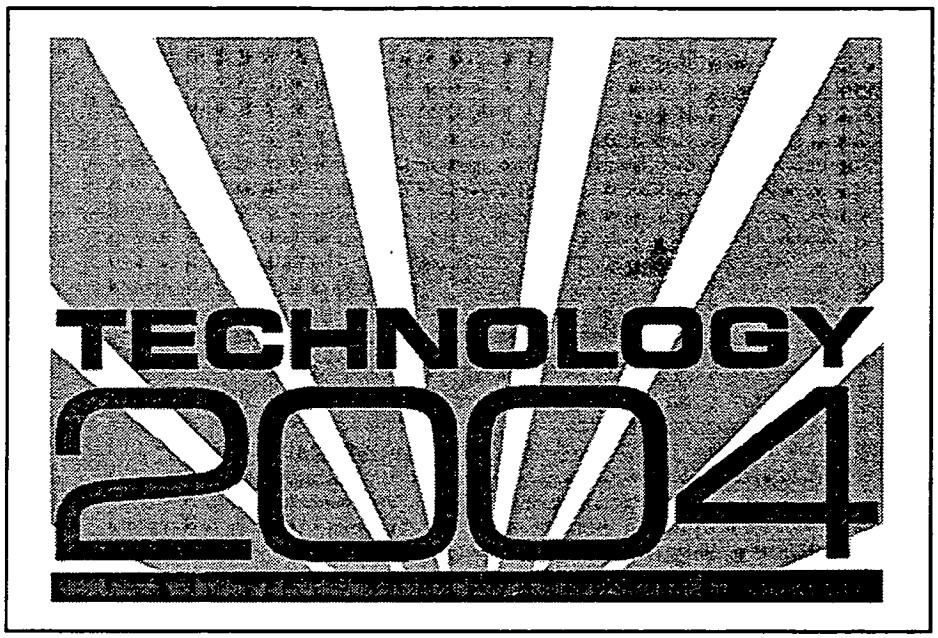
NASA Conference Publication 3313
Volume Two



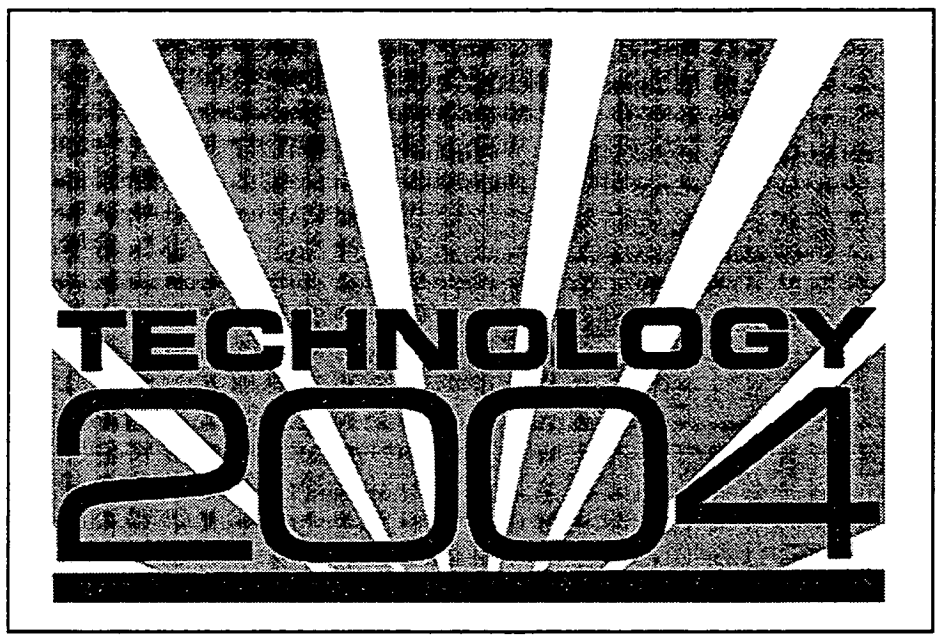
Conference Proceedings

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 Two



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 Two

Computers and Software

Virtual Reality Simulation

Environmental Technology

Pressurized Spray Stripping and Cleaning Techniques (Not available at this time)

Turbine/Brush Recycling Pipe Cleaning System

Photovoltaic Power Without Batteries For Continuous Cathodic Protection and an Alternative Photovoltaic/Ultracapacitor Combined Power Source

Video and Imaging

Flat Panel Planar Optic Display

Dual-Mode Non-Washout Liquid Crystal Display

Development of CMOS Active Pixel Image Sensors for Low Cost Commercial Applications

Enhancing the Galileo Data Return Using Advanced Source and Channel Coding

Optical Product Grade Sensor for Process Control

On-Line Analysis and Process Control Using Line-Scan Video Photometry

Automated Die and Wire Bond Inspection Using Machine Vision for Multi-Chip Module Manufacturing

3-D Digital Robotic Optical Inspection Device

Medical Technology and Life Sciences

Mammographic Computer Assisted Diagnosis Using Computational Statistics Pattern Recognition

A New Robot for High Dexterity Microsurgery

Combined Supercritical Air and Body Cooling Development

Robotics and Artificial Intelligence

Tripod Operators for Realtime Recognition of Surface Shapes in Range Images

Development of a Commercially Viable, Modular Autonomous Robotic Systems for Converting Any Vehicle to Autonomous Control

Navy Omni-Directional Vehicle (ODV) Development and Technology Transfer Opportunities

An Integrated Fault Tolerant Robotic Controller System for High Reliability and Safety

Advanced Graphical Programming Environment for Robotic Applications

Sensor Skin

Mobile Robotic System for Servicing of the Space Shuttle Lower Surface Tiles

Robot Control in Dynamic and Uncertain Environments with Known Objects

Electronics

Fiber Optic Communication Systems for Spaceflight and Avionics Applications (Not available at this time)

The Global Positioning System (GPS) Service: A Technology Ripe for Commercial Innovation (Not available at this time)

SATS: Small, Automated Tracking System—Elements of a Better System for Satellite Tracking and Telemetry

Person Locator System

High-Speed/High-Precision Analog-to-Digital Converter for All-Digital Radio/Television

High Frequency Electronic Packaging Technology

Aerogels for Electronics

A 3.5 W Output, Diode-Pumped, Q-Switched 532 nm Nd:Yag Laser Pumped by Fiber-Coupled Diode Lasers

Computers and Software

THE DEVELOPMENT OF A LOSSLESS DATA COMPRESSION TECHNOLOGY FOR NASA MISSIONS AND ITS APPLICATION TO MEDICAL IMAGING[1]

Pen-Shu Yeh and Warner H. Miller
Goddard Space Flight Center, Code 738.3
Greenbelt, MD 20771
T: 301.286.4477 F:301.286.1751 EMail:psyeh@psy.gsfc.nasa.gov
T:301.286.8183 F:301.286.1751 EMail:whmiller@gsfcmail.nasa.gov

ABSTRACT

Lossless data compression has been studied for many NASA missions to achieve the benefit of increased science return; reduced onboard memory requirement, station contact time and communication bandwidth. This paper first addresses the requirement for onboard applications and provides rational for the selection of the Rice algorithm among other available techniques. A top-level description of the Rice algorithm will be given, along with some new capabilities already implemented in both software and hardware VLSI forms. The paper provides several case study examples drawn from a broad spectrum of science instruments including the thematic mapper, x-ray telescope, gamma-ray spectrometer, acousto-optical spectrometer. The status of the technology will be given with its application to medical imaging.

INTRODUCTION

With the development of new advanced instruments for remote sensing applications, sensor data will be generated at a rate that not only requires increased onboard processing, storage capability, but imposes demands on the communication link and ground data management system. Data compression provides a viable means to alleviate these demands. Two types of data compression have been studied by many researchers in the area of information theory: a lossless technique that guarantees full reconstruction of the data, and a lossy technique which generally gives higher data compaction ratio but incurs distortion in the reconstructed data. To satisfy the many science disciplines NASA supports, lossless data compression becomes the priority for technology development in this area.

To implement a data compression technique on the spacecraft, several criteria are considered:

1. The algorithm has to adapt to the changes in data to maximize performance.
2. It can be easily implemented with few processing steps, small memory and little power.
3. It can be easily interfaced with a packetized data system without performance degradation.

There exist a few well known lossless compression techniques including Huffman code, arithmetic code, Ziv-Lempel algorithm and variants of each. After extensive study and performance comparison on the same test image data set [1],[2], the Rice algorithm originated at Jet Propulsion Laboratories [3] is selected for implementation.

The Rice algorithm is essentially a set of Huffman codes organized in a structure that does not

---

1. Part of the paper was presented in the International Geoscience and Remote Sensing Symposium, 94 and the 1994 Science Information Management and Data Compression Workshop at Goddard Space Flight Center.

require lookup tables. The set of the Huffman codes can be easily extended to the information range of the science data. It is adaptive to the changes in the statistics of the data, and can be easily implemented. The structure of the algorithm also permits simple interface to data packetization scheme without having to carry side information across packet boundary. Therefore its performance is file size independent.

In 1991, a hardware engineering model was built in an Application Specific Integrated Circuit (ASIC) for proof of concept. This particular chip set was named as Universal Source Encoder/ Universal Source Decoder (USE/USD) [1]. Later, it was redesigned with several additional capabilities and implemented in Very Large Scale Integration (VLSI) circuits using gate arrays suitable for space missions. The flight circuit is referred to as Universal Source Encoder for Space (USES). The fabricated USES chip is capable of processing data up to 20 Msamples/second and will take data of quantization from 4-bit to 15-bit [4].

A description of the Rice algorithm will be given in the next section, followed by systems issues and case study examples on remote sensing data either acquired from launched spacecrafts or simulated for future missions.

THE RICE ALGORITHM ARCHITECTURE

A block diagram of the architecture of the Rice algorithm [5] is given in Figure 1. It consists of a preprocessor to decorrelate data samples and subsequently map them into symbols suitable for the following stage of entropy coding. The entropy coding module is a collection of options operating in parallel over a large entropy range. The option yielding the least number of coding bits will be selected. This selection is performed over a block of J samples to achieve adaptability to scene statistics. An Identification (ID) bit pattern is used to identify the option selected for each block of J input data.

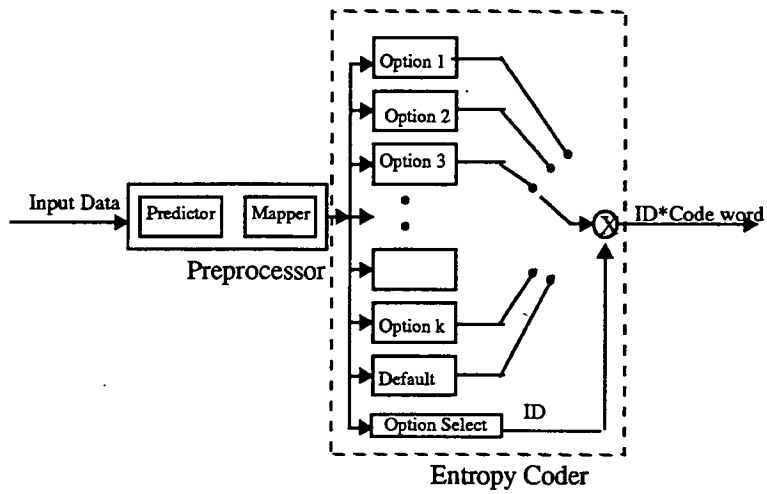


Figure 1. The encoder architecture.

The Preprocessor

The predictor in the preprocessor can be as simple as a first order predictor using previous sample, or other higher order predictors. To maintain the pipeline processing in the hardware, only a few predictor types are implemented, these include: a 1D predictor, a 2D predictor, a multispectral predictor and a user-supplied external predictor.

The function of the predictive coder is to decorrelate the incoming data stream by taking the difference between data symbols. The mapper takes these difference values, both positive and negative, and orders them, based on predictive values, sequentially into positive integers.

Entropy Coder

Most of the options in the entropy coder are called "sample-split options". These options take a block of J preprocessed data samples, split off the k least significant bits, and code the remaining higher order bits with a simple comma code before appending the split-off bits to the coded data stream. Each sample-split option in the Rice algorithm is optimal in an entropy range about one bit/sample [2]; only the one yielding the least amount of coding bits will be chosen and identified for a J-sample data block by the option select logic. This assures that the block will be coded with one of the available Huffman codes, whose performance is better than other available options on the same block of data. The $k = 0$ option is optimal in the entropy range of 1.5 - 2.5 bit/sample; the $k = 1$ option is optimal in the range of 2.5 - 3.5 bit/sample, and so on for other k values.

To improve the performance below 1 bit/sample, a new option is devised and included in the full set of options implemented in VLSI. This new option is particularly efficient over data with very low entropy values.

The default option is an option not to use any of the split-sample options or the low-entropy option. It bounds the performance of the algorithm by simply passing through the preprocessed block of data through the encoder without alteration but with an appended identifier.

CASE STUDY EXAMPLES

This section contains several compression study results for several different instruments. The compression performance is expressed as Compression Ratio (CR). It is defined as the ratio of the quantization level in bits to the average code word length, also in bits. It should be noted that the CR value is data dependent and can vary from one test data set to the next.

Landsat Thematic Mapper

Mission Purpose: The Landsat program was initiated for the study of Earth's surface and resources. Landsat-1, 2, and 3 were launched between 1972 and 1978. Landsat-4 was launched in 1982, and Landsat-5 in 1984.

Landsat Thematic Mapper (TM) on Landsat-4 and 5: The TM data represent typical land observation data. An image acquired on Landsat-4 at 30m ground resolution for band 1 in the wavelength region of 0.45 - 0.52 μm is shown in Figure 2. This 8-bit 512x512 image was taken over Sierra Nevada in California.

Compression Study: Using a 1D predictor in the horizontal direction, setting a block size of 16 samples and inserting one reference per every image line, the lossless compression gives a compression ratio at 1.83 for the 8-bit image.

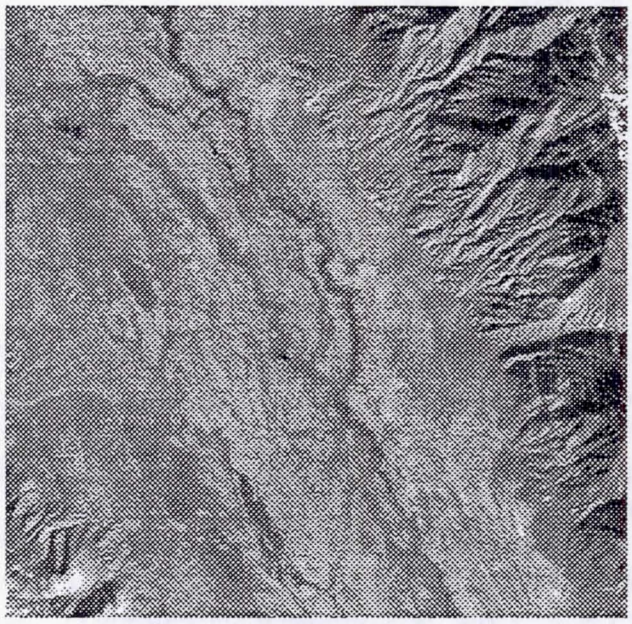


Figure 2. Thematic Mapper image

Soft X-ray Telescope (SXT) on Solar-A Mission

Mission Purpose: The Solar-A mission, renamed as Yohkoh mission after its successful launch in August, 1991, is dedicated to the study of solar flares, especially of high-energy phenomena observed in the X- and gamma-ray ranges.

Soft X-ray Telescope (SXT): The instrument detects X-ray in the wavelength range of 3-60 Angstrom. It uses a 1024x1024 CCD detector array to cover the whole Solar disk. Data acquired from the CCD is of 12-bit quantization and is processed on board to provide 8-bit telemetry data. The image in Figure 3 is an averaged image of size 512x512 with dynamic range up to 15 bits in floating point format as a result of further ground processing.

Compression Study: The test image is first rounded to the nearest integer. Then a 1D predictor is applied to this seemingly high-contrast image. A compression ratio of 4.69 is achieved.

Figure 3. Solar-A X-ray image

Acousto-Optical Spectrometer (AOS) on Submillimeter Wave Astronomy Satellite (SWAS)

Mission Purpose: The Submillimeter Wave Astronomy Satellite (SWAS) is a Small Explorer (SMEX) mission, scheduled for launch in the summer of 1995 aboard a Pegasus launcher. The objective of the SWAS is to study the energy balance and physical conditions of the molecular clouds in the Galaxy by observing the radio-wave spectrum specific to certain molecules.

Acousto-Optical Spectrometer: The AOS utilizes a Bragg cell to convert the radio frequency energy from the SWAS submillimeter receiver into an acoustic wave, which then diffracts a laser beam onto a CCD array. The sensor has 1450 elements with 16-bit readout. A typical spectrum is shown in Figure 4(a). An expanded view of a portion of two spectral traces is given in Figure 4(b). Because of the detector nonuniformity, the difference in the Analog-to-Digital Converter (ADC) gain between even-odd channels, and effects caused by temperature variations, the spectra have nonuniform offset values between traces, in addition to the saw-tooth-shaped variation between samples within a trace. Because of limited available onboard memory, a compression ratio of over 2:1 is required for this mission

Compression Study: 1D prediction between samples is ineffective when the odd and even channels have different ADC gains. Using the multispectral predictor mode, the achievable CR is 2.32.

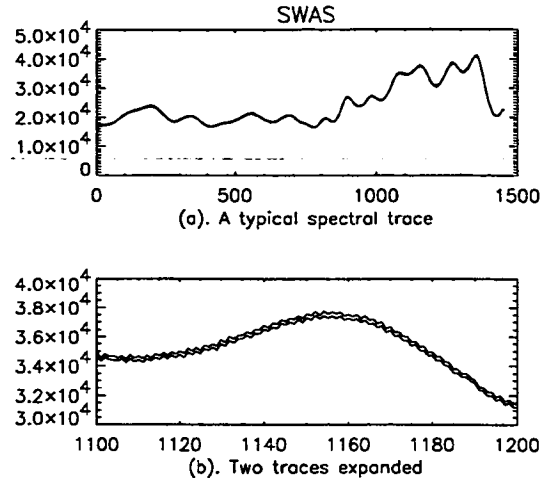


Figure 4. AOS radio wave spectrum

Gamma-Ray Spectrometer on Mars Observer

Mission Purpose: The Mars Observer was launched in September 1992. The Observer will collect data through several instruments to help the scientists understand the Martian surface, atmospheric properties and the interactions between the various elements involved. In the summer of 93, contact with the spacecraft was lost.

Gamma-Ray Spectrometer (GRS): The spectrometer uses a high-purity germanium detector for gamma rays. The flight spectrum is collected over sixteen thousand channels. The total energy range of a spectrum extends from 0.2 Mev to 10 Mev. Typical spectra for a 5-second and a 50-second collection time are given in Figure 5. These spectra show the random nature of the count. The spectral count dynamic range is 8-bit.

Compression Study: The achievable compression depends on the channel collection time. At 5-second collection time CR is over 20 and it decreases as collection increases. At 20-second collection time, CR is over 10.

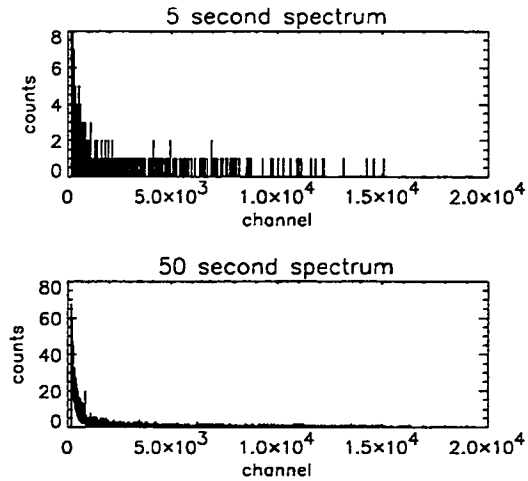


Figure 5. Gamma-ray spectrum

TECHNOLOGY STATUS

Current status of this lossless data compression technology is:

- Both hardware flight qualified silicon (VLSI) encoder chip and software in C are developed and fully tested.
- Decompression VLSI chip is currently in design phase, with delivery planned in 95.
- Goddard Aerospace Data Systems Standard for lossless data compression has been written based on the extended Rice algorithm and is currently in review cycle at Goddard Space Flight Center.
- The algorithm is implemented in software in four flight instruments: gamma-ray spectrometer on Mars Observer, SWAS on the Small Explorer, gas chromatograph-mass spectrometer on the CASSINI mission and a spectrometer on the Mars-96 mission.
- The VLSI hardware encoder is under integration on two flight missions: the "Lewis" mission of the Small Satellite Technology Initiative (SSTI) awarded to TRW, and the Solar Extreme UV Rocket Telescope Spectrograph (SERTS) built by Goddard Space Flight Center.
- The hardware encoder is under consideration by several defense agencies for potential applications.

