results immediately; we can send pricing signals or other information to the customer; remotely disconnect and reconnect service; control appliances and services within the home. All of this over the existing telephone line, using existing meters, under full control of the utility, without load recorders, laptops, on-site batteries, site visits, or proprietary software.

Real-Time Load Profiles Without Load Recorders

In addition to mountains of dial readings, for months during our field trials we obtained continuous five-minute demand profiles in real-time from ordinary $30 residential watthour meters such as the GE I-70, from an office twenty miles away equipped with a personal computer running only commercial software: ProComm and Lotus 123. Our computer accessed each customer's ordinary meter via the house telephone line every five minutes, night and day (we could have used any demand interval down to one minute); no load recorders, special meters, proprietary software nor site visits were needed. (We later repeated this using ordinary industrial meters.) Figure 1 on the next page shows some typical results from house "R3," the four plots representing successive four-hour segments of a single February day. The customer's use of the telephone was not affected and did not affect our operation, and the customer was never disturbed--not even in the middle of the night. We needed no batteries at the meter site because our data were collected directly into our remote computer: if the power failed, no more than five minutes' worth of data would be lost--only one of the approximately 288 points obtained each day from each meter.

Simultaneous profiles from the larger house across the street ("R4") showed strikingly higher peaks, making us suspect that the utility's cost of service was higher there than at R3: if so, the existing residential rate structure (total monthly energy consumption alone) would not permit recovery of the additional cost. We decided to take a quick look at how Time-Of-Day rates would affect these two houses--in effect, at whether TOD rates would encourage different consumption patterns at the higher-peaking house, and how the utility's revenue would be affected.

As our host utility had no time-differentiated rates, we used two residential TOD rate schedules and a conventional non-TOD rate kindly supplied by MetEd (Reading PA). These rates, when applied

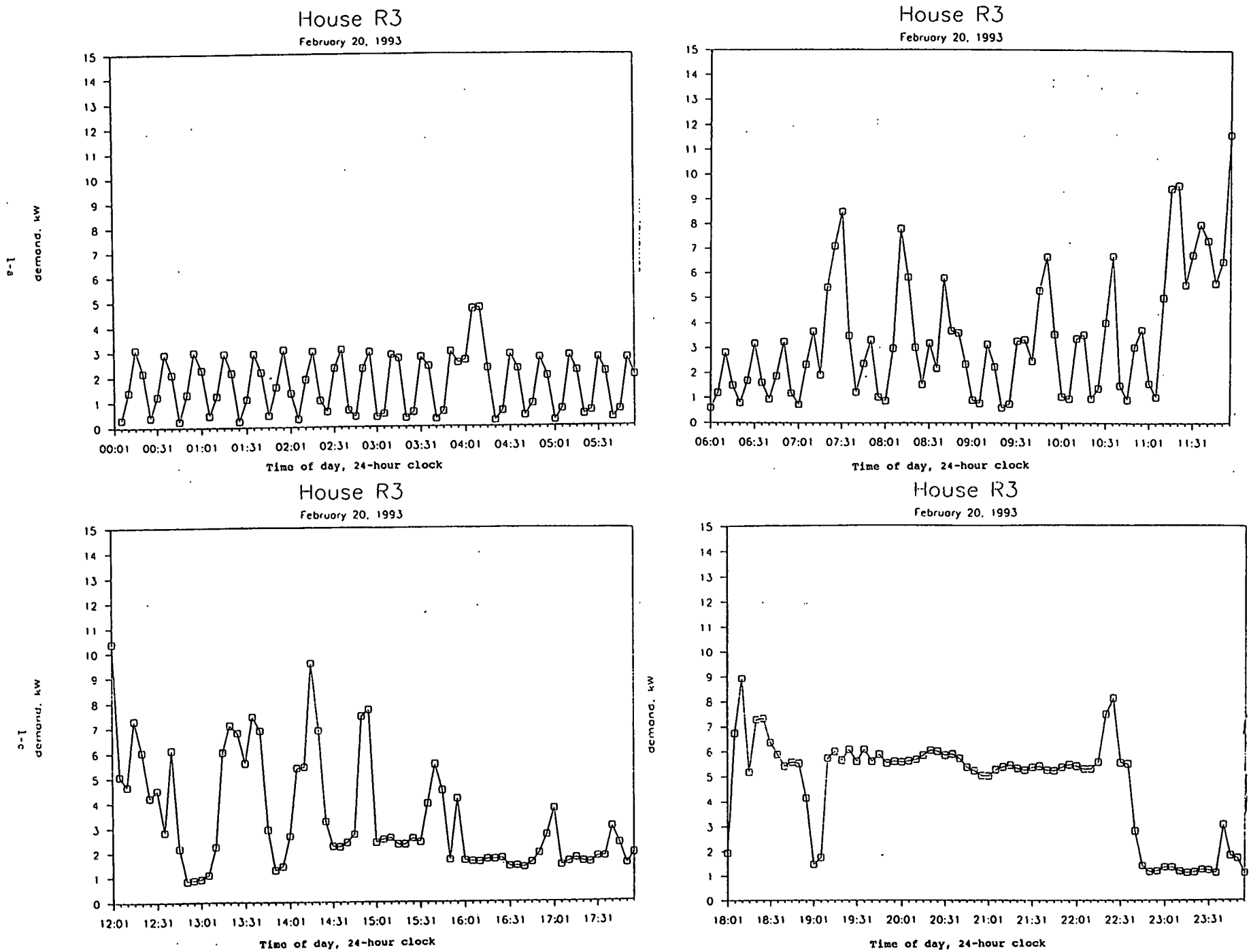


Figure 1: Real-Time Load Profiles Without Load Recorders

to our data from fourteen days in March 1993, had these effects on the monthly bills of the two houses:

	R3	R4
Seasonally-adjusted (winter) TOD	-$6.62	+$5.25
Non-seasonal TOD	+ 7.74	+23.76

This suggests that properly designed TOD rates could provide good incentives and improve the utility's cost recovery, even in cases where revenue might be slightly lower (as in the winter rate above) if this is offset by lower cost of power. Real-time telemetry will not only reveal many such opportunities by making widespread load research convenient, fast and economical: it will also permit rate structures to be flexibly administered directly from the utility office without special meters or visits to the meter site for data collection or reprogramming. It is reasonable to predict that future utility rate schedules will be custom-tailored to fit each customer's usage pattern and will be individually adjusted from within the utility's office in real time as circumstances warrant. Moreover, fully-auditable data will be available in real time to support rate design and rate-case presentations to regulators.

Automated Power Outage Monitoring, Plotting, and Mapping

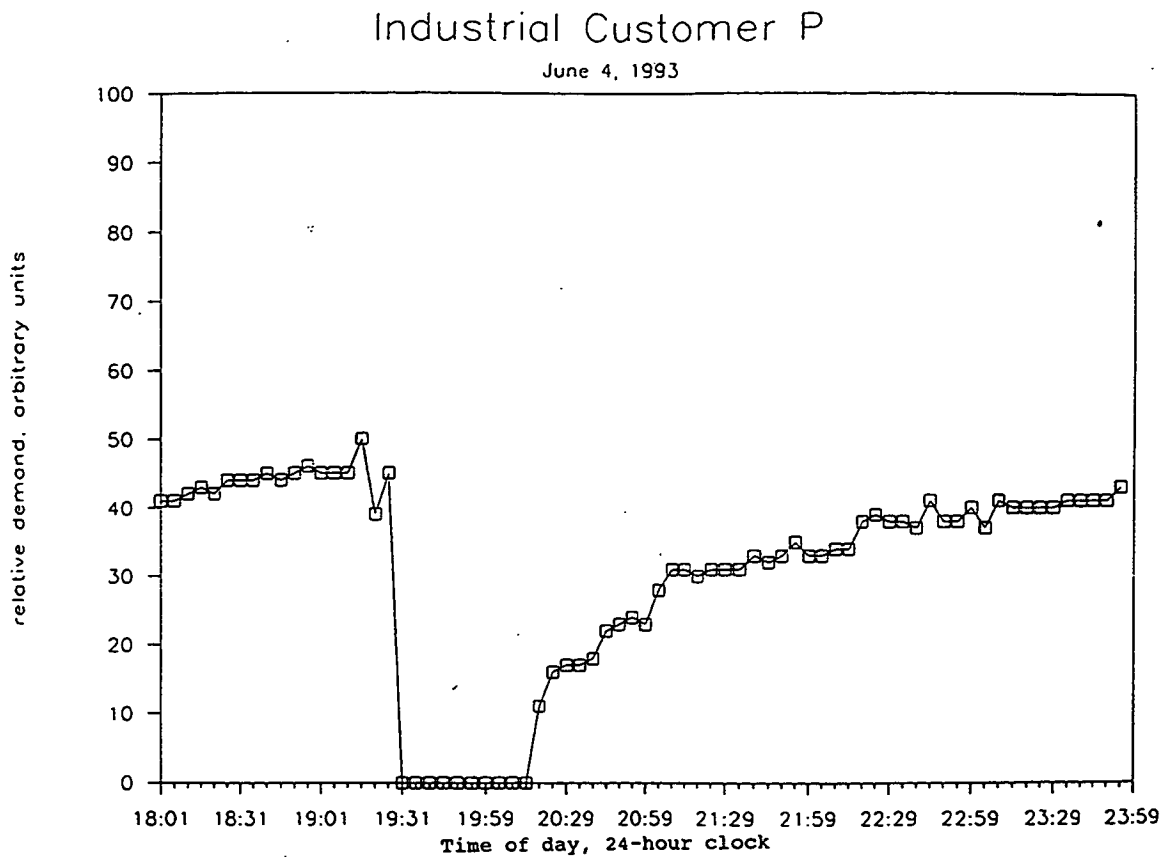
Most utilities today still map their outages by keeping track of incoming customer calls. When enough calls come in (and many can't get through, or are redundant), the utility can deduce the location of the fault and send a repair crew. The process is slow and costly, causing both lost revenue and customer dissatisfaction. Our field trials showed that a real-time system can entirely eliminate the need for customer call-ins.

On Saturday 3 June 1993 severe thunderstorms moved through eastern North Carolina, with several funnel clouds, intense lightning, golf- and baseball-sized hail, and numerous broken power poles. In one of our test towns 80% of the distribution system went out of service. But our telephone-powered modem units continued to operate normally through the outage, and in real time our computer in a neighboring town monitored the outages at three industrial plants. Figure 2 on the next page shows two of these; the third lasted fifty-eight hours and its plot, though complete, is really wide and unexciting.

With our system's outputs fed into a utility's AM/FM/GIS system, an outage map can be produced in seconds--including a fault analysis and a list of parts for the repair crew. As a useful side benefit we can immediately, accurately, and automatically measure how long each customer was without power (useful for quality-of-service analyses) and how much revenue was lost.

Real-Time Measurements of DSM Actions

Earlier we suggested that real-time communication could provide not only confirmation, but measurement of remote DSM actions. Figure 3 gives an example from an industrial plant.



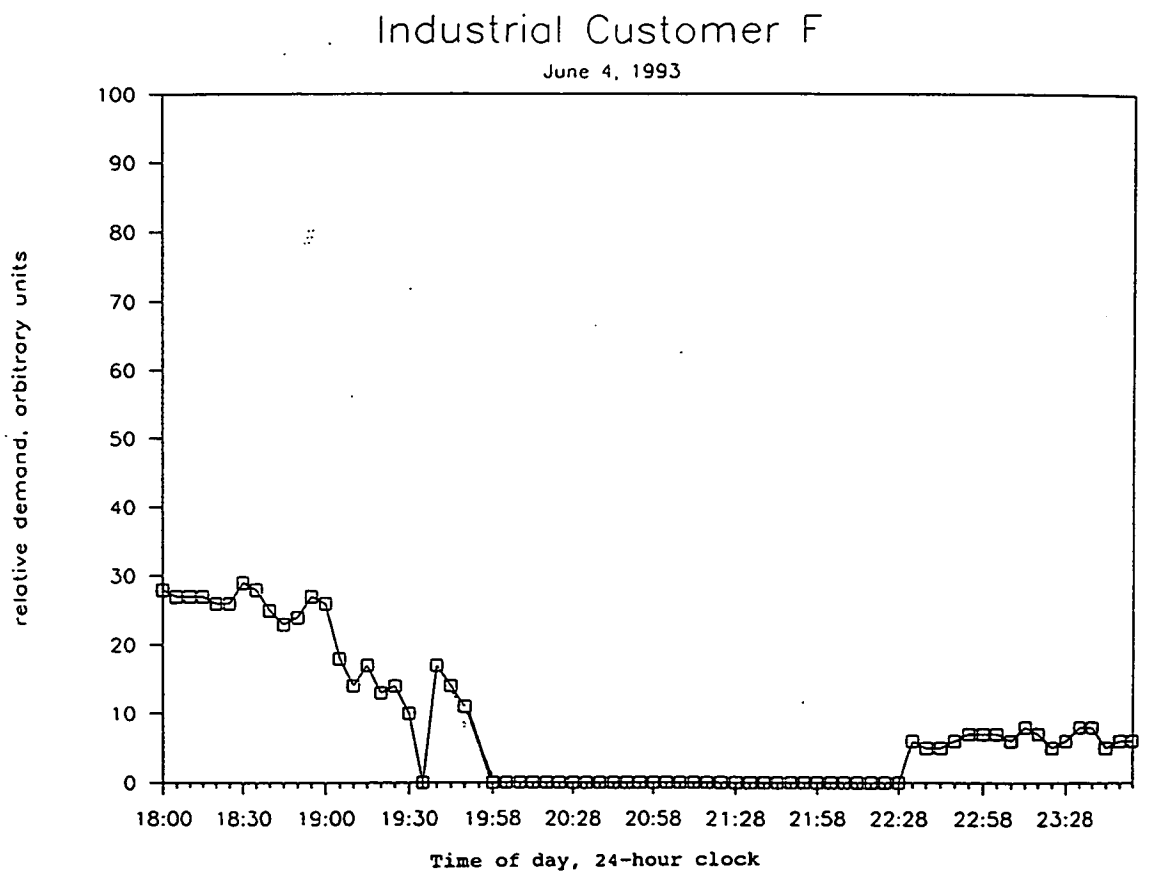


Figure 2: Real-Time Profiles of Power Outages

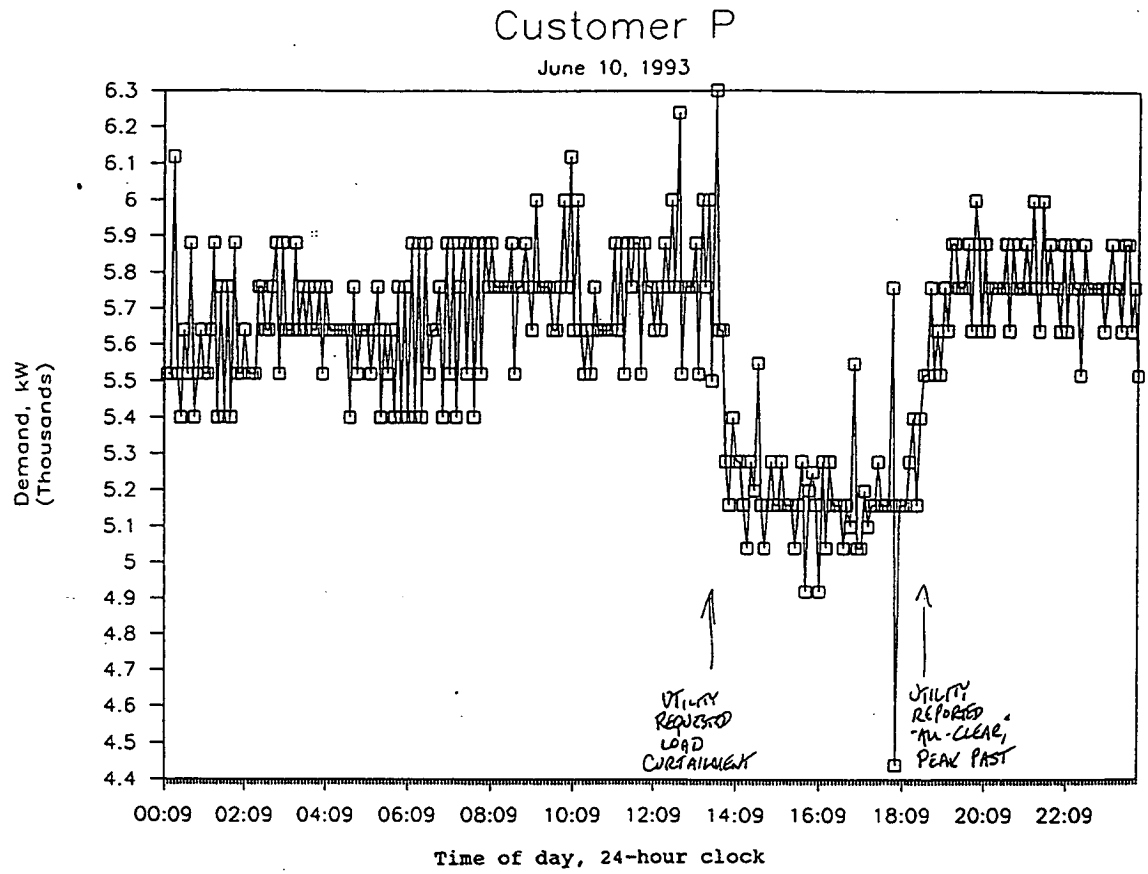


Figure 3: Real-Time Measurement of Remote DSM Results

A "coincident peak" rate gives this 5-6MW customer a strong incentive to shed load upon telephoned requests from the utility. On this day the utility called at 1:30PM; the plant reduced its load 10% within fifteen minutes, keeping it down until the utility "all clear" four hours later. Our computer watched this in real time from another town via the customer's telephone line. The data provide an immediate, objective and accurate measurement of the results and a record of the transaction, documenting both the customer's savings and the utility's avoided load.

Reliability

Despite measured internal temperatures above 130F and two years of exposure to high humidity, lightning and salt spray, our prototype system has shown communications reliability well above 99.6% for prolonged periods. Further improvements are underway.

Other Features and Projected Cost

In addition to collecting data in real time, our telemetry system can control switches at the meter site; can send commands, data or text to one or many meter sites at any time, without delay or interruption; and is under the complete control of the utility at all times. With appropriate sensors it can produce real-time maps of power factor, end-user voltage, gas pressure or other

variables. By collecting real-time profiles of gas demand, for example, it can help gas utilities minimize revenue loss and comply with "take-or-pay" contract provisions by precise management of interruptible accounts during periods of high demand.

Despite its capabilities and flexibility, in production quantities the installed cost will be below $100/point--comparable to that of many of today's "AMR-only" systems, and cost-effective even where only one utility participates.

BUSINESS CONSIDERATIONS

Although the benefits and economics of our real-time telemetry will soon be more fully defined in larger pilot projects, there is no question that it can help utilities cut costs, improve customer service and counter their growing competition. Nor is there any question that, given this new capability, telephone networks are up to the job. But how does this technology fit with the telcos' plans for converting their networks to broadband (optical fiber and/or coaxial cable)? And in a broader context, what does the future hold for utilities and telecommunications companies? Does UTILEX really expect utilities, telcos and cablecos to cooperate with each other? Indeed we do, and deregulation is creating unprecedented opportunities for them to do so for mutual profit, public benefit, and perhaps even survival.

What Utilities Need...

Already downsizing to prepare for competition, utilities need advanced telecommunications to cut costs still further and to improve customer service. As our technology is deployed and integrated into utility operations, the new data traffic it generates will steadily grow until it requires the speed and capacity of broadband networks. Indeed, utility applications are the clearest justification for such networks, as consumer markets are still unproven.

...and What They Are Being Forced to Do

But broadband networks will take several years to reach most customer sites (especially in small- to mid-sized municipalities and rural areas) and utilities can't afford to wait. With competition pressing, they urgently need to define the benefits, costs, pitfalls and regulatory issues of real-time telecommunications, then adopt it as quickly as is practical. Believing that presentday telephone and cable networks are not capable of real-time two-way operation--and knowing that their own wires are not--utilities are turning instead, often reluctantly, to wireless systems despite their cost, questionable reliability and lack of data security. (Some utilities are considering extending fiber cables of their own to each meter site; this is also slow and costly.)

What's at Stake for Telcos and Cablecos:

Often unfamiliar with utilities' needs and tightly focussed on converting to broadband systems, telcos and cablecos are in serious danger of irretrievably losing the potentially huge utility telecommunications market. With the steady erosion of this opportunity to wireless and other less suitable technologies, their futures grow increasingly dependent upon undemonstrated consumer markets--even though their own reliable, secure networks are already in place. Unable before now to use these networks for real-time two-way telemetry, they have had little alternative to their present course: building and deploying broadband systems for expected consumer markets, hoping to complete them before the utility market is foreclosed to them, and meanwhile attempting to discourage utilities from expanding their own fiber networks to compete with them. Now there is a bridge technology.

In addition to preserving their share of a strategic market there are other reasons for telecomm providers to cooperate with utilities. First, broadband systems will require several watts of electrical power at each subscriber location, and it can't be carried to the site by glass fiber. Second, utilities own most of the poles that broadband systems will need. And third, many utilities--electric and gas--have miles of their own optical fiber in place, and might offer their own competing broadband consumer services. We believe cooperation between utilities and telecomm providers is virtually inevitable: the mutual and public benefits of such cooperation are becoming too numerous and too important for them or their regulators to ignore.

But by permitting presentday copper networks to provide real-time telemetry, won't UTILEX's system compete with broadband?

UTILEX's Bridge to Broadband[tm]

Enormously fast and capacious, broadband systems are the major league of telecommunications for which our technology will serve as a system of farm clubs. To take full advantage of their virtues, and to repay their large upfront costs, broadband systems will need large amounts of data traffic. Our bridge technology permits utilities in all areas--rural, suburban and urban--to explore the uses, benefits, costs and regulatory issues of real-time telemetry now without waiting for broadband. As real-time telemetry becomes more familiar, utility data traffic will grow into a large, stable base of traffic which will help justify the expansion of broadband into otherwise hard-to-serve areas.

And because it permits existing, inexpensive utility meters to perform complex functions previously requiring costly special-purpose equipment, the UTILEX system can reduce utilities' costs while providing an economical, low-risk way to integrate real-time telemetry seamlessly into their operations.

ACKNOWLEGEMENTS

We deeply appreciate DOE, its SBIR Director Sam Barish, our technical contract officer Sam Tagore, and our North Carolina utility and telco co-hosts Rocky Mount Utilities, Brunswick EMC, Atlantic Telephone and Carolina Telephone for their help within a truly excellent SBIR program. Special thanks are due to David W. Paterson, whose high-quality work, enthusiasm and dedication are unmatched; to Ron Genova for his invaluable consultation, friendship and support; and to Delane Stanley, Don Gardner, Bill Sherk, Lewis Shaw, Tom Wason and Bob Scoggin for their help in more ways than can be described here. Finally but not least, our corporate hat is off to our excellent North Carolina suppliers Fawn Industries and Logical Design Group for the first-class prototypes they made for us, and to the Electrical Engineering Department of North Carolina State University for advice and lab space when we needed it. To those inadvertently omitted, our apologies and thanks nonetheless.

HARD COPIES OF PLOTS

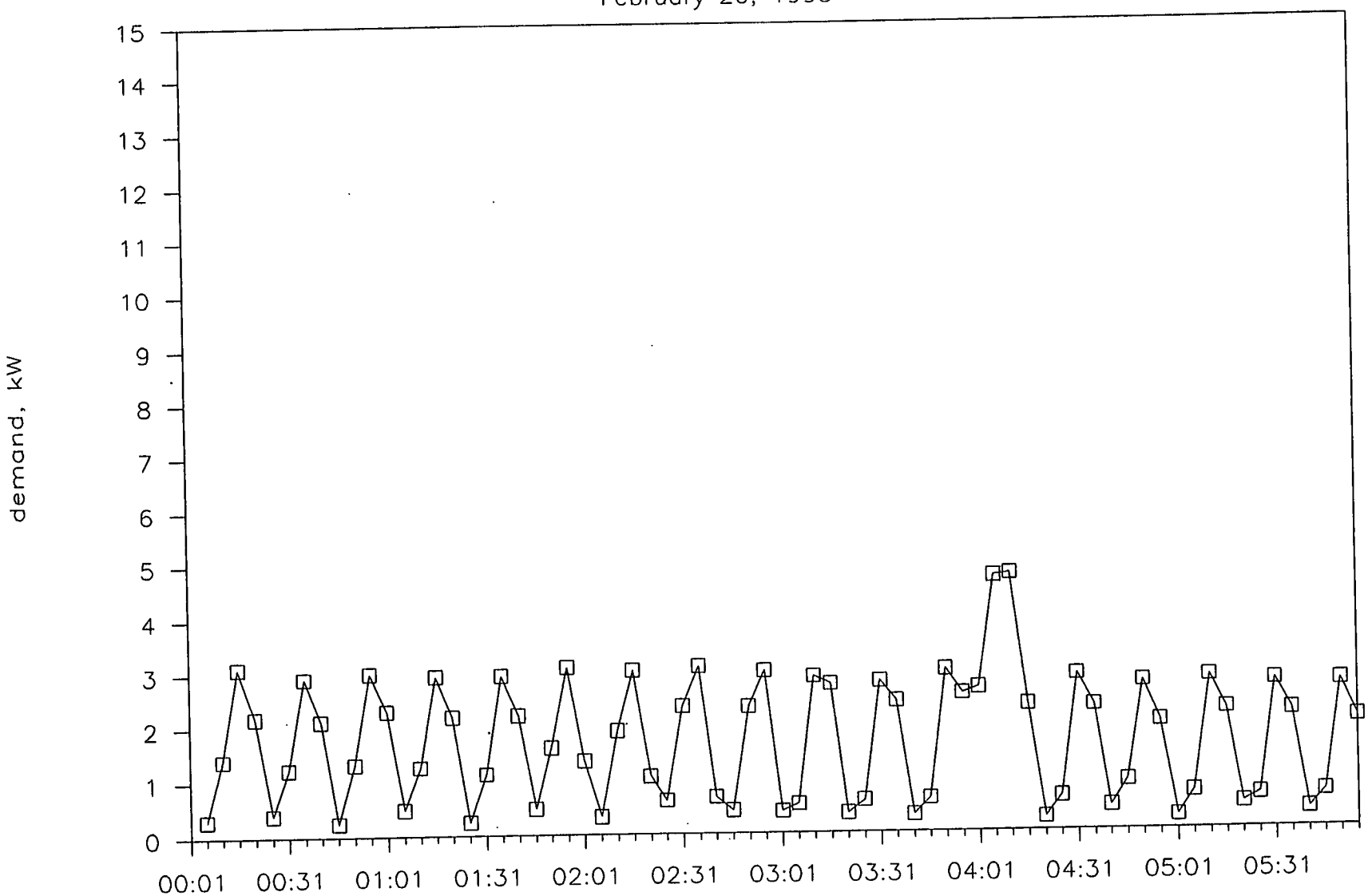
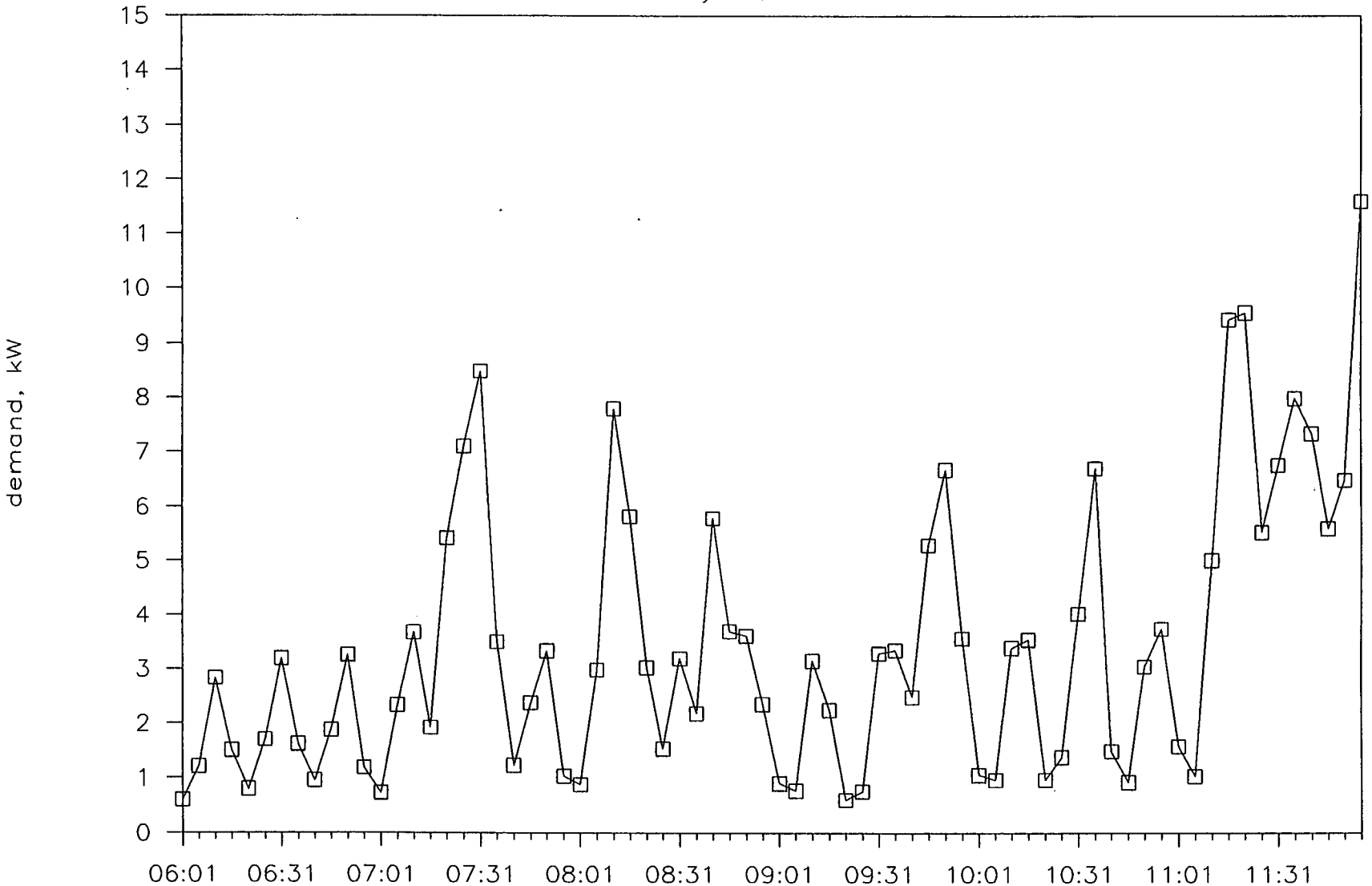

House R3
February 20, 1993
demand, kW
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
00:01
00:31
01:01
01:31
02:01
02:31
03:01
03:31
04:01
04:31
05:01
05:31