MEDICAL IMAGING APPLICATIONS

Advancement in the development of medical imaging equipment has created the need for handling large amounts of digital data. The data carries information on the pathology of patients and is acquired in different imaging modalities, such as CT, MRI, ultrasound, etc. Often, the data has to be first stored on memory device, later processed, analyzed, displayed and sometimes transmitted across network to other institution. Inherent in these processes are the requirements for massive storage device, faster device access time, communication bandwidth and affordable data transmission time. Data compression offers a viable solution to alleviate these demands.

Similar to the remote sensing data which will be processed by many science disciplines, the medical imaging data will be analyzed for various changes in chemical, physical, structural components of body parts by physicians specialized in various fields. The desire to preserve the highest quality of the data mandates that a lossless data compression technique to be applied whenever possible. A study has been performed on a suite of medical data including CT, MRI, digitized angiograms, nuclear medicine images and reported in [6]. The study compared the compression ratio performance among several commercially available techniques with the extended Rice algorithm. It clearly shows the superior performance of the latter. In addition, there are other implementation issues that would support the application of NASA's technology to the medical imaging field [7].

CONCLUSION

A lossless data compression technology has been successfully developed for remote sensing applications. This technology is based on the extended Rice algorithm. The performance of the algorithm has been established through analysis and simulation. Hardware in VLSI form as well as software are currently available for space flight missions. Over a dozen case studies have been performed on post-flight data and several new missions have adopted the technology for onboard

implementation. The lossless compression technique has been evaluated on medical imaging data including CT, MRI, nuclear medicine data, angiograms and others. Preliminary results strongly support transfer of the technology to the field of medical imaging.

REFERENCES

[1] Venbrux, Jack, Pen-Shu Yeh and Muye N. Liu, "A VLSI Chip Set for High-Speed Lossless Data Compression," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 2, No. 4, Dec. 1992

[2] Yeh, Pen-Shu, Robert F. Rice and Warner H. Miller, "On the Optimality of Code Options for a Universal Noiseless Coder," Revised version, *JPL Publication 91-2,* 1991. A shortened version "On the Optimality of a Universal Noiseless Coder," is published in the *Proceedings of the AIAA Computing in Aerospace 9 Conference*, San Diego, CA, Oct. 19-21, 1993.

[3] Rice, Robert F., "Some practical universal noiseless coding techniques," *JPL Publication 79-22*, 1979. Available from author, JPL, 4800 Oak Grove Drive, Pasadena, CA. 91109, USA.

[4] MRC, "Universal Source Encoder for Space - USES," Preliminary Product Specification, Version 2.0, Microelectronics Research Center, University of New Mexico, 1993.

[5] Rice, Robert F., Pen-Shu Yeh and Warner H. Miller, "Algorithms for a Very High Speed Universal Noiseless Coding Module," *JPL Publication 91-1*, 1991. A shortened version "Algorithms for High-Speed Universal Noiseless Coding," is published in the *Proceedings of the AIAA Computing in Aerospace 9 Conference,* San Diego, CA, Oct. 19-21, 1993

[6] Yeh, Pen-Shu and Miller, Warner, "The Development of Lossless Data Compression Technology for Remote Sensing Applications," *International Geoscience and Remote Sensing Symposium Digest*, 1994.

[7] Venbrux, J. Yeh, Pen-Shu, Zweigle, G., and Vesel, J. "A VLSI Chip Solution for Lossless Medical Imagery Compression," *Proceedings of the SPIE Medical Imaging,* 1994.

A SYSTEM DESIGN FOR A HYBRID NETWORK TERMINAL USING ASYMMETRIC TCP/IP TO SUPPORT INTERNET APPLICATIONS

ABSTRACT

Access to the Internet is either too slow (e.g. dial-up SLIP) or too expensive (e.g. switched 56 kbps, frame relay) for the home user or small enterprise. The Center for Satellite and Hybrid Communication Networks and Hughes Network Systems have collaborated to develop a prototype of a low-cost hybrid (dial-up and satellite) network terminal which can deliver data from the Internet to the user at rates up to 160 kbps. An asymmetric TCP/IP connection is used breaking the network link into two physical channels: a terrestrial dial-up link and a receive-only satellite link. This system has been designed to support any Intel 80386/486 PC, any commercial TCP/IP package, any unmodified host on the Internet, and any of the routers, etc. within the Internet. The design exploits the following three observations: 1) satellites are able to offer high bandwidth connections to a large geographical area, 2) a receive-only VSAT is cheap to manufacture and easier to install than one which can also transmit, and 3) most computer users, especially those in a home environment, will want to consumer much more data than they generate. IP encapsulation, or tunneling, is used to manipulate the TCP/IP protocols to route packets asymmetrically.

A SYSTEM DESIGN FOR A HYBRID NETWORK TERMINAL USING ASYMMETRIC TCP/IP TO SUPPORT INTERNET APPLICATIONS

Aaron Falk
Institute for Systems Research
University of Maryland, College Park, Maryland

Narin Suphasindhu
Institute for Systems Research
University of Maryland, College Park, Maryland

Doug Dillon
Hughes Network Systems
Germantown, Maryland

John Baras
Institute for Systems Research
University of Maryland, College Park, Maryland

PURPOSE OF THIS PAPER

The purpose of this paper is to describe the system design process used to build a prototype terminal which uses a hybrid satellite-terrestrial connection and "asymmetric" TCP/IP network protocols to give the computer user who has a minimum bandwidth modem connection to the Internet access to bandwidth hungry Internet applications.

INTRODUCTION

Using hybrid networking, the terminal will merge two connections, a bidirectional terrestrial link using a modem and a receive-only satellite link, so that the TCP/IP software above the device driver sees one "virtual" device. This design exploits three concepts: 1) satellites are able to offer high bandwidth connections to a large geographical area, 2) a receive-only VSAT is cheap to manufacture and easier to install than one which can also transmit, and 3) most computer users, especially those in a home environment, will want to consume much more data than they will generate.

This design supports any 80386 or 80486 processor PC, any TCP/IP package, access to any existing Internet host, and any commercial SLIP provider. These design drivers should maximize the commercial potential of the system.

In the hybrid network, a client will transmit requests over the terrestrial link to a server located somewhere on the Internet. The server will route the reply to a satellite uplink where it will be broadcast to all the clients on the system creating a virtual "ethernet in the sky." The client terminal will trap only the packets with the correct address and send them up to the application which requested them.

The potential user of this terminal would be a computer user in a home or small business without high bandwidth access to the Internet. High bandwidth access means anything higher than current modem speeds. A user with an IBM PC compatible and a modem connection to the Internet running the SLIP protocol has been used to define the user requirements.

Internet applications have been chosen as a user requirement for this design because there exists a vast number of enormous databases available on the Internet, much of it accessible by using applications such as FTP, Gopher, Archie, WAIS, or Mosaic. The Internet is the closest existing prototype of the National Information Infrastructure and if this hybrid, asymmetric link design has performance and cost advantages over other methods of accessing information on the Internet, then it may be instrumental in demonstrating satellites' significance in the development of the NII.

While the goal has been to provide the user with a terminal, it has been necessary to design an entire system of which the terminal itself is only a single subsystem in order to interoperate with the existing Internet. It would not be possible to create this design without considering the "big picture" since modifications of the hardware and physical, link, and network layers are all required to make the system interoperable with the existing hardware and protocols. Therefore, this design is an excellent example of the use of systems integration techniques to incorporate the super-system architecture into subsystem designs.

THE HYBRID NETWORK TERMINAL DESIGN

Students and faculty from the University of Maryland collaborated with engineers from Hughes Network Systems to create a prototype system that was demonstrated to offer up to 200kbps receivable data. The user terminal in the system used a Hughes device known as a BIC or DirecPC card to interface the computer with a 0.6 meter antenna.

The key problem to solve was how to force routing of packets on the Internet without requiring special modifications to the user application or the destination machine the user was trying to reach. Solving this problem meant manipulating the two primary Internet packet protocols: Transport Control Protocol (TCP) and Internet Protocol (IP) to affect the routing we wanted.

TCP/IP is a connectionless protocol where each packet is addressed with its destination and the network finds the best path each time a packet is sent into the system. Address Resolution Protocol (ARP) is used to locate machines on a local network once a packet has arrived at the gateway between the destination machine and the Internet. A typical client-server exchange includes the client creating a request packet which includes the server's IP address (the destination) and the client's own IP address (the source). Once the server has received the request and generated a reply, it merely swaps the two addresses from the request header and sends out the reply to the client.

The following sections describe in detail the design of the prototype system. The section on user requirements describes what the system must accomplish from the user's perspective. Following that are sections highlighting development of each subsystem.

The User & User Requirements

Developing the definition of user requirements is an iterative process with the goal of producing a set of system parameters which collectively satisfy the ultimate goal of the system. It is useful to explicitly (as much as is possible for a non-existent product) identify the potential user/customer, first. This will drive the definition of the user's requirements. With their experience in providing network services, Hughes Network Systems was instrumental in this task. The user for this system can be defined as the following:

> A personal computer owner located in a home or small business in the continental US who has an interest in accessing the Internet using applications such as FTP, Gopher, Mosaic, News, or Archie with the lowest delay possible. The user is capable of installing peripheral cards and software in his computer but may require professional assistance with the installation of a satellite dish. The user may be willing to spend the same amount of money on this system as for a high speed modem. The basic user hardware configuration is an IBM or IBM-compatible PC and a modem capable of at least 2400 baud. The basic user software configuration is DOS 5.0, Microsoft Windows 3.1, and a commercial TCP/IP package that includes client versions of the above mentioned Internet applications.

The user definition, in turn, generates several requirements for the system. These user requirements are system parameters that the system must satisfy in order to satisfy the user requirements. Each of these requirements are listed below:

- The system must provide significantly less delay responding to requests for large data files experienced using a modem.
- The system must work with any 386/486 33MHz machine.
- The system must work with any commercial TCP/IP package.

- The system must work with any commercial SLIP service provider.
- The system must be able to access any Internet host.
- The system must support Internet initiated connections.

Satisfying the user requirements while minimizing user cost and development time has justified the design of the different subsystems within this system.

How It Works

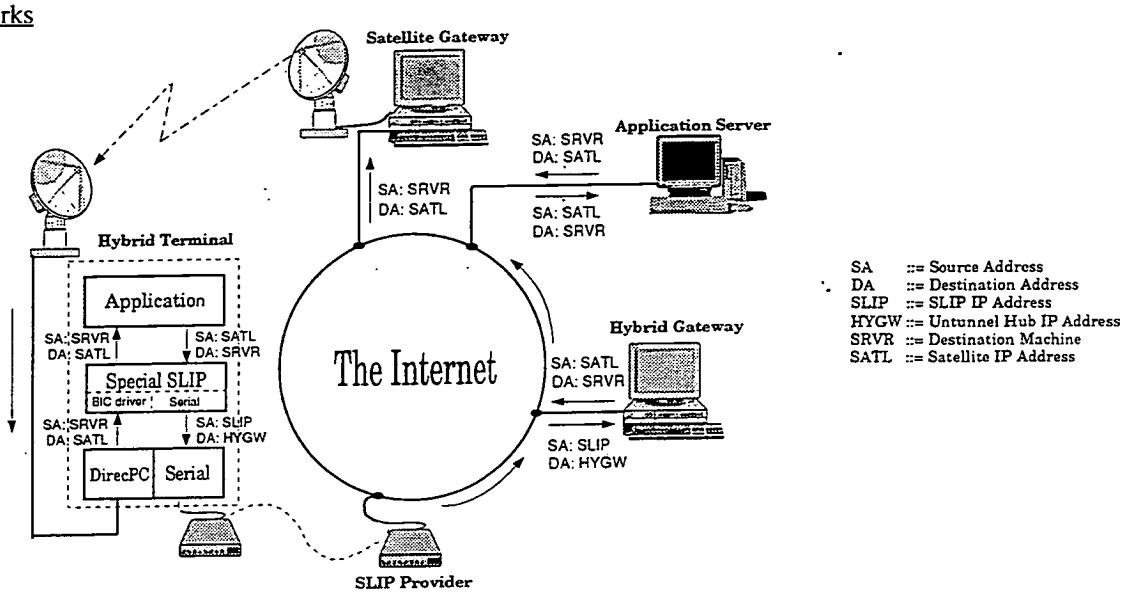


Fig. 1 The path travelled by a tunneled packet generated by the Hybrid Terminal. For simplicity, the diagram does not show that, before a packet can get from the Application Server to the Satellite Gateway, it must first be sent to the Hybrid Gateway to be encapsulated in a special satellite packet.

In this section we describe the general process of how a request from the user terminal is carried through the Internet and to a machine running a host application and how that machine's response is carried back to the user. The various subsystems are introduced here and will be described in more detail in the following sections.

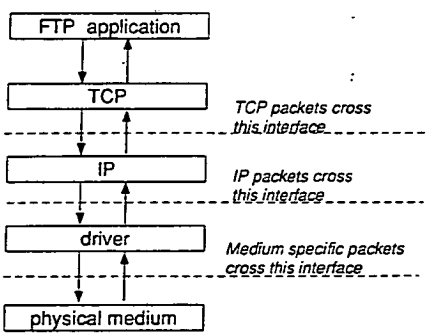


Fig. 2 The software stack when running FTP (File Transfer Protocol).

Before describing an example of how the system routes packets, it is important to point out that the user terminal is given two IP addresses. One IP address corresponds to the SLIP interface and would by typically be assigned by the SLIP service provider. The other IP address corresponds to the satellite interface and would be assigned by the hybrid service provider. These IP addresses correspond to completely different networks. Observe that the SLIP service does not need to know anything about the satellite IP address or even whether the user is using the hybrid service. If

a host somewhere in the Internet is trying to deliver a packet to the satellite IP address by using the Internet routing scheme of routers, gateways, and ARPs, the *only* way that packet can reach the satellite IP interface is to traverse the satellite by being routed to the satellite gateway.

When requesting a data transfer, say using FTP, the user sends a request to a remote machine that is running FTP server software. This software receives file transfer requests and responds to them in the appropriate fashion. If a hybrid terminal user wanted to receive a file from a machine running FTP server (we'll call it the Application Server), every packet from the user terminal would take the following path:

1) Within the User Terminal, Hybrid Host, the FTP client software generate a request and pass it to the TCP/IP module. TCP/IP would place the request first in a TCP packet then in an IP packet which would then be passed to the Special SLIP driver software. This request would have a source IP address corresponding to the satellite interface and the destination IP address of the Application Server.

2) In Special SLIP, the IP packet is encapsulated, or tunneled, inside of another IP packet and sent over the modem connection to the SLIP server host. The encapsulation amounts to adding a new IP header in front of the original one with a source address corresponding to the SLIP interface and a destination address corresponding to the machine we are calling Hybrid Gateway.

3) SLIP server receives the IP packet analyzes the tunneling header and, thinking it is destined for Hybrid Gateway, uses standard Internet routing to send the packet to Hybrid Gateway.

4) When Hybrid Gateway receives the packet it strips off the tunneling header, revealing the true header with Application Server as the destination. The packet is then sent back out into the Internet.

5) Internet routing takes the packet to the Application Server which replies with the requested file and addresses the reply to the request's source IP address, i.e. the IP address of the User Terminal's satellite interface.

6) In order to find the user terminal's satellite interface, the Internet routing protocol will send the packet to the subnet containing a router/gateway connected to Hybrid Gateway. When that router/gateway sends out an ARP for any user terminals' satellite IP address Hybrid Gateway responds and says "send it to me."

7) Once Hybrid Gateway receives the reply packet, it encapsulates it in a special packet format that is used over the satellite link and uses the satellite IP address to uniquely identify the satellite packet's destination. Then Hybrid Gateway sends the packet over ethernet to the Satellite Gateway.

8) Satellite Gateway broadcasts over the satellite link any packets it receives from Hybrid Gateway.

9) The driver in Hybrid Host that services the DirecPC card scans all packets broadcast over the satellite looking for the satellite IP address in the header. Once it identifies one, it captures it, strips off the satellite header revealing the reply IP packet, and sends it to the Special SLIP driver.

10) The special SLIP driver calls the TCP/IP package notifying it that it has received an IP packet and passes up the reply, completing the transaction.

The User Terminal

The User Terminal has required the most development. Device driver software has been developed that will appear to an off-the-shelf TCP/IP package that the computer is connected to an ethernet card when it is actually connected to a satellite dish and a modem. At the same time it must appear to the SLIP Server that the computer has a single IP address assigned by the SLIP provider, and force the Internet to route IP packet replies to a different IP address than the requests originated from. For this last task the Hybrid Gateway is needed also.

The TCP/IP package includes some of the Internet applications that the user wants to run. It also, of course,

contains the TCP and IP protocol stacks. In a normal configuration, the TCP/IP package would sit on top of a driver that would talk to an ethernet card, providing a fast connection, or a modem via the computers serial communications port, providing a slow connection. With a normal symmetric connection, TCP/IP would send and receive data over the network by passing and receiving frames across a software interface to the driver. The driver would handle the moving the frames back and forth over the physical connection to the network.

In the hybrid configuration the interface between TCP/IP and the driver doesn't change. It can't if we are going to support the user requirement of being able to use an off-the-shelf package without modification. However, instead of communicating with a single physical network, the driver for the hybrid terminal, which we are calling Special SLIP, communicates with two physical networks as shown in the following diagram:

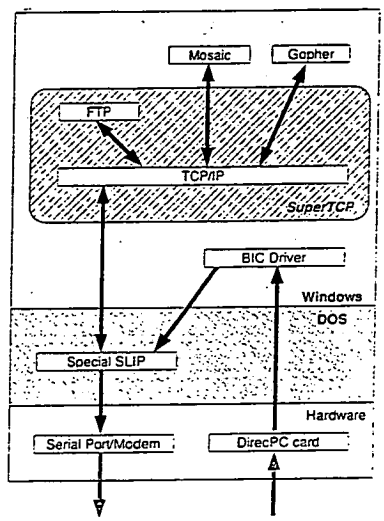


Fig. 3 The data paths within the User Terminal take data in off the DirecPC card and send data out to the modem.

Serial port handling

The serial port provides the physical connection to the modem and, through it, the terrestrial network. The Serial Line IP (SLIP) protocol will be used over the terrestrial connection. SLIP is the crudest form of IP protocol for serial lines. Its function is to delimit IP packets by inserting a control character (hex 0xC0) between them. To insure that a data byte is not mistaken for the control character, all outgoing data is scanned for instances of the control character which is replaced by a two character string. The protocol is described in more detail in [Romkey, 88].

An interrupt service routine is required to handle hardware interrupts coming asynchronously from the UART. On an interrupt, the routine should read or write a byte to the UART. This interrupt service routine (isr) needs to conform to the standard DOS calling conventions, i.e. it needs to chain off an interrupt vector stored in PC memory, service the interrupt as quickly as possible, and send an end-of-interrupt to the system to allow other interrupts to get through. This code should be configurable as which chip, connection parameters, and COM port are used and it should load at system boot time.

BIC driver call handling

The BIC driver has been developed by HNS. The driver's functions include scanning all packets transmitted over the satellite channel for one with a header corresponding to the IP address of the satellite interface, performing some error detection and correction on the packet, buffering the received packet, and passing it to the Special SLIP driver. It will call the Special SLIP driver using the DOS IOCTL_output_cmd() call and will pass the address and length of a received packet in the BIC driver's buffers. Special SLIP needs to copy the data out of the BIC's buffers as quickly as possible and pass it up to the TCP/IP package.

TCP/IP call support

There are two popular interfaces between network drivers and network software: the Crynwr/Clarkson Packet Driver Specification and the 3Com/Microsoft Network Driver Interface Specification (NDIS). The NDIS specification has greater complexity but also greater functionality and has somewhat greater industry acceptance. Also, the Frontier SuperTCP TCP/IP package that is being used for development supports the NDIS standard and not the Clarkson so that is what was used in this design.

Tunneling

One of the most innovative concepts incorporated in this design is the tunneling of IP packets to fool the Internet routing scheme. This idea was proposed by Doug Dillon of HNS and developed by the author with assistance from Narin Suphasindhu.

The reason for tunneling is this: the user terminal has two IP addresses associated with it—one for the SLIP interface which is assigned by the SLIP provider, a commercial service the developers have no control over, and the other corresponding to the satellite interface, assigned by HNS and essentially an extension of the uplink network. The way to get the Internet to route packets to the satellite interface when the request came from the SLIP interface is to set the source IP address in the request packet to be the satellite IP address. That way, when the Application Server forms its reply to the request, it addresses the reply to the source address, i.e. the satellite IP address.

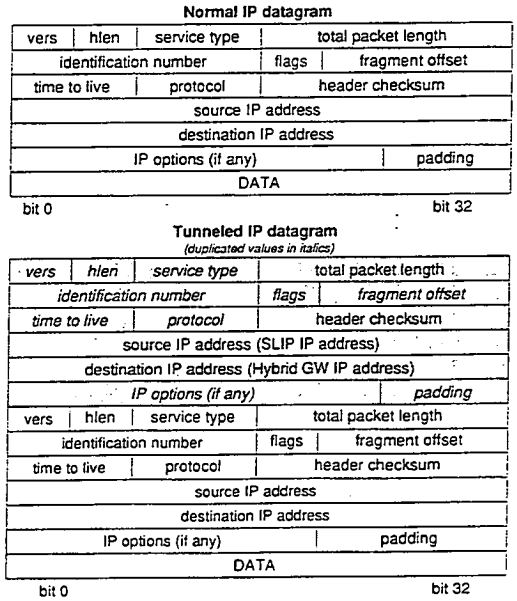


Fig. 4 When tunneling an IP datagram, a complete IP header is appended before the "true" IP header duplicating all values from the "true" header except the source and destination addresses, the total packet length, and the header checksum.

There is a complication, however. When SLIP service is purchased from a commercial provider, the provider assigns a single IP address and may pass IP packets containing the assigned address as the source address. If the SLIP provider thought that an entire network was going to connect through the account, the traffic rate would be higher and the provider could, reasonably, ask for larger fee. Therefore, it is not unreasonable to assume that the SLIP provider may be checking the source IP addresses of traffic flowing from the user terminal into the SLIP server. If the user terminal is changing the source IP addresses to match that of the satellite interface, the SLIP provider could declare a violation of the service agreement or maybe just drop the packets. Either event is unacceptable.

To cope with this possibility, IP encapsulation or tunneling is used to bypass the address checking that may occur in

the SLIP server. In tunneling, every IP packet passed to the Special SLIP driver by the TCP/IP package has the satellite IP as its source address and some Application Server as its destination address and is encapsulated in *another* IP packet which has SLIP IP as its source address and Hybrid Gateway as its destination address. The "encapsulation" really just amounts to adding a new header in front of the true one. The effect of the new header is to route every packet to the Hybrid Gateway. At the Hybrid Gateway, the tunneling header is removed and the packet is sent back out into the Internet to be rerouted to its proper destination.

In forming the tunneling header, all of the values from the old header are copied into the new one with the following exceptions. Of course the source and destination address change. Also the total packet length grows by one header length. Additionally, the header checksum needs to be recalculated since some of the fields have changed.

Now, each packet experiences some additional mileage and this scheme clearly will add some additional delay to each transaction over the network. However, it satisfies the user requirement of making the system operable with any commercial SLIP service. The added delay can be minimized if the Hybrid Gateway is "well connected" to the Internet. The main arteries of the Internet operate at very high rates and if the Hybrid Gateway has a high rate connection to a main artery of the Internet, the added delay can be minimized. In the prototype a 1.5Mb/s T1 link was acquired to SURAnet, the southeastern United States Internet provider to minimize this additional delay.

ARP handling

As stated above, the Special SLIP NDIS driver declares to the TCP/IP package that it is an ethernet card. The TCP/IP package handles ethernet routing and when it is trying to send data to a new IP address, it tries to resolve that address to a hardware or MAC address using an ARP as described in the section on Internet Routing. The SLIP connection only carries IP level communication and MAC addresses have no meaning at the IP level. So, in order to satisfy the TCP/IP package's request for MAC address, each packet sent by TCP/IP needs to be checked to see if it's an ARP trying to resolve a new IP address. If TCP/IP does send an ARP, then the driver creates an ARP reply handing TCP/IP a bogus MAC address for the ethernet header of the packet to be transmitted. The contents of the MAC address are irrelevant since the ethernet header is stripped off to send the packet over the SLIP link.

ARPs don't occur that frequently because most connections will be made outside of the subnet of satellite IP addresses. In this case the packet is automatically sent to the local router. So, at least one ARP is required but not many more than that.

Segmentation

Since the TCP/IP is configured to talk to ethernet and we want to be able to receive the largest sized packets we can, TCP/IP is configured such that the Maximum Transmission Unit (MTU) of the network is as large as possible, 1500 bytes for ethernet. This specifies the largest packet size the network can handle. Our experience has shown that SLIP servers can have a much smaller MTU such as 512 bytes or even as small as 256 bytes per packet. Usually, the application is generating small packets to send over the SLIP link, like 60 byte acknowledgments. However, the tunneling header adds about 40 bytes to each packet and occasionally the application will generate some large packets to send. To handle this situation, the driver must implement it's own segmentation procedure. In segmentation, the packet is broken into pieces the size of the SLIP MTU and the header minus one header length is copied onto each piece with an offset value specifying where that particular piece goes in the original packet. Once a tunneled packet is segmented, it is reassembled when it reaches the Hybrid Gateway. Only the tunneling header is copied onto the head of the segments.

The Hybrid Gateway

The Hybrid Gateway is allocated all the special network routing functions that must occur outside of the User Terminal. Untunneling is one of those functions but not the only one. Nonstandard packet formats are used to tunnel IP packets from the User Terminal as well as to send packets over the satellite link.

Because the Hybrid Gateway is a bottleneck through which traffic from all hybrid terminals must flow, the functions

are kept as simple as possible to maximize throughput. Each function is implemented so that the processing requirements are minimized. In the prototype situation, only a single User Terminal is active at a time. However, in an operational situation, the demand on the Hybrid Gateway could be such that a single PC could not keep up and a more powerful platform or collection of machines should be considered.

Untunneling

Every IP packet from every User Terminal is tunneled and sent to the Hybrid Gateway for untunneling. The Hybrid Gateway should have good Internet connectivity to minimize the accumulated delay from having to route every packet via this machine. When a tunneled packet is received by the Hybrid Gateway, the length of the header is read from the IP header and those bytes are simply stripped off and the packet is sent back out into the Internet.

Segmentation and Reassembly

It's possible that a tunneled packet is segmented on its way from the User Terminal to the Hybrid Gateway. This can occur if segmentation occurs within the driver as described above or if the packet traverses a network with an even smaller MTU than the MTU of the SLIP connection. Since only the tunneled header is copied onto the head of each segment, the segments must be reassembled within the Hybrid Gateway before the packet can be untunneled and sent back out into the Internet.

Reassembly involves allocating several buffers for partially received packets and filling in the segments as they arrive. A time to live value is assigned to each packet and if all the segments don't arrive before the time to live timer expires, the packet is discarded.

ARP responding

The machine that forwards packets over the satellite is on the same network as the Hybrid Gateway. This network's router will receive packets with the user terminal's satellite IP address and will send an ARP to find out what MAC address to send them to. The Hybrid Gateway needs to encapsulate these packets and so it must respond to ARPs for any User Terminal satellite IP address so as to receive them.

This is implemented by specifying a range of IP addresses that will be assigned to User Terminals and having the Hybrid Gateway respond to ARPs for it's own IP address as well as any IP address in the specified range. Once the router gets an appropriate ARP reply from theHybrid Gateway for a certain IP address, it will send all packets with that destination IP address to the Hybrid Gateway.

Satellite Packetizing

The Satellite Gateway expects IP packets to be encapsulated first in a special satellite packet and then within an LLC packet. The special satellite header identifies the downlink and contains a sequence number and the packet length. The LLC header is used to send the packet to the Satellite Gateway which is on a token ring network. The Hybrid Gateway must prepare packets for the Satellite Gateway by appending the correctly configured headers to the front of the packet. The receiver in the User Terminal does not get the LLC header and identifies packets destined for it by the least significant byte in the satellite IP address. Therefore, the six byte satellite destination address is determined by reversing the order of the bytes of the satellite IP address for the user terminal and then padding the rest of the address with zeros. The sequence number is just a counter and the length is calculated from the packet header.

The SLIP Server

A SLIP Server was configured for testing purposes but in real operation the SLIP Server will be a commercial service. Its functions are to receive SLIP encoded IP packets from a modem connection with the User Terminal, uncode them, and forward them to the Hybrid Gateway via the Internet.

For development, we configured a PC with the Frontier TCP/IP package using two interfaces: a serial port connected

to a modem and an ethernet card. The Frontier package acted as a router receiving packets from the SLIP connection and forwarding them to a router available on the ethernet connection with Internet access.

The Satellite Gateway

The Satellite Gateway has been previously developed by HNS. It consists of a PC with token ring connection and a DirecPC card configured for transmission. When the Satellite Gateway receives a token ring frame at its MAC address, it checks for the correct service access point identifier and sequence number and then sends it over the satellite.

The Application Server

The Application Server represents any server running an Internet application available on the Internet using the TCP/IP protocol suite. If the User Terminal is running a Mosaic client, the Application Server may need to support a variety of connections. The key element to this subsystem is that it should include *any* host on the Internet reachable by normal Internet communication.

CONCLUSION

In this project we have taken several existing pieces of technology, the commercial TCP/IP package, the DirecPC card, and the Internet and combined them (with a little software glue) to provide a level of data communication service previously unavailable to homes. This is an excellent example of the potential of satellites in communications as well as the field of systems integration.

Hughes Network Systems has plans to integrate Hybrid Internet Access into a group of new services/products that it will be marketing in the next year. Hybrid Internet Access will be combined with a software "package" delivery service for point-to-point distribution of off-line video or commercial software and a broadcast distribution service for sending out live, real-time video, audio or a news "ticker" service as a suite of products based on the DirecPC card technology. A PC with the DirecPC card and a modem could purchase any of these products. This fall, HNS plans a 100 site nationwide field trial.

This system is capable in its current state of producing over 200kbit/s per terminal. With the modifications suggested, we hope that the 1Mb/s barrier will be broken and true ethernet bandwidths will be available to the home user. This is sure to change the nature of computing for the general population. A growing number of Americans have computers in their homes and the utility and importance of the National Information Infrastructure will have a direct correlation to the average Americans ability to satisfactorily navigate the "net."

On a personal note, this project has meant a lot to me. Over the past two years my wife would accompany me to school and use the computers there to run Mosaic and visit art archives, opera databases, and other Internet treasures around the world. I wanted to give her (and others with her curiosity and bravery to enter the currently nerd oriented net) the ability to access this tremendous resource from our home at her leisure. I feel that there are many others like her and want to empower their curious impulses with this simple design.

SOURCES CONSULTED

3Com/Microsoft Corporation. 3Com/Microsoft LAN Manager Network Driver Interface Specification. Version 2.0.1. 1990.

Adams, Phillip M. and Clovis L. Tondo. Writing DOS Device Drivers in C. Englewood Cliffs, NJ.: Prentice-Hall, 1990.

Blanchard, Benjamin S. and Wolter J. Fabrycky. Systems Engineering and Analysis. 2d ed., Englewood Cliffs, NJ.: Prentice Hall, 1990.

Borland International, Inc. Borland C++ 3.1 Programmer's Guide. Scotts Valley, Cal.: Borland International, Inc., 1992.

_____, Turbo Debugger 3.0 Users Guide., Scotts Valley, Cal.: Borland International, Inc., 1991.

Comer, Douglas E. Internetworking with TCP/IP, Vol. 1: Principals, Protocols, and Architecture. 2d ed. Englwood Cliffs, NJ.: Prentice-Hall. 1991.
Dettman, Terry. DOS Programmer's Reference. 3d ed. Carmel, Ind.: Que Corp., 1992
Duncan, Ray, ed. The MS-DOS Encyclopedia. Redmond, Wash.: Microsoft Press, 1988.
DalBello, Richard. "The Role of Satellites in the National Information Infrastruture Initiative". Via Satellite. February, 1994. pp 48-56.
Eggebrecht, Lewis C. Interfacing to the IBM Personal Computer. 2d ed., Carmel, Ind.: SAMS, 1992.
Goodwin, Mark. Serial Communications Programming in C and C++. New York: MIS: Press, 1992.
Habayeb, A. R. System Effectiveness. Oxford: Pergamon Press, 1987.
Hunt, Craig. TCP/IP Network Administration. Sebastopol, Cal.: O'Reilly & Associates, 1992.
Intel Corporation. Intel486 Microprocessor Family Programmer's Reference Manual. Mt. Prospect, IL.: Intel Corp. 1992.
Kelly, Al and Ira Pohl. A Book on C: Programming in C. 2d ed. Redwood City, Cal.: Benjamin/Cummings. 1990.
Malamud, Carl. Stacks: Interoperability in Today's Computer Networks. Englewood Cliffs, NJ.: Simon and Schuster Co. 1992.
Microsoft Coporation. Microsoft MS-DOS Programmer's Reference (covers through version 6). 2d ed., Redmond, Wash.: Microsoft Press., 1993
Phoenix Technologies Ltd. System BIOS for IBM PC/XT /AT Computers and Compatibles. Phoenix Technical Reference Series. Reading, Mass.: Addison-Wesley, 1989.
Romkey, J. A Nonstandard for Transmission of IP Datagrams over Serial Lines: SLIP. RFC 1055, June 1988, 6 pp.

FAULT TOLERANCE SOFTWARE FOR HIGH PERFORMANCE COMPUTERS

Alice Copeland Brown, Principal Engineer
Raytheon/Equipment Division
528 Boston Post Road, Sudbury, MA 01776

ABSTRACT

Despite the increased speed of high performance computers, their instability creates risk. Sources of this problem are many and varied: e.g., a message routing chip failure can block and make ineffective all communications that pass through that node to others. Such a failure would not present an insurmountable problem if alternate routing patterns were available. This paper discusses this and other specific faults with techniques and designs to increase availability and reliability among High Performance Computers. Current fault tolerance techniques for Massively Parallel Processors (MPPs) rely primarily upon redundancy. We go beyond this, implementing novel methods: using the ARPA-funded Parallel Virtual Machine (PVM), under funding from the Rome Labs Cooperative Research And Development Agreement for Parallel Processing, we are creating fault tolerance software as an overlay to the OS. Techniques which will be employed as deamons or as part of the PVM library extensions running on each node are: Checkpointing and rollback, Triple Redundancy Synchronization, Timing Redundancy, Reset and Restart, Health Management with both Nodal and complementary Global Services, Memory Scrubbing, Redundancy of Function with Diversity of Implementation in Design, and Alternate Routing Algorithms. Fault tolerant software for high performance computers has relied primarily on redundancy of function, processor, and communications media. This paper discusses new techniques that marry redundancy with other schemes on MPPs.

****

Software development for Massively Parallel Processors (MPPs) is in its infancy. This paper discusses specific problems in reliable high performance computing and emphasizes the importance of fault identification, isolation, and recovery. Several types of MPPs exist, as described below:.

- Multiple Instruction stream, Multiple Data stream (MIMD) machines function asynchronously and independently (e.g., Cray Y-MP). Uses are CAD/CAM, simulation, modeling, and communication switches. They use either shared or distributed-memory. In a distributed-memory machine, each processor has its own individual memory store. For data to be shared, it must be passed from one processor to another by messages. Because these processors do not share memory, contention among processors is not a problem and can be readily used to support scaling up to a large number of processors.

- The Single Instruction stream, Multiple Data stream machine (SIMD), a simpler, but less flexible MPP, is one in which the same instructions are executed at the same time on different data (e.g., CM-2 with 64K processors). Applications using this architecture are image processing, partial differential equations, matrix operations, FFTs, and database searches, all of which feature fine-grained parallelism.

The programming challenges presented by MPPs are formidable, but the rewards of success are even more compelling. On one hand, we have already seen quite a few parallel processor manufacturers fail, some because their architecture was too brittle or inflexible. Each of these systems was defined to meet a specific need. As the technology matures, however, manufacturers will identify architectures that are flexible enough to meet various needs with a simple re-configuration of the intercommunication network. On the other hand, the problems solvable by MPPs are complex and we are identifying further problems that could not be solved in our lifetime on sequential computers. For example, as we map the ocean currents, we may need to simultaneously process megabytes of data on thousands of processors in the shortest possible time, while suffering only marginal degradation in performance should any of the processors expire/overheat/experience a transient problem. It is desirable for the system to detect, diagnose, and heal processing problems in real-time, with no single points of failure.

Fault tolerant high performance computers are needed wherever man finds life imperiled or untenable:

- In satellites - We at Raytheon have a contract with the government of Brazil to monitor the rain forest, noting deforestation patterns, traffic patterns, human movement, and weather patterns to provide a means for Brazil to more effectively monitor ecological and drug production activities, among other things.

- In underwater unmanned vehicles (e.g., Remotely Piloted Vehicles [RPVs]) - Such conveyances provide invaluable information in marine science, ecology, offshore oil exploration, and naval operations.

- In space probes - Inaccessible, high value machines require autonomous fault detection, isolation, and recovery. These research vehicles are able to capture and analyze more information concerning our universe for longer periods of time, Given the added advantages of high performance computers whose Mean Time Between Failures (MTBF) is made irrelevant by the switchover capabilities of the system, not even the sky is the limit to what we can learn about our environment.

- On the battlefield - Data fusion and decision support in this environment require Electrical Intelligence (ELINT) and Signal Intelligence (SIGINT), as well as Human Intelligence (HUMINT). A ruggedized fault tolerant MPP satisfies the need to provide correct, secure information under the stress and bombardment of battle by automatically processing and analyzing the data and communicating the pertinent information to the command and control center via a STU-III or other encryption device.

Inner space is also being explored. As computer components become progressively smaller, the opportunities to embed fault tolerant multi-processors inside surgical and diagnostic instruments are apparent. For example, embedding smart fault tolerant sensors within the human brain would allow us to discover more about the bio-electronic impulses that are part of the thinking process. Likewise, micro-surgery is a burgeoning field of its own, and creating computerized tools that will not fail or give false information because of a few lost bits will enable the surgeon and patient to trust instruments that were previously unreliable.

We are identifying the problems inherent in MPPs. In the late '70s we had a few parallel processors, a few vector supercomputers like the Cray, and, in our naive assumptions about their programmability and reliability, thought that if we could say "parallel processor," we could program them. The resulting code was extremely expensive and unreliable. While it would appear that the multiplicity of processing resources would make fault tolerance easier to achieve, the truth is that the complexity of interrelationships makes the failure rate even higher unless sound fault tolerance principles are employed.

In the past, most of the MPP's processors were left idle for most of the time because the weakest link in the process was the slowest process, constraining the others. Few analysis tools were available, and the parallel processors crawled off into academia to be attended by long-suffering, patient, and tenured professors and their slave graduate students. Esoteric papers were written which we engineers read, politely applauded, and then ignored, as we turned back to the "real world" of VonNeumann machines, determined to never again be disillusioned by the mirage of fault-free, speedy processing from scores of processors grinding away in perfect synchronization.

Fault tolerance comes in many different flavors, all of which must be in place to provide full fault coverage. Fault coverage refers to all of these activities, taken together:

- Fault detection
- Fault isolation
- Fault correction
- Fault recovery.

These mechanisms do not operate singly, but are interdependent. Although they may be explained generically, they interoperate in a synchronized manner, orchestrated by the Fault Manager.

Fault detection encompasses diagnostics, collection of error statistics, use of watchdog timers and heartbeats, N-version synchronization and voting, and reasonableness checks. [1] One example of fault detection is the Self-Test task, a low-priority task which runs in a background mode. Its role is to detect latent faults and, after detection, estimate the fault's location and severity. Example entities tested are memory, CPU logic, bus interface logic, and the network interface.

---

[1] Sampath Rangarajan and Donald Fussell, "Diagnosing Arbitrarily Connected Parallel Computers with High Probability," IEEE Transactions on Computers, May, 1992.

Fault isolation bounds the destructive effect of a fault. While the error must propagate to be detected, its influence must also be limited to prevent damage to the rest of the system. Proper isolation techniques require synchronization. Because message traffic and the occurrence of faults are both non-deterministic, many techniques already in place for MPP synchronization will also supply the synchronization needed for fault tolerance:

- Establishment of thresholds (i.e., number of faults which constitute a hard banishable fault)
- Establishment and restoration of recovery points
- Establishment of cross-checking points
- Redundancy - Shadowing or hot standby.

Fault correction may involve the use of parity bits (e.g., CRC) enabling the Fault Manager to restore the word to its original meaning. It may involve restarting a processor from an alternate memory source. It may be completely unnecessary if sufficient processors or functionally identical codes are present to carry on in the absence of the injured element. Each of these techniques in itself is insufficient, however. But given the problem, a combination of these techniques becomes effective in establishing a realtime strategy for dealing with faults.

A few of the most common MPP fault categories are:

- Maskable faults (Bad memory bits, latent errors caused by a cosmic ray discharging a memory cell, which produce incorrect answers during a system "READ" operation)

- Related faults resulting from a fault common to all variants or from dependencies in separate designs and implementation

- Cumulative faults which are latent occurring singly, but which become troublesome upon repetitiously reaching a certain threshold

- Compound faults caused by interrelationships of several faults occurring at once.