Fig 1.4

House R3

February 20, 1993



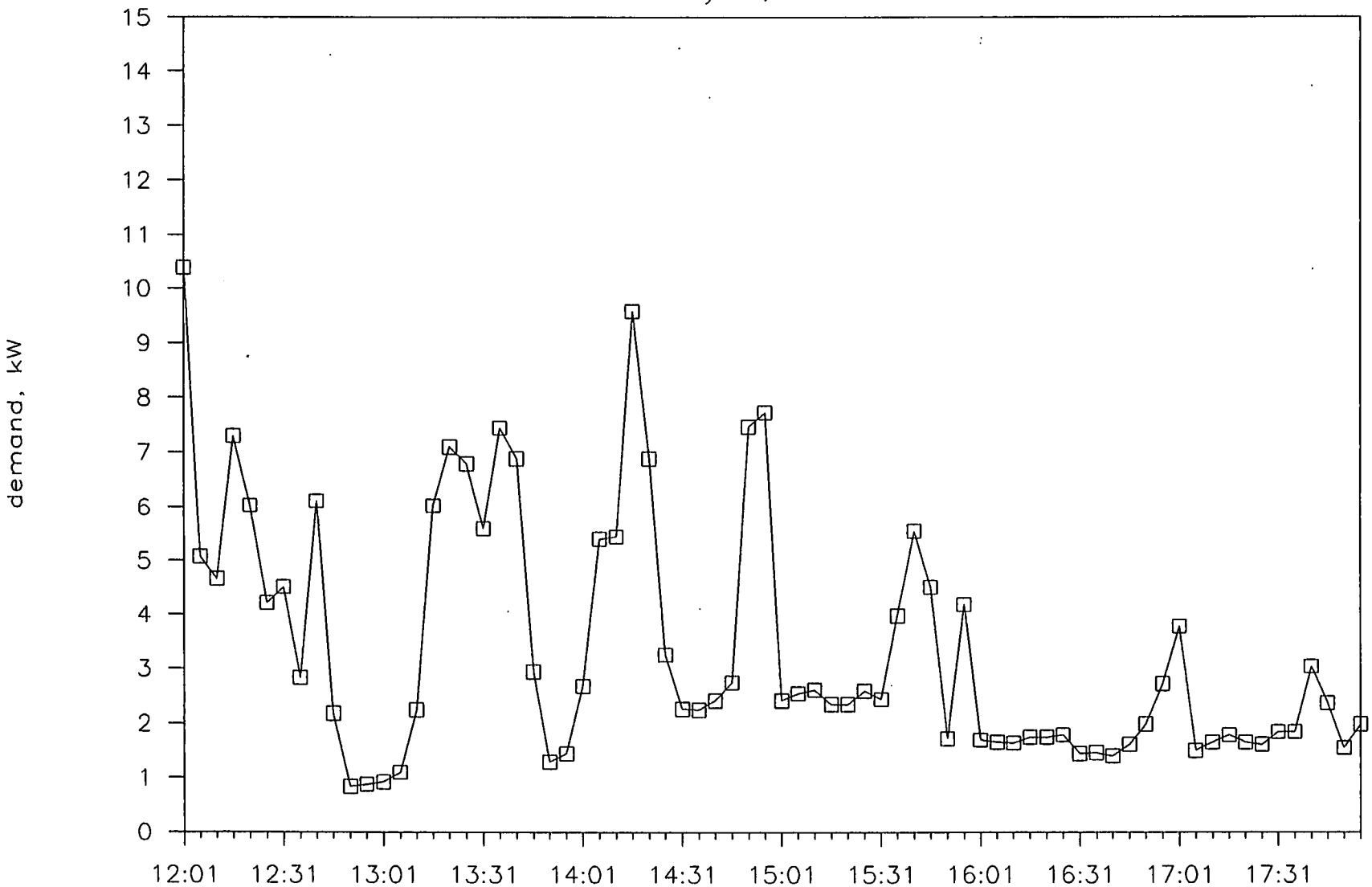

House R3
February 20, 1993
demand, kW
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
12:01
12:31
13:01
13:31
14:01
14:31
15:01
15:31
16:01
16:31
17:01
17:31

Fig 1c

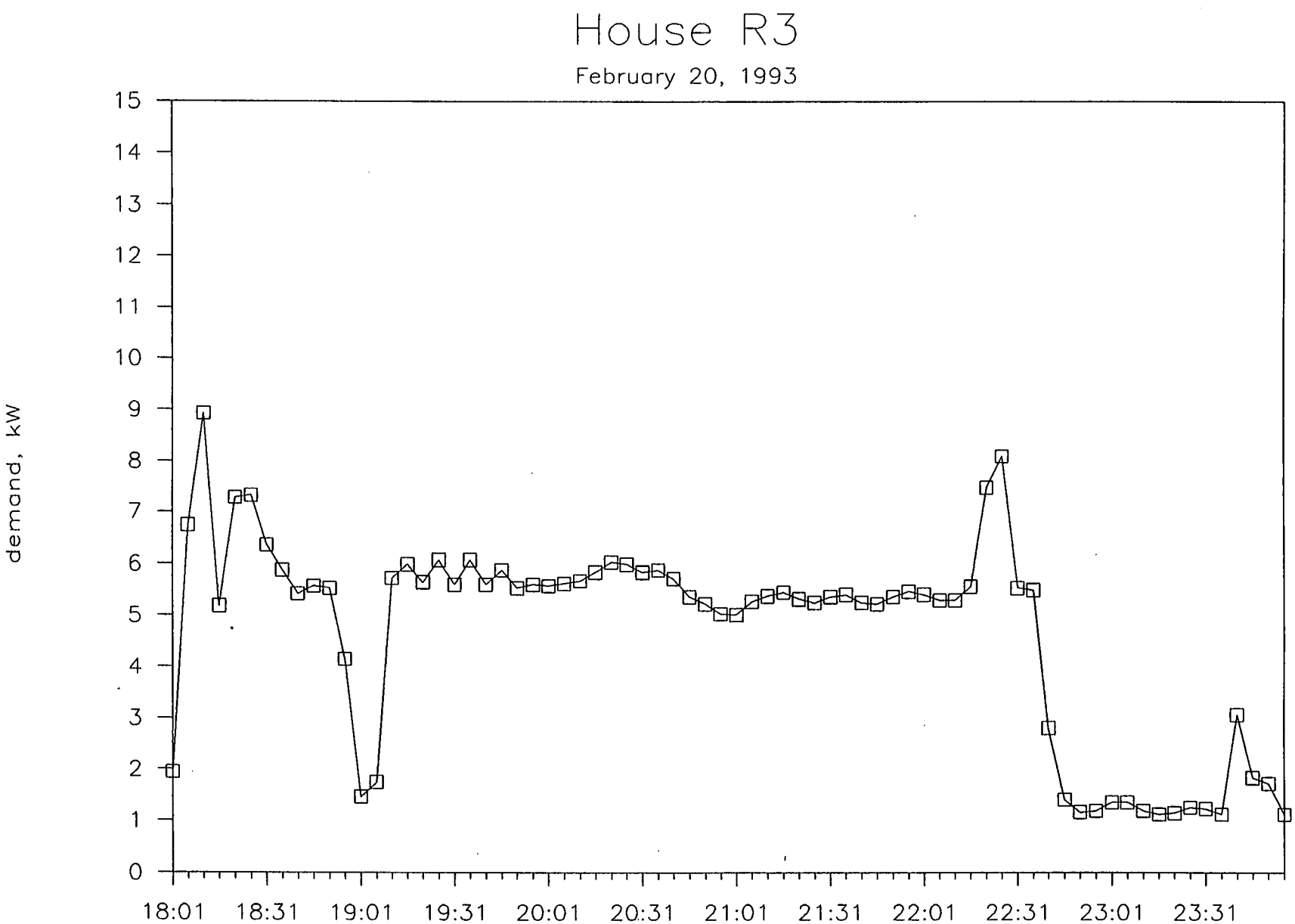

House R3
February 20, 1993
demand, kW
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
18:01
18:31
19:01
19:31
20:01
20:31
21:01
21:31
22:01
22:31
23:01
23:31

Fig 1

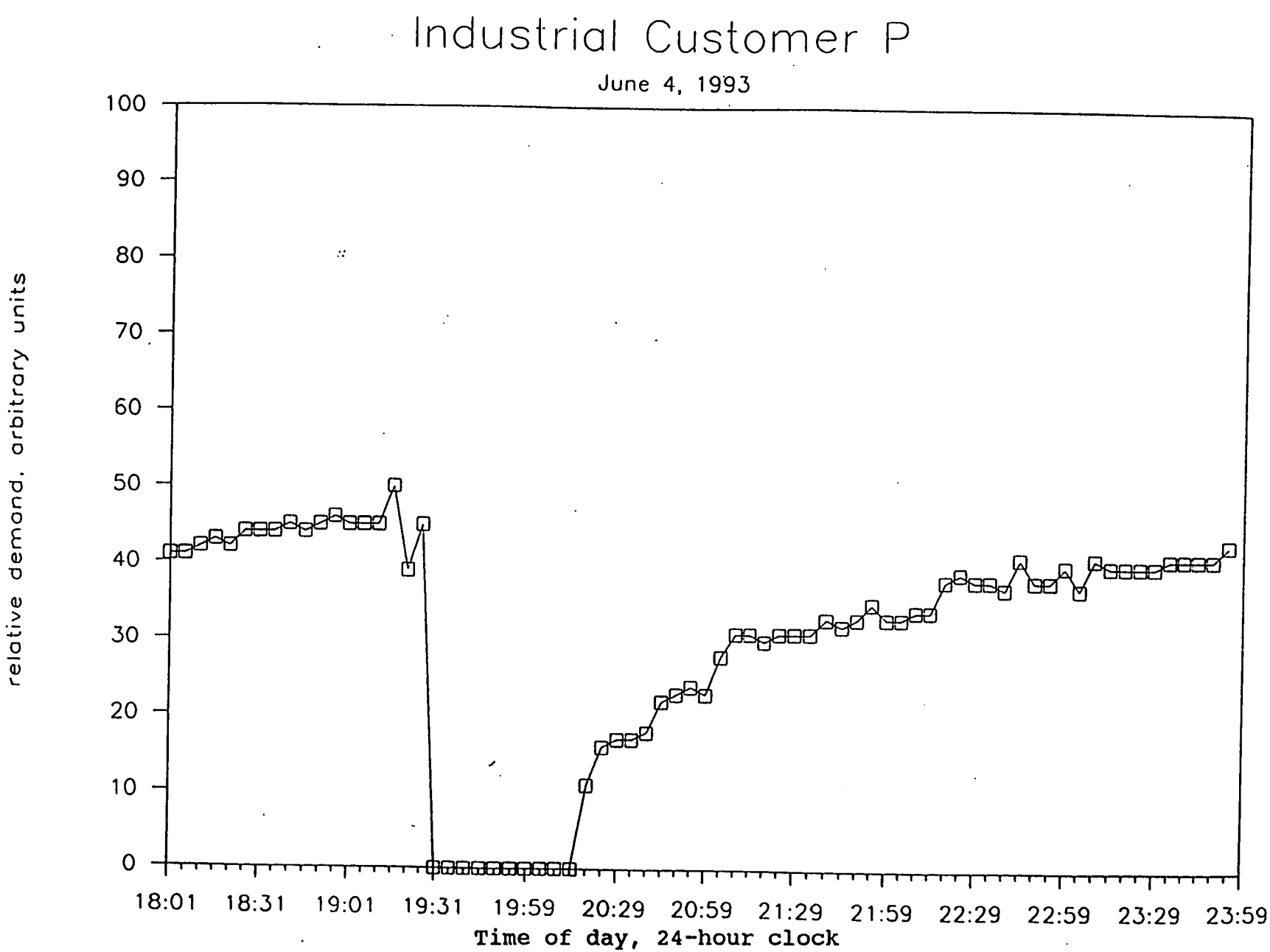

Industrial Customer P
June 4, 1993
relative demand, arbitrary units
100
90
80
70
60
50
40
30
20
10
0
18:01
18:31
19:01
19:31
19:59
20:29
20:59
21:29
21:59
22:29
22:59
23:29
23:59
Time of day, 24-hour clock

FIG 2A

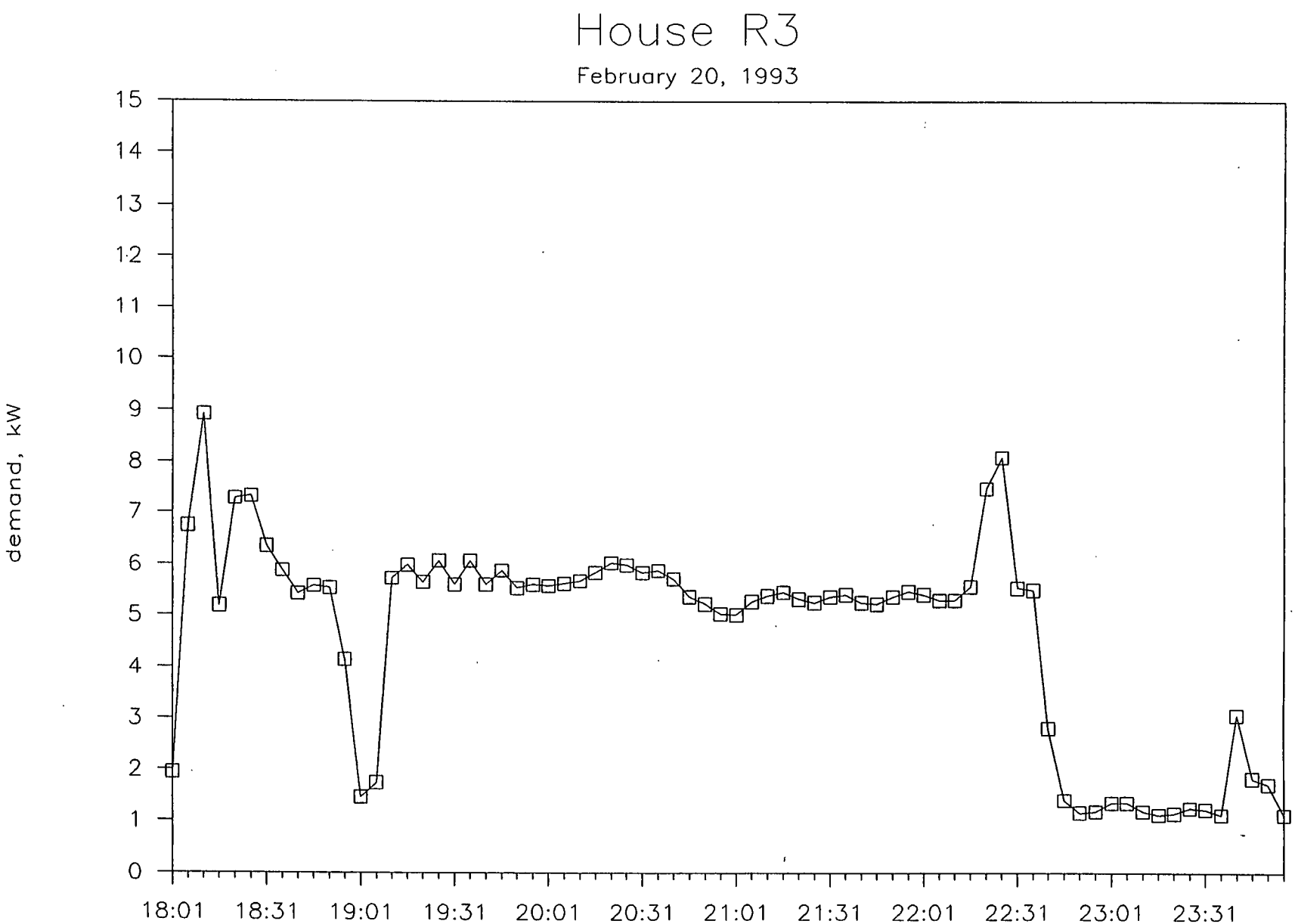

House R3
February 20, 1993
demand, kW
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
18:01
18:31
19:01
19:31
20:01
20:31
21:01
21:31
22:01
22:31
23:01
23:31

Fig 1.D

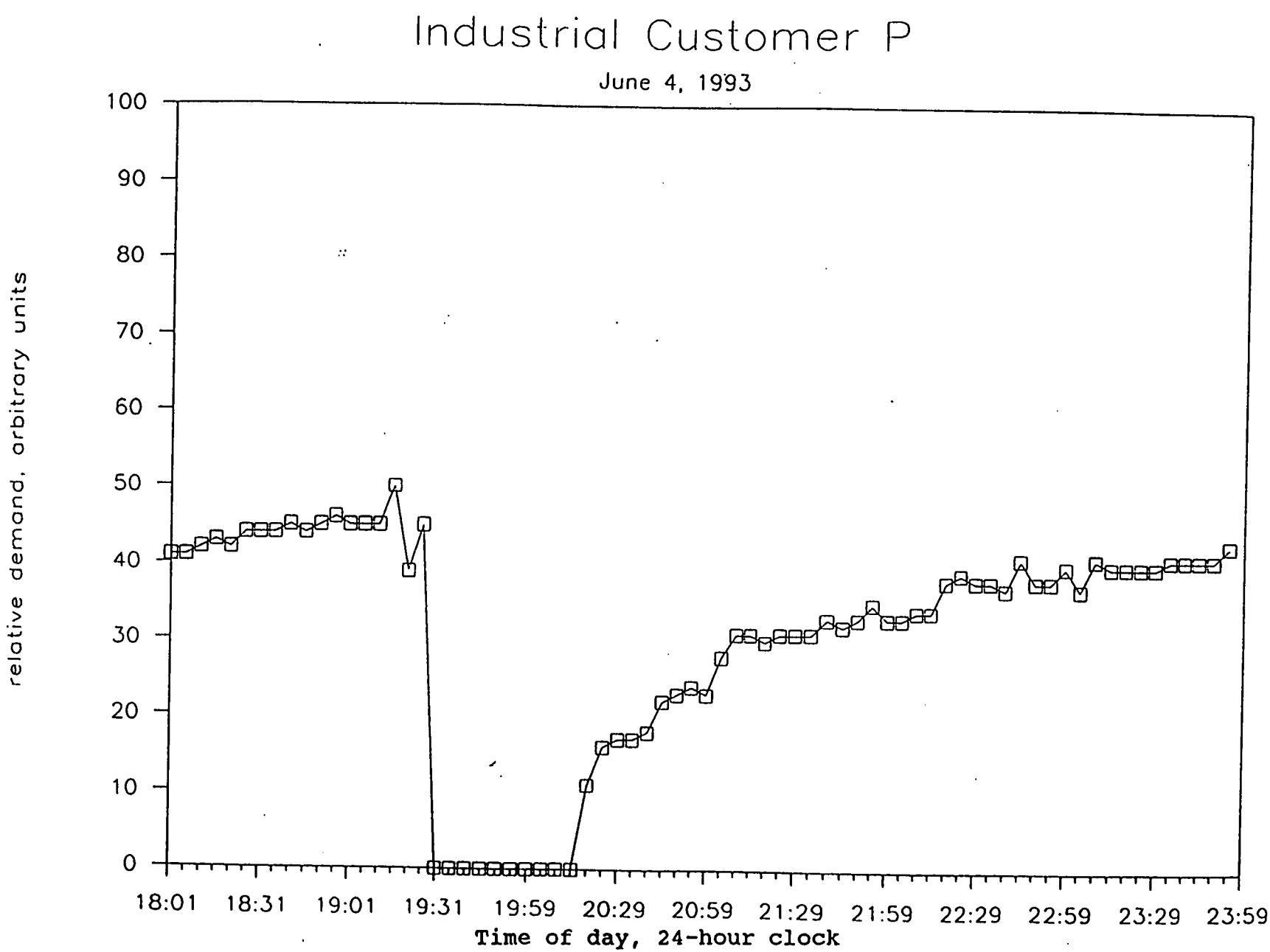

Industrial Customer P
June 4, 1993
relative demand, arbitrary units
100
90
80
70
60
50
40
30
20
10
0
18:01
18:31
19:01
19:31
19:59
20:29
20:59
21:29
21:59
22:29
22:59
23:29
23:59
Time of day, 24-hour clock

FIG 2A

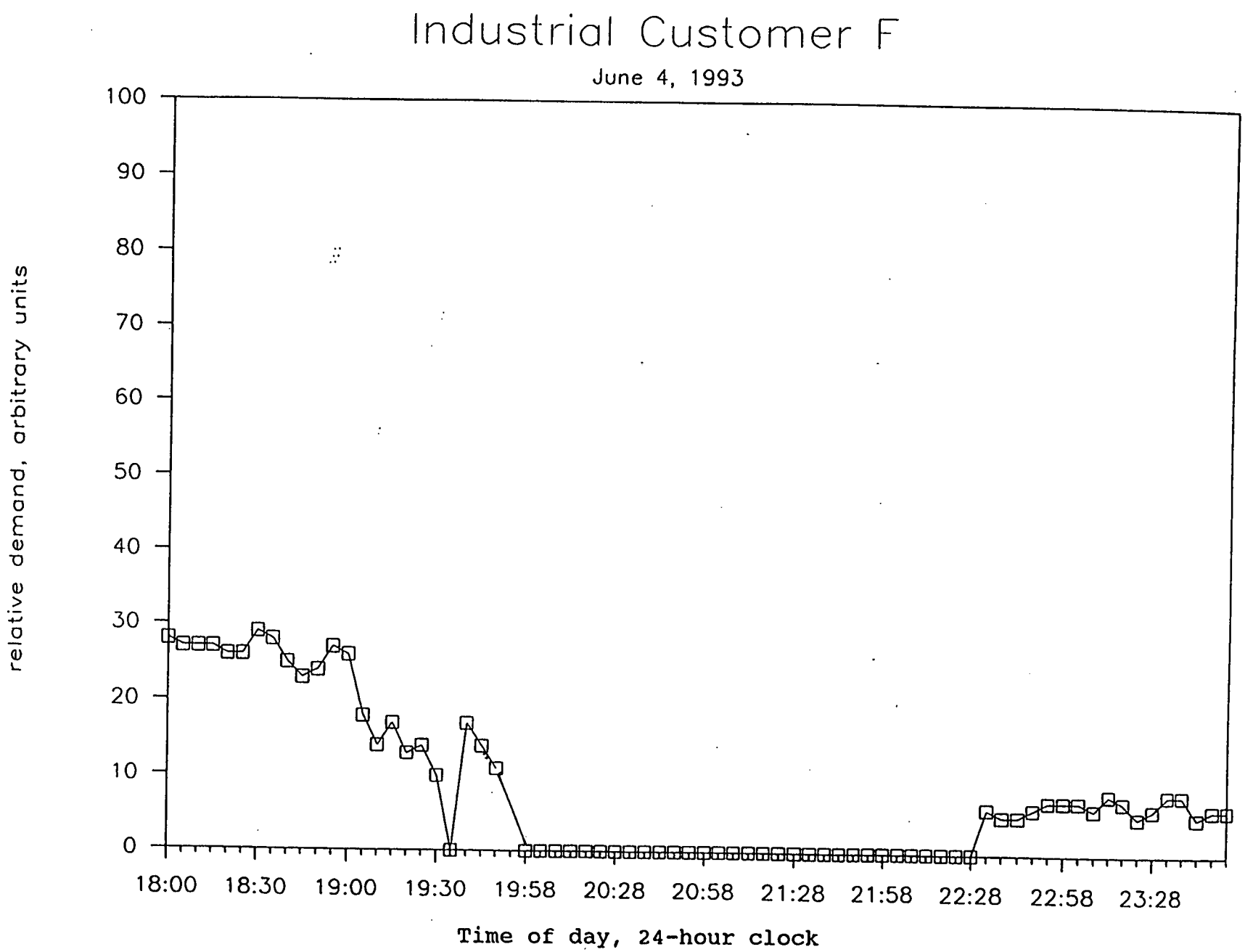

Industrial Customer F
June 4, 1993
relative demand, arbitrary units
100
90
80
70
60
50
40
30
20
10
0
18:00
18:30
19:00
19:30
19:58
20:28
20:58
21:28
21:58
22:28
22:58
23:28
Time of day, 24-hour clock

Fig 2B

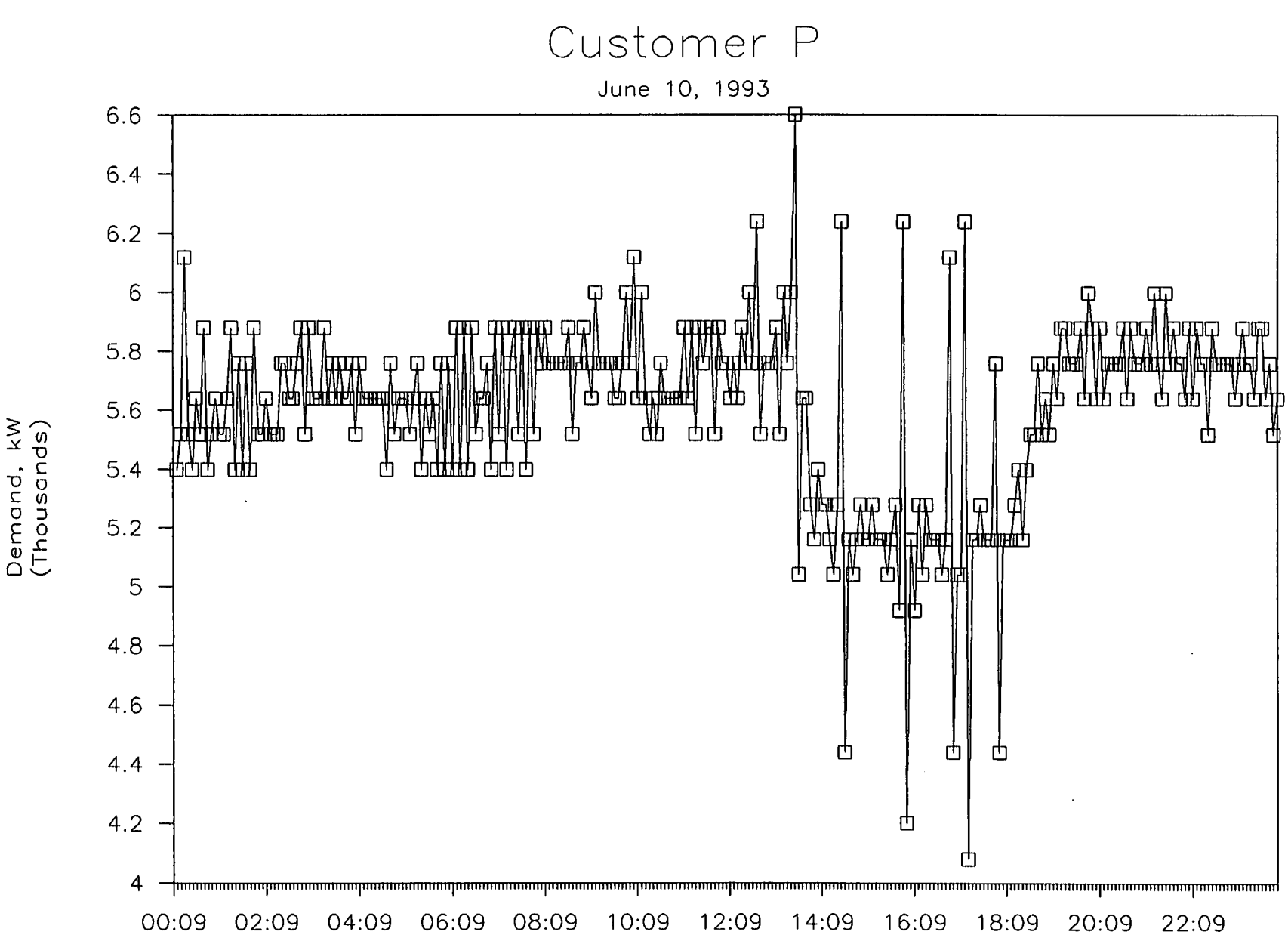

Customer P
June 10, 1993
Demand, kW
(Thousands)
6.6
6.4
6.2
6
5.8
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
00:09
02:09
04:09
06:09
08:09
10:09
12:09
14:09
16:09
18:09
20:09
22:09

FIG. 3

Test and Measurement

OPTICAL BINARY MIXTURE METER

Robert A. Cortese
Electronics Engineer
U. S. Bureau of Mines, Pittsburgh Research Center
P.O. Box 18070, Pittsburgh, PA 15236

Carrie E. Lucci
Research Physicist
U. S. Bureau of Mines, Pittsburgh Research Center
P.O. Box 18070, Pittsburgh, PA 15236

ABSTRACT

The U.S. Bureau of Mines (USBM) has developed a portable meter for measuring the composition ratio of binary mixtures whose components have different optical reflectivities. Meter operation is based on the intensity of near-infrared light reflected from the surface of a sample consisting of a mixture of highly reflective and poorly reflective components. The meter consists of an electronics package connected to a handheld cylindrical probe, approximately 5 inches long and 3/8 inch in diameter, containing a light source and detector which are located behind a hemispherical glass window at the probe tip. Laboratory studies conducted with dry mixtures of coal dust and pulverized limestone show that the measured mixture composition agrees with the rock dust content, as determined by chemical analysis, within ± 1 wt pct. Extensive studies conducted by other government agencies and commercial organizations have reported similar results on air-dried mine samples. Studies in which the meter was used to determine the progress of a chemical reaction in which carbon black is fluorinated to produce a white product were equally successful.

INTRODUCTION

U.S. Federal law mandates that the coal dust deposited in U.S. underground coal mines during mining operations be covered with rock dust such as pulverized limestone, which acts as a flame inhibitor in the event of a mine explosion. The total incombustible content of the resulting mixture must be at least 65 wt pct rock dust in the entries and ≥ 80 wt pct in the returns where the coal dust is expected to be finer. Periodically, Mine Safety and Health Administration (MSHA) inspectors take samples of coal and rock dust mixtures and send them for chemical analysis to ensure compliance with the regulations. However, the processing time for these samples is approximately two weeks during which time the mine could be operating under hazardous conditions. To provide a fast, in situ determination of the rock dust content of dry homogeneous coal and limestone dust mixtures, the U.S. Bureau of Mines (USBM) developed a handheld optical binary mixture meter (OBMM). Research on a version of this meter which corrects for the presence of

moisture in coal samples is currently being conducted [1]. The subsequent sections detail the design, theory, and evaluation of the original meter.

METER DESIGN

Meter operation is based on the measurement of the intensity of near-infrared light reflected from the surface of a dust sample consisting of a mixture of highly reflective and poorly reflective constituents. Although originally designed for the coal mining industry, this meter can be used with any binary mixture of powders or liquids whose components have significantly different optical reflectivities.

Figure 1 shows a cross-section of the probe used to measure the optical reflectance of binary samples. When the probe is inserted in the mixture, infrared radiation emitted by a gallium-arsenide light emitting diode (940 $\pm$ 20 nm), located in the optical module, is transmitted through the glass window to the mixture's surface. The light is back-scattered to the PIN photodiode also located in the optical module. The amount of radiation back-scattered from the sample surfaces increases with the concentration of the more reflective component.

The normalized reflectance, Φ, is related to the mass fraction of the sample's highly reflective constituent, f_1, by the following theoretical Equation: [2]

$$\Phi \equiv \frac{I_x - I_d}{I_1 - I_d} = \frac{f_1}{f_1 + K\cdot(1-f_1)} \quad , \tag{1}$$

where I_x is the intensity of light reflected from the binary mixture's surface, I_d is the intensity of light reflected from the poorly reflective component's surface, I_1 is the intensity of light reflected from the highly reflective component's surface. K is a dimensionless constant which increases as the particle size ratio of the poorly reflective component to the highly reflective component decreases. Equation 1, is solved to determine the mass fraction of the light-colored component, as follows:

$$f_1 = \frac{K\cdot(I_x - I_d)}{K\cdot(I_x - I_d) + I_1 - I_x} = \frac{\Phi\cdot K}{\Phi\cdot(K-1)+1} \quad . \tag{2}$$

Figure 2 shows typical results for three different coals each combined with the same pulverized limestone, and the corresponding values of K. The figure illustrates the excellent fit of Equation 1 (solid traces) to the measurements (data points).

Although analog and digital versions of the electronics for the meter have been developed (Fig. 3), both versions use the same optical sensor and transconductance amplifier. The analog version utilizes operational amplifier stages in conjunction with an analog divider chip to solve Equation 2. This version requires potentiometers for calibration. The digital unit digitizes the optical signal to enable microcomputer processing.

The theoretical equations are programmed into microcomputer memory and calibration parameters are read and stored in memory without the need of potentiometers.

METER EVALUATION

Before using the meter, it must be calibrated with the pure binary constituents and a mixture of known concentration preferably in the range in which the meter is to be used. This calibration procedure fixes the transconductance amplifier voltages corresponding to the intensity reflected from the two pure components, I_d and I_l, and the value of K in Equation 1.

Laboratory Studies

To determine the functionality of the meter, the meter was used to measure the rock dust concentration in mixtures containing known ratios of Pittsburgh Pulverized Coal (PPC) and limestone rock dust prepared in the laboratory. Twenty-one samples were prepared in 5 wt pct limestone increments from 0 to 100 wt pct limestone. The unit was calibrated, prior to making the measurements, using the pure constituents and a mixture containing 50 wt pct limestone dust. Once calibrated, the OBMM was used to measure the limestone percentage of the remaining eighteen samples. The results of these studies are shown in Figure 4, and as shown, the meter measurements agree with the results of chemical analysis within ± 1 wt pct limestone. Additional mixtures prepared using different coals were similarly tested and displayed comparable agreement [3]. Encouraged by these results, the USBM began field testing the instrument in other mining and industrial applications.