Functions developed to offset such problems are listed below and are used singly or interdependently:

- Bit error detection
- Memory scrubbing
- Redundancy in function or physical component
- N-version programming (i.e., design diversity)
- Health diagnostics
- Watchdog timers
- Statistical error counters and thresholds with related actions
- Checkpoint and rollback
- Reset
- Repetition – resending messages, re-reading disc blocks
- Fail-fast mode: module or processor immediately marked off-line
- Self-checking: range or data typing checks
- Parallel execution of two self-checking components
- Interdependent operating systems
- Firewalls – to prevent memory leaks, overwrites
- Redundant power supplies
- Fault status sampling

Problem: Debugging tools for fault tolerant HPC software are primitive. It is difficult to replay or view the exact conditions that were present when a situation that leads to failover occurs. Too many unknowns exist within a system containing multiple CPUs, communication handlers, and memories. Methods are currently being developed to show I/O usage and communication among processors, enabling us to cluster those processes or combine several processes into one task to reduce communication overhead. Critical I/O capabilities are guaranteed by common pool sparing. The I/O Control Processor can be a member of the common pool of CPU elements in the system. As Figure 1 below shows, the high performance computers of tomorrow will be even more complex than those of today.

Our visualization tools are improving; these are tools that can identify a range of values in 16,000 processors at one look, and windows that display various conditions (e.g., source code, register contents). Holograms have been designed that display data structures in 3-D and are especially helpful in Object Oriented Technology to demonstrate the relationships among objects, classes, and subclasses. Our debuggers will soon be able to reliably perform the playback necessary, providing checkpoints to which the operation can be rolled back. Our performance analysis

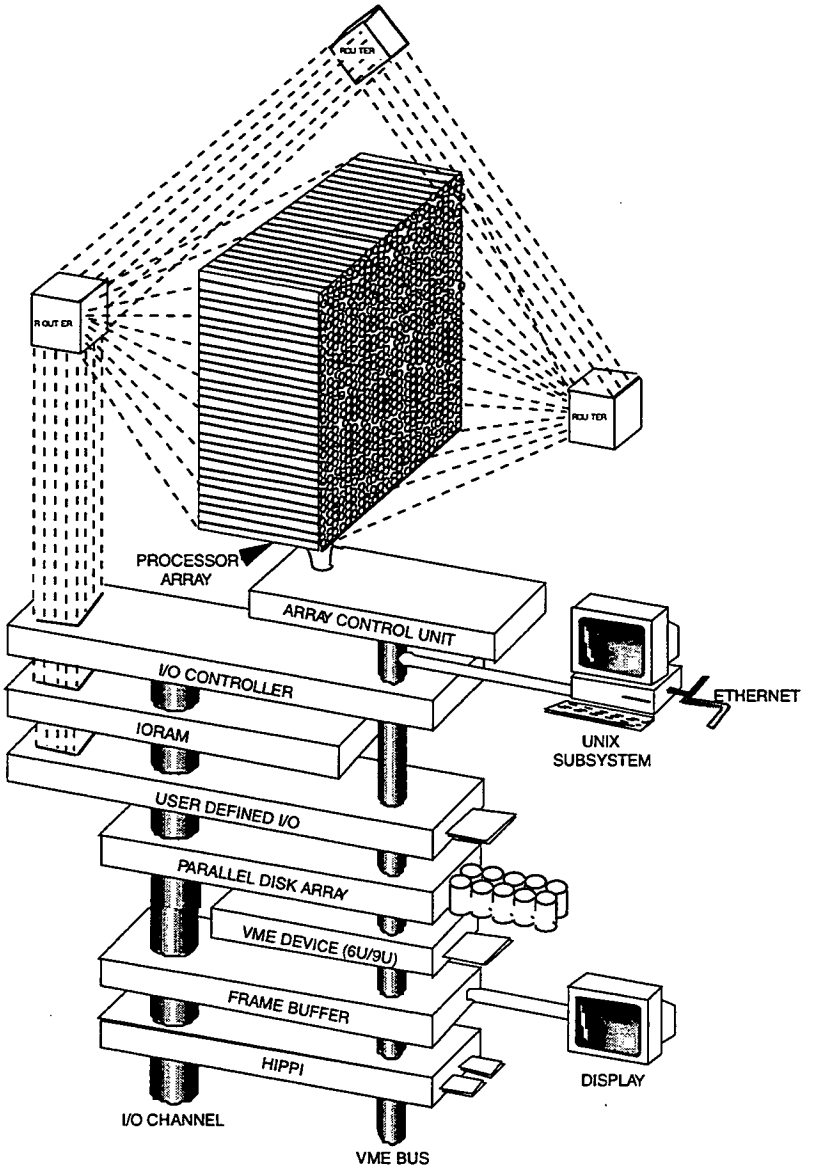


Figure 1. Sample MPP Architecture

tools will show us where both the latent stress and bottlenecks are. These primitive tools are being developed into realtime tools which can detect and solve logjam problems at runtime. Called "dynamic load management", this approach involves marking off the processors that are down, configuring an updated scheme for offloading processor loads, monitoring the data queues, and duplicating any data necessary to decrease communication time. The shelf-life of these primitive tools is short. By the time one tool has been honed for a specific architecture, the architecture is made obsolete by a better one.

As the shakeout continues, however, architectures will become more flexible and cheaper, thus decreasing or eliminating the need for a different architecture for every application. As architectures become more flexible and more easily reconfigurable, debugging tools will be built that allow fault tolerance functions to be independently debugged.

Problem: Tools to meet a variety of objectives are needed. There is a need for compiler developers to embed a strategy for creating hooks at compile time to be filled at run time with assessment of inactive processors and assignment of various processes to appropriate processors.

A tool called Optimized Mapping Alternate Routing (OMAR) developed by Dr. H.J. Siegel, et al, at Purdue University permits the user to run raw code. Using a library of various routing algorithms, OMAR tracks the routes and re-maps the code for most efficient usage. By using such tools as OMAR and their Partitionable SIMD/MIMD (PASM) machine for rapid prototyping, expensive mistakes can be avoided. Such tools will become more sophisticated, pointing out hidden areas of parallelism within sequential code, indicating and correcting dependencies that prevent automatic parallelization, and automatically duplicating data to run on clusters of processors (avoiding expensive message-passing). The alternate routing algorithms provide means for re-routing communications should a communications processor or link break. In the future, these routing algorithms will be called upon dynamically to circumvent failed processors until they are healed and returned to service.

Problem: Processors die unmourned and unknown, creating indeterminate results. The expiration might be blessedly sudden or happen bit by bit, leaving the programmer or user puzzled and confused. Multiple failures create even more problems in high performance computers. Because of this phenomenon, researchers are studying the habits of bees and geese[2]. Rather than being directed hierarchically, it appears that each is internally directed, able to form a part of a flying wedge, or build a hive, and (more important to our purpose) reconfigure if another of the group cannot deliver. Strategies that look at explicit groups of processors and develop shared responsibility for responding to requests are being developed to mitigate the risk of multiple failures. Failure of the primary processor stimulates actions by other group members. All members have updated lists of the group membership (i.e., omitting the failed processors) in order to partition the tasks appropriately.

This works well in conjunction with fault recovery processes that bring back on line processors that have experienced transients. Transients, especially as a result of alpha particle bombardment, constitute one of the chief reasons for satellite processor failures. Once back on line, message delivery announcing, "I'm up and ready to join the group, carrying my responsibility" must be assured as well as reassignment and continuation of the process. A system which can differentiate hard and transient faults will not permanently discard hardware modules on the basis of a single transitory fault and will have higher availability and reliability than a system which is unable to make such a distinction. It is important, however, to be able to differentiate between transient and intermittent faults. A processor failing intermittently should not be restored, as results will be indeterminate. Fault status sampling is not performed cyclically, as too much time would be wasted on continual efforts to restore subsystems with hard or intermittent faults. Instead, the time delay between successive samplings is a function of indications that the fault is still present. For every sampling that indicates a fault, the time is doubled before that processor is sampled again. An upper threshold is set, which indicates how many retries will be permitted. Once that threshold is reached, the fault is marked "hard" and no further retries are indicated.

Upon detection of a permanent failure in, for example, one of the active memory devices, spare memory devices are also included and reconfigured into the memory array. High performance computers have the advantage here because of the cheap processors available for cold standby until reconfigured into active service upon failure of primary Memory Units.

Problem: Synchronization becomes an important consideration when using explicit groups of processors. The order in which actions are taken must be synchronized [3] because group members are acting independently as a result of dynamically changing but shared information. Via these schemes, fault tolerant services can be implemented by a group of programs that adapt transparently to failures and recoveries.

By operating groups of processors as redundant clusters, high reliability is afforded. Inter-processor connectivity can thus be shared among many PEs. [4] This shared connectivity, in conjunction with functional synchronization of processes executing on redundant processors, makes the system easily reconfigurable. When a processor fails in one

---

[2] Mitchel Resnick, "Turtles, Termites, and Traffic Jams: Explorations in Massively Parallel Microworlds," MIT Press, Cambridge, MA, May 1994.

[3] Kenneth P. Birman, "The Process Group Approach to Reliable Distributed Computing," Communications of the ACM, Dec 1993.

[4] Richard E. Harper and Jaynarayan H. Lala, "Fault Tolerant Parallel Processor," Draper Lab, AAIA Journal of Guidance, June 1991.

group, a processor from another group can be used to replace the failed one. Function criticality is matched with the appropriate redundancy level, making most efficient use of processing resources. I/O device interfaces will possess a degree of redundancy determined by the criticality of the sensors or shooters with which they interface. For example, it would be vital that navigation on a space probe be considerably more accurate than mission management. These redundant groups of processors communicate internally within the group and externally with other processor groups. The design, called "Byzantine resilience," must handle concurrent faults as well and could include the intermittent case in which a processor starts and restarts execution at some future time, sending conflicting information to different destinations and generating other problems. This is the reason for marking a processor offline, and its function's last result as lost. When the processor is brought back online, the process is begun anew at the proper place in time.[5]

In the future, we will have sophisticated health monitors which will use various polling strategies. One strategy has each processor performing heart-beat checks on its neighbor. If the software stethoscope does not detect a pulse, the processor goes into degraded mode, performing a routine that may involve:

- Reaching a checkpointable stage in its own process
- Marking the processor disabled
- Reaching back to the last checkpoint in the dead processor's past and continuing with its processing
- Determining which resources were locked by the dead processor's task and unlocking them
- Attempting to resuscitate by issuing a "Restart"
- Performing diagnostics on the processors
- Commanding the Configuration Fault Manager to continue to heal the processor and, if done, mark the processor back on line
- Return processing to the revived processor
- Continue to process a smaller volume of data, if necessary.

A necessary precondition for designing robust software for MPPs is: know the architecture of your platform. This is key to determining the design of the Fault Manager. During the pathfinding phase, the problem should be investigated deeply enough to match the problem to the architecture. How are the processors interconnected? What is the granularity of parallelism:

- Fine-grained (instruction level)?
- Medium-grained (cooperating concurrent task level with process synchronization every few hundred machine instructions)?
- Coarse-grained (synchronization occurring every few thousand machine instructions, practically independent)?

The type of granularity involved helps dictate the fault tolerance strategy.... where the algorithms will be implemented and at what intervals. Synchronization design places a strain on the designer because it is a new challenge and because it involves so many processors and constraints. Has one of the voting processors failed since the last synchronization? Has the membership of the cluster changed? If so, this information must have been received by the synchronization messenger. Effective fault tolerance synchronization implies that the maximum difference between fastest and slowest execution rates of the redundant processors must be known. It then becomes a simple process of message transmission and reception.

---

[5] ibid.

Problem: MPP hardware is fragile and expensive. We will also have cheaper hardware, which means we can design our systems with redundant bus interfaces to provide for switchover (buses usually being a single point of failure). The SCSI or HIPPI interface is made fault tolerant through dual redundancy. To provide optimum fault tolerance, a combination of two different communication strategies should be used. For example, two Ethernets might fail for the same reason, while an Ethernet and a Token Ring could provide backup for one another because they use different resources and schemes. Fault detection of bus activity is provided by monitoring the bus and/or software protocols used by network managers to continually provide status not just of failure but of failing health.

Will the processors be I/O bound? If pipelining is used, the pipelines must be kept full of instructions and data. Toward that end, high speed buffers can be placed between the main memory and the functional units. The four-part buffers feature a perpetual read/write activity as the data alternately fills a buffer and is then read out. As with memory, buffers must also be continually scrubbed to determine if and where bad bits might exist or if memory leaks have occurred.

Disk farms capable of 80 Gigabit transfer speeds have been designed that enable the "ribbonning" of data arrays into the processors. As our input speeds approach the limit of the speed of light, buffering will become more and more vital to the throughput needed.

We can design the system with enough processors in reserve to provide full service for all loads, or all loads except peak loads. Under dire environmental conditions or for systems seldom used, it is advisable to cycle processors so that all spares are exercised periodically. Using this design, a non-coverable error in any spare will cause immediate failure as soon as it is used. In conjunction with this strategy, health monitoring should include spares as well as active processors. In our SDI studies, we discovered that the systems we were designing to provide an umbrella of defense would seldom be needed except at peak capability, which required contingency plans and degraded mode designs. For similar systems, cycling of spares is worth the extra effort and time required to perform the periodical reconfigurations.[6] Mode switches, however, must be carefully analyzed so that such cycling could not occur during Attack mode, for example.

Problem: Deterioration Caused by Inactivity. In peacetime, a missile system under attack may detect and identify a marginal number of objects at any one time to demonstrate the point, including a number of false alarms. However, the radar system in Attack mode must actually handle over 1M objects, taking into consideration that some incoming missiles may explode into a large number of pieces, each of which must be tracked in order to ensure that a kill has taken place. In a fault tolerant system, this leaves thousands of processors dormant for most of their lives. It thus behooves the designer of a satellite-based system to implement diagnostics that periodically test the health and welfare of these processors. Disuse itself may cause certain elements to deteriorate. An alternative is to periodically shift the peacetime processing from processor to processor, ensuring that all communication paths and processors are used over a small period of time.

In designing a fault tolerant system for high performance computers, the goal is to invest in the minimum hardware necessary. Towards that end, the fault management subsystem must include interprocessor communication, fault detection, error logging, application management, resource management, reconfiguration, and restart functions. Systems without the benefit of human maintenance must have the 99% sparing capability, which we are used to keeping in a back room, onboard and automatically available. Without this capability, we will not easily be able to unlock the mysteries of the ocean floor, outer space, polar regions, nor provide the decision support necessary on tomorrow's battlefield.

Another fault tolerant scheme is the voting paradigm, where two processors performing identical operations are monitored by a comparator. In some versions, the third processor is also a task processor but can be fully dedicated to the arbitration role. The comparator either performs task processing or the relative test for one pair of processors. Algorithms exist to identify all faulty components (processors and comparators). [7]

---

[6] Joe Stella, "Complexity vs. Reliability in Fault Tolerant Computing Systems," Internal Raytheon Memo #EMS88-0070, Nov. 1988.

[7] — "Reliable Distributed Sorting Through the Application-Oriented Fault Tolerance Paradigm," IEEE Transactions on Parallel and Distributed Systems, July 1992.

Problem: Our HPC compilers today are unreliable, allowing runtime errors because of lax data typing, inability to perform load or data management, and the creation of stacks and queues that are easily "blown" with no fallback flexibility.

Tomorrow we will be able to suggest to the FORTRAN 90D compiler precise processors to use to eliminate time wasted in transporting data from one processor to another. By providing a memory pool and warning levels, we can first be warned of stacks and queues approaching a critical level and, through pointers, expand the entity at risk. This will be done at runtime rather than at compile time, a limited capability which is currently available.

Fault tolerance in HPCs is based upon a locally robust system with successively larger regions of local redundancy. In principle, a single healthy node is capable of resuscitating an entire partition or machine that is otherwise sick. Isotropy must also be designed in, allowing the generalization of conceptual solutions across diverse architectures and across failure modes.

Problem: Communications among processors form perhaps the most fragile link in the chain. Redundancy of processors can help only if communications reporting the outage and substitution or failover are timely and reach all processors involved. Moreover, if a communication link fails, other processors will be uncertain regarding another's status, and others may be totally unaware of the problem. This can result in chaos or, at least, inconsistency. One particular MPP crashes when attempts are made to communicate either South or East of the downed communications chip. No routing can go in either of these directions. When channels break, it is unwise to assume either end point failed. Data may be lost in the process. This problem would be even more crucial in a distributed system, where various workstations' clocks are not closely synchronized. Operating systems introduce unpredictable software delays, processor execution speeds vary widely, and scheduling is often unpredictable. Comm problems may derive from messages which are:

- Lost in transit
- Arriving out of order
- Duplicated
- Discarded due to inadequate buffering capacity.

One fault tolerance scheme is to consider a process, once gone, as absent from the system and any state it may have recorded to be irrelevant. Once the process is recovered, it is treated as a completely new process. It is easier to treat faults at a higher level. To use a hardware analogy, component level is harder to maintain than board level. Another approach is to embed a context record within each message, which fault detection will use to determine order of arrival and will delay delivery until prior messages arrive. Priority of messages and priority queuing can be used to handle notification of changes in group membership, which would preempt normal messages. To insure that a multi-cast is received by all processors affected, an AND synchronization barrier should be used . Once all messages have been delivered, all bits are set and the Boolean variable is set to TRUE. Again, "processor down" messages have precedence and would create a set bit.

With an ever-increasing number of processors to continually diagnose in a realtime environment, the challenge is to create a fault tolerance strategy the allows any faulty component to misbehave in any manner, obviating the need to prevent such behavior. If conflicting information is given, the system is able to nonetheless function by performing fault containment (firewalls). [8] This also includes limiting the connectivity and amount of synchronization and implementing interchannel communication protocols. The protocol consists of ordering the processing of any message which changes system state data by sending a copy to the backup process, sending a message acknowledgment with copy to backup process, checkpointing the data to disk, and distributing the data to other processes. The rationale behind this is that the acknowledgment is sent only after data is changed so that a positive response is not sent when data can't be stored. The acknowledgment is also sent only after the data is backed up so that a single failure won't cause loss of data. Checkpointing is done only after data is sent to the backup so that if the primary fails, the backup becomes primary, and the failed process is restarted and restored, then the revived process will not be ahead of the new primary. Distribution to users occurs after backup and checkpointing to minimize the possibility of a user having data not held by the master. By buffering all the outputs and processing them in the correct order, they can be discarded in case of error. Redundancy of software and hardware in groups

[8] Harper and Lala.

creates a level of trust because detection and isolation of faults, the resultant masking, and the reconfiguration of resources are done by the fault detection manager and are transparent to the applications programmer.

If possible, an architecture should be used that features local memory that can also be shared under certain conditions. The downside to this approach is that the firewalls are removed, making the processor vulnerable to corruption. This shared local memory feature makes the system more robust, however. A good fault tolerant design also ensures that "garbage collection" is performed periodically. As software ages, the damage caused by lack of diligence to this important task can break the system. Memory leakages may occur. If the BIT diagnostics discover bad memory bits, these must be masked off. However, retests should periodically be performed as indications are that the most common errors are a result of transients which are self-correcting.

A technique known as Rate Monotonic Analysis (RMA), developed by the Software Engineering Institute[9], allows the engineer to predict with a high degree of confidence the performance of his code. Using a "worst case" approach, the engineer can design his MPP system to mitigate the greatest volume of asynchronous interrupts his system may accept. At the least, RMA can help you predict where the system will "break" and take appropriate steps to mitigate the consequences. One can guarantee that the set of tasks assigned to the system is schedulable (i.e., will meet its timing constraints or deadlines) by meeting a set of much more manageable constraints involving cycle time, processor utilization, resource conflicts, etc. This replaces the currently pervasive method of drawing a timeline by hand and scheduling each task's use of resources individually. From a fault tolerance aspect, this means that fewer bottlenecks, which lead to memory leaks and stack blowouts, will occur, lessening the stress on system resources.

As RMA becomes more sophisticated, future implementations will allow a more realistic application where asynchronous interrupts are concerned. To "schedule" an asynchronous interrupt for the smallest interval possible for this specific interrupt is impractical. Future applications will allow for the usual case, providing a shift into the background for other functions when the asynchronous high-priority interrupts reach saturation point.

Adaptive routing is another essential part of a robust architecture for a fault tolerant system. By "adaptive routing" we mean the ability to bypass "busy circuits" and use those routes that, while not shorter, provide a faster arrival time for messages. This implies that the network manager will at all times have knowledge of the length of each message (or packet), the number of packets, and the "ack/nak status" (number of packets that have gotten through in good shape to their destination as compared with the number of error conditions reported). A time-out scheme must be included in any fault tolerant communication design to allow the message to be re-sent under the following conditions:

- Specified elapse of time since an acknowledgment
- Specified number of non-acknowledgments (naks).

The latter condition can create deadlock as the number of error messages accumulates, blocking the delivery of legitimate messages. After a pre-determined number of "naks" (depending upon the message priority), all packets must be discarded and the message re-sent.

Message-passing communications structures designed with each processor having sole responsibility for its own memory (i.e., loosely coupled processors) prevent processors from unknowingly corrupting another's memory.

Cyclic redundancy checks are also used, with the hardware providing an automatic retry. If a second failure occurs, the message is automatically rerouted and the processor is set "off." Messages are sent to all clustered processors to report the loss of the processor and are sent again once the processor is "resuscitated."

Any HPC which handles extensive computational loads requires fault tolerance capabilities to prevent the loss of valuable work and time (e.g., stock market analysis). The checkpointing capability provides this security: files are kept in non-volatile memory on the node to protect data during power outages. The system state information, known as "application checkpoint files," are maintained through re-initialization. Power outages may be provided for by supplying alternate sources of energy to spare power supplies: solar, as well as chemical. Again, we are planning for no single point of failure.

---

[9] Klein, Ralyia, Pollak, Obenza and Harbour, A Practitioner's Handbook For Real-Time Analysis, Kluwer Academic Publishers, 1993.

Problem: Failures caused by design faults outnumber hardware faults by 10:1. By applying independent failure modes, software and design repair, fail-fast modules, and concepts of modularity, the user can tolerate these faults. Software and design repair can be done in the background or, in some cases, repaired and rebuilt elsewhere and installed remotely. Fail-fast modules are immediately marked off-line upon identification. By providing firewalls, damage can be limited to the immediate module or cluster of modules

Modularity

Designing firewalls into the system is highly important because faults must be contained. Without independent modules and sanity checks across interfaces, faults can become infectious, corrupting other processes. When modules are grouped together, failure in one forces all interdependent modules to be switched out.

Self-checking

Range and type checks of inputs, outputs, and data structures – sometimes called "defensive programming" – helps satisfy the integrity assertion. Logical checks across interfaces prevent further corruption. If the integrity assertion is not passed, the program fails fast or attempts repair. Independent processors called "watchdogs" can also observe the process and raise a flag, erase the state, or repair it.

N-version Programming

Because redundancy is not sufficient to eliminate faults resulting from common causes of failure, a fail-safe but very expensive solution is n-version programming. Choosing a set of processors to perform the same function, each cluster is given a different set of instructions on a MIMD machine. Duplicate processes also mask hardware faults. If one processor fails, the other processor can continue. Most errors caused by design failures are transient and will disappear if the process runs later in a different context. These "benign bugs" are sometimes easier to ignore than to fix, and by having two processes performing the same function, the transient errors will become transparent and irrelevant. The primary process performs, while the backup process provides a cold backup, passively observing. Checkpoint messages are sent to the backup at appropriate times. However, when an inconsistency is detected in the process state, it fails and notifies the backup process, which then takes over. Unfortunately, the takeover process is very difficult to write, test, and maintain.

Because n-version code is so expensive, it is wise to determine where in a program such precautions should be taken. One version of this is to design process pairs, in which each system carries half the load during normal operations. A comparator examines their results and declares a fault if the outputs are not identical. If one fails, the other serves all the users. This allows online software and hardware updates and masks most operational, maintenance, hardware, software, and environmental faults. However, independent diversity can result in independent groups making the same mistake or a common mistake arising from the original specification.

Memory Scrubbing

To help offset the loss of bits, it is suggested that a memory scrubbing technique be used, writing and reading the data in and out of memory over at least an hour's time. Undetectable errors are mostly high order multiple bit errors, as shown by an analysis of all possible non-recoverable-error bit patterns for a 16-bit ALU.[10] Only slightly lower coverage was shown for a 32-bit ALU. The memory protection code should provide the following:

- Detection and correction of burst errors: This provides on-line correction of Single Event Upset (SEU) errors, as well as on-line correction of whole chip failure.
- Detection of double burst errors: This provides on-line detection of additional SEU-induced errors which may occur in the same data word across two memory devices at the same time.
- Detection of address errors: This is performed by adding three redundancy symbols or 12 redundancy bits.

---

[10] Stella.

In order to minimize the accumulation of latent faults in memory, each word corrected on-line by the error code should be rewritten into the faulty memory location. This enables the fault as well as the error (the erroneous bit in the word) to be corrected. In addition to this level of scrubbing, off-line scrubbing is also performed which will periodically read/write information in memory locations that are accessed infrequently by normal system operation. Obviously, this effort is time-intensive, and careful analysis should be used in determining where the technique should most effectively be used (e.g., safety-critical software).

Recovery offers several choices: remove the faulty component, reload (which is appropriate for a transient fault), or reconfigure. Reloading can be achieved either by checkpointing or by synchronizing with one of N healthy, replicated processes. Checkpointing is practical when the interval of uncertainty is sufficiently large that infrequent checkpoints are adequate. In safety-critical applications, the tolerable interval of uncertainty can be very short, in which case replicated processes with N-version synchronization and voting offer a viable solution. Reconfiguration requires a dynamic routing capability.

An example of the interrelationship among various fault tolerance mechanisms is shown when a power supply goes down. This immediately eliminates the processing capability of as many processors as are associated with the power supply in question. Attaching power supplies to each cluster of processors provides a guarantee that, short of catastrophic failure, processing can continue. The fault identification capability of the fault manager detects the fault through its "health and welfare" polling, the diagnostic capability determines that the power supply is one source of the problem, and the fault manager immediately reconfigures the system, using checkpoint and rollback that has been in place to provide continued processing. A distributed operating system becomes the system controller, portions of which are allocated to separate processors. A short pause occurs while the unaffected portions of the operating system manage a controlled restoration of service. Or redundant processes continue their work uninterrupted. However, the configuration manager must now put into action a fallback scheme for insuring continued redundancy. And, if failing to possess the resources for that capability, another scheme must be formulated that allows for degraded but continuing performance. A journal is kept of all recovery events for later use by maintenance personnel at a status display. In the meantime, the healing process is activated, attempting a restart. Because statistically it has been found that transient errors, such as bombardment by alpha-bits in space, cause 80% of the failures, there is good reason to attempt a re-start. When this is accomplished, the fault manager returns control to its former state, again using the checkpointing and rollback stratagem.

In this paper, I have discussed a variety of faults that beset Massively Parallel Processors and suggested various mechanism to detect, isolate, heal, and recover from them. MPPs become more unstable as the number of processors increases, and yet the larger number of processors available make fault tolerance an accomplishable reality. The promise of Fault Tolerance and MPPs is approaching fruition. However, much more work is required before MPPs reach a point of stability and robustness of software. Using some of the mechanisms and tools discussed herein, the software industry can create a system of high reliability in which the failure of any one module or component will not degrade full functionality.

REFERENCES

[1] — "Reliable Distributed Sorting Through the Application-Oriented Fault Tolerance Paradigm." IEEE Transactions on Parallel and Distributed Systems. July 1992.

[2] Birman, Kenneth P. "The Process Group Approach to Reliable Distributed Computing." Communications of the ACM. December, 1993.

[3] Harper, Richard E. and Jaynarayan H. Lala. "Fault Tolerant Parallel Processor." Draper Lab. AAIA Journal of Guidance. June, 1991.

[4] Klein, Ralyiabjectives are needed, et al. A Practitioner's Handbook for Real-Time Analysis. Kluwer Academic Publishers, 1993.

[5] Rangarajan, Sampath and Donald Fussel. "Diagnosing Arbitrarily Connected Parallel Computers with High Probability." IEEE Transactions on Computers. May, 1992.

[6] Resnick, Mitchel. Turtles, Termites, and Traffic Jams: Explorations in Massively Parallel Microworlds. MIT Press, Cambridge, MA, 1994.

[7] Stella, Joe. "Complexity vs Reliability in Fault Tolerant Computing Systems." Internal Raytheon Memo #EMS88-0070. November, 1988.

ANALYSIS AND MANAGEMENT OF SCHEDULE AND COST RISK FOR PROJECT PLANNING AND EXECUTION

Robert Fung
Lumina Decision Systems
4984 El Camino Real, Suite 105
Los Altos, CA 94022

Max Henrion
Lumina Decision Systems
4984 El Camino Real, Suite 105
Los Altos, CA 94022

Timothy Barth
MS TP-PED-1
John F. Kennedy Space Center, NASA
Kennedy Space Center, FL 32899

ABSTRACT

The Schedule and Cost Risk Analysis and Management system (SCRAM) is an innovative tool for modeling and analyzing the schedule and cost risks associated with complex projects and schedules. The SCRAM prototype builds on the Demos product developed by Lumina Decision Systems and the Schedule Publisher product from Advanced Management Solutions. Since it is based on commercial products, SCRAM is able to provide powerful and well-tested general facilities, and offers low technical risk as a foundation for developing innovative concepts. We anticipate that SCRAM's combination of powerful and innovative features will be appealing for a variety of sophisticated scheduling applications in government and industry. SCRAM was built to support the analysis of ground processing operations for the space shuttle. We have worked with NASA personnel to develop a risk model of ground processing that centers around a categorization of the common delays that effect ground processing tasks. SCRAM evaluates the importance of those delays and tasks in terms of their contribution to overall schedule and cost risks, and thereby identifies those tasks and delay types where the ground processing staff can most fruitfully concentrate their efforts at process improvement.

BACKGROUND AND MOTIVATION

Managing and accounting for the risks due to uncertainties is a major function of project planning and scheduling and few tools are available to support this difficult task. The Schedule and Cost Risk Analysis and Management system (SCRAM) is an innovative tool for modeling and analyzing the schedule and cost risks associated with complex projects or schedules.

Almost all projects[1] are subject to significant risks. These risks can be low probability, high impact events (e.g., extremely bad weather, unexpected equipment failures) as well as more common, low impact events (e.g., the delay in the availability of the appropriate personnel for a task). Risk analysis for project planning proceeds as with all other risk analyses:

1. Categorize the risks, and identify variables of interest
2. Model the interaction between risks and their effect on the variables of interest
3. Quantify the risks, and finally
4. Assess the (cumulative) effect of these risks on the variables of interest

For project planning, the variables of interest are relatively stable from application to application and include project length and project cost and might include other variables such as the variability in resource usage, etc. However the risks to a project and measures of the effect (i.e., importance) of the risks to the variables of interest are project-specific.

Since their inception, project planning tools have attempted to deal with the uncertainty in forecasting and managing uncertainty in project task lengths and costs. However project risk tools could be improved in certain key areas.

PERT (Program Evaluation and Review Technique) and CPM (Critical Path method) are the classical technologies for project planning. The original version of PERT prescribed using three estimates of the task length: a most likely estimate, an optimistic estimate, and a pessimistic estimate. A critical path is computed using the most likely estimate of each task. To compute a probability distribution for schedule length, PERT makes two assumptions. First, PERT assumes the critical path does not change with perturbations in task lengths. Second, PERT assumes independence between the probability distributions between each task length. A third minor assumption is that the three task length estimates specify a Beta distribution. Given these assumptions, the three task length estimates for each task on the critical path can be used to estimate the mean and variance of the overall project length.

Recently, several commercial project planning tools have developed add-on risk analysis modules that make use of Monte Carlo techniques to improve upon the PERT analysis. In these tools, the probability distribution for each task length can be of arbitrary form. For each Monte Carlo trial, task length values are chosen for each task from its task length probability distribution. A critical path is computed for each trial. At the end of the trials, a project length and a project cost distribution can be computed as well as the probability that a task is on the critical path. Thus the Monte Carlo approaches have addressed several of the weaknesses of the PERT method: the poor assumption that the critical path does not change and restricting task length distributions to be Beta distributions.

However both PERT and current Monte Carlo techniques suffer from a major drawback – they are entirely focused on the modeling of the project elements (e.g., tasks, milestones) themselves and do not provide for the representation of risks that are external to project elements. Because they do not model external risks explicitly, they must model the impacts of these risks through assessments of the risks on individual project elements.

---

[1] For simplicity we will use project planning terminology in this paper even though SCRAM is equally applicable to schedule analysis.

There are two major drawbacks to this approach. The most important and obvious drawback is that since external risks are not represented, analysis of the importance of these external factors cannot be performed. It seems this is a critical function in many project planning situations (as it is in the Shuttle ground processing), For example, the manager of a construction project could not easily analyze the effect of the delay in a regulatory approval based on his uncertain beliefs about the delay length. Second, if the relationship between the external risks and the project elements is complicated, then attempting to summarize that relationship in individual, independent task length distributions will be a very approximate model of the relationship. For example, if a risk effects more than one task, the lengths and costs of the effected tasks can become highly correlated. These cannot be modeled at all either in PERT or with the current Monte Carlo tools.

The Schedule and Cost Risk Analysis and Management system (SCRAM) is an innovative tool for modeling and analyzing the schedule and cost risks associated with complex projects or schedules. SCRAM provides the functionality to specify an arbitrary risk model for a project. The model specification is performed through an intuitive graphical process. In addition, SCRAM brings both new project and state-of-the-art risk analyses to bear on understanding the importance, and prioritizing the risks facing a project.

SCRAM

The SCRAM architecture is modular and has made use of existing commercial tools to the maximum extent. SCRAM includes three components: a risk modeling and analysis module, a standard project planning tool, and a system executive.

SCRAM Risk Modeling and Analysis Module

The risk modeling and analysis module is implemented in the Demos, a commercial system developed by Lumina Decision Systems. Demos is a graphical environment for creating, analyzing, and communicating probabilistic models for risk analysis. Standard modeling tools, such as spreadsheets and simulation languages, often hinder effective analysis and communication because they do not have methods for understanding a model's organization, and do not provide for the explicit representation or analysis of risks. Demos is designed specifically to focus on these areas so as to provide insights to quantitative models, as quickly and effectively as possible.

To develop and display a model's organization, Demos uses hierarchical influence diagrams, enabling the user to see and edit the model structure visually, and at various levels of detail. Figure 1 shows four levels of the hierarchy for the Ground Processing application of SCRAM. In addition, Demos uses multidimensional spreadsheet views instead of a two-dimensional standard spreadsheet view to organize the data in a model.

Finally, Demos provides state-of-the-art risk analysis functionality, providing for the modeling of both discrete and continuous risks, and Latin hypercube and Monte Carlo simulation. Figure 2 shows a Demos-computed distribution on project completion length for one of the shuttle's main subsystems.

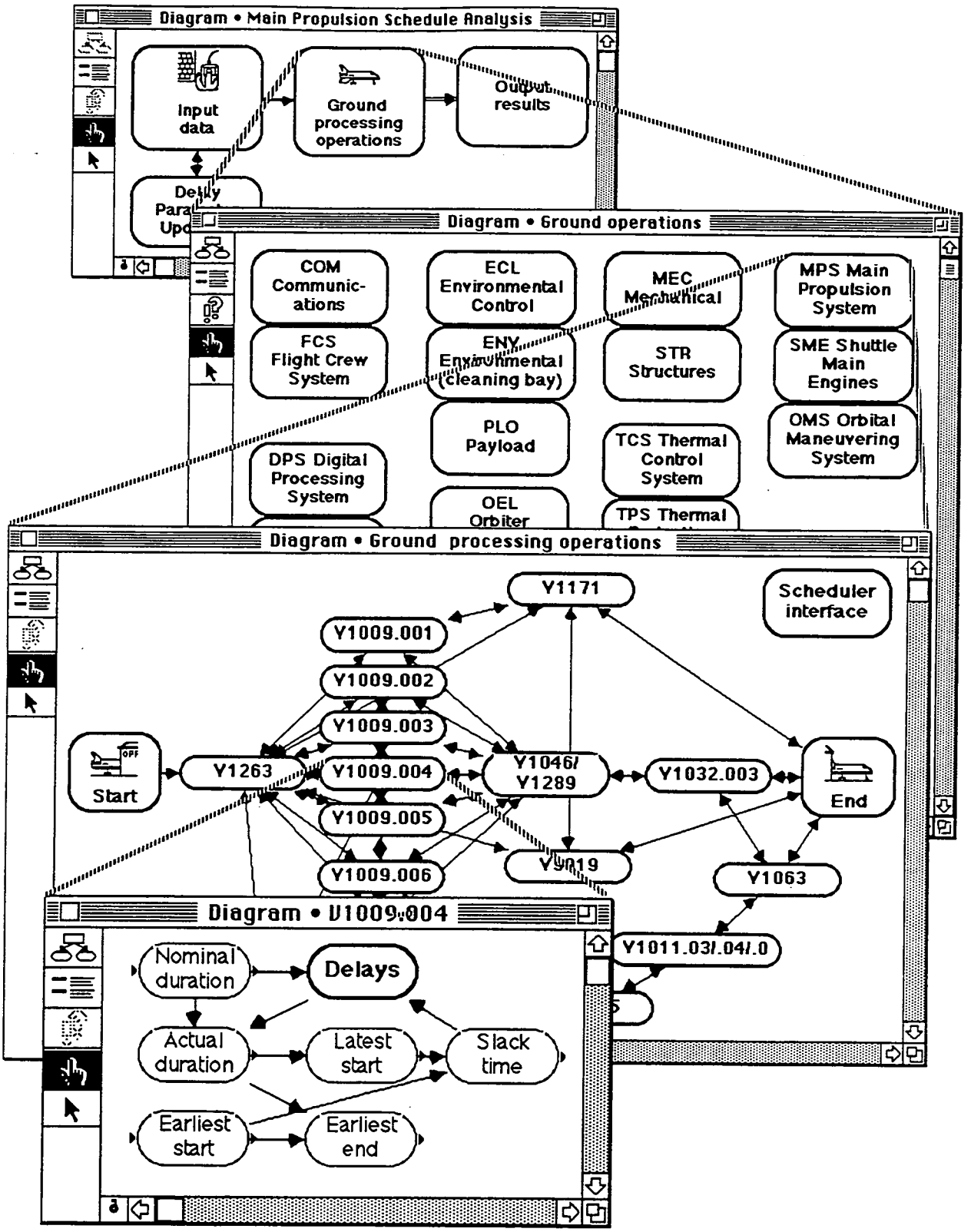


Figure 1: Four levels of influence diagram hierarchy of the SCRAM: Ground Processing Application

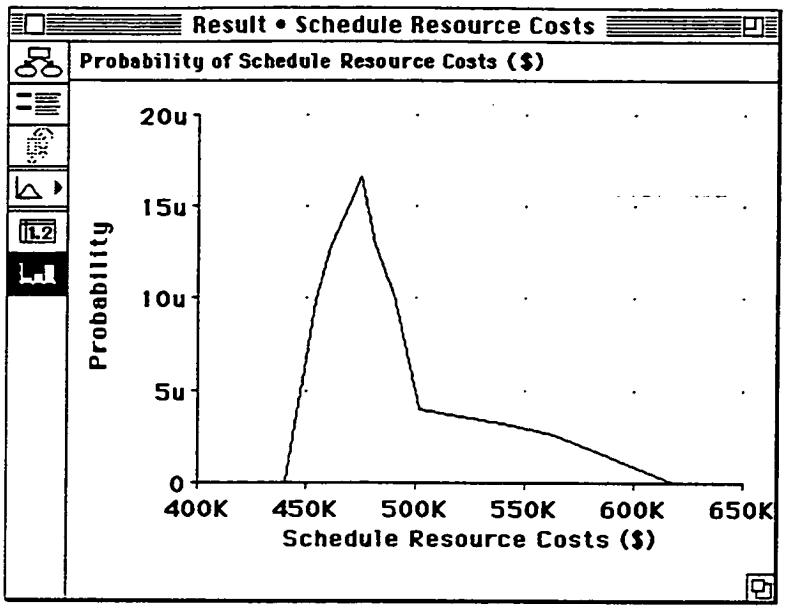


Figure 2: Probability Distribution of the Total Schedule Cost

SCRAM Project Planning Module

The project planning module of SCRAM provides three main functions. Most importantly, the project planning module provides the standard project planning analysis techniques including critical path analysis, earned value task analysis, and resource-leveling. Second, the project planning module provides a convenient and familiar means for specifying scheduling tasks and resource usage, and displaying schedules in standard PERT and GANTT chart forms. Third, the planning module provides a means for input or output of scheduling information from other standard scheduling software.

The project planning module in SCRAM is based on the Schedule Publisher product of Advanced Management Systems (AMS) of Yucipa, California. Schedule Publisher is a widely-used project planning tool. AMS initiated the development of Schedule Publisher in 1988. Based on experience with designing and building an earlier product and in close association with end users trying to schedule their projects, AMS has designed and built unique graphical planning and scheduling functions into Schedule Publisher. Focusing on portability, Schedule Publisher is available on a number of platforms including PCs, Macintosh, and various Unix computers.

The primary modification to Schedule Publisher for SCRAM was the design and implementation of an algorithm for computing task float times when resource leveling had been applied to the project. Consider the simple example network shown in Figure 3. Let Task V100 have a duration of 7 days and V200 have a duration of 8 days. Let both of the tasks use one unit of the technician resource and let only one technician be available during the projects execution. The resource-leveled schedule for this simple situation is clear – the tasks must be executed in series.

We modeled this network in several commercial project planning tools. All gave the same answer. The slack time for V200 was 0 as it should be, but the slack time foe V100 computed by all these project scheduling systems was 8 days – the length of V200! Clearly, both tasks are critical (i.e., the slack time for V100 should be 0) since they are necessarily sequenced due to

resource constraints! To address this problem, we have developed concepts and an algorithm for computing slack times (and task criticality) under resource constraints.