Mine Samples

At the request of the National Office of Coal Mine Safety and Health, an OBMM was provided to the Industrial Safety Division (ISD) of MSHA. The ISD evaluated the meter for in situ measurement of the limestone content of mine samples collected by MSHA inspectors in underground mines. Before measuring field samples, the meter was calibrated using field samples whose composition had been previously determined through chemical analysis. Samples from six underground mines were measured with the OBMM and then sent to the MSHA laboratory. The measurements from one of the six mines studied is presented in Figure 5 [4]. The results shown in the figure are typical for all six mines.

MSHA evaluators concluded that the OBMM was a valuable supplement to laboratory analysis, and recommended that the meter be used as a screening tool to reduce the number of samples which require chemical analysis to those within ± 5 wt pct rock dust of the federally mandated limit. In addition, and more importantly, obtaining an in situ measurement of the limestone content reduces the hazards of operating in unsafe conditions while waiting for laboratory results.

An OBMM was also loaned to the USX Cumberland Coal Mine in Pennsylvania. Ninety dust samples (approximately 100 g) were collected from all parts of the mine.

USX employees analyzed these samples using both the OBMM and alcohol volumetric analysis [5]. Comparison of these results is shown in Figure 6. The OBMM results agreed with the results of volumetric analysis to within ± 2 wt pct limestone. USX estimated that the meter would reduce the cost of sampling and analyzing mine samples by 65 pct. Similar testing by the Mathies Coal Mine in Washington County, Pennsylvania, produced comparable results.

Fluorination of Carbon

The Allied Chemical Corporation (ACC), which operates a carbon fluorination facility that converts carbon black into a white fluorinated product, requested an OBMM to monitor the quality of their product during various stages of the reaction [5]. Twelve samples, including the pure constituents and mixtures, known to within ± 0.1 wt pct fluoride, were sent to the USBM for analysis with an OBMM. Following this initial study, the USBM loaned ACC an OBMM to monitor the progress of the fluorination process. Figure 7 shows a graph of the fluorocarbon content measured by ACC as compared to OBMM measurements. The data shows that the meter measurements and ACC chemical analysis agreed to within ± 1 wt pct.

CONCLUSION

The USBM has developed a portable optical binary mixture reflectivity meter for rapid in situ determination of the composition of homogenous binary mixtures. It has been demonstrated through laboratory studies and field testing that the infrared reflectivity of a binary mixture of constituents with significantly different optical reflectivities can be used to measure the mixture composition. The device is simple to use and requires only a small representative dry particulate or liquid sample (approximately 30 ml).

The adoption of the meter for use in the coal industry will benefit both the government inspection agencies and the mining industry. A prototype of the instrument was successfully used in the field by USX Cumberland Mine, Mathies Mine, and by the ISD of MSHA in six different mines. Response from these studies suggests that the OBMM would enable an inspector to decide when to collect mine samples for laboratory analysis. Samples whose OBMM measured rock dust content falls within ± 5 wt pct rock dust of the federal requirement would be sent for chemical analysis thus reducing the number of samples taken. It would also provide the mine operator an alternative to visual inspection to determine when to rock dust the mine which would aid in maintaining safe working conditions.

The current version of the meter is susceptible to moisture which lowers the optical reflectivity of the sample, and consequently results in a lower rock dust reading than would be obtained if the sample were dry. To alleviate this problem, the USBM has designed and is currently evaluating a moisture correction module which simultaneously measures the sample's optical reflectivity and electrical conductance to determine and correct for the moisture content of a sample [1].

Originally designed to measure the limestone concentration in coal mines, the meter has been found to be useful in a variety of other applications including the chemical industry. Favorable results were reported by Allied Chemical Company in determining the progress of a carbon black fluorination reaction.

REFERENCES

1. Cortese, R. and H. Perlee. Electrical Conductivity of Moist Coal Dust/Rock Dust Layers. Paper in Proceedings of 11th West Virginia University International Mining Electrotechnology Conference, Morgantown, WV, July 29-30, 1992.
2. Perlee, H., J. Pinkerton, and M. Sapko. Monochromatic Reflectance Meter. Paper in Proceedings IEEE Annual Meeting, Denver, CO, September 29 to October 3, 1986, Vol. II.
3. Woods, T., M. Sapko, H. Perlee. Analog Optical Rock Dust Meter. Paper in Proceedings, Mining Industry Committee of IEEE Industry Society 23rd Annual Meeting. Pittsburgh, PA, 1988.
4. Hoffman, W. A. An Evaluation of the OCENCO Portable Rock Dust Analyzer. Mine Safety and Health Administration, Industry Safety Department, Report No. 04-183-90, July 2, 1990.
5. Sapko, M., N. Greninger, and H. Perlee. Optical Rock Dust Meter Field Evaluation. USBM RI 9385, 1991.

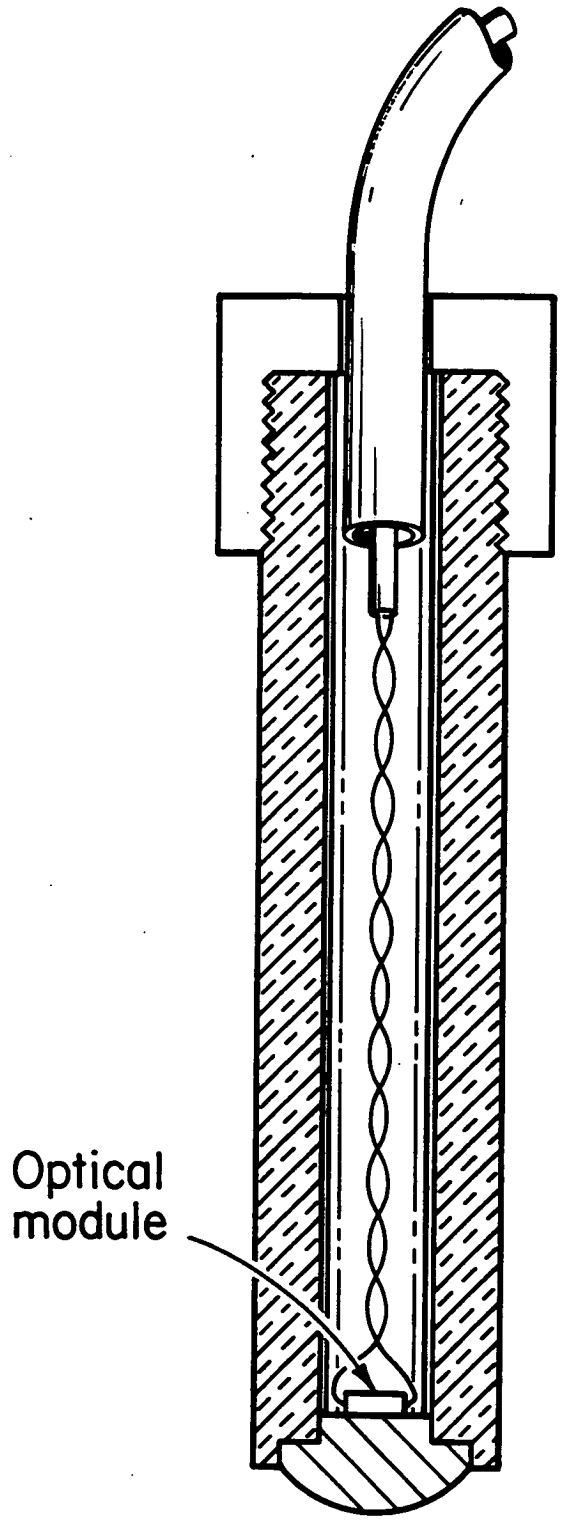


Figure 1. Cross-section of optical probe

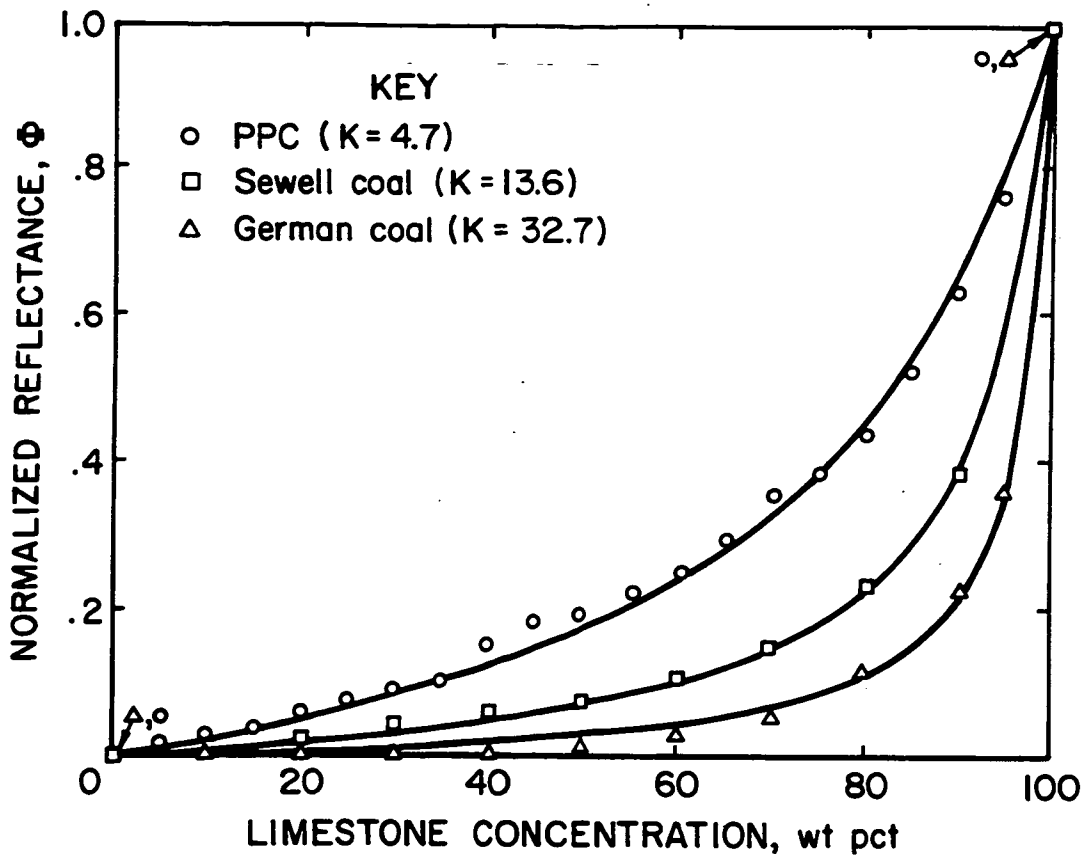


Figure 2. Calibration data for three coals with limestone rock dust. Solid traces are theoretical fits.

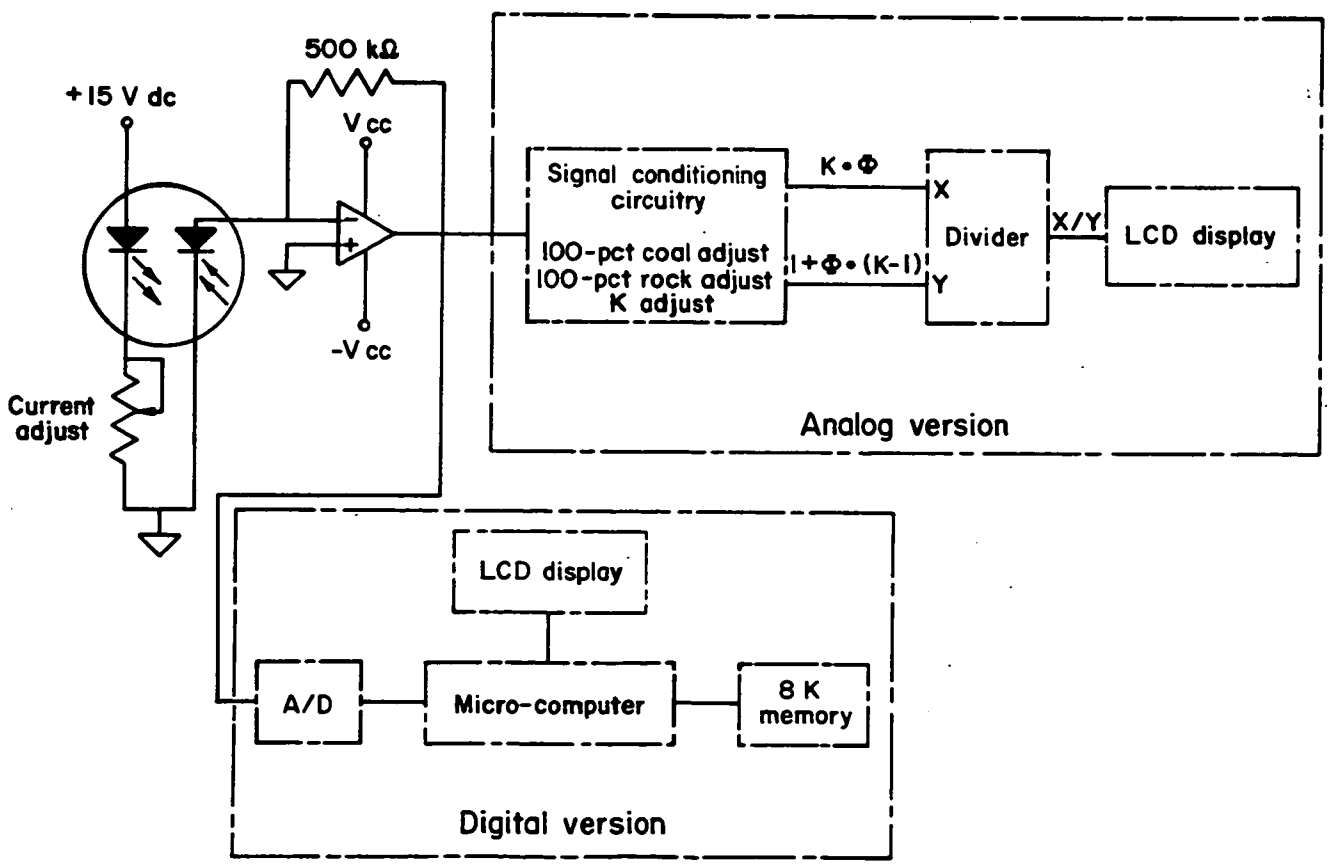


Figure 3. Block diagram of the analog and digital versions of the OBMM.

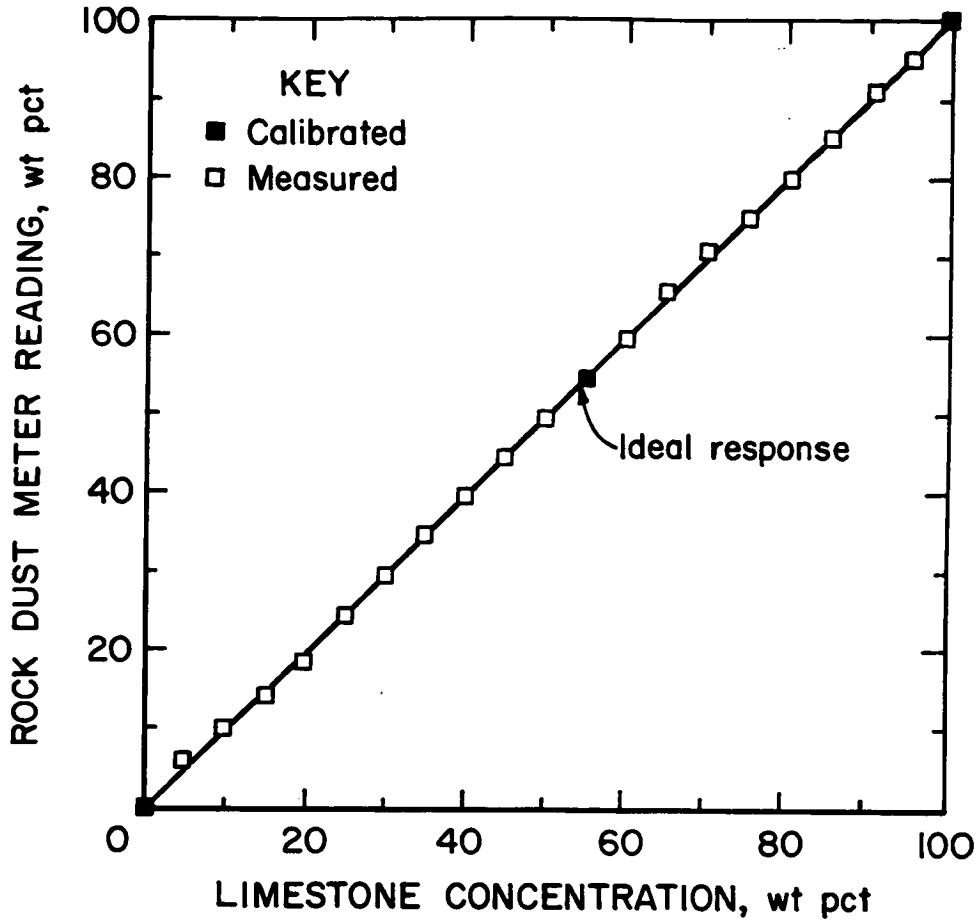


Figure 4. Comparison of actual limestone concentration and OBMM readings for PPC laboratory prepared samples.

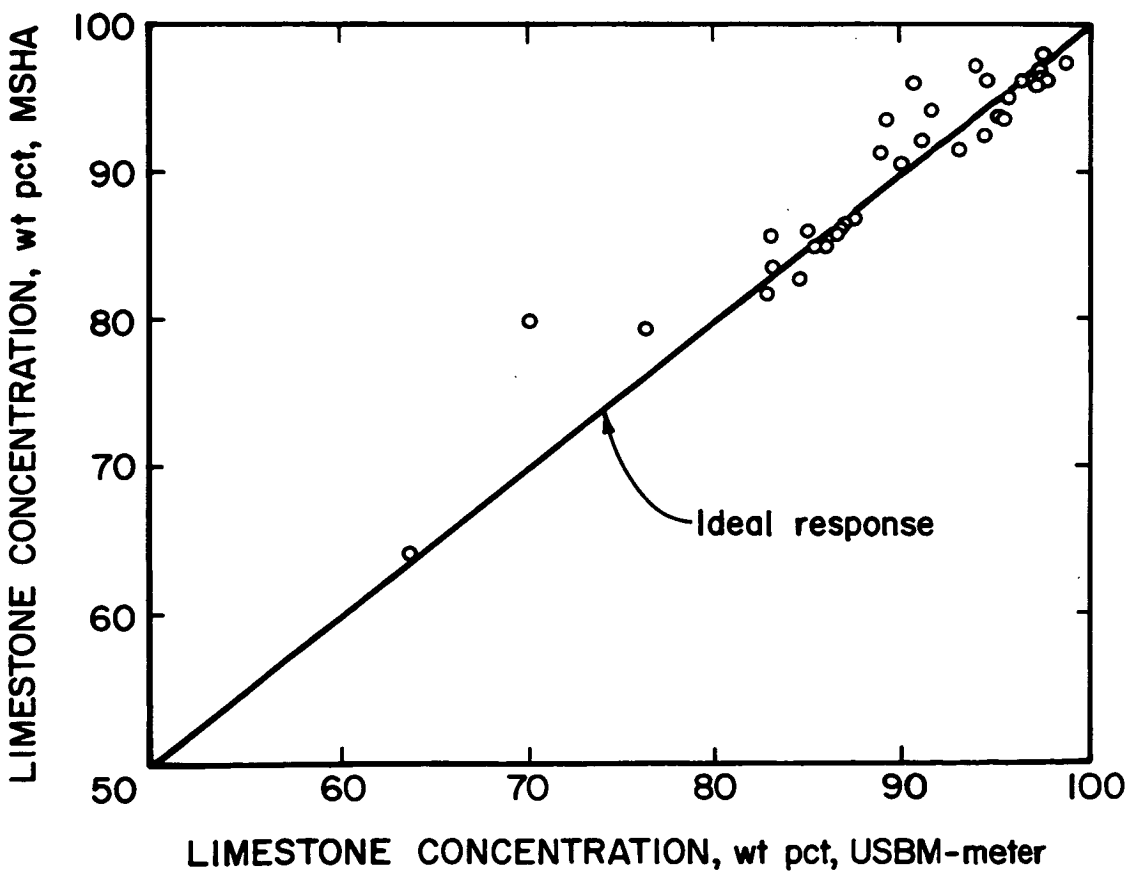


Figure 5. Comparison of MSHA chemical analysis and OBMM readings for one of the six mines in the MSHA study.

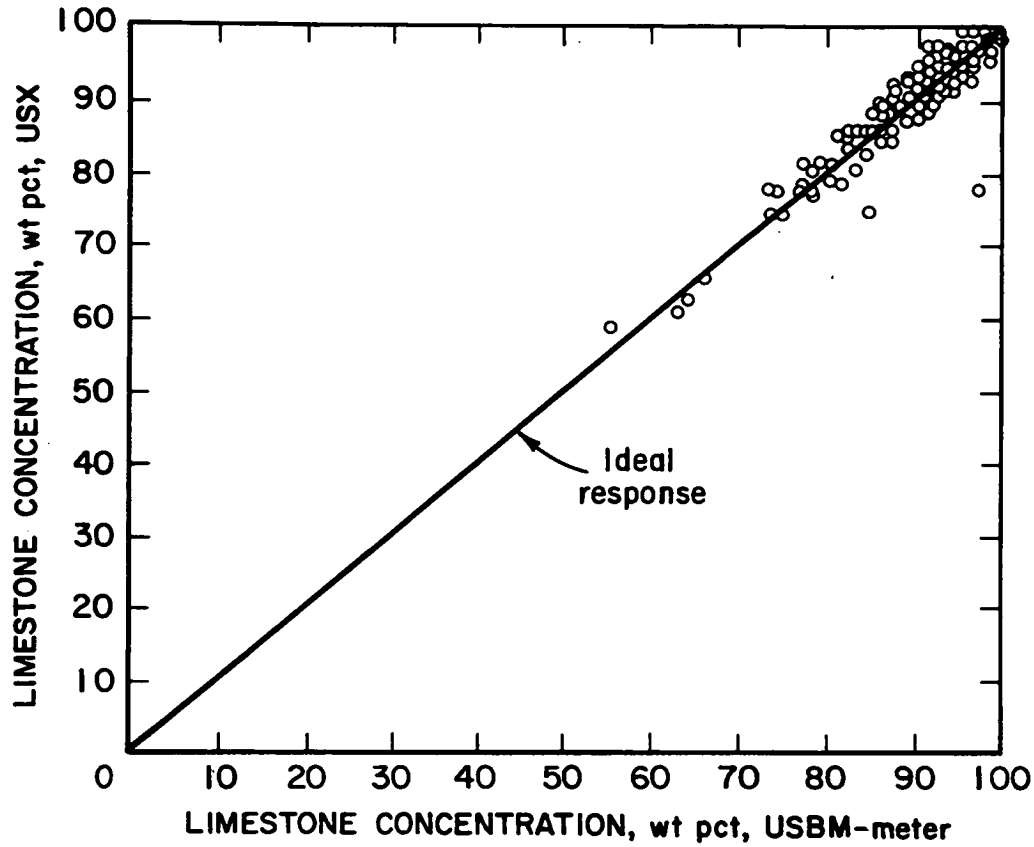


Figure 6. Comparison of USX volumeter and OBMM readings for samples taken from USX mines.

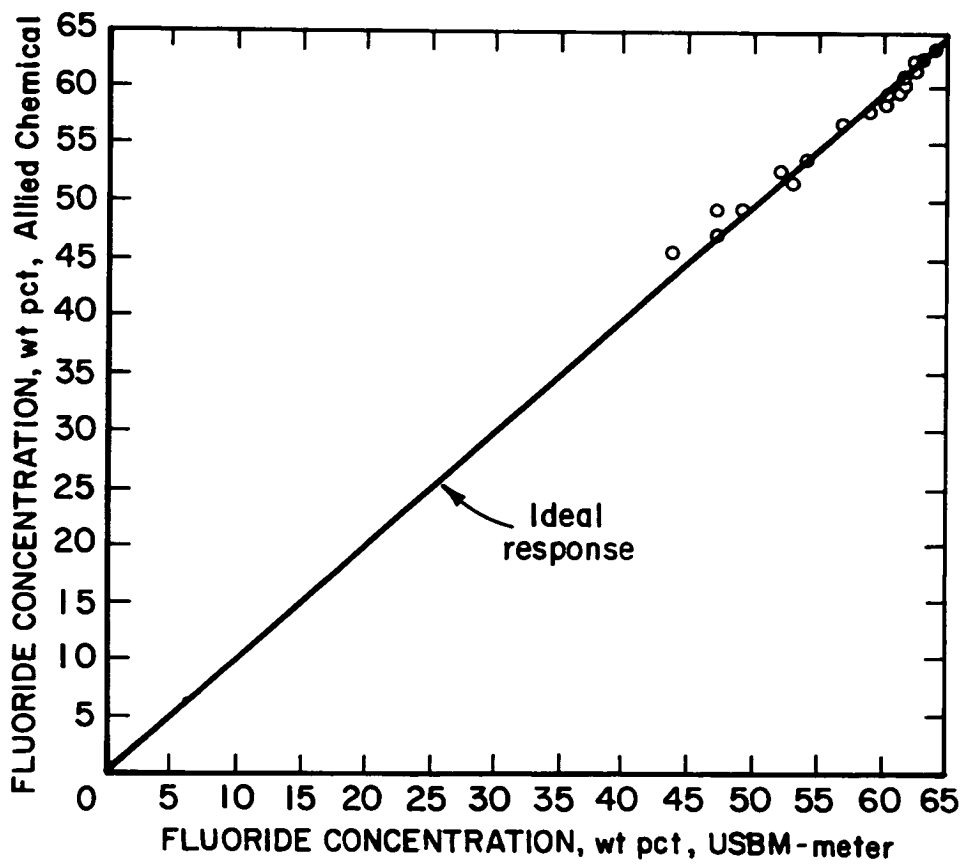


Figure 7. Comparison of Allied Chemical analysis and OBMM values for fluorinated product.

ADVANCED MATERIAL CHARACTERIZATION USING LASER INDUCED EDDY CURRENT IMAGING

E. James Chern
Code 313 / Materials Branch
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

Quality is the most important engineering factor that dictates the safety, reliability, performance and competitiveness of a product. To ensure a superior product, various quality assurance measures such as mechanical testing, nondestructive evaluation, and life-cycle testing have to be employed to fully characterize properties of materials and structures[1]. It is thus imperative to develop advanced evaluation techniques to improve the understanding of materials. We have developed an innovative laser induced eddy current imaging system for advanced material characterization[2-3]. The basic operating principle of this technique is the monitoring of the relative changes in material electromagnetic properties induced by laser effects. In this paper, we review the fundamental principle of eddy current sensing of laser induced thermal and elastic strains, describe the hardware and software components of the system, and present test results of laboratory experiments. Potential applications are for example, in material homogeneity, joint integrity, defect detection, damage assessment, and probe characterization. Specific materials of interest include thin films, micorelectronic wafers, and coatings.

BACKGROUND

Advancements in computer, electronic and photonics technology have facilitated the development and improvement of many test and measurement systems in recent years. These newly developed evaluation systems encompass a broad field of applications such as material selection, properties characterization, defect detection, damage assessment, process qualification, process monitoring, quality assurance, in-service sensing, and failure analysis, etc. The versatility of a system is often limited by physical principles and operational constraints of the system, and geometry and service environment of the product. Advanced approaches and systems are thus constantly being sought to overcome these constraints.

For most of the systems, excitation and detection mechanisms are based on the same physical principle as that of a stress-strain testing system: mechanical strain gauges are used to measure the strains induced by the mechanical loading of the specimen. Other systems use cross excitation and detection schemes such as a Scanning Electron Microscope (SEM) system: a scanning focused electron beam induces secondary electrons and x-rays from the material, scintillators and lithium doped silicon detector are used to sense the scattered secondary electrons and x-rays. Cross interactions are generally higher order materhial effects that provide pertubative thus more sensitive information of the specimen.

Continuing advancements in photonics technology have enabled the development and broadened the applications of lasers. We have developed a cross-interaction material characterization system based on laser induced eddy current imaging. Modulated laser pulses impinge on the surface of a specimen and generate thermal strain and elastic strain that propagate in the material. Eddy current sensing probe is used to monitor the changes in the material electromagnetic parameters induced by the thermal and elastic strains. By modulating the laser excitation pulse and scanning the beam with an optical scanning device, an eddy current image can be produced based on the sample coordinates of the laser impingement.

PRINCIPLE OF OPERATION

The thermal strain and elastic strain generation mechanism is relatively simple and straightforward. A modulated high power laser is used as the excitation source. The temperature gradient between the on/off pulse cycle of the laser will cause expansion and contraction which induce elastic mechanical waves. These elastic waves along with the thermal waves will propagate into the material by diffusion effects.

The eddy current detection mechanism is somewhat complicated. From fundamental theories of electromagnetism the impedance of the eddy current search coil can be expresses as

$$Z = 1 / I^2 \int (E \times H)\ ds$$

where I is the current in the probe, E and H are the electric and magnetic fields induced by the eddy current coil, and s is the area enclosed by the electromagnetic field. the coil impedance Z is normally expressed as a complex quantity, $Z = R + jX_L$. the resistance R, is the real component and the reactance X_L, is the imaginary component of the impedance Z. R and X_L can be measured by a precision LCR impedance meter.

The impedance change due to laser induced thermal strain and elastic strain in a material can be expressed as

$$\Delta Z = 1 / I^2 \int \left({\partial\sigma}/{\partial T}\, E^2 + j\omega\, {\partial\mu}/{\partial T}\, H^2 \right) dT ds$$

where $\partial\sigma/\partial T$ is the change in electric conductivity σ and $\partial\mu/\partial T$ is the cgabge ub magnetic permeability μ of the material due to induced thermoelastic effect T. For example, the coil impedance Z of a 100 kHz pancake eddy current probe at ambient temperature is experimentally measured to be: resistance R = 14.3 Ω and reactance X_L = 82.5 Ω. the coil impedance change ΔZ, due to laser induced thermoelastic effects to the material is measured to be approximately 3 Ω or 20% for resistance and 4 Ω or 6% for reactance. Coordinating the impedance change with respect to specimen position, will allow advanced characterization of thermoelastic and defect properties of the material[4].

SYSTEM CONFIGURATION AND EXPERIMENTS

The laser induced eddy current imaging system consists of four major hardware components: a modulated laser source, an X-Y scanning mirror assembly, an eddy current measuring system, and a data acquisition system controller. The block diagram of the system is shown in Figure 1. A continuous wave (CW) laser is used with a mechanical chopper or a cyclic function generator to modulate the CW laser beam. A μ-sec pulse laser also can be used as the excitation source. A mechanical indexer drives an X-Y mirror to steer the laser beam to the point of interest on the specimen surface. A flat-geometry pancake eddy current sensing coil is mounted on the opposite side of the specimen from the laser.

In operation, the eddy current is generated by the secondary electromagnetic field induced by the primary field of the driving eddy current probe. The eddy current in the specimen is perturbed by the stress waves and thermal waves generated by the impinging laser beam. The eddy current instrument monitors the eddy current signal. The optical scanner is commanded to scan across the specimen surface in a raster fashion. By coordinating the eddy current signals and the laser impinging positions, an eddy current image can be produced.