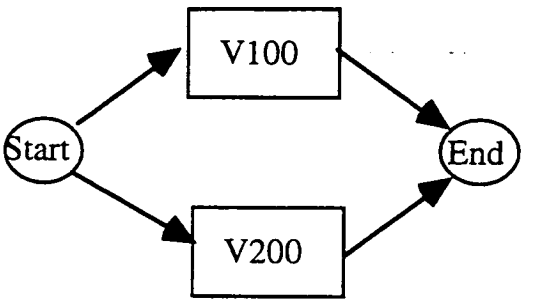


Figure 3: A simple project with two parallel tasks that contend for the same resource.

SCRAM System Executive

The SCRAM system executive provides the end-user with a simple and intuitive graphical user interface for editing both schedule data and probabilistic data, and viewing analysis results, as well as simple controls for communicating commands to the risk analysis or planning modules to carry out various analyses. Finally the system executive has as a submodule a database which contains all system data. Figure 4 shows the starting screen of the Macintosh version of the System Executive. Figure 5 shows the analysis screen of SCRAM in which the user can request a number of results as well as control the number of Monte Carlo trials with which the results are computed

Applying SCRAM

Use of SCRAM can be separated into two major phases: initialization and analysis. In the initialization phase, there are two major steps: the development of the schedule, and the construction of the risk module. The user typically populates the schedule data of the system by constructing the project plan using the plan module interface or he may use a preexisting plan in another format. The plan is then read into the system executive database, so that editing can be done from the system executive user interface. Similarly, the risk models are entered through the risk module user interface and read into the system executive database. Since the form of the risk models can change from application to application, the user interface requires some tailoring from application to application.

In the analysis phase, the user can define multiple variations of the plan, make changes to the risk model parameters etc.. Based on a particular plan and set of risk parameters, the user can through the system executive request analyses to be performed. Analyses are performed through Monte Carlo simulation. SCRAM simulates the execution of a plan by simulating, in time order, the execution of each task in the schedule. Simulation of a task involves simulating all those variables that effect the task's duration. For example, in the SCRAM Ground Processing application, the delays associated with a particular task are simulated. The simulation includes whether the delay occurs or doesn't occur on the task and for how long. The delay lengths are then added to the basic task length to simulate the effect of the delays on the task. After each

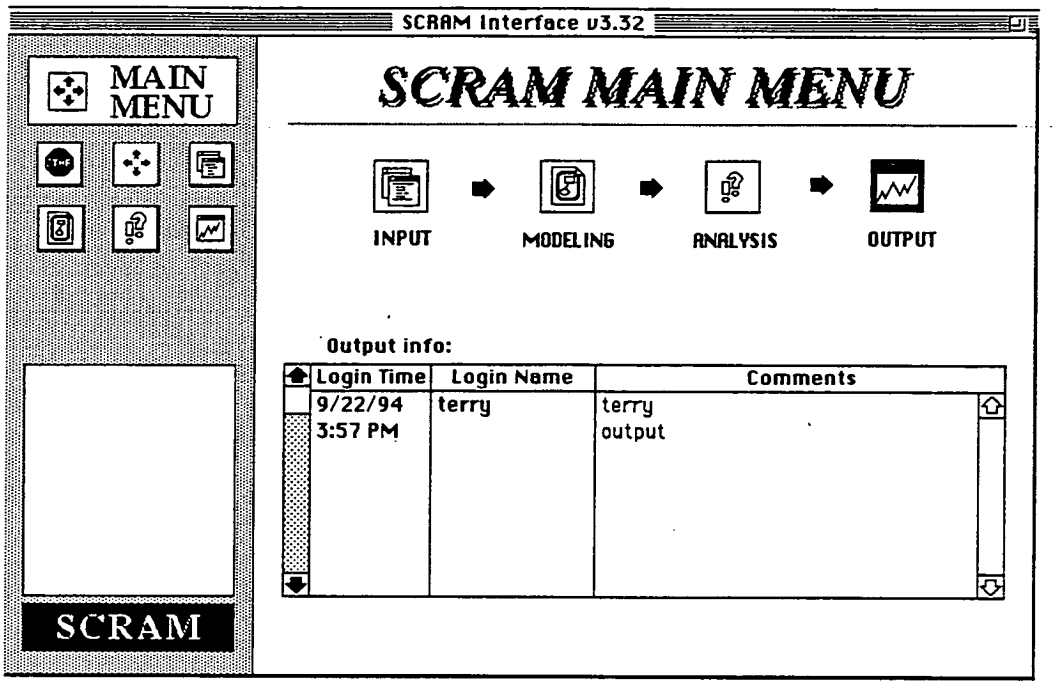


Figure 4: The starting screen of the Macintosh version of SCRAM

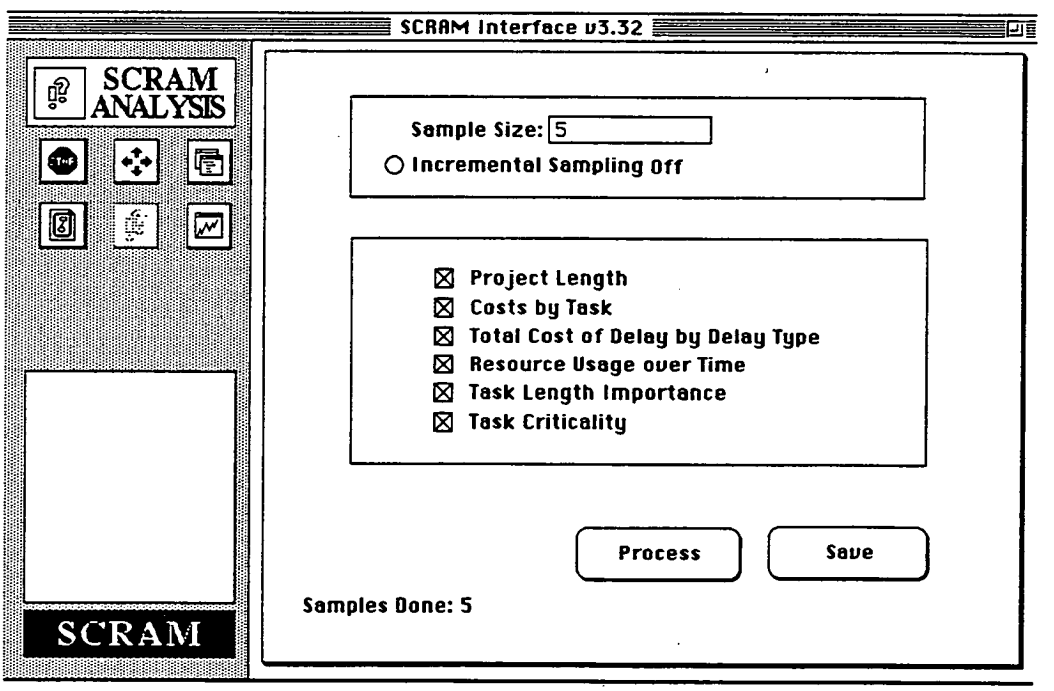


Figure 5: The SCRAM analysis screen in the Macintosh version

task is simulated, SCRAM adjusts the schedule to meet the resource and temporal constraints between tasks. The simulation of schedule execution is performed a number of times randomly selecting values from the distributions of the input parameters to obtain a Monte Carlo sample of output values. Analyses of schedule length and costs are computed from the simulation results.

SCRAM Shuttle Ground Processing Application

SCRAM was designed originally to support operations analysts and the shuttle ground processing staff. SCRAM evaluates the importance of tasks and external events in terms of their contribution to overall schedule and cost risks, and hence identifies those tasks and events where the ground processing staff can most fruitfully concentrate their efforts at process improvement.

Based on discussions with NASA personnel it was determined that for shuttle ground processing, the appropriate uncertain events to model are delays to task completion since a categorization of the types of delays tasks are subject to was available, as was a system for collecting the data on the frequency and characteristics of delays. Table 1 shows a partial list of the delay categories used by the SCRAM Ground Processing model.

A. Paper
A1. Paper Unavailable
A2. Constraints list/running addendum problems
B. People
B1. Unavailable
B2. Quality
C. Parts
C1. Orbiter Parts

Table 1: Partial List of Delay Types

Figure 6 shows how delays are modeled to effect tasks. Delays have a frequency of occurring on particular tasks. The frequencies are labeled as "x out of 10 times". For example, the probability that delay B11 occurs on task V1263 is 30% (3 out of 10). The min, mode, and max columns represent the minimum, mode and maximum delay times in hours if the delay occurs.

The primary goal of the SCRAM Ground processing application is to provide a schedule-based analysis of the "importance" of the delay types, so that steps can be taken to reduce the delays that are causing the most problems. By "importance" we mean the impact of a delays category on schedule and cost. Previous analyses of delays have only taken into account the frequency and length of delays. Clearly, this could be improved since delays on critical tasks increase schedule length and therefore should be counted more heavily and since the costs of delays were not considered at all.

To compute the "importance" of each delay type, for each trial, and for each delay occurrence, the length of the delay is weighted according to whether or not the delay was associated with a task on, near, or off the critical path. For each delay type, we then aggregate the costs and the weighted length of every delay of that type that occurred during the schedule

simulation. This gives an overall balanced analysis of the importance of each delay to schedule length and cost. Figure 7 shows an example graph of the delay importance analysis for SCRAM.

Result • U1263 delay table

Mid Value of Y1263 delay table

Relevant delay types

Delay Parameters

	Frequency	Min	Mode	Max
B11	3	0.25	1	8
B12	4	0.25	1	4
A1	2	1	6	10
C21	3	0.25	0.75	2
D22	5	1	2	8
A3	2	0.5	2	3
B15	1	1	2	4

Figure 6: Portion of *Delay Table by Task*

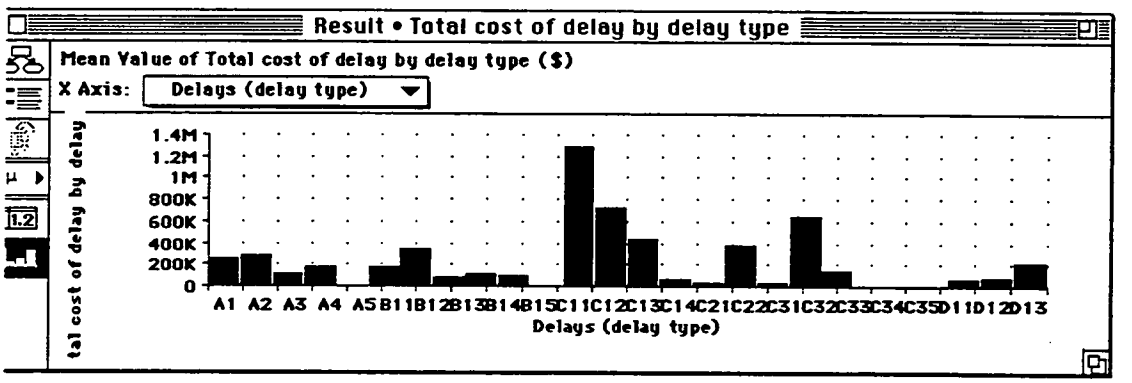


Figure 7: Example delay importance analysis for SCRAM Ground Processing Application

Current tools for analysis of the timing and delays in ground processing are based on deterministic analysis of historical delay data. They summarize the total number of delays, length of delays, and time averages (e.g., 3-week, 6-week, and 12-week averages) by directorate, contractor, and delay type. Because existing tools are oriented toward historical analysis rather than prediction of future performance, they cannot assist in forecasting the effects of proposed improvements to the ground processing. For this reason, they do not directly support analysis of the decisions for improving ground processing. Finally, the existing tools are oriented only at schedules and timing, but do not support analysis of the effects on overall costs, which is a major concern in these times of constrained budgets.

DISCUSSION

In summary, the first generations of project risk analysis tools have been focused on representing uncertainties directly on project attributes (e.g., task length and task cost). While this minimizes the extra information that is needed to perform such an analysis, such systems cannot analyze the effect of external risk factors to the project.

We have developed SCRAM to allow project managers to substantially improve their understanding of the impacts of the risks to their project. The application of SCRAM involves modeling how the risks that face the project, impact individual project elements (e.g., task length). Standard scheduling techniques are then used to propagate the individual element impacts to impact on schedule length and cost. An application of SCRAM also involves the customization of the basic SCRAM interface to take into account the project-specific risk factors. A working SCRAM system has been completed for Macintosh computers and a PC version will be completed by the end-of-the-year. SCRAM has made use of two commercial software products and the relatively new trend towards inter-application communication and operation.

Possible commercial applications include any project planning situation in which uncertainties pose a serious threat to the successful completion of the project and where the economic stakes are high enough to justify the resources it takes to do such an analysis. Shuttle ground processing certainly fits these criteria as might large construction projects where financing or regulatory approvals might pose significant risks. Other applications include large, critical software projects and R&D planning for companies (e.g., pharmaceuticals) for whom this is a critical function.

Virtual Reality Simulation

Virtual Environments for Training and Education

NOT AVAILABLE AT THIS TIME

SIMULATION VIRTUAL MACHINE

Kenneth Hill
Computer Engineer, NASA-DK
Johnson Space Center, Houston, Texas 77058
(713) 244-7250

Rob Sturtevant
Systems Engineer, CAE-Link
2224 Bay Area Blvd., Houston, Texas 77062
r_sturtevant@hso.link.com

ABSTRACT

This paper highlights the "Simulation Virtual Machine" (SVM), a software executive layer running on top of a COTS real-time operating system that provides distributed, fine-grained message communication, rate-monotonic scheduling, and other capabilities in support of real-time training simulation. SVM was developed at the Johnson Space Center in Houston, Texas and will be used to execute software models which simulate the Space Shuttle and the International Space Station both in multi-system, high fidelity training, and in smaller, single-system training. Because of its use in a wide range of simulators at the Johnson Space Center, SVM will also have the potential for a diversity of applications outside the space industry.

INTRODUCTION

Simulation Virtual Machine (SVM) is a software executive layer for training systems. It consists of a set of real-time executive Ada modules and an Ada architecture which eases the effort involved in real-time software development and integration, and significantly reduces the cost of software maintenance and sustaining engineering. SVM lowers development costs and enhances code reuse opportunities by influencing structural consistency throughout a project using a documented object-oriented software architecture. SVM's executive lowers integration costs by shielding the application from the computer hardware specifics and by providing a standard interface to applications. SVM's executive is designed to support a distributed hardware environment (multi-task, -CPU, -node, -platform) and provides parallel, rate-monotonically scheduled execution threads, thread-to-thread time-consistent data-homogeneous message communication, and other executive functionality such as moding, datastore, and data term visibility. SVM is also based on standards (Ada, POSIX) and therefore is easy to rehost to different computer platforms. SVM's structural consistency and the executive's flexibility lowers maintenance/upgrade costs by reducing the amount of work required to understand the software, change the structure, and add new capabilities.

The following aspects distinguish SVM as a leading technology in simulation design:

Rate Monotonic Scheduling: Although the traditional flight simulator uses a frame-based approach to scheduling the execution of software models, CAE-Link has chosen an innovative scheduling approach for SVM known as Rate Monotonic Scheduling (RMS). Developed from a recently derived scheduling theory, RMS provides automation beyond the extent of a traditional frame-based scheduler. One advantage is the automatic allocation of software models to a CPU in a multitasking environment. This allocation mathematically assures the completion of models within their deadline. A second advantage is that RMS allows execution of models with non-harmonic rates, which could not be done without certain modifications using a frame-based approach. Thus, a software model can be incorporated into the system regardless of its execution rate.

Software Backplane: The software backplane, which provides fine-grained messaging between models, is another benefit of SVM. The communication path (local, shared, and reflected memory or across a LAN) is transparent to the models. This allows models to be easily relocated within the system as changes are needed. The software backplane also guarantees messages are delivered in a time-consistent manner in a distributed environment.

Object-Oriented Design: The SVM design is based upon object-oriented methodology. CAE-Link has developed a Software Architecture Standard which defines the coding standards and object-oriented principles for the system. Object-oriented applications tend to be easy to maintain because software components are easily mapped to real-world objects, providing a logical and understandable structure for the system. In addition, code changes and system

updates are localized, minimizing effects to other parts of the system with which they interface. Object-oriented design also maximizes the potential for code reuse. Finally, as opposed to the traditional data pool configuration, software models in SVM hold their data locally, minimizing the possibility of data corruption during execution.

Ada Programming Language: SVM is programmed in the Ada language, which inherently provides several benefits. When proper coding standards and techniques are used, Ada programs are easier to maintain as well as to port to other hardware platforms. Because of its structure and Ada's standardization, SVM has been quickly ported from Silicon Graphics computers to Solbourne, VAX, and Apollo workstations and a version is executing in the Rational development environment.

SVM's Environment: SVM was developed for the Space Station Training Facility (SSTF) at the Johnson Space Center. Figure 1 shows the distributed SSTF hardware environment. It consists of 2 multi-node, multi-CPU Silicon Graphics Challenge Session computers with interfaces to Space Station flight-equivalent units (FEUs) (like MDMs), 14 multi-display instructor / operator stations, a network training system for the Tracking and Data Relay Satellite System and ground station training, USA crew stations, Japanese / European / Russian crew stations, associated visuals (out-the-window, closed-circuit TV), and an interface to the shuttle simulation. It also includes an operation support computer that supports load builds, input collection (for mission-specific initialization points), and program download and execution monitoring. The computers are time-synced via GPS-based Central Timing Equipment (CTE). Operationally, different assets (hardware devices on a LAN) can be used with either Session computer and can be added or dropped on-the-fly. SVM creates a virtual computer allowing models to execute without regard to their physical location in this distributed environment.

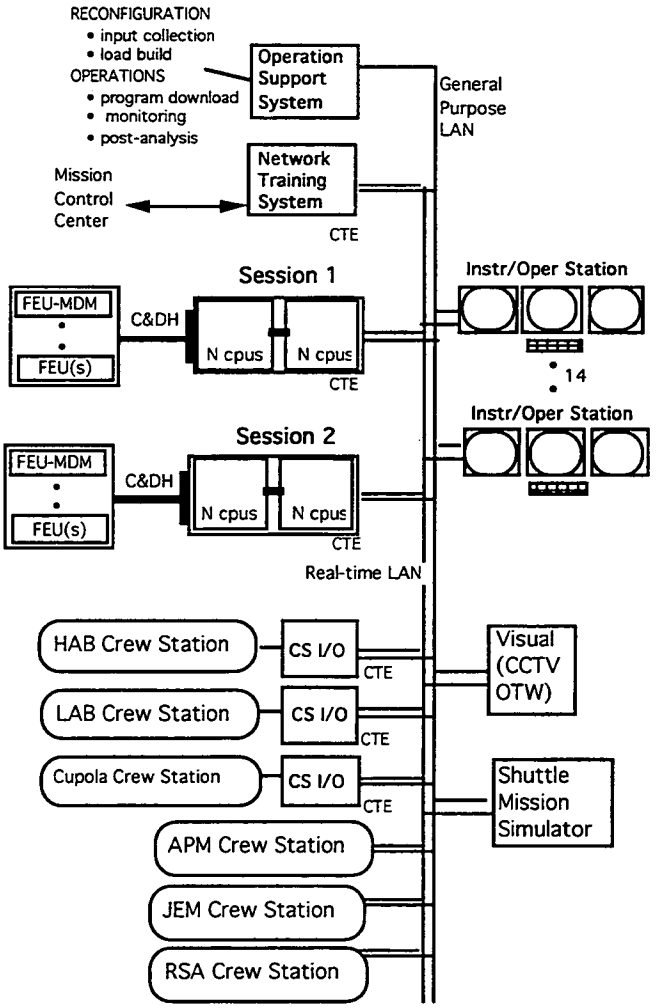


Figure 1 - Space Station Training Facility Hardware

SVM DESIGN

A real-time system controls an environment by receiving data, processing it, and taking action or returning results quickly enough to affect the functioning of the environment within the allotted time envelope. Real-time models cannot execute too fast or too slow - they must run within a specific time period (e.g. 40 hertz) in a deterministic fashion.

In the marketplace, there are many real-time operating systems (OSs) available (such as VxWorks, LynxOS, VRTX, and IRIX) that provide generalized scheduling, basic communication, real-time response to interrupts, and support to other low-level OS features. Some are compliant with POSIX real-time extensions or other open systems standards. For real-time simulation, a layer must be built on top of the COTS real-time kernel to provide domain-specific services. For example, in a training system an executive layer is created to provide simulation moding (initialize,

run, freeze, step-ahead, datastore, safestore, hold), additional scheduling capabilities, and specialized communication mechanisms based on the requirements of the system. In the SSTF, the SVM layer is built on SGI's IRIX OS (figure 2). SVM consists of:

- rate-monotonic scheduled executive
- software backplane providing communication for threads across a distributed hardware execution environment
- moding and control system providing distributed, coordinated moding and timing
- logical symbol specification, "Distributed Identifier Specification", for IOS terms, datastore, safestores, and datalog
- ability to read values for logical symbols for display or logging (on a term-by-term request)
- datastore and safestore capability
- interface agents to other LAN assets
- real-time I/O capability
- device drivers
- Ada software architecture.