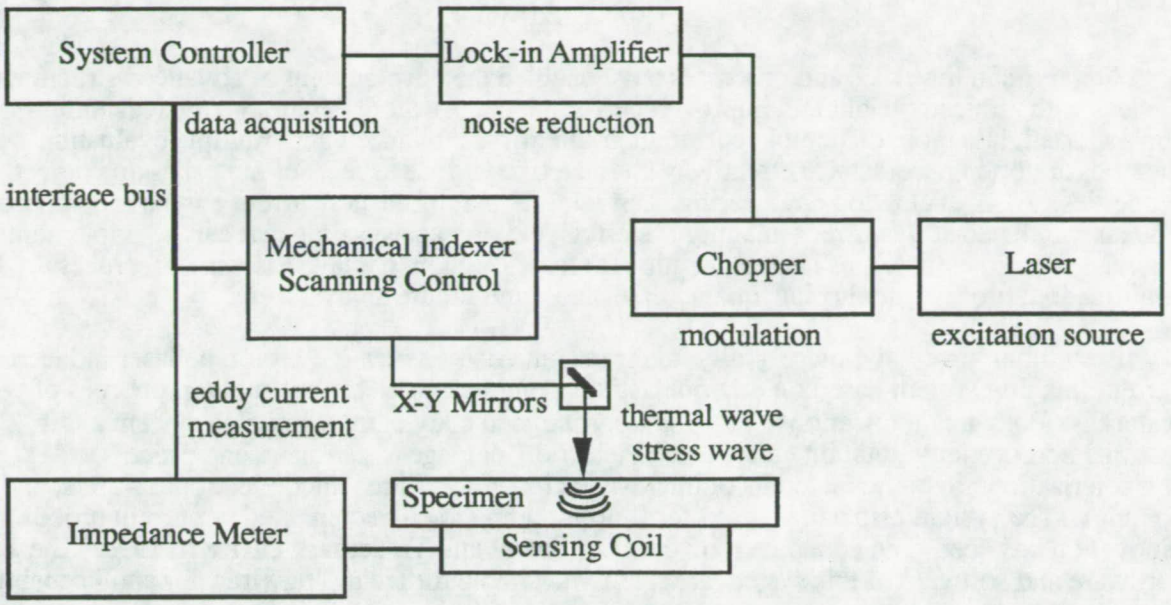


Figure 1. Block diagram of the laser induced eddy current imaging system.

To demonstrate the laser induced eddy current imaging, a simple experiment was conducted on a 2cm x 2cm thin aluminum foil with a pin hole. The experimental procedure was as described previously. Eddy current signals along with the laser beam positions were acquired as the laser beam scanned across the foil surface. The resistance component of the laser induced eddy current image is shown in Figure 2. The image clearly demonstrated eddy current sensing of thermal and stress wave.

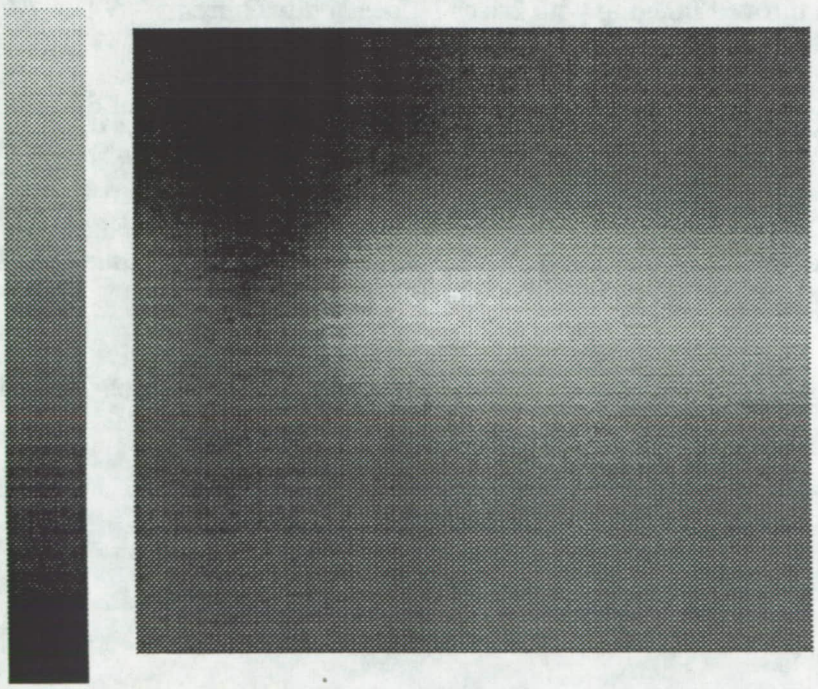


Figure 2. An laser induced eddy current image of an aluminum foil with a pin hole.

CONCLUSION

Advances in materials and processes have enabled the development of advanced structures and devices with reduced weight and improved performance. As these devices get increasingly sophisticated, it is more difficult to ensure their integrity. Advanced and multiple evaluation methods are often necessary to effectively characterize various aspects of materials and defect properties. Also, evaluation engineering is an integral part of the concurrent engineering approach. The laser induced eddy current imaging system is one of these systems that can be implemented for material selection, properties characterization, defect detection, damage assessment, process qualification, process monitoring, quality assurance, and failure analysis, etc.

In summary, we have successfully integrated and demonstrated a prototype laser induced eddy current imaging system based on thermoelastic dependence of electromagnetic properties of materials. Potential applications[5-6] of this laser induced eddy current imaging system are in material homogeneity, joint integrity, defect detection, damage assessment, and probe characterization. Specific materials of interests include (thin films, micorelectronic wafers, and coatings. The system also can be used for field mapping to characterize eddy current probes. This innovation has been granted a patent by the U.S. Patent and Trademark Office in 1992. The hardware and software of this system are readily available for technology transfer and licensing.

REFERENCES

1. E. J. Chern, "Concept of Nondestructive Evaluation," *Materials Evaluation*, Vol. 49, No. 9, pp. 1228-1235 (1991).

2. E. J Chern, "Method for Advanced Material Characterization by Laser Induced Eddy Current Imaging, " U. S. Patent #5,124,640, assigned to NASA Goddard Space Flight Center, June 1992.

3. E. J. Chern, "Improved Imaging with Laser-Induced eddy Currents," *NASA Tech Briefs*, Vol. 17, No. 3, pp. 63-64 (1993).

4. E. J. Chern and A. L. Thompson, "Eddy Current Imaging for Material Surface Mapping," *Review of Progress in Quantitative NDE*, Vol. 6A, Plenum Press, New York, pp. 527-534 (1987).

5. J. C. Moulder, M. W. Kubovich, J. M. Mann, M. S. Hughes and N. Nakagawa, "Applications of Photoinductive Imaging," *Review of Progress in Quantitative NDE*, Vol. 9A, Plenum Press, New York, pp. 533-538, (1990).

6. S. Mitra, G. C. Ojard, N. Nakagawa and J. C. Moulder, "Photoinductive Imaging Studies of Cu-Ni Diffusion Bond," *Review of Progress in Quantitative NDE*, Vol. 11B, Plenum Press, New York, pp. 1379-1385, (1992).

MICROWAVE ANTENNA HOLOGRAPHY SYSTEM

David J. Rochblatt
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109

ABSTRACT

Microwave Antenna Holography System (MAHST) has become a leading technique for increasing the performance of the large NASA / JPL Deep Space Network (DSN) antennas. The MAHST provides an efficient and low cost technique to optimize and maintain the performance and operation of large earth station antennas, thereby fulfilling today's requirements for link budget performance increases of commercial systems. The MAHST offers high-resolution and high-precision antenna imaging with a 100-micron standard deviation and 10-20 micron panel adjustment accuracy. The system provides far-field amplitude and phase pattern measurement with 90-dB dynamic range and a fast subreflector position optimization which increases the antenna performance capacity and pointing accuracy. Other outputs of the system are: aperture amplitude and phase functions, gravity deformation characterization and analysis, and directivity computations at other frequencies. The MAHST is a portable system that can be shipped to any antenna around the world and easily interfaced with its encoders and antenna drive systems. The MAHST was designed utilizing much "off the shelf" commercially available components. The remaining parts were designed and built at JPL guided by a friendly "man machine interface." The MAHST has been successfully tested and demonstrated in the NASA / JPL DSN.

MAHST DESCRIPTION

Microwave holographic metrology is considered to be a key technique for achieving improved performance from large reflector antennas, especially at the shorter wavelengths. The technique utilizes the Fourier Transform relationship between the complex far-field radiation pattern of an antenna and the complex aperture distribution (Figure 1.) [1.]. Resulting aperture phase and amplitude distribution data are used to precisely characterize various crucial performance parameters, including panel alignment, subreflector position, antenna aperture illumination, directivity at various frequencies, and gravity deformation effects. The MAHST provides performance improvement to the earth station antenna which increases its signal to noise ratio (SNR), and therefore its channel capacity or information processing rate. A block diagram of the MAHST is shown in Figure 2. Strong Continuous Wave (CW) signals obtained from geostationary satellite beacons are utilized as far-field sources. Strong CW beacon signals are available on nearly all satellites at Ku-band, X-band, C-band, and S-band. A portable 2.8-meter reference antenna is used as a phase reference and provides the signal to the receiver phase lock loop (PLL) channel. The IF section of Hewelett Packarad (HP) Microwave Receiver HP8530A and an external JPL PLL enable amplitude and phase measurement of the ground antenna sidelobes with 90-dB dynamic range. The far-field data is collected by continuously scanning the test antenna against a signal from a geosynchronous satellite, following a two-dimensional grid. The measured far-field amplitude and phase are shown in Figure 1.A and 1.B respectively. The angular extent of the response that must be acquired is inversely proportional to the size of the desired resolution cell in the processed holographic maps. The data processing [2.] provided with the system computes the desired information. Figure 1.B and 1.C show the derived test antenna aperture amplitude illumination and surface error map respectively. It is the information in the surface error map that is used to compute the adjustments of the individual panels in an overall main reflector best-fit reference frame. The amplitude map provides valuable information about the energy distribution in the antenna aperture. Figure 3. shows a high-resolution surface error maps of the NASA / JPL DSN 34-meter beamwaveguide (BWG) antenna designated DSS-24 for which the MAHST was utilized. As shown in the figure, the antenna main reflector rms surface error was reduced from 0.50-mm to 0.258-mm after applying one (1) iteration of panel setting. Figure 4. shows the same antenna far-field pattern after correcting its subreflector position. The data acquisition time for the high-resolution maps required for panel setting takes 6.5-hours. The data processing for obtaining panel setting information takes 8-hours. It took an additional 8-hours to actually reset the panels of the antenna. The measurement and data processing time required for subreflector position correction for a 34-meter antenna is approximately 2-hours. The resulted performance improvement for the antenna at Ku-band (12-GHz) is 0.7-dB, and at Ka-band (32-GHz) is 5.0-dB. The correspondence data rate increase at Ku-band would be 17%.

THEORY

The mathematical relationship between an antenna far field pattern (T) and the antenna induced surface current distribution (J) is given by the exact radiation integral relationship [1,2,3]:

$$\vec{T}(u,v) = \int\int_s \vec{J}(x',y') \exp^{jkz'} [\exp^{-jkz'(1-\cos\theta)}] \exp^{jk(ux'+vy')} dx'dy' \quad (1)$$

Where in equation (1), Z'(x',y') defines the surface S, (u,v) are the direction cosine space, and θ is the observation angle. For a small angular extent of the far-field pattern, this expression reduces to:

$$\vec{T}(u,v) = \int\int_s \vec{J}(x',y') \exp^{jkz'} \bullet \exp^{jk(ux'+vy')} dx'dy' \quad (2)$$

Equation (2) is an exact Fourier Transform of the induced surface current. To derive the residual surface error, geometrical optics ray tracing is used to relate the normal error ε, to the axial error and phase in a main reflector paraboloid geometry.

$$1/2\Delta PL = 1/2[P'P+PQ] = 1/2[\frac{\varepsilon}{\cos\varphi}+\frac{\varepsilon\cos 2\varphi}{\cos\varphi}] = \varepsilon\cos\varphi \quad (3)$$

$$\text{Phase}(\Delta PL) = \frac{4\pi}{\lambda}\varepsilon\cos\varphi \quad (4)$$

and for a paraboloid:

$$\cos\varphi = \frac{1}{\sqrt{1+\frac{x^2+y^2}{4F^2}}} \quad (5)$$

Where F is the antenna focal length.
Allowing for the removal of a constant phase term and substituting Eq. (4) into Eq. (2) yields:

$$\vec{T}(u,v) = \exp^{-j2kF} \int\int_s |\vec{J}(x',y')| \exp^{j4\pi\frac{\varepsilon}{\lambda}\cos\varphi}] \exp^{jk(ux'+vy')} dx'dy' \quad (6)$$

For processing sampled data, the associated Discrete Fourier Transform (DFT) is utilized:

$$T(p\Delta u, q\Delta v) = sx\ sy \sum_{n=-N1/2}^{N1/2-1} \sum_{m=-N2/2}^{N2/2-1} J(nsx, msy) \exp^{j2\pi(\frac{np}{N1}+\frac{mq}{N2})} \quad (7)$$

where :

N1XN2: Is the measured data array size
sx,sy: Sampling intervals in the aperture coordinates
n,m,p,q: Integers indexing the discrete samples
$\Delta u, \Delta v$: Sampling intervals in u,v far-field space

Since the magnitude of the far-field pattern is essentially bounded, the Fast Fourier Transform (FFT) is usually used for computation. The solution for the antenna residual surface error in the normal direction is therefore:

$$\varepsilon(x,y) = \frac{\lambda}{4\pi}\sqrt{1+\frac{x^2+y^2}{4F^2}}\ \text{Phase}[\exp^{j2kF}(FFT)^{-1}[T(u,v)]] \quad (8)$$

The spatial resolution in the final holographic map δ, is defined here at the -3-dB width of the resolution cell [4]:

$$\delta = \frac{D}{kN} \tag{9}$$

Where:

D: Main reflector diameter
N: The square root of the total number of data points
k: Sampling factor, usually $0.5 < k < 1.0$.

The lateral resolution is inversely proportional to the number of sidelobes measured. For a 34-meter diameter antenna for example, a resolution of 0.32-meter in the final holographic map can be achieved with a data array size of 127x127 (16,129) and sampling factor of 0.84. For a 34-meter antenna constructed of 348 panels, this measurement will enable imaging of each panel by 33 resolution cells.
In Figure 1, the fat-field amplitude and phase, Fig. 1(a) and 1b), respectively, are measured on rectangular coordinates of 127x127 with sampling intervals of 34.0 mdeg (the sampling factor is 0.84). Figures 1(c) and 1(d) show the aperture amplitude and surface error function with lateral resolution of 0.32-m.

The accuracy in each resolution cell of the final holographic map is [5]:

$$\sigma = 0.082\frac{\lambda D}{\delta SNR} \tag{10}$$

λ: Wavelength
SNR: Beam Peak voltage signal to noise ratio.
σ: Standard deviation (accuracy) in recovering the mean position of a resolution cell.

The accuracy across holographic maps varies with the antenna aperture amplitude taper illumination. Results are better at the center of the dish and gradually becomes worse toward the edge of the dish, for a uniformly illuminated dish, accuracy stays relatively constant through most of the dish and becomes quickly worse just at the edge where the illumination falls off rapidly. From equation (10) it is noticed that the accuracy is inversely proportional to the spatial resolution. This is due to the larger averaging area available at the larger resolution cell, as expected.
For an holographic measurement receiver incorporating a multiplier integrator or a divider integrator (ex. HP8530A), the effective signal to noise ratio SNR_e can be expressed [5]:

$$SNR_e = \left[\sqrt{\frac{1}{SNR_t^2} + \frac{1}{SNR_r^2} + \frac{1}{SNR_t^2 SNR_r^2}}\right]^{-1} \tag{11}$$

Where:

SNR_t: Test channel SNR.
SNR_r: Reference channel SNR.

Phase errors introduced during the measurement due to pointing and subreflector position errors are removed via a best-fit paraboloid program. The best-fit paraboloid is found by least squares fitting the data (residual surface error function), allowing six degrees of freedom in the model [1,2]. This algorithm insures that the minimum adjustment (distance) is computed for the screw adjusters. The least squares fit is computed by minimizing S, the sum of squares of residuals, path length changes:

$$S = \sum_{i=1}^{N^2} \Gamma(\Delta PL_i)^2 A_i \tag{12}$$

Where:

A_i: Amplitude weighting factor associated with the i^{th} data point.

Γ: The masking operation which is antenna type dependent.
ΔPL_i: Path length change at point (x_i,y_i,z_i).

It is correct to apply the best-fit paraboloid algorithm to either the conventional cassegrain paraboloid-hyperboloid or dual-shaped reflector systems even though the latter does not use a paraboloid as the main reflector. Either of the reflector antenna design is, overall, a planewave-to-point source transformer, differing only in the intensity field distribution.

The resultant aperture function at the end of this process is defined here as the effective map [2] since it includes all phase effects that are contributing to the antenna performance. These frequency-dependent effects include the subreflector scattered feed phase function and struts diffraction effects. Removal of the frequency dependent effects results in a mechanical map [1]. By deriving panel adjustments based on the effective map, the surface shape will conjugate the phase errors, optimizing the performance of the antenna at a single frequency, while degrading the performance of the antenna at all other frequencies. For antennas operating at a single frequency, this procedure is advantageous. However, many antennas operate at several different frequencies and require wide-bandwidth performance response. For these antennas, the mechanical map must be used to derive panel setting information.

From the mechanical map, surface tolerance efficiency can be computed at frequencies other then the measured frequency by scaling the residual aperture phase error (which are now due only to surface deviations) to other frequencies [1]:

$$(K)_{surface} = 20 * \log_{10} \frac{\sqrt{\left[\sum_{i=1}^{N^2} 10^{ampdb_i/20} \cos\left(\phi_{m_i}\left(\frac{\lambda_m}{\lambda_k}\right)\right)\right]^2 + \left[\sum_{i=1}^{N^2} 10^{ampdb_i/20} \sin\left(\phi_{m_i}\left(\frac{\lambda_m}{\lambda_k}\right)\right)\right]^2}}{\sum_{i=1}^{N^2} 10^{ampdb_i/20}} \qquad (13)$$

In this computations it is assumed that the aperture amplitude illumination is frequency independent. The error introduced in this assumption is thus negligible.

These polynomials can be used to separates systematic (low-frequency) errors from random (high-frequency) error components by filtering. The above coefficients are the core in the computation of the Jacobi-Bessel series expansion, which is an efficient computational tool for the antenna radiation integral
To simplify the discussion on panel settings, the normal component of the residual surface error (E_n) is modeled here as having only two components, one which is due to panel misalignment - rigid body motion , and a second which is due to surface error due to panel bending.

$$E_n = E_b + E_p \qquad (14)$$

Where

E_n: Total surface normal error.
E_b: Normal error due to panel bending.
E_p: Normal error due to panel misalignment.

To improve the antenna surface error due to panel misalignment, panels are allowed to move as rigid bodies, with 3-degrees of freedom. The panel position correction is computed by least squares fit. The derived motion of the panel is then used to compute the needed adjustment at the exact location of each screw on the panel. Only the pixles (resolution cells data) projected on the panel are considered in the computation, with the center of the pixle taken as the criteria of its location. This criteria provides some averaging near the panel edges, flaring it somewhat with its neighbors. In the panel rigid motion algorithms, three degrees of freedom are allowed; a translation at a reference point and two rotations (tilts) about the radial and circumferential axis.
Screw adjustments at point q_i is computed via :

$$E_{p_{q_i}} = -\left(S + d_i * \tan(\alpha) - \left(e_i/\cos(\gamma)\right) * \tan(\beta)\right) \qquad (15)$$

Where

$$\gamma = \arctan (R_{Q_r} / 2 * F)$$

and

F : Focal length of the best fit paraboloid

R_{Qr}: Radial distance from dish center to panel coordinate center.

Figures 5. and 6. shows the MAHST menu driven data acquisition computer interfaces and displays [6].

REFERENCES

1. D. J. Rochblatt, "A Microwave Holography Methodology for Diagnostics and Performance Improvement for Large Reflector Antennas," JPL/TDA Progress Report 42-108, pp. 235-252, February 15, 1992.

2. D. J. Rochblatt, "A User Manual, Data Processing Software for Microwave Antenna Holography, Computer programs for Diagnostics, Analysis and Performance Improvement of Large Reflector and Beamwaveguide Antennas," JPL Internal Document JPL D-10237, January 15, 1993.

3. D. J. Rochblatt and B. L. Seidel, "Microwave Antenna Holography,"IEEE Trans. MTTs Special Issue on Microwaves in Space, VOL. 40, NO. 6,PP.1294-1300 June 1992.

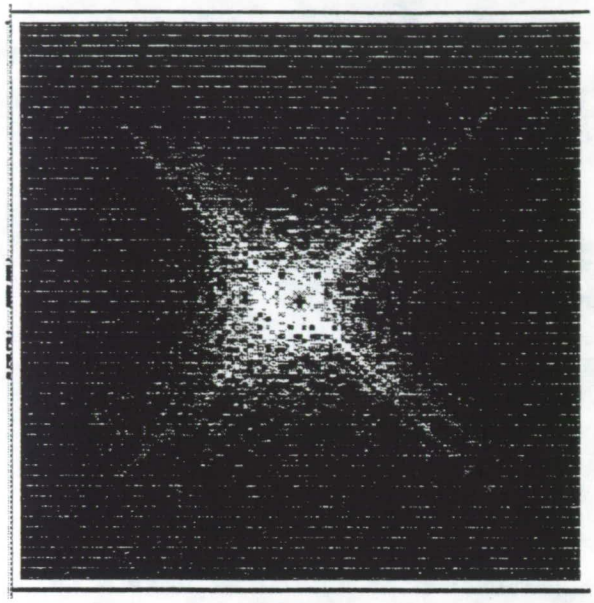
4. D. J. Rochblatt, B. L. Seidel, "DSN Microwave Antenna Holography," JPL/TDA Progress Report 42-76, pp. 27-42, October - December 1983.

5. D. J. Rochblatt and Y. Rahmat-Samii, "Effects of Measurement Errors on Microwave Antenna Holography," IEEE Trans. Antennas Propagat., VOL. 39, NO.7, pp, 933-942, July, 1991.

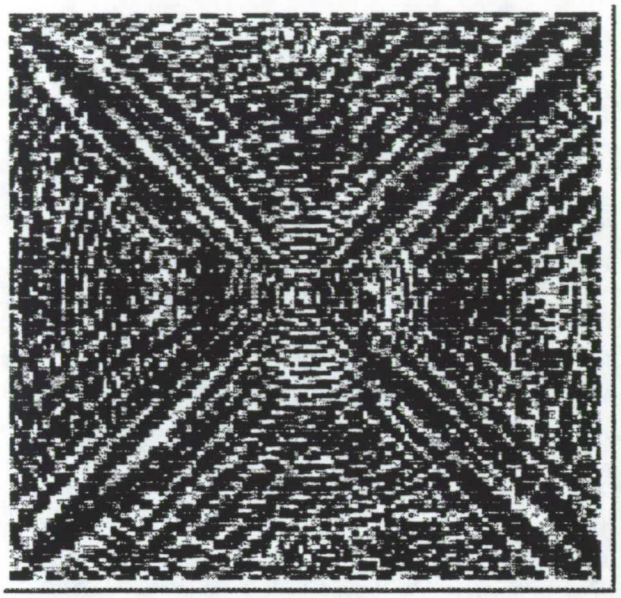
6. D. J. Rochblatt, B. L. Seidel, P. M. Withington, "JPL-DSN Microwave Antenna Holography System," JPL IOM 3328-94-004, January 14, 1994.

7. D. J. Rochblatt and B. L. Seidel, "Performance Improvement of DSS-13 34-m Beam-Waveguide Antenna Using the JPL Microwave Holography Methodology," JPL/TDA Progress Report 42-108, pp.253-270, February 15, 1992.

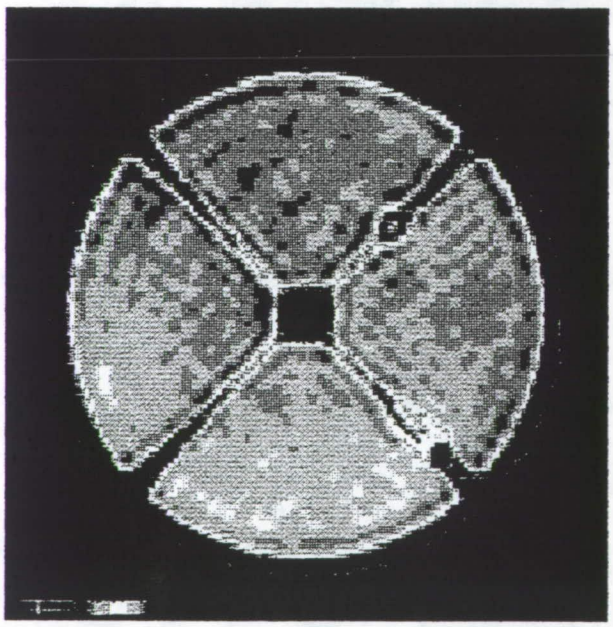
8. D. J. Rochblatt, " Microwave Holography Helps Improve Performance of Large Antennas," LASER Tech briefs, Publication of NASA Tech Briefs, September 1993 Vol. 1 No. 1, PP 81.



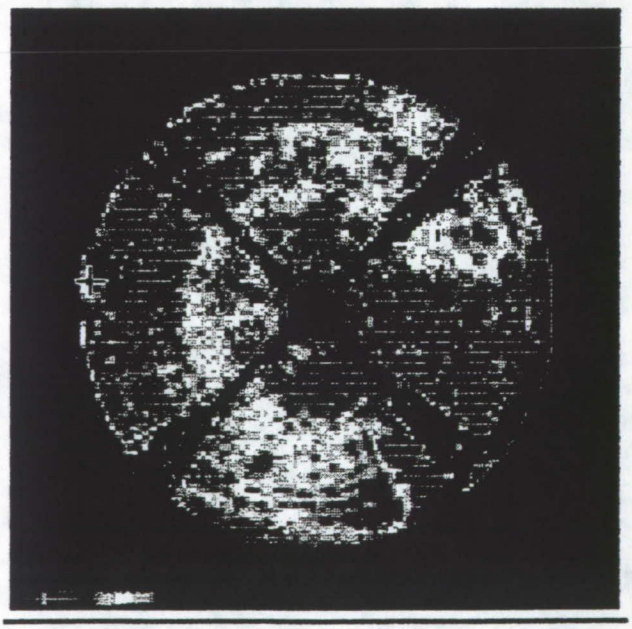
A. Far-Field Measured Amplitude



B. Far-Field Measured Phase



C. Derived Aperture Amplitude Illumination



D. Derived Aperture Surface Error Map

Figure 1. Complex (Amplitude and Phase) Fast Fourier Transform (FFT) Relating the Antenna Far-Field and Aperture Functions

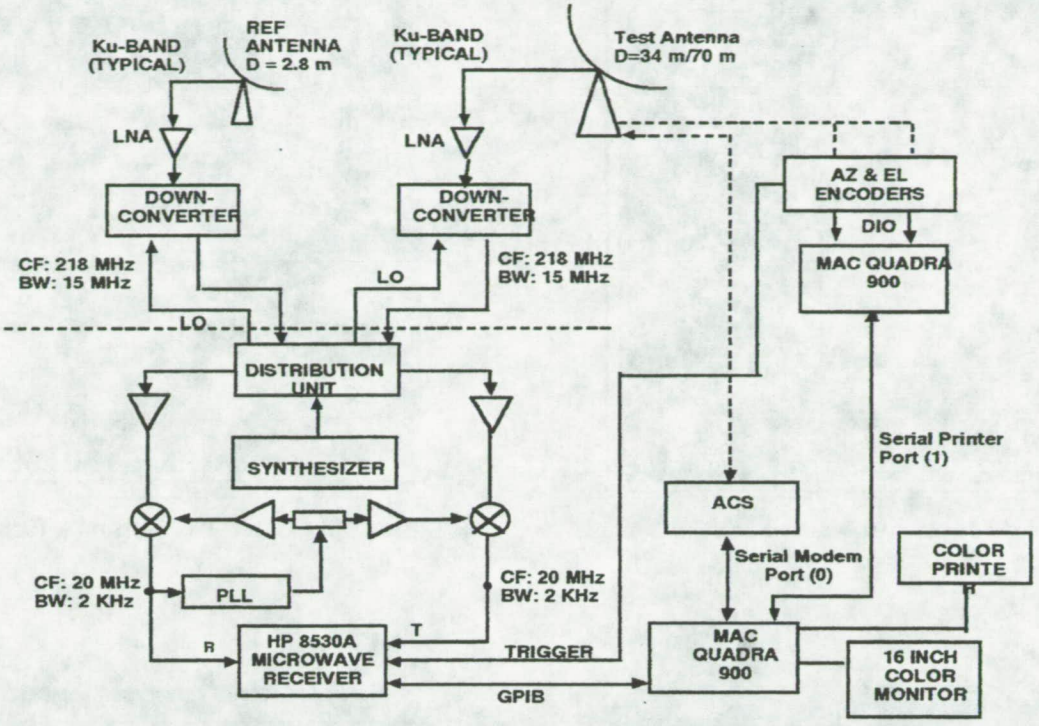
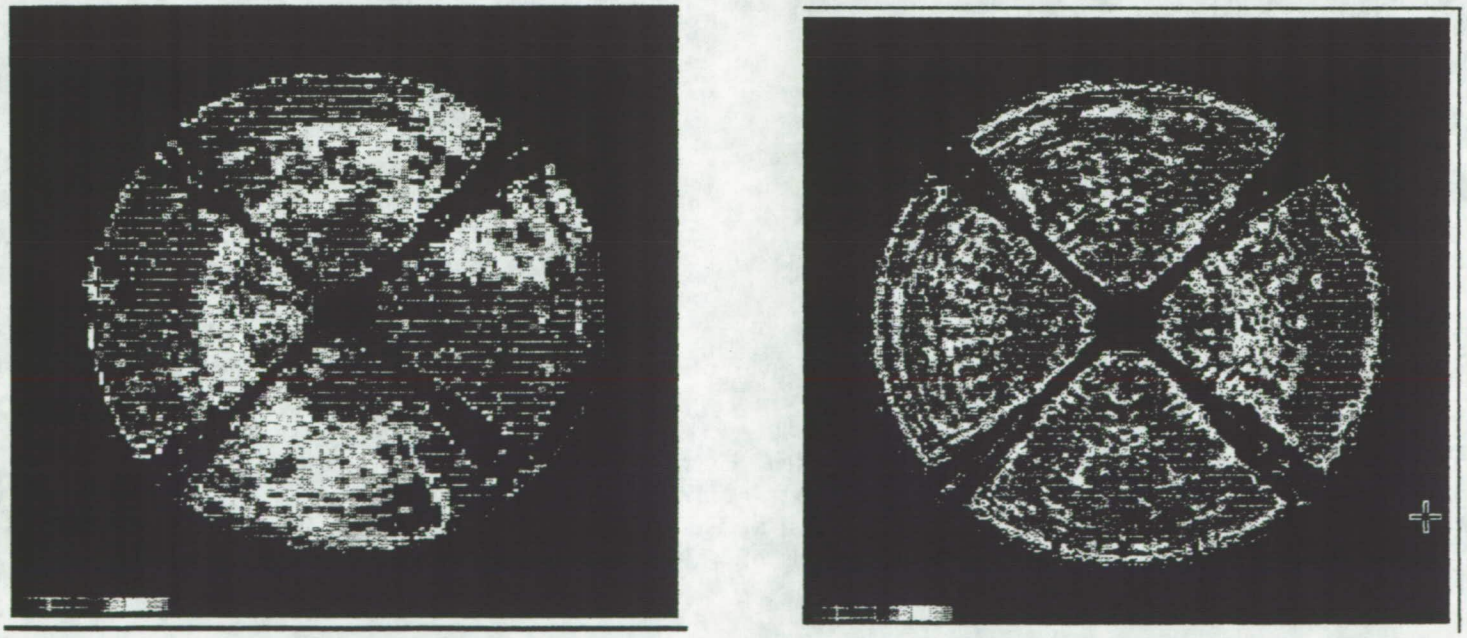
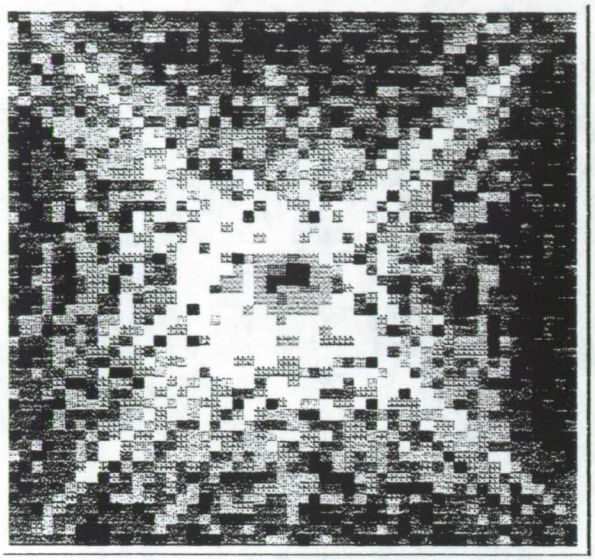


Figure 2. Microwave Antenna Holography System Block Diagram

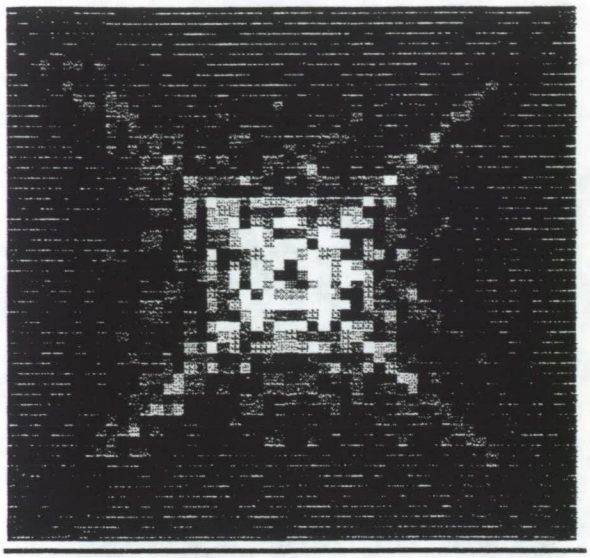


DSS-24 Before: 0.50-mm rms **After: 0.26-mm rms**

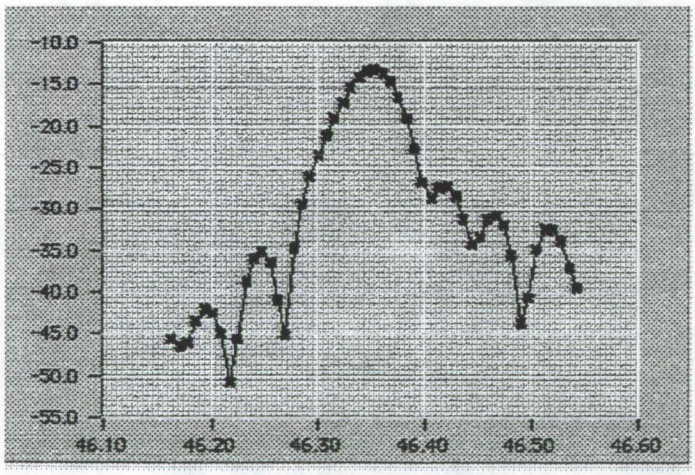
Figure 3. Improved Main Reflector rms Surface Error From 0.50-mm to 0.26-mm After Applying one (1) Iteration of Panel Setting (Potential Of 0.20-mm).



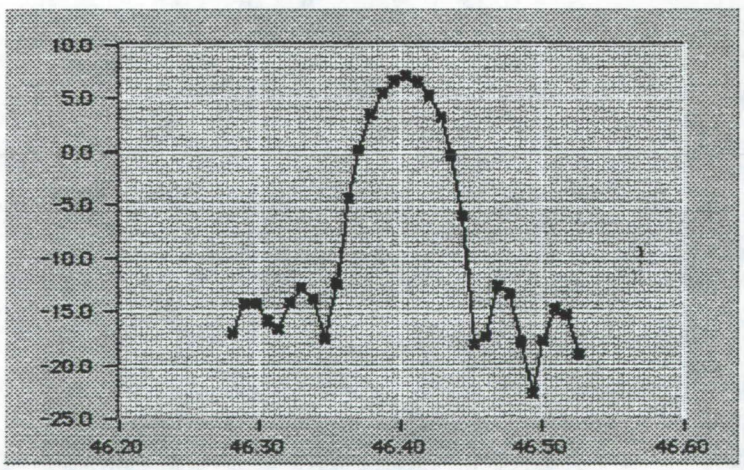
A. As Found - unfocused



B. After Correction - focused

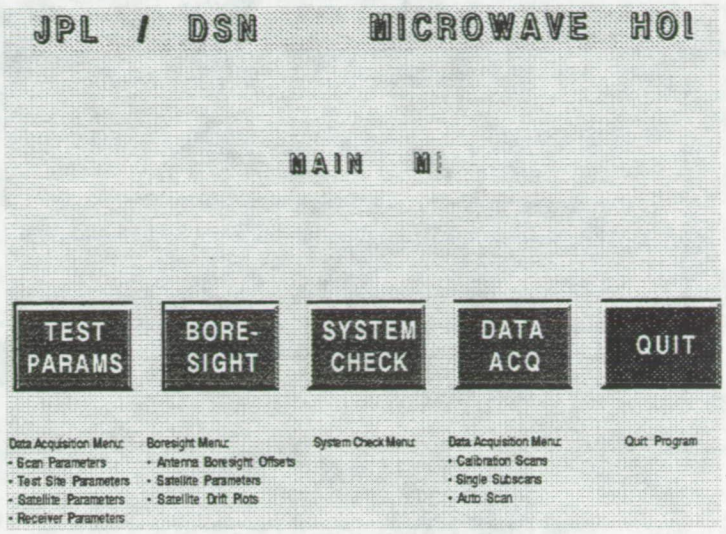


C. One Sidelobe Formed 13-dB Below Main Beam (Elevation Cut)

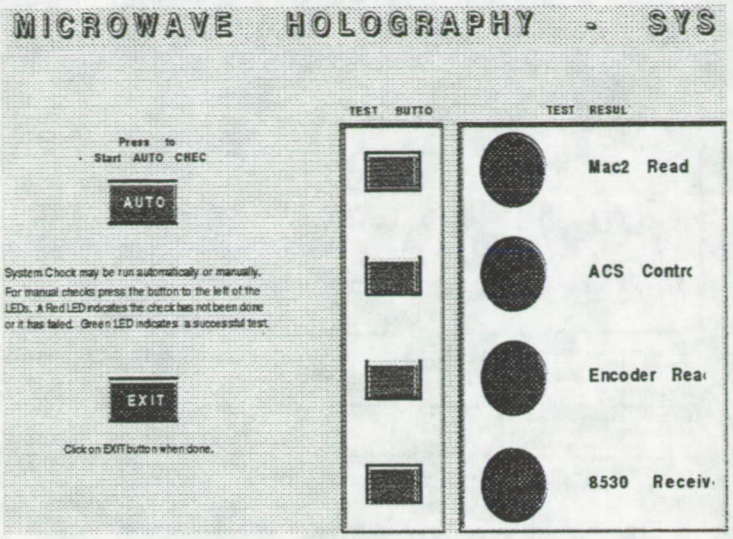


D. Two Even Sidelobes Formed 19-dB Below Main Beam (Elevation Cut)

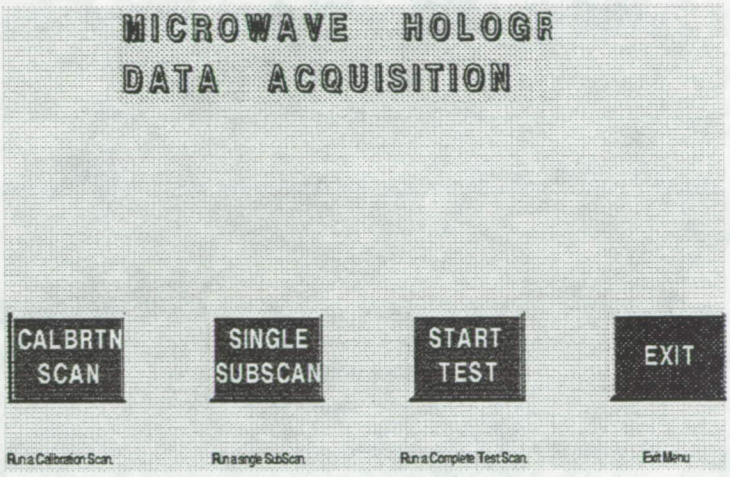
Figure 4. Accurate Subreflector Position Correction. Derived Subreflector Offsets: ΔX=-0.516",ΔY=+0.375",ΔZ=+0.135" Far-Field Amplitude Pattern at Ku-band (12-GHz), at 45-deg. Elevation, Cassegrain Focus of a 34-m BWG Antenna



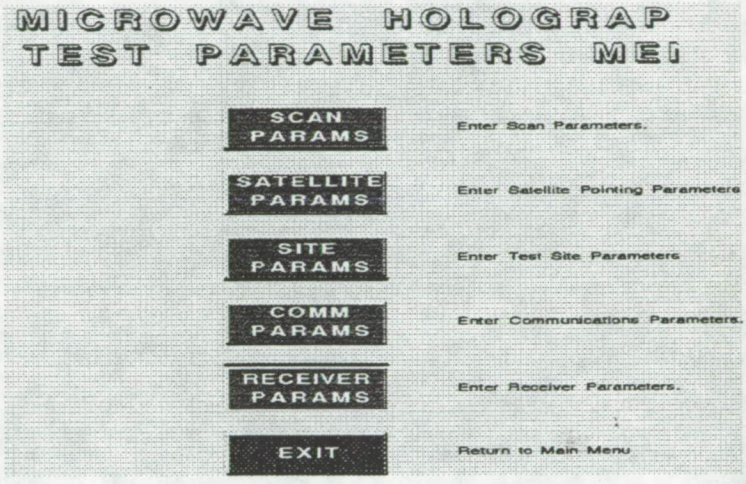
Main Menu



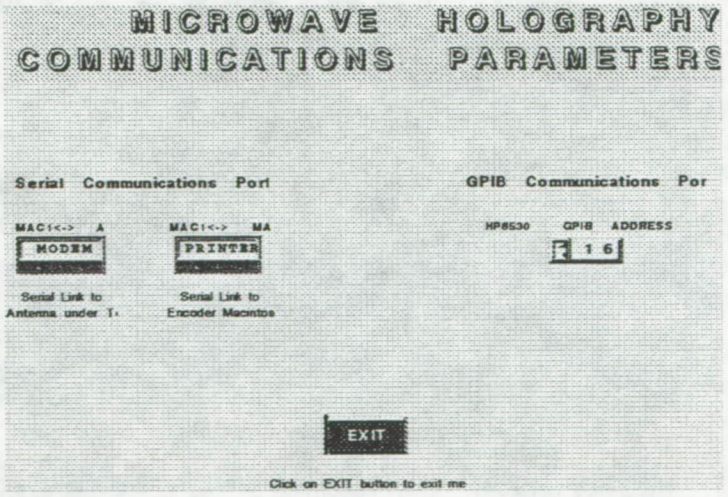
System Check Menu



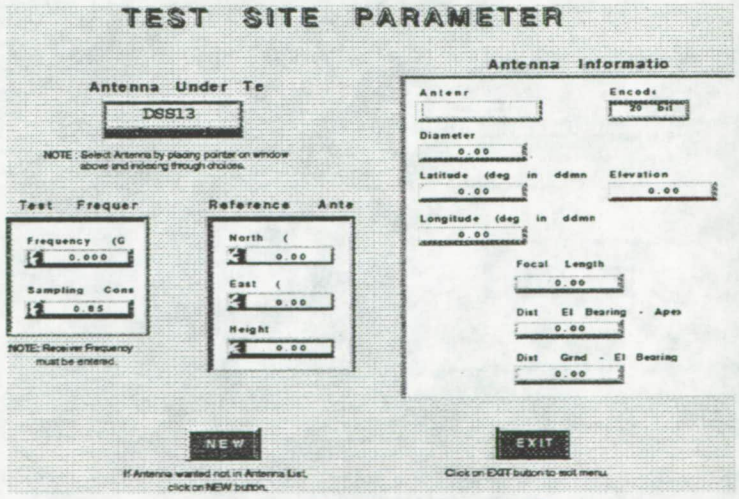
Data Acquisition Menu



Test Parameters Menu

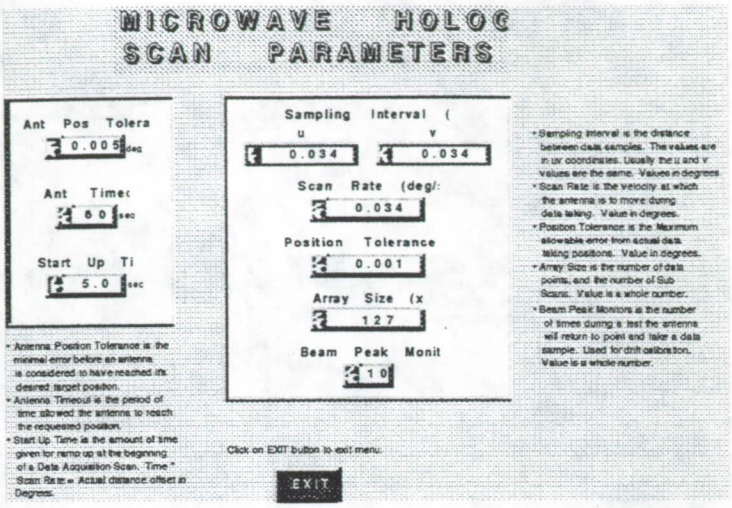


Communication Parameters Menu

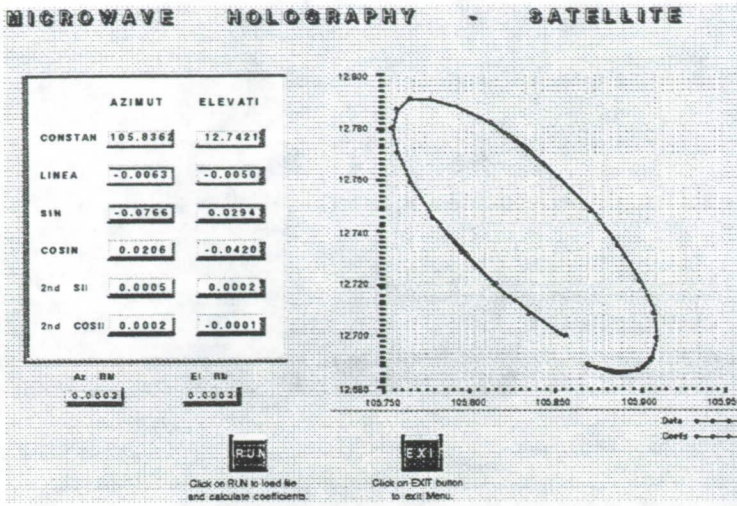


Test Site Parameters Menu

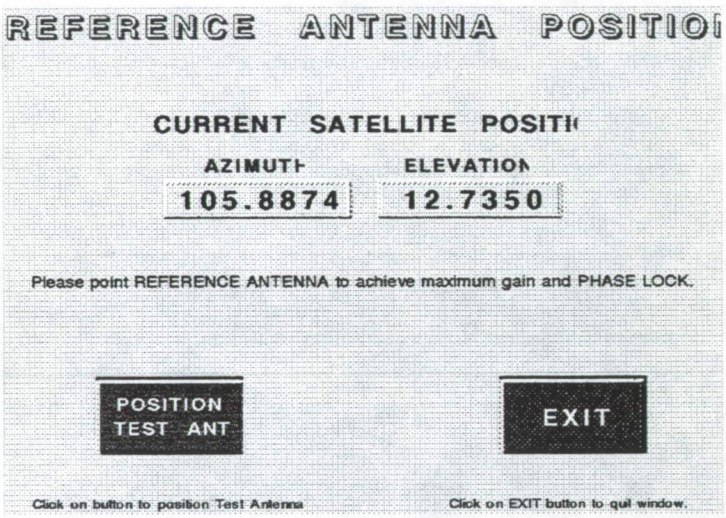
Figure 5.



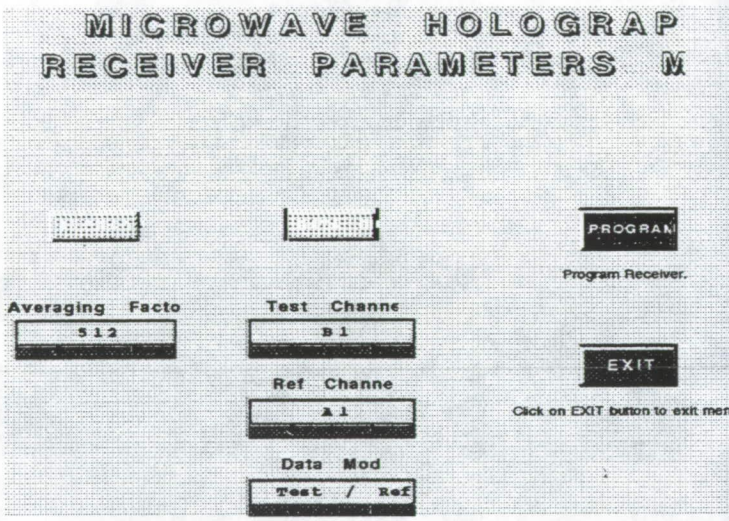
Scan Parameters Menu



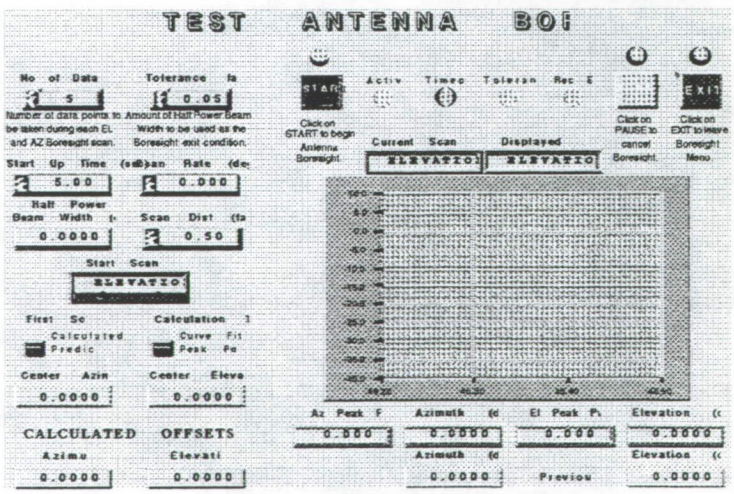
Satellite Parameters Display Panel



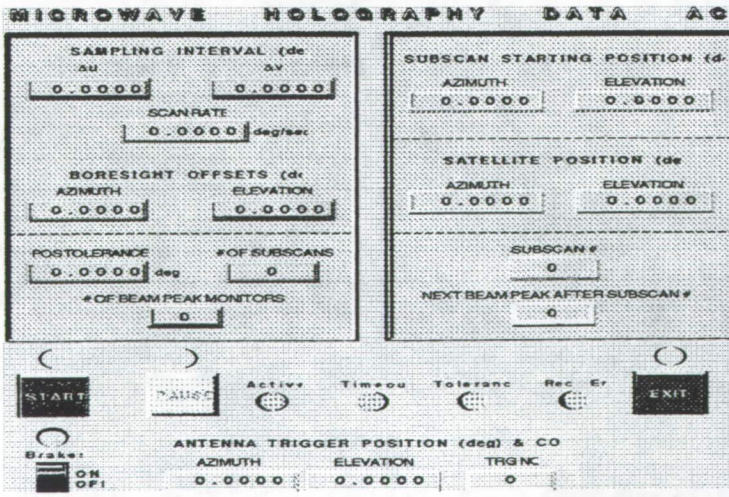
Reference Antenna Position Display Panel



Receiver Parameters Menu



Test Antenna Boresight Display Panel



Data Acquisition Display Panel

Figure 6.

SELF-CONTAINED MAGNETIC FIELD MEASURING SYSTEM

Pedro J. Medelius and H. James Simpson
I-NET Space Services
INI-10, Kennedy Space Center, FL 32899

Jim Cecil
NASA
DL-ICD-T, Kennedy Space Center, FL 32899

ABSTRACT

Electronic Instrumentation is very sensitive to overvoltages caused by induced magnetic fields. An electronic conductor, when exposed to a varying magnetic field will develop currents and voltages that could damage sensitive equipment. Most lightning strikes have currents in the order of tens of kiloamps, with very fast risetimes, and therefore generate large magnetic fields. A self-contained instrument which records varying magnetic fields has been developed TO capture the induced waveforms. The instrument is equipped with batteries that will allow unattended operation for up to 2 weeks. The instrument provides information regarding the peak magnetic fields and the peak rate-of-rise of these fields for all three orthogonal axis (x, y and z). This information can be used to assess the potential damage to sensitive instrumentation which for applications at the Kennedy Space Center will help decide whether flight hardware or cargo requires to be retested.

DESCRIPTION

Thunderstorms are a very common source of damage to electronic equipment. A typical lightning strike carries currents in the order of several tens of thousands amperes, with risetimes even faster than one microsecond. A lightning strike usually consists of a first stroke, followed by one or more subsequent strokes at 20 to 200 millisecond intervals. These fast varying currents result in voltages induced on conductors. When conductors such as power or communication lines are attached to electronic equipment, the induced voltage can cause failures in the equipment.

The self-contained magnetic field sensor is an instrument which records the peak values of both the magnetic field and the magnetic field derivative. These peak values include the most positive and most negative magnetic field strengths received on all three orthogonal antennas (x,y and z) during each lightning return stroke. It also stores one 50 μs waveform corresponding to a preselected antenna. The data is stored in non-volatile memory inside the sensor and is later retrieved by an operator using a portable computer through an EIA-232 port. A block diagram of the sensor is shown in Figure 1.

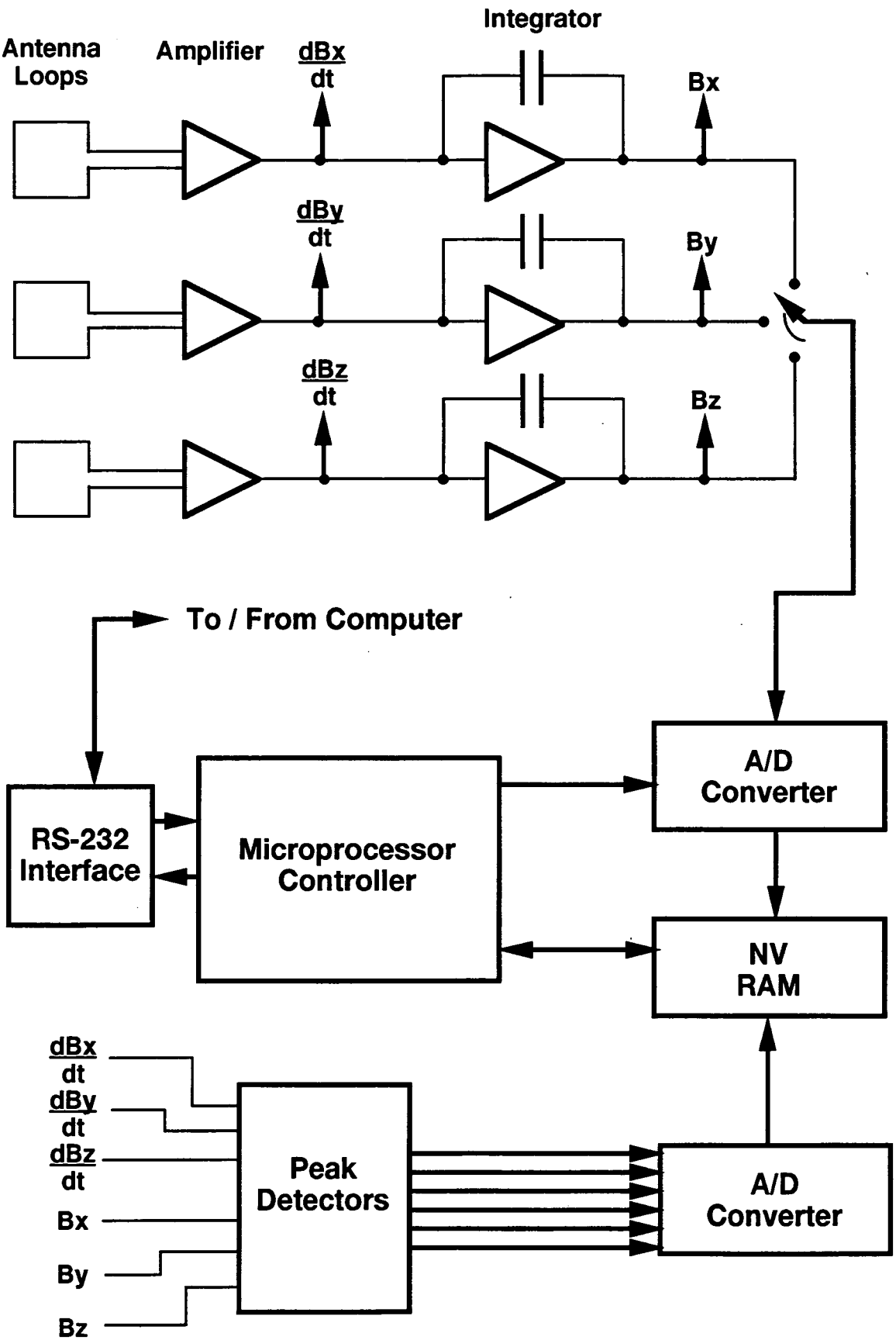


Figure 1 Block Diagram

Loop Antennas

The sensing elements in this instrument consist of three orthogonal loop antennas. The antennas are placed apart from each other in order to minimize interference or cross-coupling among them. The voltage developed across the terminals of each antenna is proportional to the rate-of-rise of the magnetic field, and is given by:

$$V = nA\mu_0 \frac{dH}{dt} \cos\Theta$$

where:

V = Output voltage
A = Area of the loop
n = Number of turns
$\mu 0$ = permeability
dH/dt = rate-of-rise of the magnetic field
Θ = Angle of the magnetic field with respect to the normal to the plane of the loop.

Since the output voltage is directly proportional to the rate-of-rise of the magnetic field, the measurement of dH/dt requires only an amplifier. In order to obtain a voltage proportional to the magnetic field H, the dH/dt signal has to be integrated.

By measuring three components of the magnetic field, the total magnetic field can be found by:

$$H = \sqrt{H_x^2 + H_y^2 + H_z^2}$$

Amplifiers and Integrators

The amplifiers and integrators were designed to achieve fast response, since lightning-generated fields can have risetimes faster than 1 μs. The frequency response of the amplifiers and integrators was limited to about 3.5 MHz to prevent aliasing when the waveforms are digitized, as will be explained later. The magnetic field sensor was designed to have two selectable gains, which will determine the sensitivity to the magnetic fields. This gain change capability was provided to allow the instrument to perform when installed at buildings with different electromagnetic radiation shielding.

The electronic integration of the magnetic field waveform is achieved by means of operational amplifiers with a capacitor/resistor parallel combination in the feedback loop. Each integrator has a frequency domain transfer function of the form:

$$\frac{K}{S+\frac{1}{\sigma}}$$

where K is a constant. The pole at $1/\sigma$ is needed to make the integrator relatively more sensitive to the rapid lightning signals of interest by attenuating the low frequency field oscillations that are common in the thunderstorm environment. The prototype circuit boards were designed to minimize high frequency coupling among different integrators. The electric characteristics of the feedback capacitors were especially important as any departure from the impedance of an ideal capacitor over the frequency range of interest could result in an incorrect waveform.

Data Acquisition and Storage

A Motorola 68HC11 microprocessor is used to control the operation of the magnetic field sensor. Two analog-to-digital converters (A/D) are used in the instrument. One A/D is dedicated to sample for 50 μs at a 10 MS/s rate every time a pre-determined threshold is reached. The highest frequency that can be digitized without introducing frequency aliasing is called the Nyquist frequency, and is equal to one half of the sampling rate. Since the A/D converter is sampling at 10 MS/s, the frequency response of the amplifiers and integrators was limited to 3.5 MHz, so that the energy at 5 MHz would always be below the noise level of the system. The second A/D is continuously sampling the peak values of both the magnetic field and its derivative. The peak values are stored in memory only when the threshold is exceeded, otherwise the data is discarded. In order to prevent accidental loss of information, the magnetic field sensor stores the magnetic field waveforms and the peak values in non-volatile memory (NVRAM). Data retrieval is done by means of a portable computer via an EIA-232 interface. The rechargeable batteries can be changed in the field to allow for uninterrupted operation.

POWER SUPPLY

Power to the magnetic field sensor is supplied by six 6-volt, 1.3 A-h batteries. Voltage regulators provided the +5 and -5 volts required by the circuitry. A reference diode is used to provide the reference voltage for the analog-to-digital converters.

OPERATION

The magnetic field sensor requires minimal maintenance for operation. Once the unit is installed, it can operate for up to 2 weeks on fully charged batteries, and a replacement battery pack can be installed at the field. The unit can store up to 8 waveforms and 48 peak H and dH/dt values. The data retrieved by the computer includes the date and time of the lightning strike, the peak positive and negative values of both the magnetic field and its derivative (for all three axis), and one 50-μs long waveform corresponding to one of the magnetic fields.

COMMERCIAL APPLICATIONS

The commercial applications of this instrument include all areas where sensitive and/or sensitive electronic equipment are housed. Potential uses include remote (outdoor) sensing equipment, computer rooms, broadcast equipment, high-tech control rooms, air traffic control towers, and nuclear plant control rooms.

The current design of the magnetic field sensor requires an operator to retrieve the data after a thunderstorm. However, the instrument can be modified to allow remote downloading and continuous monitoring of the data even while a thunderstorm is in progress.22

ADVANCES IN COMMERCIAL HIGH POWER PULSED & CW DIODE LASERS

Kurt J. Linden
Spire Corporation
Bedford, MA 01730-2396

Michael Sanfacon
Spire Corporation
Bedford, MA 01730-2396

ABSTRACT

Designs for high-power, 780 to 810 nm diode laser arrays capable of optically pumping solid-state lasers have been broadened to include cw laser array bars with various optical fill factors. Long-pulse, quasi-cw diode laser arrays have been demonstrated, and a non-destructive diagnostic tool for identifying potential reliability problems for millisecond-pulse operation has been developed. A simplified thermal model for such millisecond pulse operation has been investigated, and thermal time constants determined. Diode laser array packaging advances and multi-bar lensing schemes are discussed.

CW DIODE LASER ARRAY BAR DESIGN AND PERFORMANCE

The need for cw diode laser arrays is increasing as the demand for solid-state lasers with high average-power levels continues to grow. The primary problem becomes one of heat removal from the active, emitting regions of the diode laser array bars. To meet these needs a variety of optical fill-factor diode laser bars have been designed. Tradeoffs between average bar brightness and thermal heat dissipation have been investigated. Conservative designs are used for long-term reliability.

Cw Diode Laser Bar Designs

A cw laser bar design which provides a 25% optical fill factor is convenient for coupling into optical fibers. This design is shown in Figure 1.

The 1-cm wide bar consists of twelve optically active emitter sections, each 200 μm wide, located on 800 μm centers. Each of the twelve emitter sections consists of a phase-locked array of 20 individual emitters 5 μm wide located on 10 μm centers. There is enough space between adjacent emitter sections to enable the use of lens arrays for fiber coupling. An individual bar can also be easily cut into twelve individual emitter dice for use as 1 watt cw lasers. Bar design is such that an optically dead region approximately 500 μm wide has been left on each end of the laser bar to enable easier handling and reduce the susceptibility of device failures resulting from edge damage. This type of cw emitter bar is typically rated at 12 to 15W cw output. Bars of this type have been operated at up to 25 W/bar cw output, but for reliability values in the thousands of hours, 12 to 15W cw operation is recommended. Another cw bar design with a similar (20%) fill factor, but smaller optical emitter width is shown in Figure 2.

This type of bar design is convenient for coupling into 100 μm to 200 μm core optical fibers. This type of bar can also be cleaved into individual emitter dice. In this case, the cw output power rating is 500 mW/ emitter section.

A third type of cw bar design, one with a 50% optical fill factor, has been used for both pulsed and cw operation, and is illustrated in Figure 3. The particular bar shown has a 9 mm width to enable its use in the Lawrence Livermore microchannel cooler package. This type of bar, when operated cw, has emitted approximately 20 to 23W of power, in either the bar-in-groove or the microchannel cooler laser package.