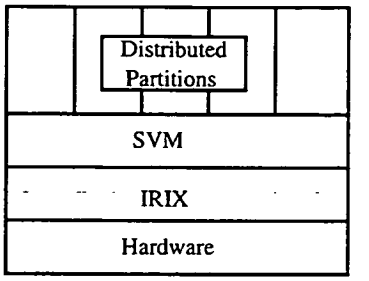


Figure 2 - Layers

SVM provides several innovative capabilities [2]. Key innovations summarized below are:

- Rate-Monotonic Scheduling
- Distributed Model Communication
- Object-Oriented, Ada Structural Definition

Rate-Monotonic Scheduling

Most traditional flight simulators use a frame-based approach to scheduling the execution of software models. In this approach, a minor frame is selected (e.g. 40 Hz or a 25ms period) (see figure 3). All models must start and end within the minor period. Models that run slower (10 Hz) must divide their execution across 4 frames or complete in 1 frame leaving 3 frames unused. In the latter, 3 other 10 Hz models can be run at different frames to provide an execution balance (maximizing the use of the CPU). Execution rates are limited to harmonics of the base rate (40, 20, 10, 5, 2.5, 1.725 Hz).

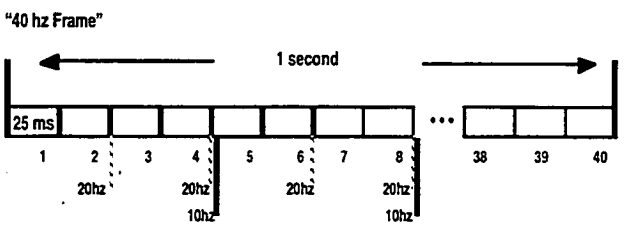
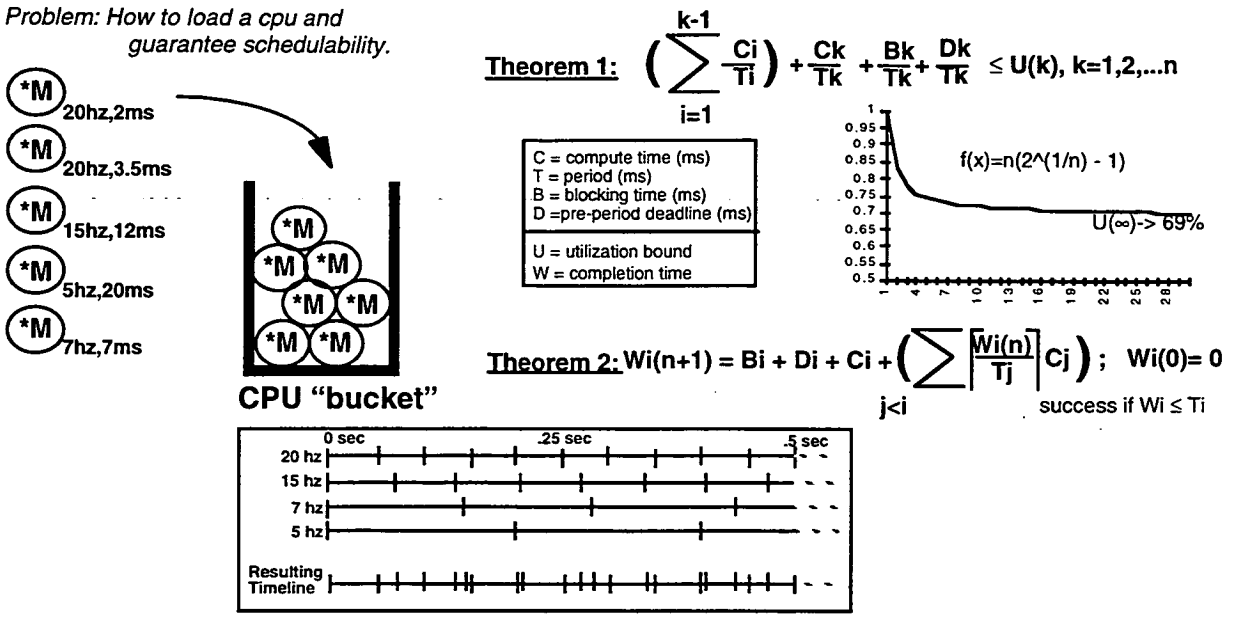


Figure 3 - Frame-based Scheduling

SVM uses a non-frame-based, innovative scheduling approach known as Rate Monotonic Scheduling (RMS). Developed from a scheduling theory in 1973 [3], RMS provides automation beyond the extent of frame-based schedulers. RMS is a fixed priority scheduling method that uses the rate of a periodic task as the basis for assigning priorities to periodic tasks. Tasks with higher rates are assigned higher priorities. RMS deadlines can be verified using a mathematical analysis known as rate-monotonic analysis (RMA). In RMS a minor frame is not required (as in a frame-based system) and therefore models do not have to execute within the minor frame boundary. Any periodic rate may be specified which allows models to run at their natural periodic rate. Through the preemptive environment and services provided by the kernel OS (task priorities, task preemption), models execute in a rate-monotonic fashion.

Rate-monotonic analysis is summarized in figure 4. On the top are models that have to be scheduled on a CPU with their given execution rate and time requirements. To verify the schedulability, the analysis is performed by running algorithms for theorem 1 and theorem 2. If the utilization bound is under the curve on the graph for theorem 1, theorem 2 does not have to be run. If the analysis passes, then the load is schedulable. The resulting execution timeline is shown at the bottom of the figure (which is a sum of all the rates supported). Note the tick marks are not harmonically divided as in a frame-based scheduler.



* a "model" consisting of logically related code plus a thread with attributes: period (ms), time (ms), pre-period deadline (ms), blocking (ms).

Figure 4 - Rate Monotonic Analysis

RMA is valid only on a single CPU basis. Therefore, an important addition to RMA is SVM's schedule assist tool which automatically sorts models onto CPUs and then performs the RMA. The schedule assist tool takes as input all the models (scheduling parameters) that execute on an asset plus the hardware configuration. The tool automatically sorts the models onto CPUs. RMA is then performed on each CPU set. The models are automatically adjusted across the CPUs until there is a schedulable set (RMA passes) on each CPU. The tool then generates the Ada mains for the systems with no human intervention required.

The internal executive real-time execution structure is shown in figure 5. An external clock provides the real-time interrupts to affect the timeline. The interrupt handler signals the master executive on each CPU which releases via Ada rendezvous all model threads that are to be scheduled. Models create a thread by instantiating a generic model Ada package in the body of their top-level model package. Aperiodic and asynchronous threads are also supported.

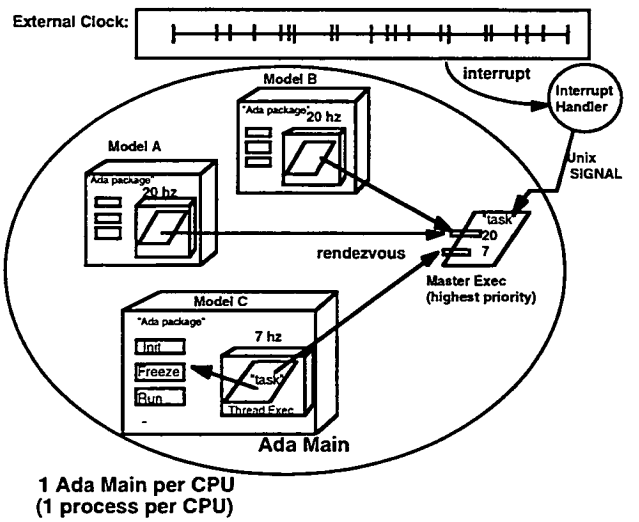


Figure 5 - Structural Scheduling Mechanism

SVM's RMS executive:

- is based on rate-monotonic scheduling approach
- supports any execution rate
- supports periodic, aperiodic (event) , and asynchronous scheduling of threads
- encapsulates application models in threads (Ada tasks). Applications instantiate the generic model for the type of scheduling required.
- synchronizes each application thread via a master executive

- has one master executive per CPU using normal Ada tasking or 1 master per node (when Ada tasks mapped to UNIX s-procs)
- synchronizes all threads with a timing source
- provides GMT and SGMT time functions and corresponding operations
- automatically configures to model threads (no set limits, not hard-coded)
- detects and statuses period overruns
- provides error status
- automatically creates schedulable Ada mains. The schedule assist tool performs sorting of application models across multi-CPU,-node,-platform systems, applies rate-monotonic analysis to each CPU, and organizes the application set into Ada mains.

Distributed Model Communication

Most traditional flight simulators use global memory to store model state data and model-to-model interfaces. If the models cannot be located in a single global memory area (for example: a multi-node system), then special services must be created to fetch the data across the multiple systems and populate the datapool. For large scale, multi-CPU systems, time-consistent and homogenous data becomes very difficult to manage with a datapool structure since models may be reading/writing at any time. Usually, the execution jumplist is maintained by hand in an attempt to keep the datapool coherent.

SVM's software backplane (SWBP) provides message-based model-to-model communication. Application models send/receive any size of messages to/from each other instead of relying on a common datapool. This fine-grained (higher resolution than general UNIX interprocess communication) model messaging service uncouples models from one another and relieves concerns of hardware specifics. The SWBP sends messages over a variety of hardware paths (local memory, shared memory, reflective memory, or LAN) using standard services of the real-time kernel OS (shared memory services, socket communication) (figure 6). The most efficient path is automatically determined by the SWBP based on the location of the sending/receiving models. The path is transparent to the models allowing them to be easily relocated within the system as changes are needed. This transparent, message-based communication provides a distributed virtual machine to applications.

Another feature of the SWBP is its ability to deliver homogeneous (whole) messages at a consistent rate (based on time) to receiving models. Unlike a global datapool where data can be read/written at any time, the SWBP buffers data between senders and receivers and

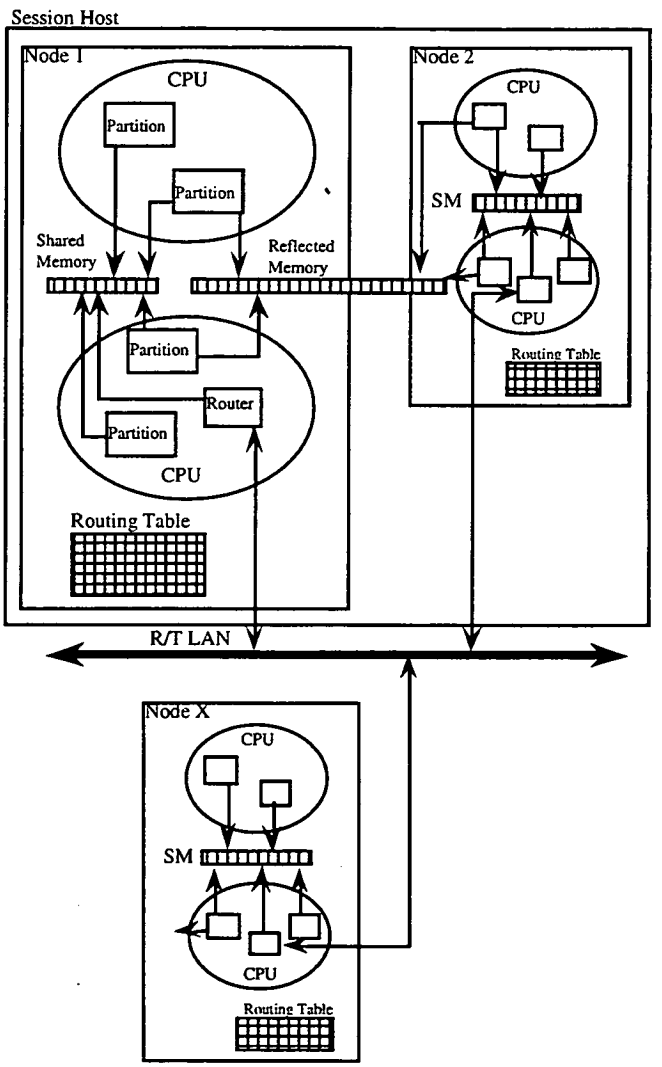


Figure 6 - Software Backplane Hardware Paths

then provides them with time-consistent whole messages. This makes communication completely deterministic across a distributed environment and, additionally, relieves communication issues associated with parallel execution of models running rate-monotonically.

The Software Backplane:

- provides fine-grained message communication between parallel threads
- always sends complete (data homogeneous) messages
- time-buffers messages between threads. Receiving threads get time-consistent messages based on their execution rates.
- automatically determines the most efficient hardware path to send messages (local, shared, reflective memory, or LAN). Path is transparent to applications.
- requires a model-to-model interface definition. Applications define Ada types for messages in interface definition packages. Senders and receivers of the message use the type to know the format of the message.
- provides non-queued and queued messages (1 sender -> many receivers OR many senders -> 1 receiver).

Object-Oriented, Ada Structural Definition

SVM's Ada architecture is a recent evolution of Ada architectures for flight simulators as shown in figure 7. [4,5,6]. This architecture uses code templates to define the structure of a class and the structure of a thread (partition). Domain-specific interfaces (like the interface to SVM) are also specified. The architecture is documented in CAE-Link's Software Architecture Standard [1].

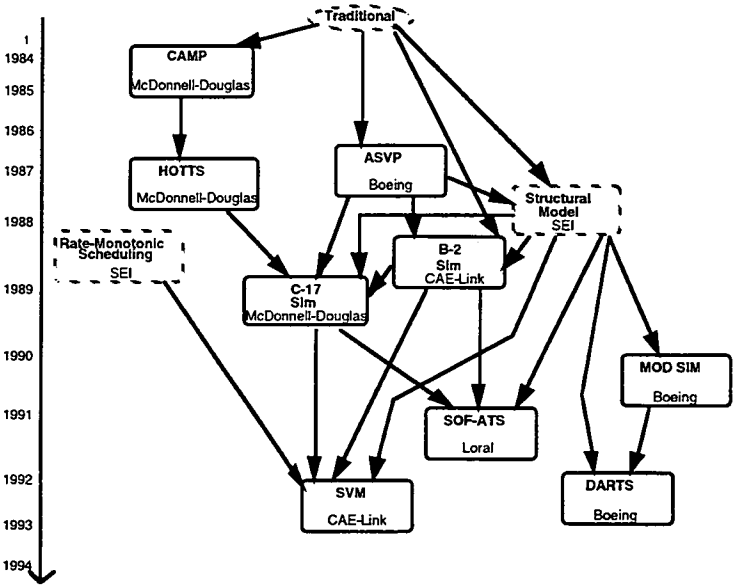


Figure 7 - Ada Structural Architecture Lineage

As an example, figure 8 shows a schematic of a real-world hydraulics system. Using an object-oriented software design approach, a desired object structure is shown in figure 9. Note the one-for-one mapping of real-world components to software classes. The classes are created in a standard form as defined by the Software Architecture Standard and are built up as shown in figure 10. The top-level package (like the "Hydraulics Control System") makes the instances (objects) of the classes and creates an executive thread as described earlier. Partition to partition messages are sent via the software backplane.

SVM's Ada Architecture:

- defines how object-oriented design is applied on SSTF
- documents Ada architecture for classes, partitions, and system interfaces
- promotes consistent design across a project
- defines systems as a build-up of classes. Generally applications are constructed of layers of classes instanced into a partition.
- defines classes as the normal object-oriented class structures defined by C++ or Ada
- defines partitions that hold the instantiation of the thread executive (a partition is a thread or active object)
- specifies partition communication via the software backplane
- defines the partition as the unit of distribution.

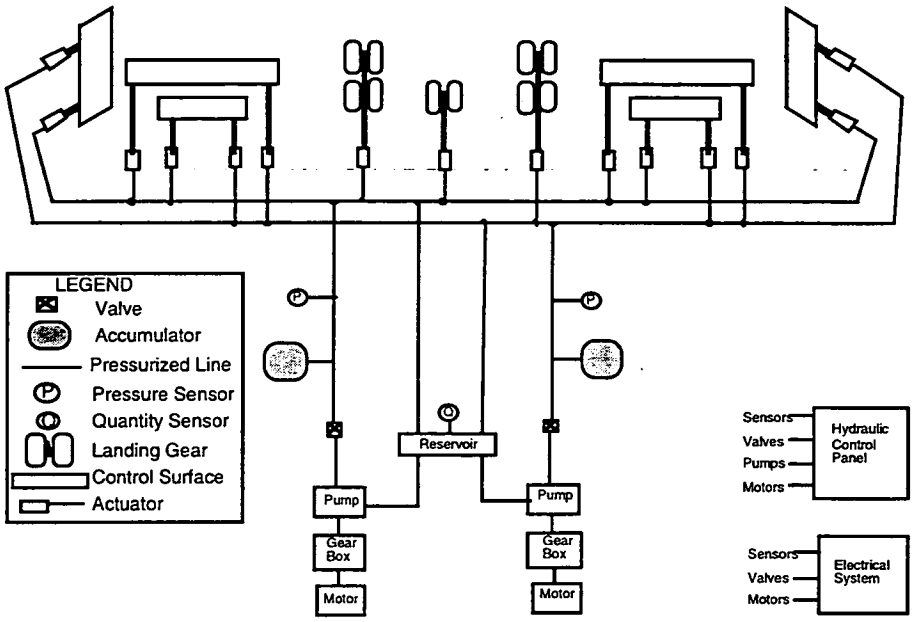


Figure 8 - Real-World Hydraulics System Schematic

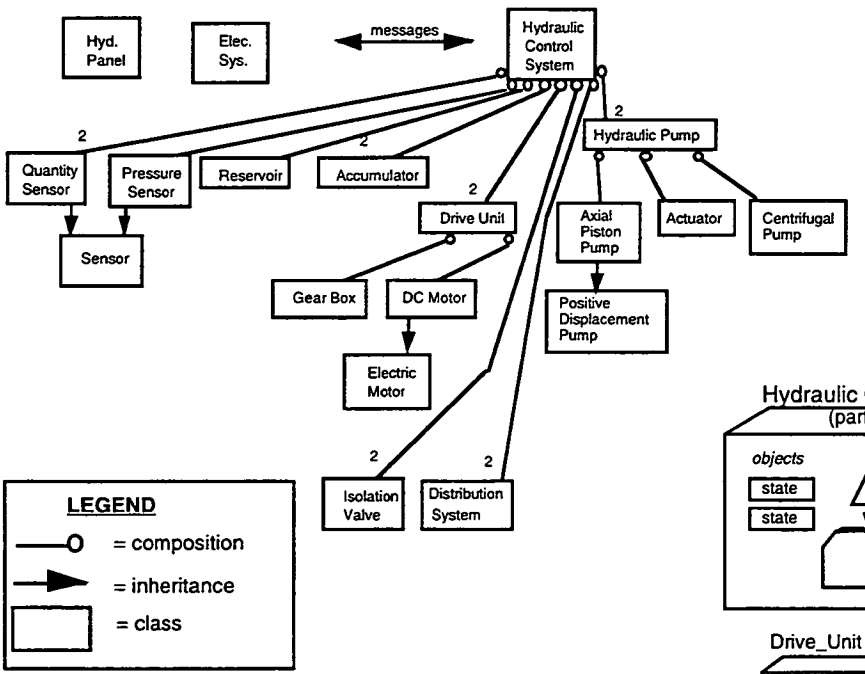


Figure 9 - Object-Oriented Decomposition

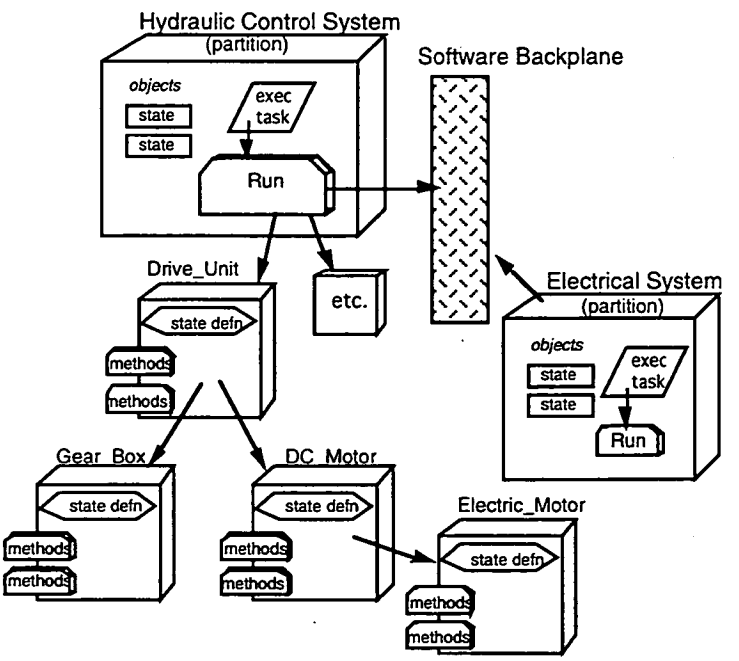


Figure 10 - Ada Structural Architecture

SVM AT THE JOHNSON SPACE CENTER

The Mission Operations Directorate at the Johnson Space Center has extensive plans for SVM. The system's portability and flexibility make it ideal for a wide range of training simulators. This section will discuss SVM's use in the SSTF--its original target--then plans for migration into other simulators, such as the Space Station's Part-Task Trainer.