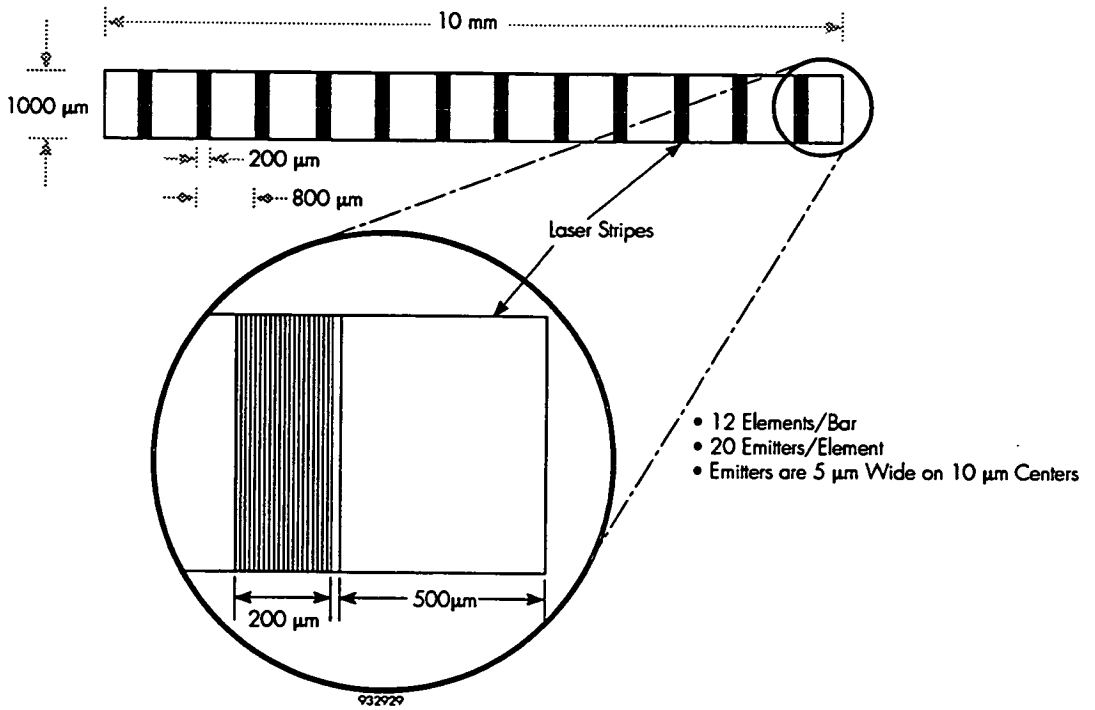


Figure 1. **Schematic diagram showing the electrical contact and optical emitter configuration of a 25% optical fill factor laser array bar with 200 µm wide emitters on 800 µm centers.**

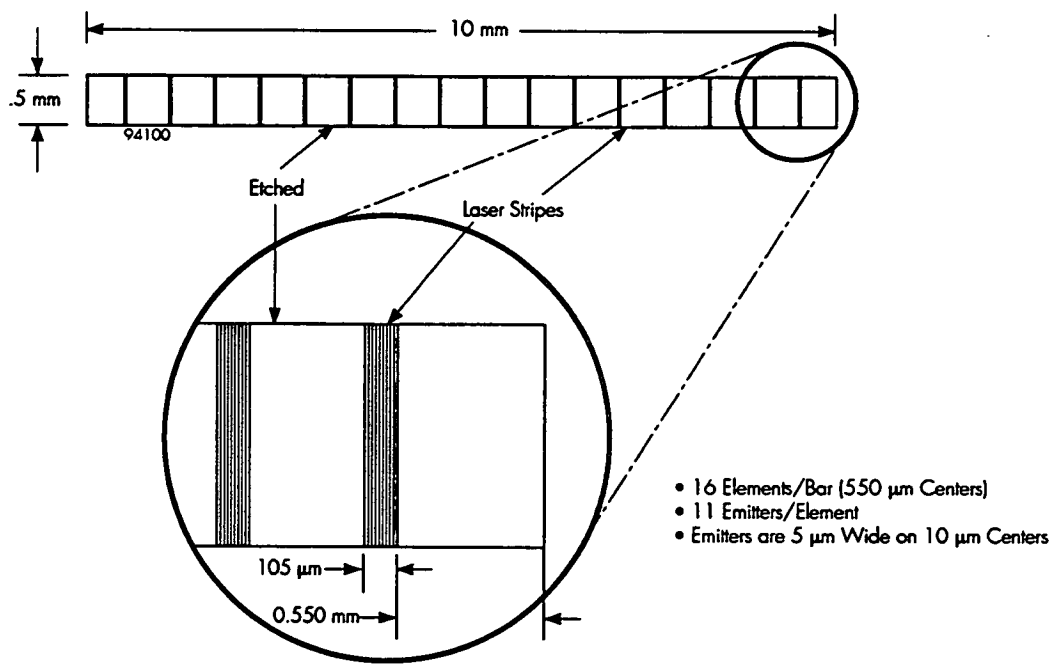


Figure 2. **Schematic diagram of a 20% optical fill factor, cw diode laser array bar with 105 µm wide emitter sections on 550 µm centers.**

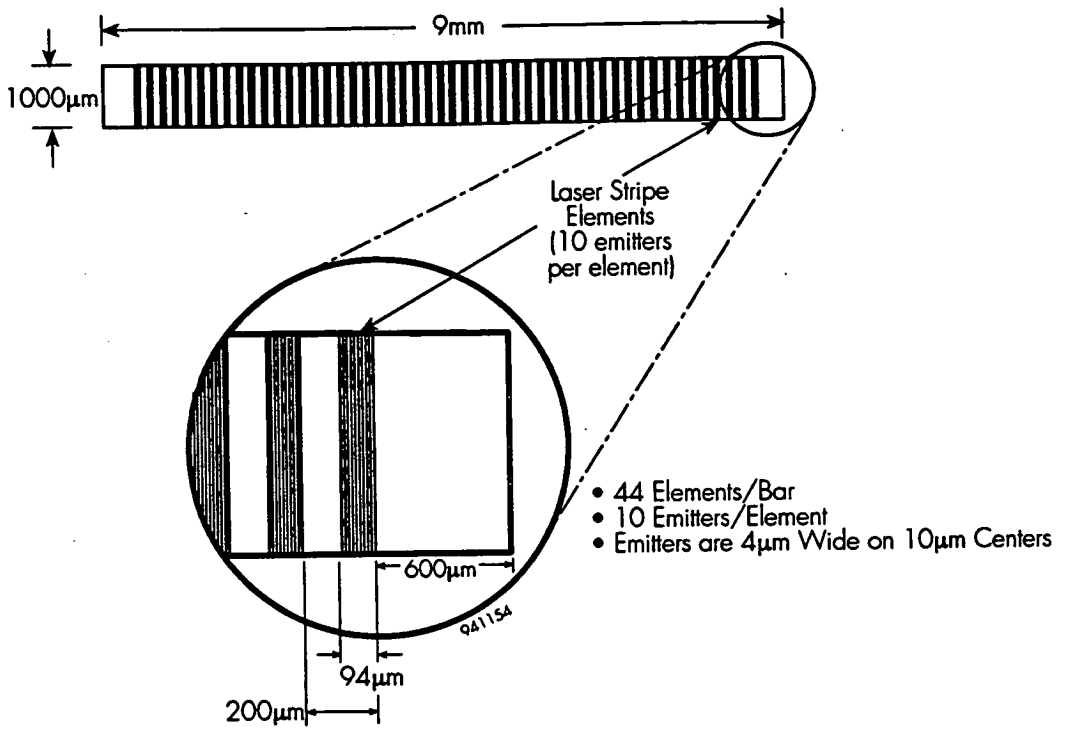


Figure 3. **Schematic illustration of a 9 mm wide diode laser array bar with 94 µm wide active regions on 200 µm centers. This bar design is well suited for the Lawrence Livermore microchannel cooler package.**

Cw Single-Bar Laser Performance

Cw bars have been tested in bar-in-groove packages and in microchannel cooler packages. The bar design tested was that shown in Figure 3 (94 µm active regions on 200 µm centers, giving a 50% optical fill factor). This type of bar design is favored for high average power applications where fiber coupling is not required. Bar widths were 9 mm, and cavity lengths were 1 mm. Each 94 µm wide active section consists of ten 5 µm wide emitters located on 10 µm centers. These ten emitters act as one phase-locked section, but the forty-four sections on each 9 mm wide bar operate independently of one another.

The single-bar laser cw output power as a function of laser current was measured for four laser bars mounted in four type 460, actively cooled, bar-in-groove laser packages fabricated by LDAI. The output power was measured with a Scientech model 365 calorimeter-type power meter. A recirculating water cooler was used to maintain a constant laser package temperature of 25°C. Due to the high power density on the calorimeter surface, the instrument was calibrated by first capturing all of the laser power, and then backing the laser away so as to reduce the power density on the calorimeter by a factor of two. This prevented the calorimeter from overheating. The results of these measurements are shown in Figure 4.

Each laser tested is from a different epitaxial wafer, thus illustrating the high degree of consistency between different epitaxial wafer growths. Each epitaxial wafer is 50 mm in diameter and thus capable of producing a maximum of approximately 100 laser bars. Slope efficiency values ranged from 0.76 to 0.87 W/A. All lasers had threshold current values of between 9 and 10A. Laser drive currents were limited to approximately 35A for these bar-in-groove packages.

A typical cw laser bar was mounted in a Lawrence Livermore microchannel cooler package and tested for P-I and V-I characteristics. The P-I results are shown in Figure 5.

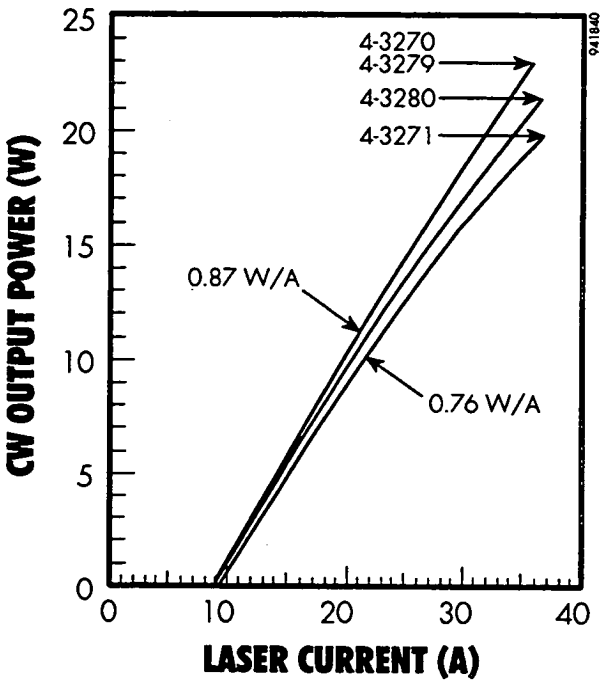


Figure 4. Cw output power as a function of laser current for four 9 mm-wide, single-bar diode lasers. Each laser is representative of a different epitaxial wafer. These bars were mounted in an LDAI type 460, water-cooled laser package.

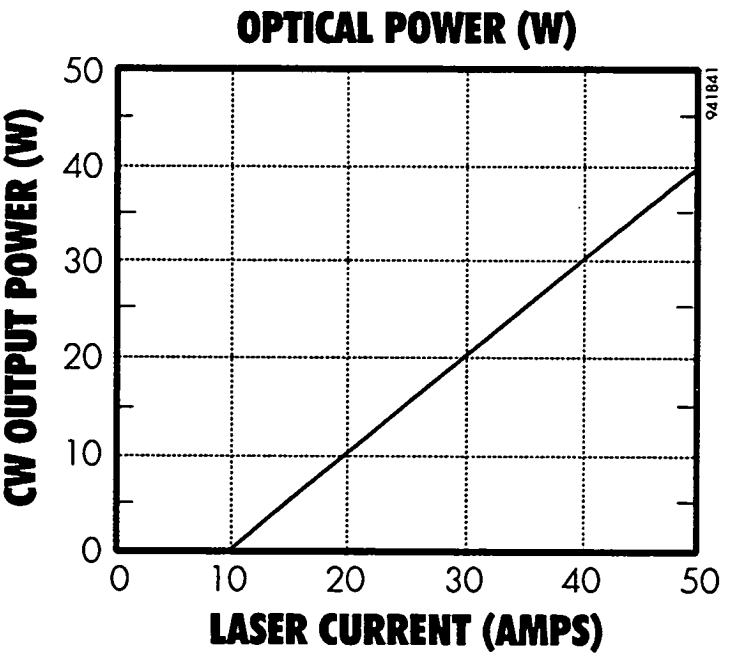


Figure 5. Cw output power as a function of laser current for a 9 mm-wide, single bar diode laser bar mounted in a Lawrence Livermore microchannel cooler package.

This device exhibited a slope efficiency of 0.98 W/A. The maximum drive current was limited to 50A resulting in a bar output of 40W. The threshold current was 9.5A, and the device exhibited an electrical series resistance of 4.5 milliohms.

100W Cw Multi-Bar Array

A cw multi-bar array was fabricated from laser bars of the geometry shown in Figure 2. Whereas the standard width of these bars is 1 cm, a group of bars of this type was cleaved to 9 mm width leaving 15 optically active sections per bar. Eight bars of this type were mounted in an LDAI water-cooled package. The bars were mounted in groups of 4, on a 1.2 mm pitch. The array was operated at an optical output power of 100W cw. A photograph of the near-field pattern of the array just above threshold is shown in Figure 6.

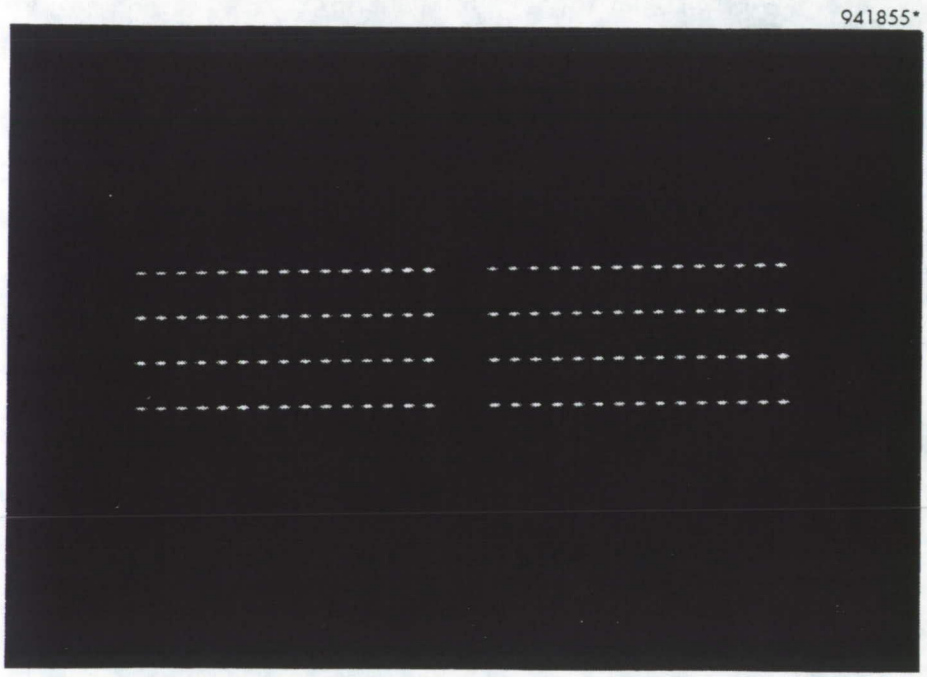


Figure 6. **Photograph of the near-field pattern of an 8-bar cw diode laser array. The photograph was taken just above threshold. This array was operated at 100W cw output in the water-cooled LDAI package.**

Because of the accurate positioning of each array bar, it is possible to fit this type of package with a fast-axis collimating lens assembly so as to produce a beam with a divergence of approximately 3 to 4 degrees along the fast axis (*i.e.*, normal to the bar width) and 7 to 10 degrees along the slow axis (in the plane of the bar width). This is further described in a later section of this paper.

LONG-PULSE, QUASI-CW DIODE LASER THERMAL MODELING

High values of pulse energy are possible by extending the pulse length of quasi-cw diode laser arrays into the millisecond range. For example, a 20-bar diode laser array, with a rated bar output of 50 W/bar pulsed, is capable of emitting 1 Joule of energy with a pulse length of 1 ms. Quasi-cw arrays have been fabricated and operated at pulse lengths of up to 5 ms, but their output power is generally derated to ensure reliable operation. To prevent excessive heating, the duty cycles must be kept low, resulting in slow pulse repetition rates. For example, for 2 ms pulse lengths the repetition rate is kept below 5 Hz. Multibar, quasi-cw arrays are generally derated to 40 W/bar at 5 Hz repetition rates when 2 ms operation is required.

A simplified thermal model has been developed to help determine the thermal time constant of this type of array. Using Dynatemp software (Hines Consulting Services, 44 West Cedar Street, Boston, MA 02114) a one-dimensional model of the type shown in Figure 7 was analyzed.

Layer	Layer Description
5	*Upper clad layer: AlGaAs (60% AlAs), 1.8 μm (approximation)*
4	*Heat-generating region: GaAs QW, 0.01 μm*
3	*Lower clad layer: AlGaAs (60% AlAs), 1.8 μm (approximation)*
2	*Contact layer: GaAs, 0.3 μm*
1	*Indium solder region: 5 μm*
0	*BeO heatsink*

941842

Figure 7. Laser array bar-mounting model for thermal-dynamics calculations using Dynatemp software.

In this model the laser array bar is mounted on a BeO plate, which is assumed to be of infinite extent, using a 5 μm thick layer of indium solder. The active quantum well (QW) region of the laser is placed on a 1.8 μm thick layer of $Al_{(0.6)}Ga_{(0.4)}As$. This thickness includes the cladding layer as well as the graded index layer. Whereas this model is somewhat of a simplification of the actual structure (for which the laser bars are mounted into grooves cut into a BeO plate which is soldered to a copper heat sink), the results are thought to be reasonably accurate because the thermal contact area between the BeO plate and the copper heat sink (1 cm^2) is very large compared to the contact area of the laser bar (0.05 cm^2).

The model assumes a bar output power of 40W, with an efficiency of 30%, resulting in heat generation of 93W in the vicinity of the QW region of the device. The temperature rise above ambient at various layer-interface points of the laser package structure was calculated for three different pulse lengths (2 ms, 5 ms and 25 ms). The computed temperature rise in the vicinity of the QW region of the device as a function of time for these three pulse lengths is plotted in Figure 8.

Several features are apparent from these calculations. The device structure has a thermal time constant of approximately 5 ms. This means that for pulses of 1 to 2 ms duration, the device junction temperature rise is proportional to pulse length. For pulses of duration longer than approximately 5 ms, the junction temperature is constant and the device is essentially operating in the cw regime. For the present example, a 2 ms pulse causes a junction temperature increase of approximately 10°C. This temperature increase is expected to result in an average spectral shift (chirping) of approximately 2.5 to 3 nm during the duration of the pulse. Experimental measurements have verified this spectral shift, as further discussed in the following section.

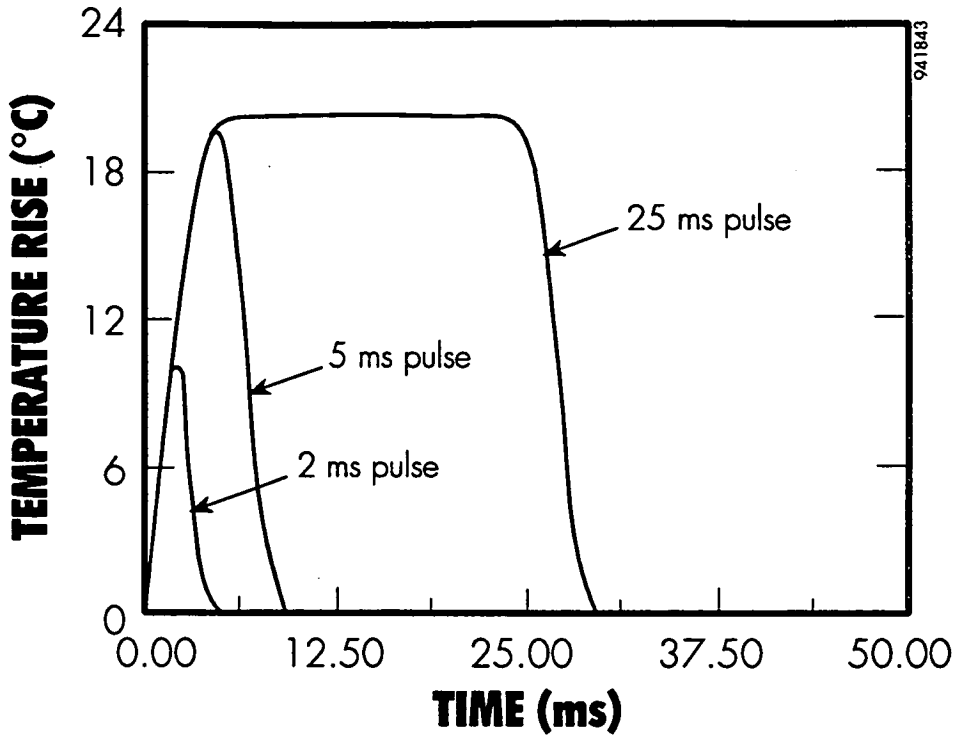


Figure 8. **Calculated temperature rise above ambient as a function of time for three different pulse lengths, each corresponding to a heat load of 93W (which corresponds to a 40W bar with a 30% efficiency).**

NON-DESTRUCTIVE TOOL FOR DIODE LASER ARRAY QUALIFICATION

The thermal model for pulsed diode laser array bars mounted in bar-in-groove packages has been discussed in the previous section. This model appears to accurately predict device junction temperature increases similar to those observed in operational devices. This observation suggests a diagnostic technique for analyzing the thermal properties of mounted diode laser array bars. If the laser bar emission spectrum is analyzed at various times during the current pulse, it should be possible to track this spectral shift. Any more rapid spectral shift would indicate excessive heating of the laser bar, and therefore facilitate screening of potentially suspect devices.

Experiments were carried out to analyze diode laser bar emission spectra over gated portions of the pulse length by use of a boxcar integrator (gated amplifier). These experiments were carried out on a single-bar laser array mounted in a SPI-LASE™ 50 type laser package. A stepping motor-controlled 0.5m spectrometer equipped with a photomultiplier tube detector was used to measure emission spectra. Gate lengths of 100 µs were used, with gate delay times varying from zero to 1.9 ms. A single scan takes approximately five minutes to complete. The spectral emission spectra of a typical single-bar device, as measured at gate delays of 0.1 ms and 1.9 ms are shown in Figures 9a and 9b. The pulse shifts approximately 2.2 nm during the 2 ms pulse, indicating a junction temperature rise of approximately 8°C.

A similar measurement was made on a specially selected diode laser bar with a thermal mounting flaw. The spectral emission results for this device are shown in Figure 10. Figure 10a and 10b show emission spectra details at 0.1 ms and 1.9 ms times into the 2 ms pulse, respectively. A distinctive feature of these spectra is the fact that a small, secondary peak develops as time increases, and that the spectral shift of this secondary peak, which occurs on the long-wavelength side of the main emission peak, shifts much more rapidly with time-delay than the main spectral peak. At 1.9 ms into the pulse, this secondary peak is separated from the main peak by about 7 nm, corresponding to a temperature difference of about 25°C. A plot of the spectral shifts of both the main peak and the secondary peak as a function of time-delay during the 2 ms pulse, as taken from the data in Figure 10, is shown in Figure 11.

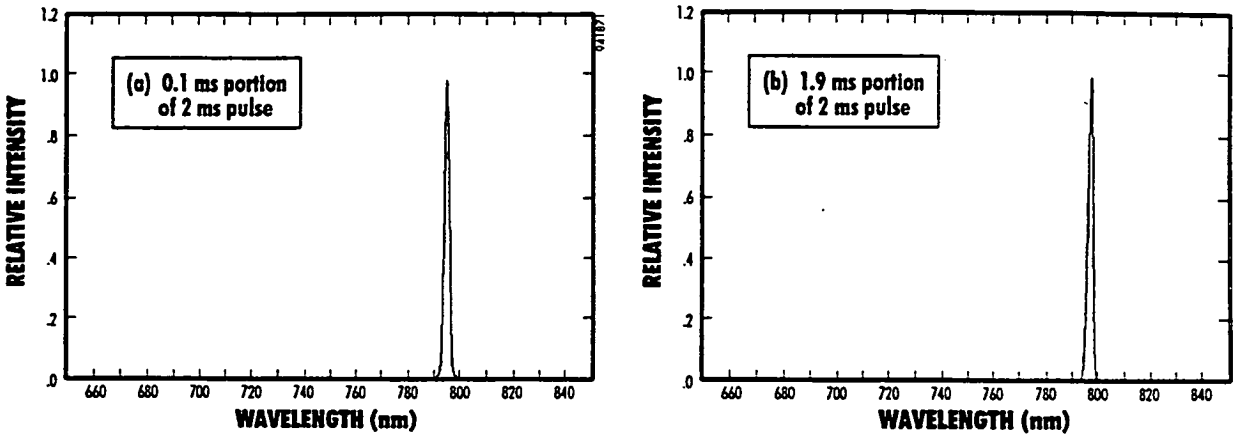


Figure 9 Emission spectra measured from typical diode laser bars at (a) 0.1 ms and (b) 1.9 ms delays of a 2 ms pulse-operated device.

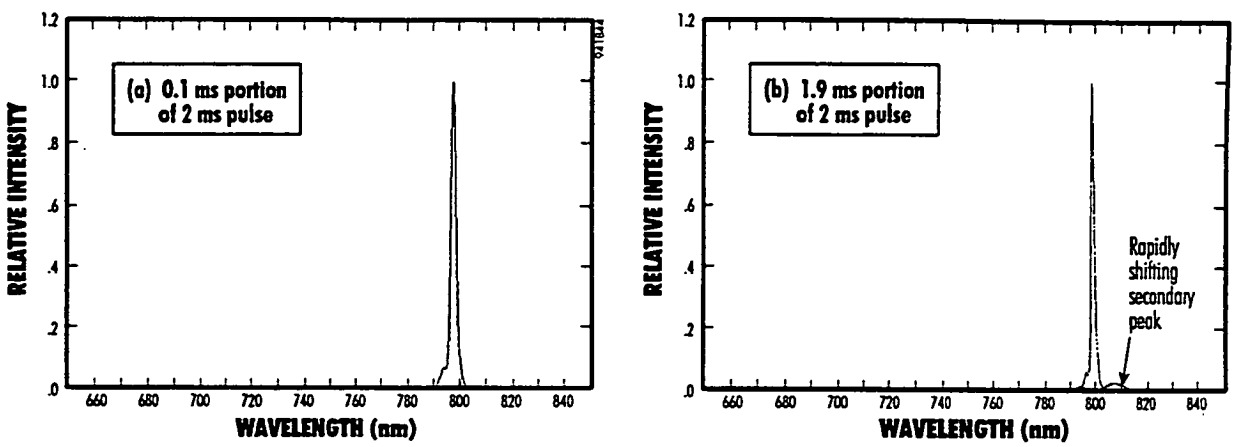


Figure 10. Emission spectra measured from an improperly mounted diode laser array bar at (a) 0.1 ms and (b) 1.9 ms delays of a 2 ms pulse-operated device.

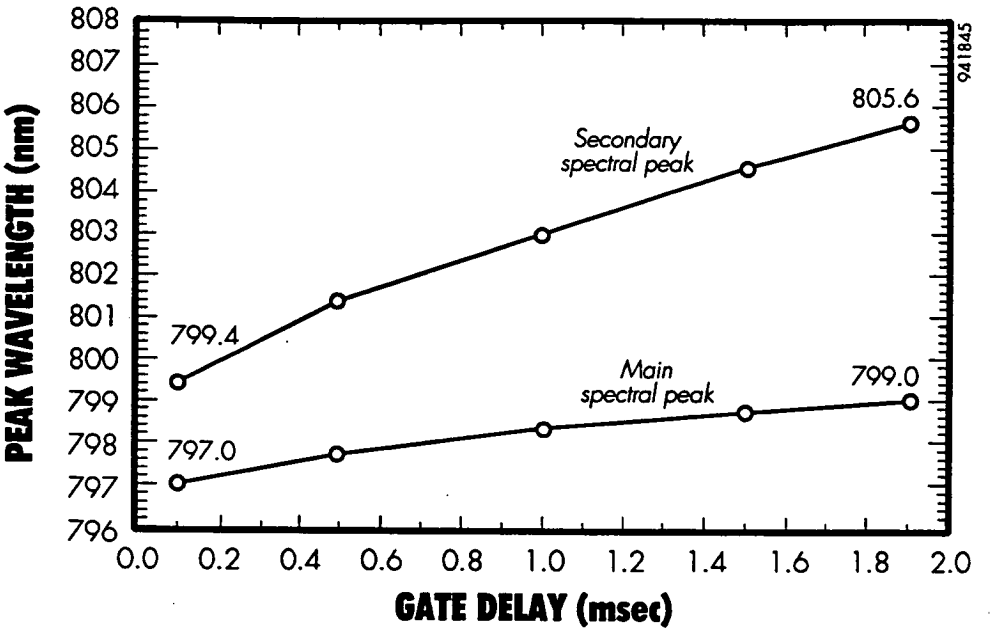


Figure 11. **Measured spectral shifts of the main peak and the secondary peaks of an improperly mounted laser array bar during a 2 ms pulse.**

The main peak has shifted 2 nm (corresponding to an estimated temperature increase of approximately 5 to 6°C) while the secondary peak has shifted 6.2 nm (corresponding to an estimated temperature increase of approximately 21 to 25°C). These data are suggestive of a possible "hot spot" on one or more regions of the laser bar. A movable slit was placed in front of the laser and the emission spectra were studied as a function of slit position. It was found that the secondary spectral peak originated at a small section of one end of the laser array bar. Subsequent SEM analysis revealed the presence of solder voids at this same location, thus substantiating the "hot spot" hypothesis. It thus appears that this spectral analysis technique may present a good screening test for identifying inadequate solder mounts for high power diode laser arrays.

DIODE LASER ARRAY PACKAGING AND LENSING

A number of multibar diode laser packages of the LDAI, bar-in-groove type have been investigated. These packages can be configured either as passively cooled (no water cooling) or actively cooled (with water cooling). Laser array modules fabricated in this package configuration generally contain between 5 and 25 1-cm wide bars, with pitches that can vary from 0.4 mm to 1.2 mm, the latter being favored for cw device operation with water cooling. A useful feature of this type of package is the ability to utilize cylindrical lens arrays with lens pitch identical to the laser bar pitch. Ten-bar, passively cooled, pulsed laser arrays (400 μm pitch) have been fabricated, together with matching, ten cylindrical-lens arrays, each lens consisting of a 0.4 mm diameter fiber. The individual fiber lenses were dropped into pre-slotted grooved ceramic lens holders which are capable of being placed in front of the ten-bar arrays. A photograph of this device is shown in Figure 12.

The mounting of such a lens assembly onto the diode laser array is a simple process, involving the placement of the lens array in front of the laser, active alignment of this lens array and then fixing it in place with a UV-cure epoxy. This lens array installation process is simple and can be completed in less than 1 hour. The typical far field pattern observed from such a lensed, 500W laser array has a rectangular cross-section beam with a divergence of approximately 7 to 10 degrees (slow axis) by approximately 3 to 4 degrees (fast axis).

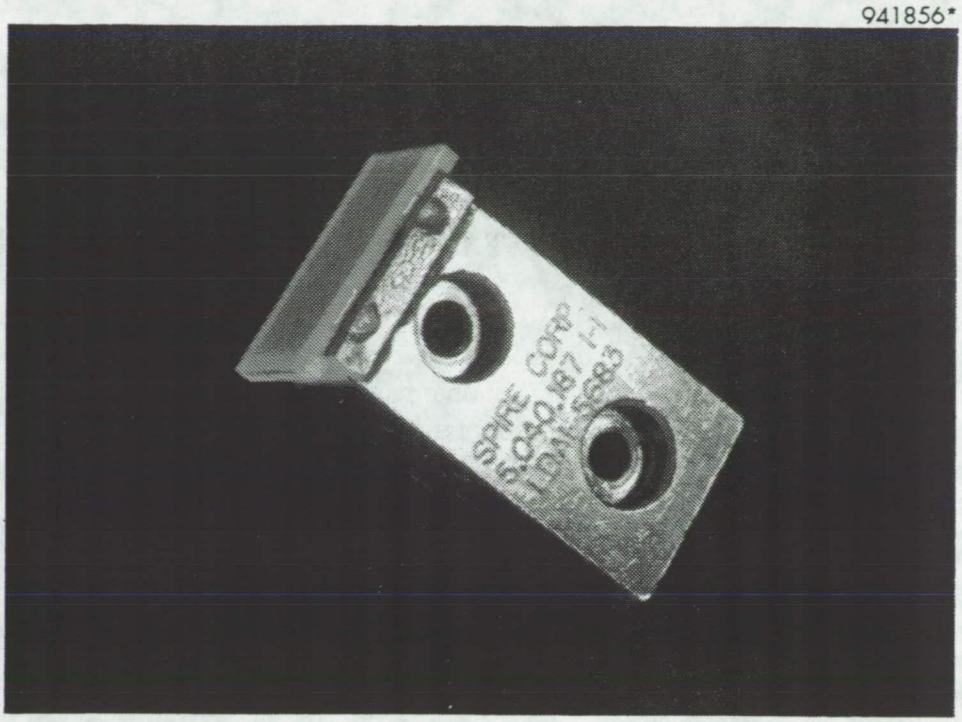


Figure 12. **Photograph of a passively cooled, 10-bar diode laser array with an integrated cylindrical collimating lens array fixed in place with UV-cure epoxy.**

SUMMARY

Recent advances in high power diode laser arrays have resulted in the commercial availability of multi-bar, cw and quasi-cw modules capable of emitting 100W cw and several KW and higher pulsed power levels in the 780 to 810 nm spectral region. A variety of diode laser bar designs are available. For cw operation, optical bar fill factors of from 20% to 50% with output power levels of up to 40 W/bar were demonstrated. Cw slope efficiencies of 0.7 to 0.98 A/W have been measured. Quasi-cw diode laser arrays have been operated with 2 ms long pulses. Thermal modeling has shown that laser temperatures are proportional to pulse length in this time-regime, with operation beyond 5 ms being basically equivalent to cw operation. An analytical technique capable of identifying laser arrays with substandard solder regions has been developed. This non-destructive technique measures the laser output spectrum at various gate times during 2 ms pulse operation, and looks for spectral regions which exhibit faster than expected spectral shifts. Multi-cylindrical lens assemblies have been shown to be readily integrated into LDAI multi-bar diode laser arrays resulting in optical output beams with divergence angles of approximately 3 x 8 degrees.

2-PHASE QUALITY/FLOW METER

Rudy J. Werlink P.E.
NASA DM-ASD
Kennedy Space Center, Fl 32899

Dr. Bob Youngquist
I-NET INI-14
Kennedy Space Center, Fl 32899

ABSTRACT

The Quality-flow meter is based on measurement of the dielectric constant variation between the phases of a 2 phase mixture. As the ratio of the two phases changes the overall capacitance changes. Monitoring capacitance changes to measure a physical phenomenon is not a new concept. Indeed, a range of sensors have been developed which utilize capacitance measurements. The present system is not significantly more sensitive than those available off-the-shelf, but demonstrates a significantly higher dynamic range and quick response than those of which this author is aware. Two prototype sensor designs have been fabricated. The First is a compact Quality meter with real time display and remote data capabilities (**Quality Meter I**). The circuit is optimized for high sensitivity, and 1 second response. The second Prototype is a Quality/Flow Meter (**Quality Meter II**)- really two sensors in series for establishing the flow rate and the quality of two phase mixtures. It's circuit is optimized for fast response. Initial testing of the sensors has been successful.

Introduction:

In the 60's a Quality meter was developed at Beech Aircraft to measure the liquid to gas ratio of cryogenic fluids[1]. The limitations of the electronics which were required to measure very low capacative levels with high accuracy, precluded practical sensor development. Currently, use of Capacitance to measure quality has already been developed as cryogenic liquid level probes where the percent of immersion is proportional to the capacitance.
One company Capacitec[2] specializes in these types of sensors and offers a variety of them. Most off-the-shelf systems assume that a small probe can be easily brought near a ground plane allowing a system to be fabricated where small changes in capacitance (e.g. 0.1 fF) are measured on a small capacitor (0.1 pF). In the present application a relatively large capacitor was required to obtain a large signal and to allow for the feed through hardware (e.g. 100 pF). Yet with the innovative microprocessor control circuit design, changes in capacitance on the order of 2 fF can be measured. This measurement corresponds to a factor of 50 improvement in dynamic range in

---

[1]KILLIAN,W.R. and SIMPSON,J.O. (1960) Measuring Vapor-Liquid Ratios During Flow by a capacitiance Method in *Advances in Cryogenic Engineering*, Vol. 5, pp. 505-508.

[2] Capacitec Corporation
87 Fitchburg Rd.
Ayer, Ma. 01432

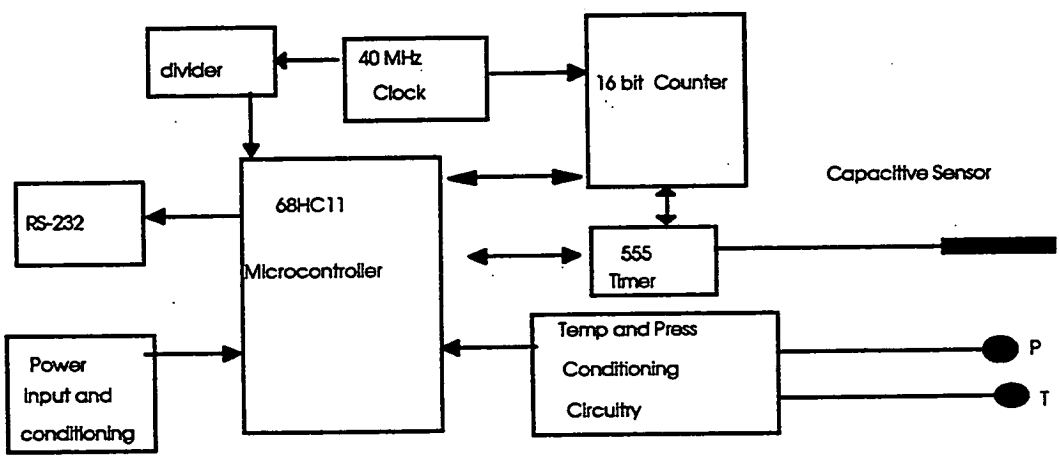
the performance of such a system. (Note that an alternative published approach simply counts the periods of an oscillator whose frequency is set by the capacitor. This approach can lead to high sensitivity and high dynamic range but suffers from a slow response time (i.e. high speed phenomena will be missed).) The present approach appears to be an ideal choice for obtaining speed, accuracy, and dynamic range in the measurement.

Electronics Operation:

The present system electronics operate by accurately and reputably measuring the time taken to charge a test capacitor. As shown in **Figure 1** this is accomplished by using a microprocessor and high speed counter to monitor a timer attached to the capacitance sensor. The microprocessor controls this process causing it to be carried out repeatedly, clearing the counter and resetting the timer before each charging cycle, and then saving the number of clock cycles needed to charge the capacitor. This information is then sent to a user via an RS-232 link. Usually, the capacitance of the sensor is relatively small so that the charging time is less than 0.1 millisecond[3]. In these cases all the data can not be shipped via an RS-232 link and consequently, a number of consecutive data samples are often averaged before shipping.

Calculations indicate that the capacitor under will change capacitance by about 13 picofarad as the fluid flow changes from gas to liquid. The sensor electronics are capable of making a capacitance measurement every millisecond to a resolution of +/-50 femtofarad $(1 \bullet 10^{-15})$ Consequently, the system can be used to monitor rapid changes on the order of one part in 260 or after averaging for one second (i.e. averaging 1000 samples) can monitor changes on the order of one part in 8000. This performance is significantly higher than that required for the present application. This enhanced performance was sought for two reasons: first, it is believed that the system performance will be limited more by thermal effects (i.e. changes in the measured capacitance due to thermal change in geometry, material dielectric constant, electronic component performance drift) than by short term noise so maximal signal was sought to help minimize these effects, and second, future applications exist for measuring flow velocity and dynamics requiring a sensitive, high speed, monitoring of capacitance.

FIGURE 1 ELECTRONICS BLOCK DIAGRAM



[3] T= R*C
where T is the time constant
R is the resistance of the circuit
C is the capacitance including the capacitance of the mixture.

Sensor Design

The unique electrical circuit was incorporated in two sensor designs:
The Quality Meter I uses a compact 12 inch long 3 inch diameter flow cross section to measure quality , pressure and temperature at a update rate of once per second. The data is displayed on a LCD readout and sent over a RS-232 connection for remote monitoring (**see Figure 2**).
A innovative capacitance measuring circuit measures capacitance with very high sensitivity. The sensing electrode uses three parallel plates held in the center of the flow stream by Teflon insulators. the electronics measure capacitance of the center plate to the side plates as the mixture passes through. Quality, pressure and temperature sensors provide real time values on the LCD display at a relatively slow rate (once per second) , this unit is a simple and rugged designed for field use where velocity is not required
The Quality/Flow Meter II is 32 inches long with a 2 inch diameter flow cross section. A center probe, 21.4 inches long is split in the center to provide two probes in series separated with a Phenolic spacer (**see Figure 3**).
The center probe is electrically isolated and held in place radial using set screws into a Phenolic pads spaced at 120 degrees around each end of the probe. The meter has an increase in inner diameter around the probe to keep the flow cross-section constant around the probe and has a vacuum jacket to minimize heat losses. The two circuits are used with each measuring the average capacitance of the mixture around each end of the probe. The mixture capacitance must be stable while flowing over the series probes to enable correlation of the capacitance readings. A Personnel Computer is used to store and analyze the readings to get the delta time of a moving 2-phase mixture. This delta time equates to velocity and flow-rate. The circuit update rate was under 1 kilohertz with averaging however the RS-232 data link only allow a 100 samples per second transmission rate. The data interface will be improved in the future to take full advantage of the circuit update rate. These sensors are designed to be used on cryogenic flows such as Liquid nitrogen, oxygen and hydrogen. Also measured on the meters are pressure and temperature of the mixture. With these parameters , density can be calculated. The equation used to arrive at the quality is :

$$\varepsilon = \frac{cap \bullet \ln \bullet (\frac{D_i}{D_o})}{2 \bullet \prod \bullet L \bullet \varepsilon_o}$$

ε = dielectric constant from mixture

cap = measured capacitance, **Do** = outside probe diameter

Di = inside diameter of pipe, ε_o = dielectric constant free space 2.249E-13 faraday/inch

$$(1-y) = \frac{(\varepsilon_f - \varepsilon)}{(\varepsilon_f - \varepsilon_g)}$$

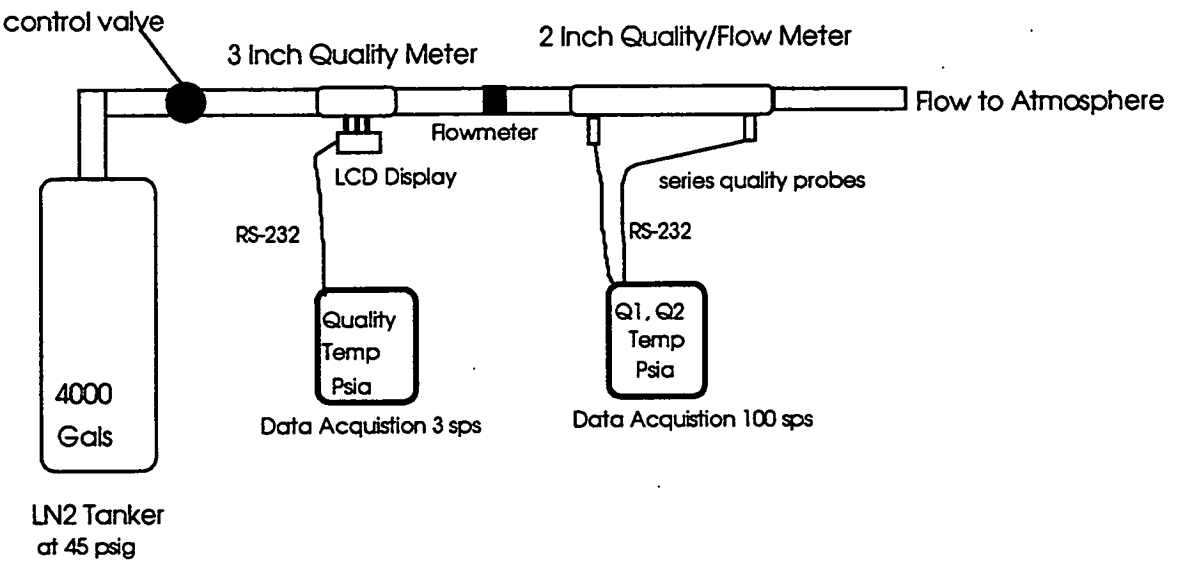
where y= liquid volume ratio, ε_f = dielectric constant of liquid

ε_g = dielectric constant of gas

Test Set-up:

The two configurations of quality meters were tested in series using a 4000 gal tanker of liquid nitrogen at 45 psig to generate a wide range of flow-rates to obtain 0-100% liquid to gas ratios. The tanker line was a 3 inch diameter Flex-hose with a control value upstream of the quality meters to regulate flow-rate. A smooth transition was used to reduce the line to 2 inches for the Quality/flow meter. The data was collected at 3 sps (samples per second) and 100 sps for each Meter separately using Labview 3.0 software and Macintosh FX computers. One computer was used for each quality meter with the system clock synchronized. The Flow-rate was varied to produce 0-400 GPM based on tanker volume (the Flow meter failed). Labview was programmed to produce a near real time cross-correlation plot for the Quality /Flow meter (Q1 and Q2) to establish the valve position and flow-rates where high correlation was indicated.

Figure 4 Quality/Flow Meter Test Setup



sps= sampling rate per second of computer

Results

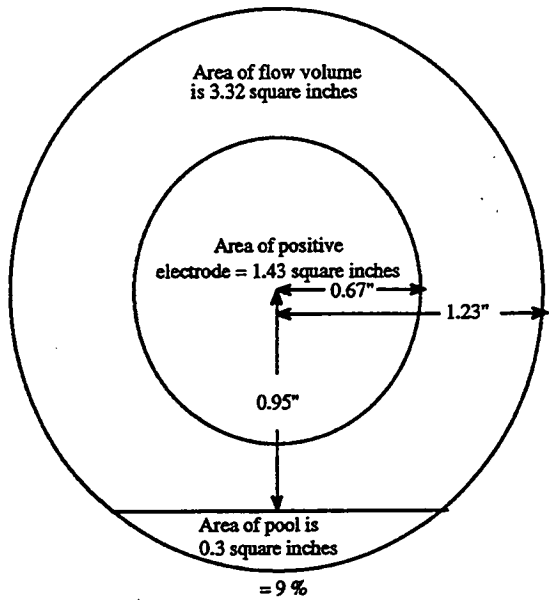
The Quality Meter II was tested as a single probe to demonstrate and verify the unit as a measurement for Quality. The Quality Meter II was tested using an LN2 Tanker with a typical data plot shown in Graph 1. The unit was zeroed at liquid nitrogen temperatures and no liquid. The span was set under full flow conditions -about 250 GPM which is assumed close to 100% by observation of the outlet flow.

Graph 2 shows Capacitance raw data vs. time plots of the Quality/Flow meter II while being tested using flowing liquid nitrogen. The upstream and downstream sensors are clearly showing a very similar response with the downstream values shifted in time. The time shift represents the average flow-rate of the liquid in the gas/liquid mixture. Graph 3 is the cross correlation of the two signals showing a 0.07 second time shift. The final Graph 4 is the two signals with the vertical axis scaled 0-1 representing the quality of the mixture from 100% gas to 100% liquid.

Explanation:

The control valve was highly restricting the flow rate while this data was taken. At higher flow-rates and hence quality the Q1 and Q2 sensors did not show related data. This was expected due to the low sampling rate of 100 sps. The quality readings remained at about 10% under no flow which we believe was caused by a pooling effect with liquid nitrogen remained in the bottom of the Quality meter II at its lowest point. Figure 5 shows a cross-section drawing of the effect. The Graph 3 readings are varying from 12 to 50% quality with a average of 30%. Subtracting the static pool leaves about 20% ratio of liquid to gas. with a measured flow velocity of 166 inches/second (.07 seconds with the Q1 and Q2 probe centers at 11.6 inches.) The liquid flow rate was then 166• 166 in/s * 3.32 (cross-section)/231 cubic inches per gallon * 60 seconds/minute * 20% volume correction = 29 GPM. This agrees with the observed exit flow and the pipe pressure of less than 1 psig.

Figure 5 Cross Sectional of Quality/Flow Meter II



Recommendation

Initial testing of the Quality/Flow meter I and II is very encouraging additional testing is planned to verify the results and extend the useful quality range for flow-rate measurement. Planned improvements of the Quality Meter include:

- Redesign of the output data rate to take maximum advantage of the Capacitance response rate.
- Optimization of the averaging of the Capacitance response circuit for the best combination of sensitivity and response rate.

- Provide Microprocessor enhancements and on board memory to conduct cross-correlation on board. This would result in flow velocity, % quality outputs directly.
- Sloping the Quality meter to prevent liquid nitrogen pooling during flow testing.

Future applications:

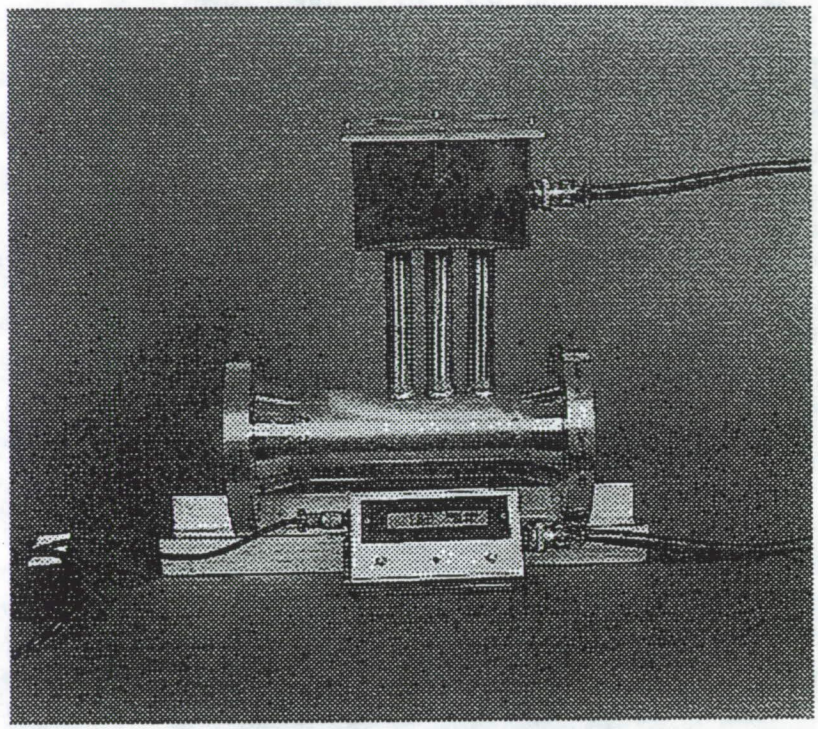
- Bi-directional Flow rate and quality near real-time with no moving parts or flow obstructions useful for Steam Plants where maintenance of high quality vapor is important.
- Food processing were the max. velocity of a solid/ water mixture is required to determine optimum sterilization times.
- Level indicator for use in modular cryogenic tanks. For example, that of the Space Station Cryo Carrier. The fill level or over-fill condition can be monitored electronically using a simple compact quality meter this is the I-NET developed configuration .

- Possible use as a humidity indicator for HVAC in commercial processes where humidity control is very important. In Shuttle and future vehicle fueling, continuous determination of dew point prior to the introduction of cryogenic propellants, precedes the costly current method of dew point determination using Sampling and off-line lab analysis.

- Fill rate (velocity) and efficiency (quality) for existing Shuttle Propellants would provide cryogenic propellant savings and time savings by determination of when 100% liquid is present during chill-down. . This is especially important in determining maximum recovery times after a revert of the fill flow cause by equipment problems.

The meter can be used for diagnosis for increased heat leaks (Insulation/vacuum piping failures). Advanced Shuttle replacement vehicles using slush hydrogen (solid-liquid phases) require density, phase and flow monitoring , which can be provided by the meter.

FIGURE 2 QUALITY METER I



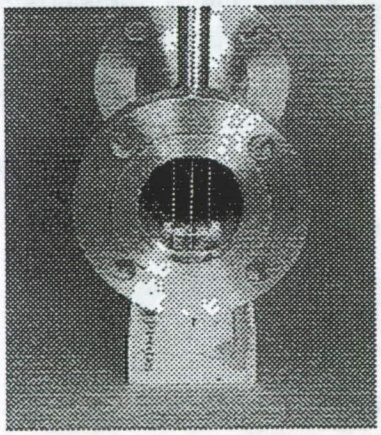
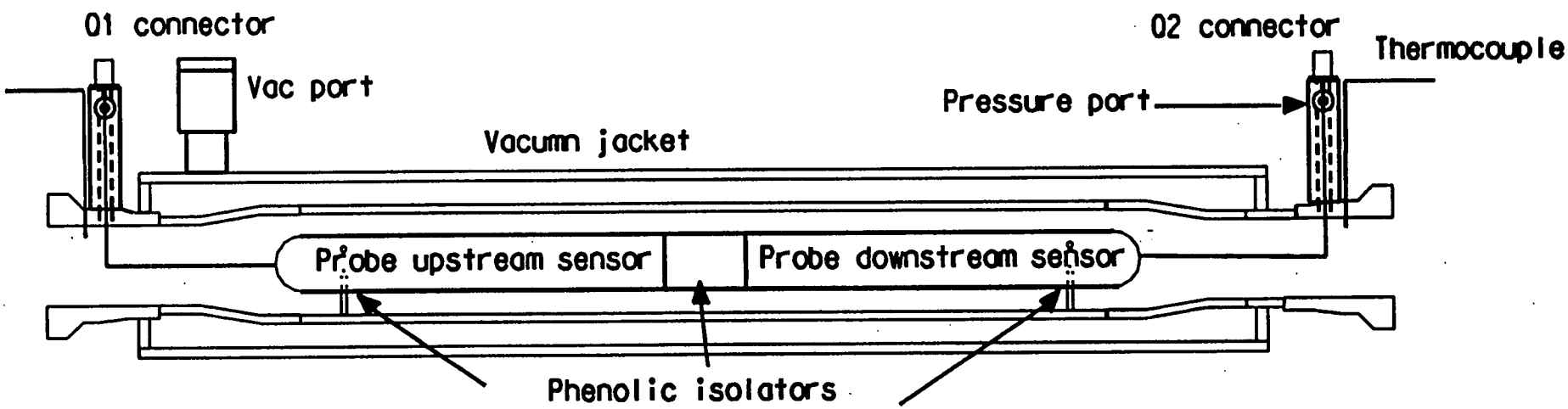
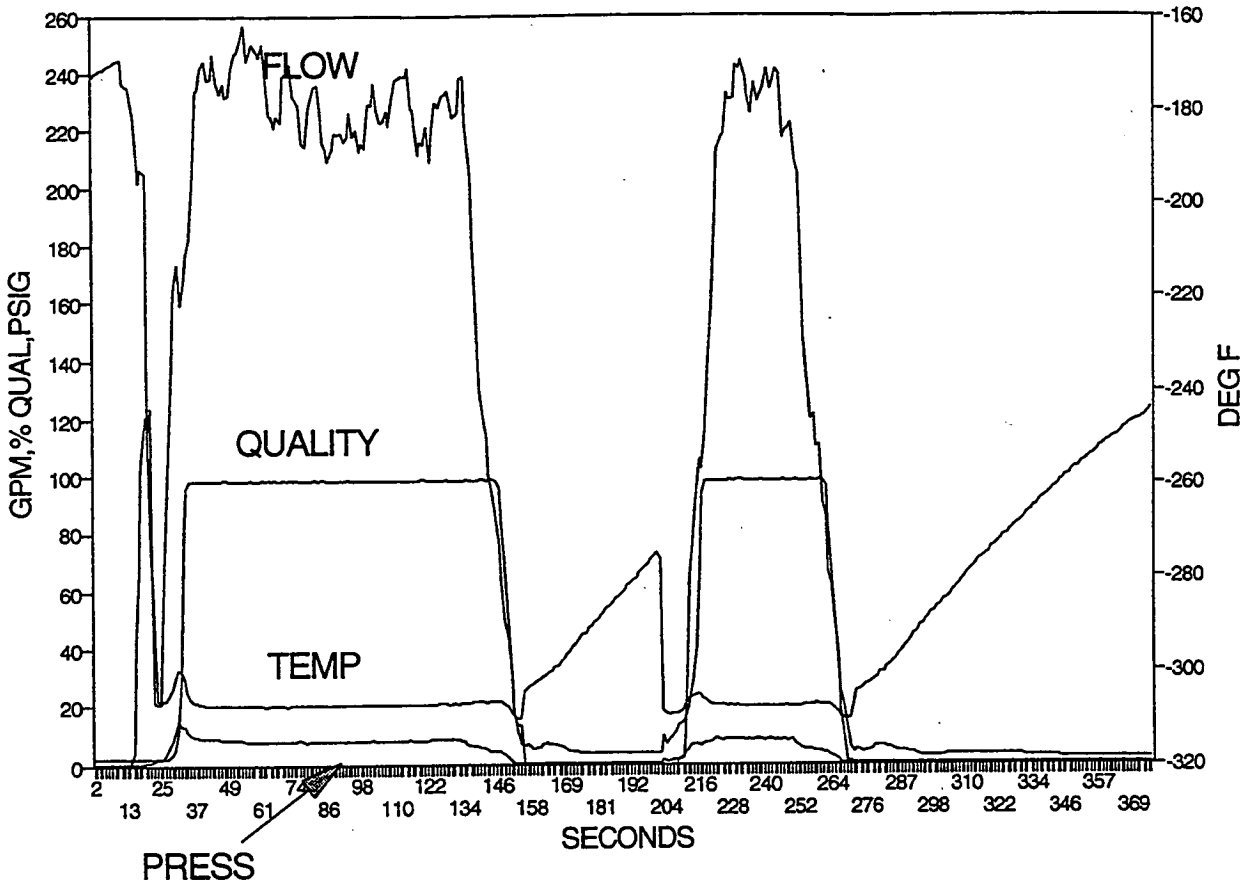


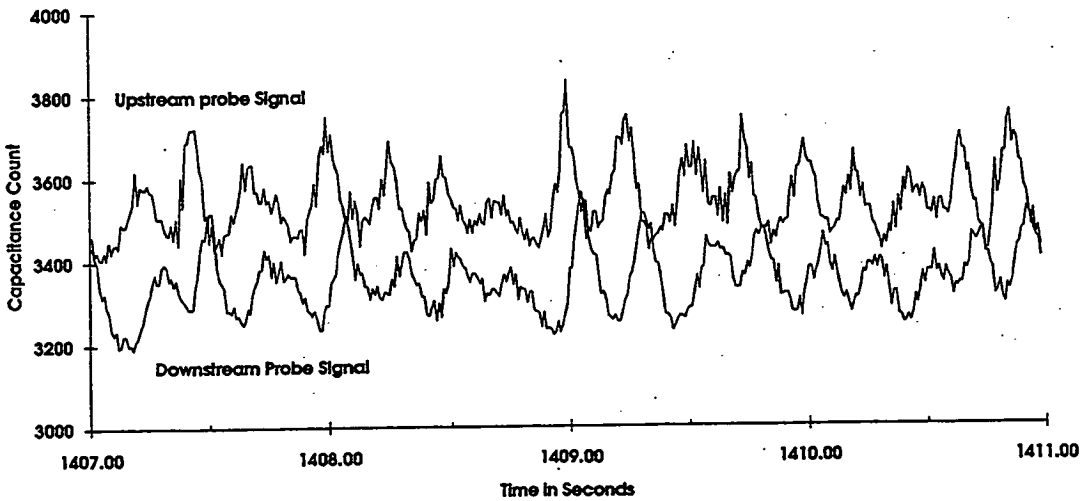
FIGURE 3 QUALITY/ FLOW METER II



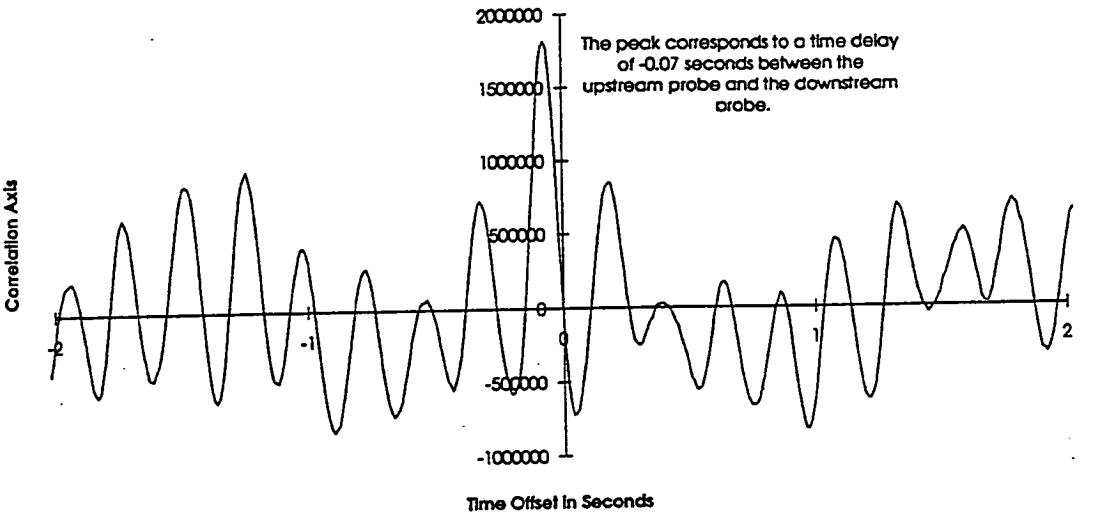
GRAPH 1 QUALITY METER II LN2 TEST
SINGLE PROBE CONFIGURATION



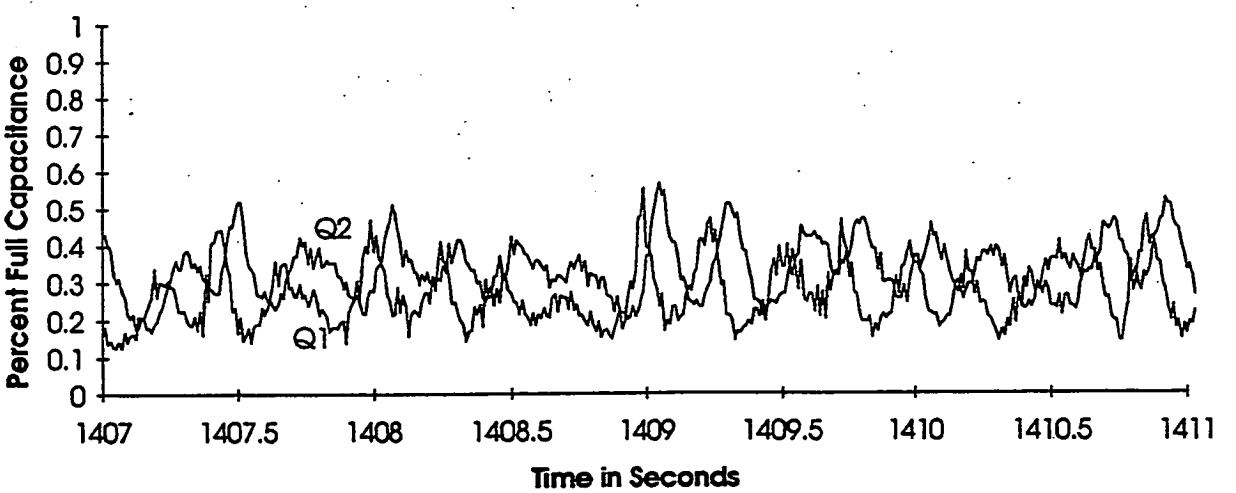
Graph 2 Quality/ Flow Meter II Q1 and Q2 with LN2 Flow



Graph 3 Cross Correleration of Q1 and Q2



Graph 4 Percent Liquid/ Gas Ratio Quality/ Flow Meter II



Spindle Thermal Error Characterization and Compensation System for Machine Tool Accuracy Enhancement

NOT AVAILABLE AT THIS TIME

A FAMILY OF SMART ACCELERATION MONITORING SYSTEMS

John C. Cole and Douglas F. Braun

Silicon Designs, Inc.
1445 NW Mall Street
Issaquah, Washington 98027
Phone: 206-391-8329

ABSTRACT

Silicon Designs has developed a family of surface micromachined accelerometers. In this paper we describe the digital accelerometers and how they can be used to build low-cost application-specific subsystems with custom ASIC and microprocessor based signal processing.