The primary purpose for development of SVM has been for use in the SSTF. This facility is a high-fidelity simulator used to train astronauts and flight controllers in the operation of the International Space Station. In addition to training in system familiarization, monitoring, and commanding, the facility will also be used for verification of real-world procedures in a low-risk environment. SVM is at the core of this facility, providing the environment for execution, communication, timing, and control of software models.

CAE-Link has adopted an incremental development approach for the SSTF. This approach consists of six-month development increments where a subset of the final capabilities is designed, coded, reviewed, and tested during this relatively short period of time. This set of capabilities is demonstrated to NASA at the end of each increment. This approach has several benefits. The users of the system are provided frequent and early reviews of development progress, giving opportunity for input as the various requirements are implemented. Also, a level of confidence in the functionality and viability of the product is established with every increment as the capabilities are added to previous increments.

Currently, as the third development increment is being completed, SVM's core capabilities are in place and have been through acceptance tests and demonstrated to NASA. SVM can now time and execute software models, provide communication between models, provide model parameters for display at remote workstations, and accept and respond to moding commands (initialize, run, freeze, datastore, and terminate) from remote workstations.

The target SSTF host computer is a Silicon Graphics, Inc. (SGI) Challenge series. However, portability has been kept at a high priority during development and this has already paid dividends. A version of SVM is currently executing in the SSTF development environment, which is Rational's APEX executing on a Solbourne workstation. In addition, SVM has been ported to VAX and Apollo workstations with minimal changes to the source code.

For the next two years, two main efforts will be under way to accomplish a delivery of the SSTF. Space Station software models will continue to be developed in-house and imported from external sources to run under SVM. Also, SVM will be extended to provide the unique performance requirements and interfaces to remote platforms needed for this particular facility.

Because of SVM's flexibility and scalability for large or small simulators, SVM is being considered for other simulators at the Johnson Space Center. One example is the Part-Task Trainer (PTT).

The PTT is a small, single-system trainer designed to train astronauts and flight controllers in a one-on-one, instructor/student environment. Although the PTT has been delivered, it was determined that savings could be achieved by converting to a common software architecture with the SSTF. SVM was ported from the SSTF's SGI Challenge Series computer to the PTT's smaller SGI Power Series. The PTT has different models at a lower fidelity than the SSTF, but now shares a common software environment, resulting in savings in sustaining engineering costs.

Other possibilities for SVM application are being considered. One possibility under consideration is the current operational simulator for the Space Shuttle, the Shuttle Mission Training Facility. This is another large, multi-system simulator that runs in real-time. Another consideration has been the software verification facility for the Space Station. In this facility, real-world flight software is tested in a simulated Station environment to obtain a level of confidence in the software. This simulated environment must respond in real-time in order to give an accurate signature of the system under question.

SVM APPLIED TO PRIVATE INDUSTRY

Real-time simulation occurs today where operating or testing a real-world system is dangerous or cost prohibitive. Airline pilots become familiar with the behavior of an airplane before they ever take the controls of the real plane. Astronauts fly the simulated Space Shuttle before ever going to the launch pad. Software for an extensive system is run in a simulated environment for test and verification. A few applications for SVM outside the space industry include:

- Commercial Airlines
- Military Aircraft
- High-speed Passenger Trains
- Real-Time Software System Test/Verification
- Virtual Reality
- Any Real-Time Simulation

REFERENCES

[1] Software Architecture Standard for the Space Station Training Facility, NAS9-18181, CAE-Link, April 1994

[2] Object Specification for the Real-Time System Software for the Space Station Training Facility, NAS9-18181, CAE-Link, March 1994.

[3] The Handbook of Real-Time Systems Analysis: Based on the Principles of Rate Monotonic Analysis, Software Engineering Institute, July 1992

[4] State of the Practice Research Paper on Simulations, CAE-Link, April 1991

[5] An OOD Paradigm for Flight Simulators, K. Lee, M. Rissman, R. D'Lppolito, C. Plinta, R. Van Scoy, Technical Report CMU/SEI-87-TR-43, Software Engineering Institute, December 1987

[6] Ada Structural Model and Ada Design Methodology for the C-17A Weapons Systems Trainer, McDonnell Douglas Training Systems, February 15, 1990.

Kenneth Hill is a system manager for the Space Station Training Facility. He has contributed in all aspects of software development projects from requirements definition, development, and sustaining of operational systems. Projects include Parameter Information Interface and Near Real-Time Telemetry for Space Shuttle, and Software Support Environment and Space Station Training Facility for the International Space Station. He holds a B.S. in Applied Mathematical Sciences from Texas A&M University.

Rob Sturtevant leads the SVM effort at CAE-Link. He has 10 years experience in Ada, object-oriented design, real-time executives, and training system applications. He has worked on numerous programs including Common Ada Missile Packages, Harpoon Operator Team Training System, C-17 Weapons System Trainer, and Space Station Training Facility, and has consulted on several programs including the A-12 WST, Cruise Missile Planning System, and Tanker/Transport Training System. He holds a B.S. in Computer Science from Texas A&M University.

ULTRA-HIGH RESOLUTION MINIATURE COLOR CRT FOR VIRTUAL REALITY APPLICATIONS

Bernard K. Vancil
FDE Associates
Beaverton, OR 97007-8739

Edwin G. Wintucky
NASA Lewis Research Center
Cleveland, OH 44135

ABSTRACT

A novel miniature color CRT has been developed for helmet mounted display and virtual reality applications that is also applicable to a variety of other display applications where compactness, high resolution, high brightness, low power consumption and low cost are required or advantageous. Initial performance data are presented for a 2 inch by 1.5 inch prototype that employs a single electron gun, a screen composed of sets of red, green, and blue phosphor lines, and a movable, slotted shadow mask interposed between gun and screen. Color fields are written by moving the shadow mask via piezo-electric actuators inside the tube so as to uncover the desired color. This approach gives comparable resolution but much higher luminance than color shutter tubes of similar size and format. It provides much higher luminance and resolution than beam index tubes. Power consumption is under three watts. Improvements in design that offer enhanced performance and greater compactness are described.

INTRODUCTION

It is well known that present helmet and head mounted displays are inadequate in many ways. They are too dim, consume too much power, cost too much, and -- above all -- have insufficient resolution. Brightness (measured as luminance) is important because a one- or two-inch display must be expanded by an optical system to a size that covers the eye's full field of view. This can mean a 20 - 50 fold increase in projected area and, consequently, a 20 - 50 fold decrease in projected luminance. For virtual reality applications, brightness should reach that of ambient sunlight for outside scenes. Power consumption must be minimized to avoid a temperature rise inside the helmet and to allow its ultimate use in tetherless systems. For many commercial applications, such as virtual reality, cost is also an important constraint. Finally, the present resolution limit of 600 - 700 lines is woefully inadequate when one considers that the resolution of the eye is on the order of 4000 lines.

This paper describes a miniature color CRT utilizing only one electron gun and employing a moving, slotted shadow mask interposed between gun and screen to select colored, light-emitting phosphor stripes [1]. A cross section drawing is shown in Figure 1 and a cutaway drawing is shown in Figure 2. The device was built substantially in accordance with the invention described in our patent *(US No. 5,198,730)* [2]. It is expected

that the approach described therein and presented here is capable of realizing higher resolution and brightness (along with excellent power efficiency and low cost) than present art using other approaches is capable of producing for one to three inch diagonal displays.

Other Display Approaches

The conventional three-electron-gun approach leads to insurmountable difficulties of beam convergence and registration in miniature color displays. Because of these difficulties, a number of single gun alternatives have been introduced.

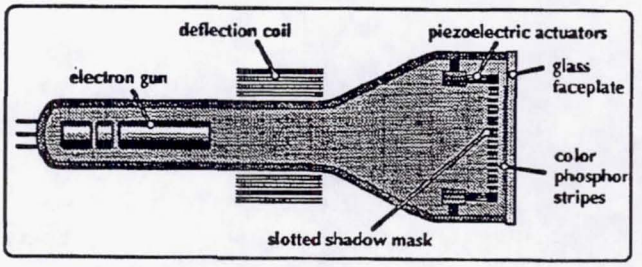


Figure 1: Cross section of moving shadow mask color CRT

Beam Index tubes [3], for example, utilize only one electron gun, wherein the screen includes four phosphor components -- the three primary colors plus one which produces ultra-violet light detected by a sensor. This sensor relays timing data to the gun allowing it to turn on at the correct time to write the correct color phosphor element nearby. This approach leads to a requirement that phosphor stripes must be less than half the width that they otherwise would have been for the same resolution were a shadow mask employed. This is because there must be a dead space on either side of each phosphor stripe for the beam to start and stop on. Otherwise it would write two colors simultaneously. This leads to formidable problems with phosphor deposition technology at high resolution. Also, luminance is substantially reduced, by a factor calculated to be at least 3, compared to the moving shadow mask tube.

Color Shutter tubes [4] also employ one gun and avoid the problem of applying stripes of different color phosphor material by covering the entire screen with a single phosphor that emits three primary colors. This approach has the drawback that the unwanted colors must be filtered out by the color shutter. In doing so, this filter attenuates all the light. Light output is reduced by 90% or more. The luminance is calculated to be only about 7% that of the moving mask tube. Consequently, these displays have inadequate luminance for many applications.

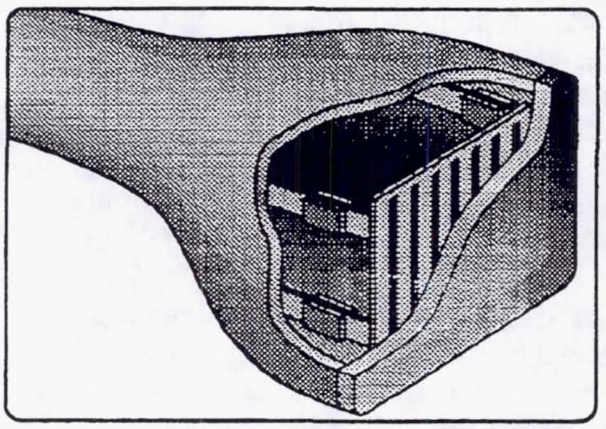


Figure 2: Cutaway drawing of moving shadow mask color CRT

The main problems with active matrix flat panel displays are low brightness and high cost, both of which become more severe as pixel size is reduced. In fact this is an inherent physical drawback of the AMLCD, where the light transmittance goes rapidly to zero for pixel sizes below 60 μm and disappears below 30 μm, even with the most advanced technology, because of the

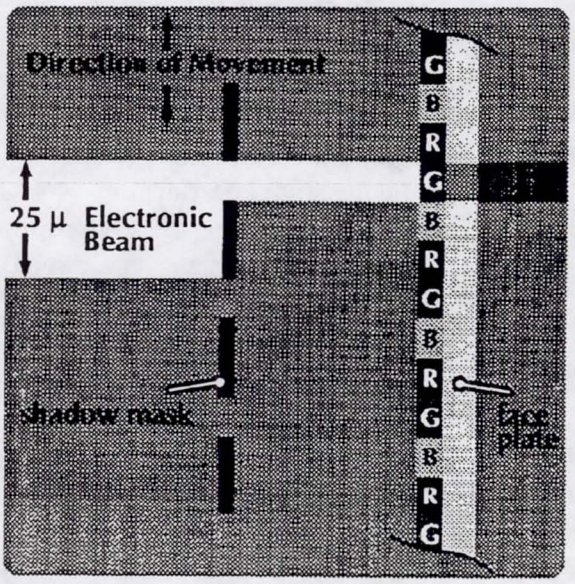


Figure 3: Electron beam/shadow mask/face plate region (top view)

through the shadow mask openings. During operation of the CRT, the shadow mask openings can then be precisely aligned over all the phosphor stripes of a given color, while the other color stripes are covered by opaque regions of the mask. So, for example, alignment can be effected with the green stripes by applying the correct voltages to the piezo-electric actuators. The electron beam then writes the entire green color field. Then new voltages are applied to the actuators and the mask is shifted so as to uncover the red stripes. Then, the red field is written. In similar fashion the blue field is written. The beam must be slightly larger than the mask pitch to avoid moire fringing [10]. The beam is swept in a direction perpendicular to the slots in the mask.

Shadow Mask/Color Screen Construction

The prototype tube described here, built for a proof-of-concept demonstration of the moving shadow mask principle, has a 640 x 480 line resolution in a 2-inch horizontal by 1.5-inch vertical display. The phosphor lines were made one mil wide. The slots in the shadow mask are 0.9 mils wide. Consequently, the mask must be able to move one mil from its center position . In addition, there must be enough additional movement that center position misalignments such as skewing and translation errors can be corrected. To correct these, DC offset voltages are applied to the actuators when the tube powers up. A timing diagram for the actuators is shown in Figure 4. The shadow mask is made of 0.8-mil thick electroformed nickel. The bars are a little more than twice as wide as the slots. The mask is tensioned in the direction of the bars over a titanium frame and welded to it. Both mask and screen are flat.

The phosphor lines are deposited using a wet PVA process. Stripes of photo-resist are first developed by passing light through the shadow mask -- the actual mask to be used with the tube. This prevents registration errors due to slight variations between masks. The mask is held fixed in relation to the faceplate by three locator posts attached to the screen. Although the mask is only five mils away from the screen, parallax problems still arise due to the fact that the beam has about a 20° angle relative to screen normal at the corners of the display. Therefore, it is necessary that the light which passes through the mask also have a 20° angle at the corners as well as all intermediate angles down to zero degrees as one approaches the center. To do this, a

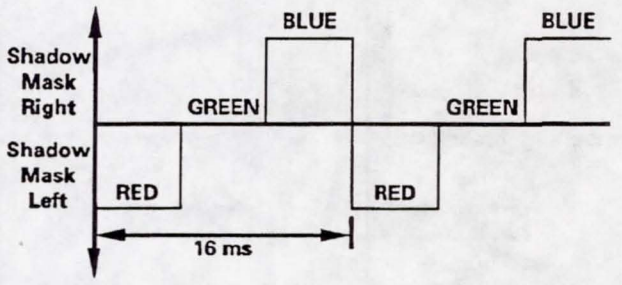


Figure 4: Timing diagram for the actuators

diverging lens is interposed at the correct distance from the screen that gives the correct divergence angle of light. An Oriel lighthouse was used. Once photo-resist stripes are developed to transparency, the mask is removed and a given color phosphor can be deposited along the stripes using a PVA process. The PVA is developed by passing light through the faceplate. The phosphor-loaded PVA is developed only in regions lacking photo-resist.

Adjacent phosphor lines of different color are then deposited by repeating the aforesaid process, except shims are used to offset the shadow mask, first one mil and then two mils away from the locator posts. No black surround was used in the prototype device to enhance contrast.

Piezo-Electric Actuators *(Benders)*

The actuators are constructed in bimorph fashion. That is, a metal center vane, 0.008 inches thick, has PZT ceramic bonded to it on both sides. The bonding material must be capable of withstanding the 400° bake cycle that the tube experiences during pumping. A glass frit formulation with appropriate thermal expansion characteristics was selected to meet this requirement. Although unimorph actuators, on the other hand, (benders with only one ceramic bonded to the metal strip) are simpler, they are very sensitive to temperature changes and hence not suitable. A diagram of one of the bimorph benders is shown in Figure 5. The principle of operation is that when a voltage is applied between the outside electrode of one ceramic and the center vane, the ceramic will either expand or contract depending on whether the imposed electric field is parallel or anti-parallel to the polarization of the ceramic. The entire device will then bow in much the same way as a bimetallic strip when heated. Applying a voltage to one ceramic that causes it to expand and applying a voltage to the other one that causes it to contract will have the effect of increasing the movement. The actuators are wired for series operation. That is, the center vane is always grounded. Then when the voltage applied to one side is positive, it is negative on the other side and vice-versa.

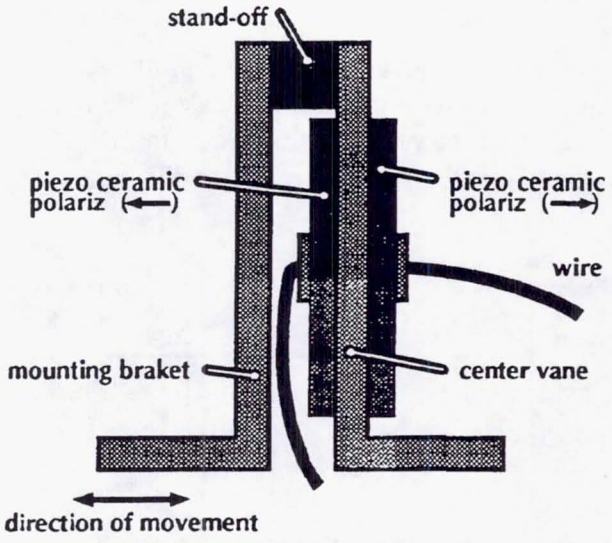


Figure 5: Diagram of bimorph actuator

There are four bimorph actuators -- one at each corner of the mask. These are operated in pairs. The top two can be operated at one phase, amplitude, and offset voltage and the bottom two can be operated at a different phase, amplitude, and offset voltage. In this way, skewing and translation errors between the mask slots and phosphor stripes are correctable. The actuators must be capable of translating the mask by one mil in one direction and one mil in the other. In addition, they must be capable of correcting center position skew and translation errors. They must also have enough margin to compensate for actuator aging -- nearly all of which occurs in the first 24 hours. Also, the actuators must be

existence of one and often two transistors of irreducible size in each cell, . The standard AMLCD has a luminance of 70 nt [5], i.e., 23 fl, whereas the moving mask CRT reported here has a luminance about 30 times higher. Electroluminescent (EL) displays are even worse, where the luminance is quoted at 17 fl [6]. Also EL displays, which operate at low voltages, lack an efficient blue phosphor and hence are not full color.

Potential Applications

As stated in the Introduction, a very promising application for the moving shadow mask miniature color CRT is in helmet or head mounted displays for virtual reality applications. Virtual reality, a system in which the user interacts with a computer-generated environment, is undergoing continuing development for military and a variety of other applications, including entertainment, education and training, architecture, medicine, design, maintenance, manufacturing and marketing [7,8]. The commercial potential for virtual reality is enormous; its realization depends on advances in a number of technologies, one of which is the user /system interface. A very important issue for head mounted displays is the need for user comfort. The moving shadow mask CRT addresses this need not only by means of compactness and low power, but also in high quality viewing (high resolution and brightness).

There are other important commercial display applications where the compactness, high resolution, high brightness, low power consumption, and low cost offered by the moving shadow mask CRT are desirable or advantageous. In particular, the compactness, low power consumption and high resolution/ brightness characteristics make it an excellent candidate for display applications in portable and hand held devices such as TVs, monitors for VCRs, and view finders for camcorders, especially for outdoor use. For comparison, consider the camcorder recently brought to market that uses a 4" wide AMLCD display for both view finder and playback monitor. Although advertised as having super high brightness, the actual luminance is only 75 fl [9]. The resolution is 266 x 200 lines and the pixels are 15 mils wide. Scaling down leads to a further loss of brightness by a factor of two or more. There is clearly an opportunity here for improvement in the viewing quality of the display. The prototype moving mask CRT reported here has a measured luminance of at least 300 fl and a resolution of 640 x 480 in a 2" by 1.5" display. The moving mask approach is feasible for displays up to at least 6" in diagonal. Because of its high resolution and brightness, another possible application is in the rapidly growing area of projection displays.

CONSTRUCTION AND PERFORMANCE OF PROTOTYPE CRT

Our approach also uses a single electron gun. It utilizes a slotted shadow mask mounted very close to the inside of the screen. The shadow mask is fastened to the funnel at each corner via piezo-electric bimorph actuators which translate the mask sideways -- i.e., in a direction perpendicular to the slots in the mask. Beneath the slots are triads of color phosphor stripes. A top view of the shadow mask-face plate region is shown in Figure 3. The width of each mask opening is approximately the width of the phosphor line beneath it. In fact, the lines are created by using a photographic technique in which light is passed

repolarized *in situ* after tube processing since the 400° C processing temperature exceeds the Curie temperature of the PZT material. Although this repolarization usually results in some degradation of bender performance, enough margin can be supplied to overcome this.

Other Tube Features

Table I shows the basic tube characteristics. The gun is a standard Southwest Vacuum TV gun suitable for 13mm neck installation, which was modified at FDE to allow