1.0 INTRODUCTION

Acceleration measurement is important in a wide range of military, industrial and commercial applications. Accelerometers for general instrumentation applications are used a few at a time in each unique application. They usually have an analog output that is amplified, filtered, digitized and stored for subsequent, off-line signal processing by a general-purpose computer. Many accelerometers may be multiplexed together for processing by a single general purpose computer. Since the accelerometers are often reused in other instrumentation applications, their cost is not a major driver in their selection.

In contrast, other accelerometer applications require accelerometers that are used for a single application, often permanently mounted on a printed circuit board. The accelerometer and its related subsystem performs a single, dedicated function, usually connected to one microprocessor or application-specific integrated circuit (ASIC) chip in order to make a decision in real-time based on recent acceleration history. Cost of the entire subsystem, its size and power consumption are often the major factors in the design of this class of subsystem. It is this class of subsystem that this paper addresses.

Silicon Designs developed and is producing the first digital, surface micromachined accelerometers, the Models 1000 and 1010, for OEM subsystems designed to be produced in large quantities at a low-cost. [1,2] They have a low power consumption (1.0 mA at 5 volts DC) and operate over a wide temperature range (-55 to +125 deg C). These accelerometer is currently available in standard packages similar to military integrated circuits, hermetically sealed leadless chip carriers or J-leaded packages.

2.0 DIGITAL ACCELEROMETER

The capacitive sense element, depicted in Figure 1, is a thin-film structure suspended above the surface of a substrate by two torsion bars. Below the sense element, on the substrate surface, are two fixed capacitor plates located symmetrically on both sides of the torsion bar axis. The sense element and the two fixed capacitor plates form two air-gap variable capacitors with a common plate.

Figure 2 shows an exploded view of the accelerometer microcircuit assembly in which a capacitive sense element chip and an integrated electronics chip are both housed in a hermetically sealed, surface mount, 20 pin leadless-ceramic-chip-carrier package. The package is about 9mm square by 3mm high.

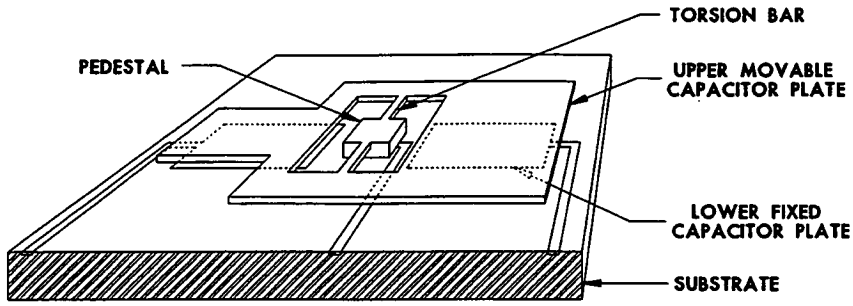


FIGURE 1: BASIC SENSE ELEMENT

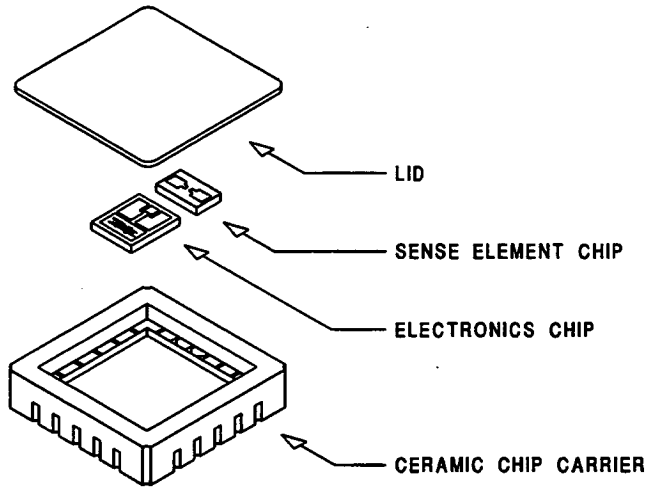


FIGURE 2: EXPLODED VIEW OF ACCELEROMETER MICROCIRCUIT ASSEMBLY

Figure 3 shows a block diagram of the digital accelerometer. Two sense elements and their corresponding plates on the surface form a fully active capacitance bridge. The accelerometer electronics measures the capacitance unbalance of the bridge and converts it into a stream of pulses with an on-chip sigma-delta A/D converter. The scaling of the output pulse frequency, or more correctly pulse density, is determined by an external digital clock input.

The pulse density output of the accelerometer is referred to as the count output or CNT. It consists of a sequence of positive going pulses that are synchronized to the clock input. During each clock period, CNT will either follow the clock signal or remain at the zero state. The pulse density of CNT, the number of pulses on the CNT output over a period of time divided by the number of clock cycles over that same period of time, is a measure of the average acceleration over that period of time. Figure 4 shows the relationship of the pulse density of the CNT output vs applied acceleration.

Examples of the various outputs available from the accelerometer are shown in Figure 5. As described before, the CNT output is a pulse synchronized to the clock. The direction or DIR output is a non-return-to-zero signal that is a logic one during clock periods when the CNT output is a pulse, and is a logic zero during clock periods when CNT stays at a logic zero. The $\overline{\text{DIR}}$ output is the complement of the DIR signal.

The following Figure 6 shows the DIR and CNT outputs for various fractions of full-scale acceleration.

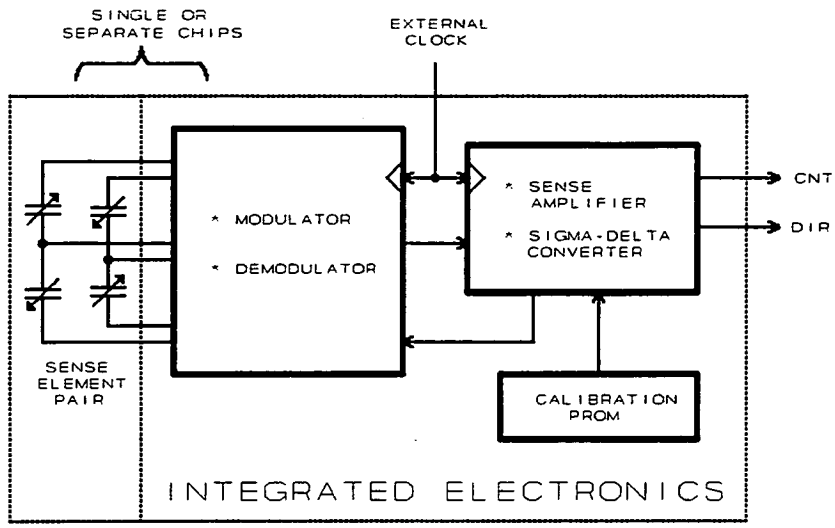


FIGURE 3: BLOCK DIAGRAM OF DIGITAL ACCELEROMETER

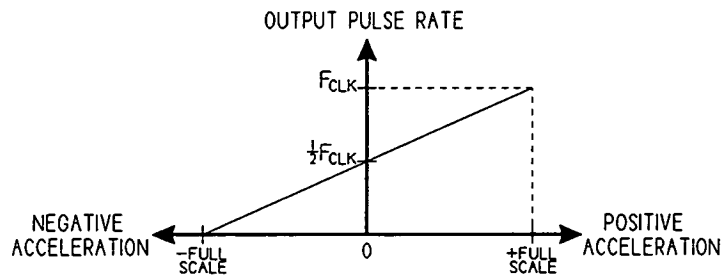


FIGURE 4: OUTPUT PULSE DENSITY OF CNT vs ACCELERATION

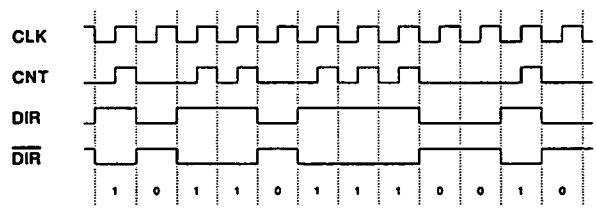


FIGURE 5: ACCELEROMETER OUTPUT WAVEFORMS

3.0 LOW-COST ACCELEROMETER SUBSYSTEMS

A minimum accelerometer subsystem is comprised of an accelerometer plus a signal processing element that performs a dedicated function. Two approaches are to use 1) a single chip microprocessor or microcomputer, and 2) an application-specific integrated circuit (ASIC).

Counters are a good way of measuring the pulse rate or pulse density of a pulse train.

One approach is to connect the CNT output of the accelerometer into the count input of a unidirectional counter which records the number of positive or negative transitions from an initial zero value. Since we are using a unidirectional counter, we can only record a positive number of events. During a time interval, the counter counts the number of pulses output from the accelerometer. The change in the counter value corresponds to the average

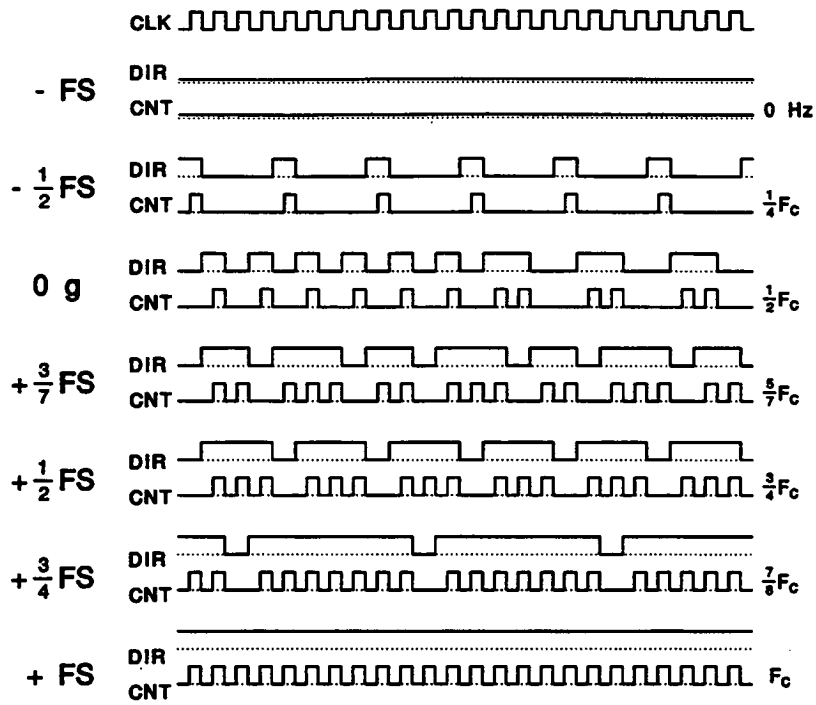


FIGURE 6: ACCELEROMETER OUTPUTS FOR VARIOUS ACCELERATIONS

acceleration over that time interval plus an offset. Since at zero acceleration, the accelerometer puts out half the clock frequency, the offset is equal to half the clock frequency times the length of the time interval.

In the second approach we can connect the DIR accelerometer output to the up/$\overline{\text{down}}$ input of an up/down or bidirectional counter, which can count up and down from an initial zero value. The clock of the up/down counter is connected to same clock as the accelerometer or its complement, depending on the phasing required. This has the advantage over the first approach that the counter reflects the signed value of the average acceleration over the time interval between readings and does not have an offset.

Because the accelerometer puts out a pulse train proportional to acceleration, when the pulse train drives a counter which counts the number of pulses, the counter acts as an integrator. The value stored in the counter is the integral of the acceleration or change in velocity, delta velocity, since the counter was reset. If the counter is read and reset periodically, then the sequence of counter values correspond to the average acceleration over the time period. If the counter is reset when the acceleration is zero, then the counter value each time it is read (without being reset) is proportional to the current velocity.

3.1 ASIC-Based Subsystems

The advantages of an accelerometer with a digital output show up well when a dedicated ASIC is used to implement the processing algorithms. With analog accelerometers the ASIC would need to have an A/D converter or use analog circuits for processing. When the accelerometer output is already in digital form, special purpose processing can be done with low-cost gate array or standard cell circuits from many digital IC vendors. The functions can be built as separate chips installed separately or in the same package as the accelerometer or added to the accelerometer electronics chip. Use of such chips and available CAD software can cut many months off the development time compared with analog or combined analog/digital ASICs.

Figure 7 shows the generic block diagram of an accelerometer subsystem using an ASIC circuit. The ASIC circuits we show in the following sections make maximum use of counters and a little known but highly useful

digital circuit, the binary rate multiplier (BRM), for performing arithmetic on digital pulse data.

A BRM takes a parallel word, which represents a binary fraction, and a clock signal as inputs and produces a digital pulse stream out whose pulse density is equal to the clock frequency multiplied by the binary fraction.

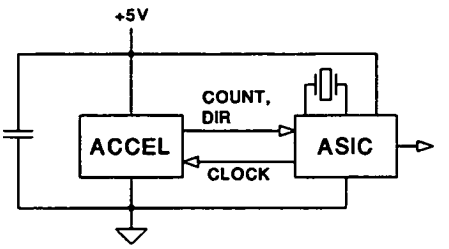


FIGURE 7: BLOCK DIAGRAM OF ASIC SUBSYSTEM

Figure 8 shows by example how a BRM works. The BRM contains a binary counter, a 3-bit counter in this example. The counter is driven by a clock input. The first three columns represent the counter value as it is driven through its eight states counting from 0 to 7.

COUNTER VALUES			ZERO TO ONE TRANSITION			OUTPUT FOR 5/8 VALUE
0	0	0	0	0	0	0
0	0	1	0	0	1	1
0	1	0	0	1	0	0
0	1	1	0	0	1	1
1	0	0	1	0	0	1
1	0	1	0	0	1	1
1	1	0	0	1	0	0
1	1	1	0	0	1	1

FIGURE 8: EXAMPLE OF A 3-BIT BINARY RATE MULTIPLIER OPERATION

The next three columns show the clock cycles where the corresponding bit makes a 0 to 1 transition. Note that the least significant bit makes four transitions from 0 to 1, the second bit two transitions and the most significant bit one transition; also note that no two bits have 0 to 1 transitions on the same clock cycle. If we were to generate a pulse train for each counter bit, where a pulse occurs when the counter bit makes a 0-to-1 transition, then we have three pulse trains with pulse densities equal to 1/2, 1/4 and 1/8th the clock frequency. If we wanted a pulse train with a density equal to 5/8 of the clock frequency (where 5/8 is the binary fraction 0.101), then we can get such a pulse train by taking the logical OR of those pulse trains where the binary fraction bit is a 1 (the first and third columns for a binary $5_{10} = 101_2$). The last column shows the OR of the first and third columns.

A discrete part that performs this function is the CD4089, made by Harris, National and SGS-Thomson. The 7497 is also a BRM.

3.1.1 Full-Scale Threshold Accelerometer

In looking back at Figure 4, we see that the pulse density of the CNT output increases as the acceleration increases. As we apply an acceleration closer to the full-scale value, we get more and more pulses in a row with only a few intervening periods without pulses. When we are at full-scale acceleration, a pulse occurs during every

clock period. Looking at DIR, when we are at full-scale, we see that DIR is always high.

We can build a simple threshold accelerometer with logic that looks for an unbroken string of ones on the DIR output. When more than a defined number of ones in a row occurs, we know that the acceleration is near or above the accelerometer full-scale value. By adjusting the full-scale value of the accelerometer during calibration, we can change the threshold acceleration value to any value within the calibration range. With the present Model 1000 and 1010 accelerometers, this threshold value can be any value from 5 to 200g. We expect to widen this range with future accelerometer versions.

Figure 9 shows a logic circuit that detects when 15 or more ones are detected on the DIR signal without an intervening zero. When this occurs, it generates a logic one on a threshold output that indicates that the threshold has been reached. The output returns to a logic zero when a zero is detected on the DIR signal. This output can be used to set a latch that generates a signal to a processor that the acceleration has exceeded a threshold until it is cleared by the processor.

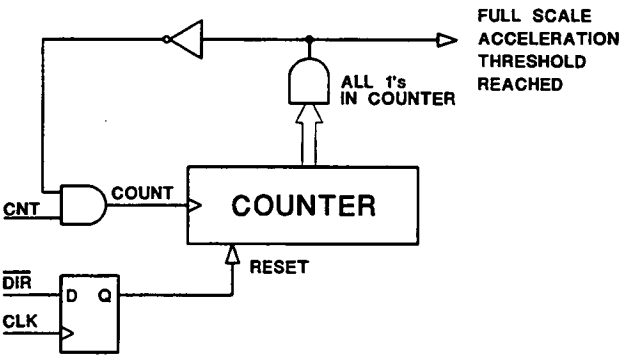


FIGURE 9: BLOCK DIAGRAM OF THRESHOLD DETECTION LOGIC

The advantages of this approach over using an analog accelerometer together with an analog comparator are its low cost, size and power requirement, since it uses only a fraction of a digital ASIC chip, and its use of digital logic that is not temperature sensitive to set the threshold value. The digital logic could also be integrated into the accelerometer itself, resulting in a threshold accelerometer the same size as the Model 1000.

3.1.2 Multiple Threshold Accelerometer

Suppose that we want to divide the acceleration range up into N non-overlapping regions with N digital outputs that indicate when the acceleration is in the corresponding range. Using analog techniques, this would require (N-1) comparators and some digital logic to combine the comparator outputs.

Figure 10 shows a simple way of performing this function with the digital accelerometer and a small amount of digital logic. Over some time interval we need to count the number of clock periods that have pulses on the CNT output or for which DIR is a logic one. At the end of the time interval, we transfer the value in the counter to a register and zero the counter for the next time period. By decoding several high order bits in the register, we can determine between what two values is the count stored in the register. If we want to combine some ranges, we can do this by combining their respective outputs with an OR gate.

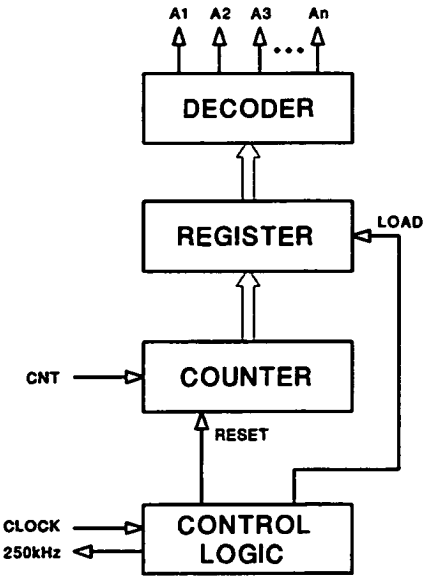


FIGURE 10: BLOCK DIAGRAM OF MULTIPLE THRESHOLD ACCELEROMETER

3.1.3 Distance Measurement Subsystem

A major use of accelerometers in the military is to determine the velocity and/or position of a vehicle by numerically integrating the acceleration. Likewise, by integrating the velocity we can determine the distance travelled. This function can be done easily in software by a microprocessor; however, it is also possible to perform it in an ASIC using counters and a BRM.

Suppose we have a vehicle that is initially at rest. Suppose we want to know the velocity and distance that a vehicle has travelled in real time as it accelerates along its longitudinal axis, and we want to generate an output when a specified distance has been reached. A digital circuit which performs this function is shown in Figure 11.

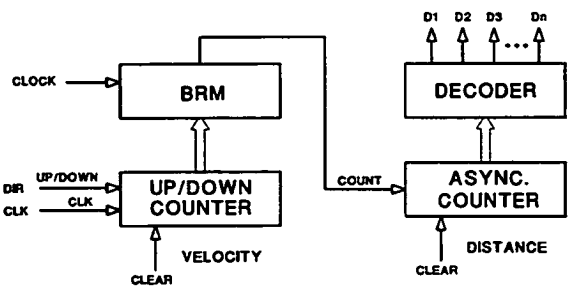


FIGURE 11: BLOCK DIAGRAM OF DISTANCE MEASUREMENT LOGIC

The instantaneous velocity is relatively easy to obtain. As described before, an up/down counter connected to the DIR output of the accelerometer is a way of measuring a change in velocity since the counter value was reset. By resetting the counter value when the vehicle is at rest, the counter value at each instant of time will be a measure of the vehicle velocity. Each bit of the counter corresponds to a velocity of

2 * (Accelerometer Full-Scale) / (Clock Frequency)

If we can produce a pulse train where the pulse density is proportional to velocity, we can use a counter to accumulate distance similar to the velocity counter above. This counter will then accumulate the distance travelled. A BRM with its binary input connected to the velocity counter will produce such a pulse stream. The BRM drives a counter where each bit of the counter corresponds to a distance travelled of about

$$2*(\text{Accel. Full-Scale})*2^{(\text{no. of velocity counter bits})}/(\text{Clk. Freq.})$$

3.2 Microprocessor-based Smart Subsystems

The advantages of an accelerometer with a digital output also show up well when a microprocessor is used to implement the processing algorithms. The model 1010 digital accelerometer is well suited for use with processors that include an on-chip event counter. Microprocessors having such a counter include the Intel 8051 and similar units from Siemens, Signetics, National, Motorola, Texas Instruments and Microchip. Figure 12 shows a 3-axis crash sensor system that uses three Silicon Designs' model 1010 accelerometers and three Microchip PIC17C42Q 8-bit microcontrollers. It is important to note that no interfacing circuitry is required between the accelerometer and the microcontroller.

In this design, the on-chip counters within the microcontrollers accumulate the acceleration information and perform the necessary signal processing to implement the crash detection function. Once the crash event is detected, one or more of the three channels will assert their IMPACT logic output to trigger the firing device. One advantage of this microcontroller based system is that self test and reasonableness checks can be easily implemented in software. One drawback of using such as system is that it is difficult to "prove" the safety of the software for all cases that can occur. A hardware based algorithm to perform the crash detection function is described in the next section. This hardware based algorithm can be implemented in software.

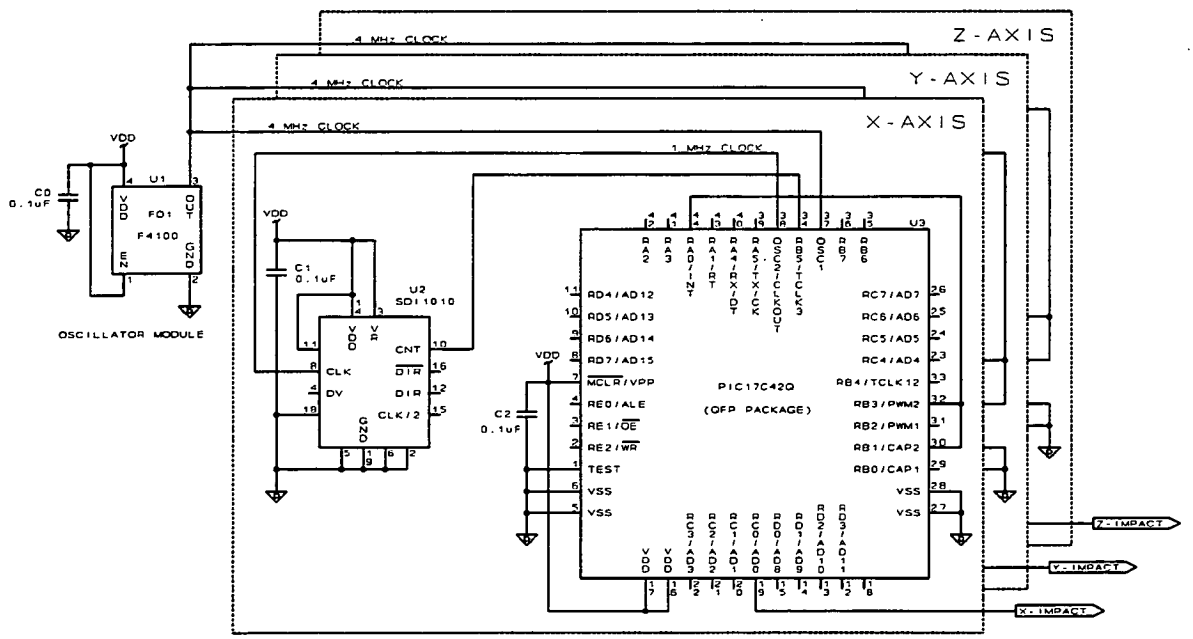


FIGURE 12: THREE-AXIS, MICROCONTOLLER BASED, CRASH SENSOR

3.2.1 Air Bag Algorithm

The basic requirement for air bag deployment is to trigger reliably on crashes into a barrier for velocities greater than 18 miles per hour (MPH) and not trigger when the crash velocity is less than 10 MPH. The air bag also must not trigger on vibration or sudden shock by a hammer blow. It is desirable to trigger as early as possible to minimize the distance that the drivers head has moved when triggering occurs.

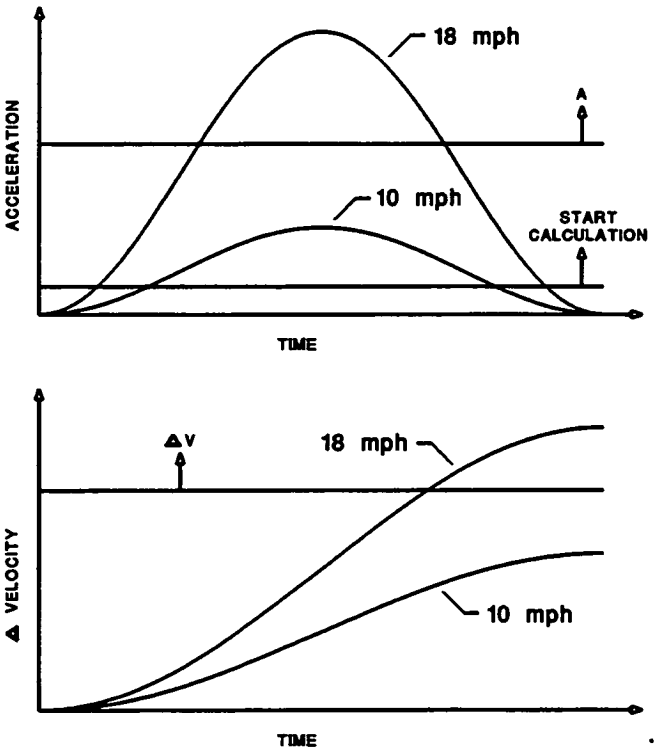


FIGURE 13: CRASH EVENTS – ACCELERATION AND CHANGE IN VELOCITY vs TIME CURVES

A simple acceleration model for an automobile crash is an acceleration vs. time curve that is in the shape of a haversine. The peak acceleration is reached in half the time it takes the vehicle to come to a complete stop and is an inverse function of the crush distance of the particular vehicle. Figure 13 shows the acceleration and change in velocity curves of 10 and 18 MPH crashes, where the crush distances of the 10 and 18 MPH crashes are 9 and 16 inches respectively; typical values for a passenger car. In both curves, deceleration is shown as positive.

A useful way of depicting the algorithm choices for a crash event is to plot the acceleration as a function of the change in velocity. Figure 14 shows such curves for 10, 14 and 18 MPH crashes. Clearly crashes at lower than 10 MPH have curves that fall inside the 10 MPH curve, and crashes at above 18 MPH have curves that are above the 18 MPH curve.

If we use these two parameters -- instantaneous acceleration and change in velocity -- in our crash detection algorithm, the algorithm can be represented by showing the region of this space which results in an air bag deployment. The 14 MPH crash curve is interesting because it separates the 10 and lower MPH crashes from the 18 and higher MPH crashes.

In practice, the region at low velocities and acceleration can not be used because of errors in the measurement and resonances of the body structure. A suitable algorithm might be to choose the earliest point on the 14 MPH crash curve which provides enough separation from the 10 and 18 MPH crash curves to accommodate system tolerances, such as body resonances and accelerometer bias and scale factor errors. Deployment would occur when

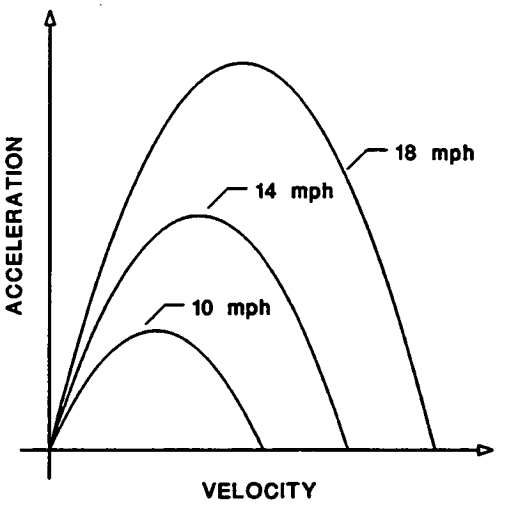


FIGURE 14: CRASH ACCELERATION vs CHANGE IN VELOCITY

both the change in velocity and acceleration have exceeded the selected values during an event when acceleration exceeds a minimum threshold. We have picked the peak value of the 14 MPH curve as being a reasonable value that provides about a 30% margin between threshold acceleration value and the 10 and 18 MPH values. Figure 15 shows the region which would correspond to air bag deployment.

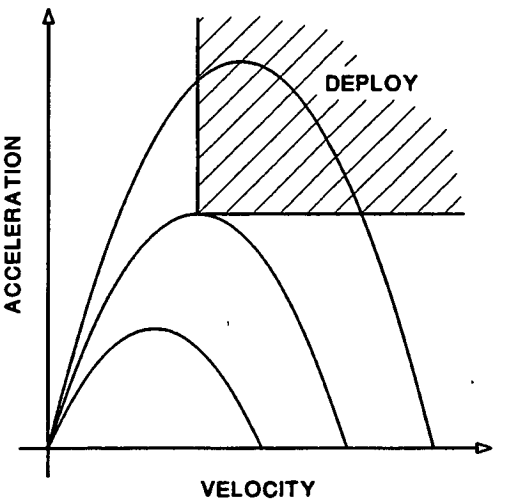


FIGURE 15: DEPLOYMENT REGION OF ACCELERATION/ CHANGE IN VELOCITY SPACE

The deployment algorithm is then implemented by looking at the values of acceleration and change in velocity in the two counters as a crash event unfolds, looking for those values that have been selected for the type of vehicle selected. When both values have been equalled or exceeded, then the air bag is triggered.

BIBLIOGRAPHY

[1] Cole, John C., A New Capacitive Technology for Low-Cost Accelerometer Applications, Proceedings - Sensors Expo 1989, p. 106B.

[2] Cole, John C., A New Sense Element Technology for Accelerometer Subsystems, Proceedings of International Conference on Solid-State Sensors and Actuators (Transducers '91), IEEE, pp 93-96.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1995	Conference Pubication

4. TITLE AND SUBTITLE

Technology 2004, Volume 1

5. FUNDING NUMBERS

6. AUTHOR(S)

Compiled by the NASA Center for AeroSpace Information

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA CP-3313, Vol. 1

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified/unlimited
Subject Category 99

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Proceedings from symposia of the Technology 2004 Conference, November 8-10, 1994, Washington, DC. Volume 1 features papers on advanced manufacturing, biotechnology, computer aided design and engineering, materials science, power and energy,and test and measurement.

14. SUBJECT TERMS

Manufacturing, biotechnology, computer aided design, ceramics, lubricants, fuel cells, electric motors, flowmeters, holography

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited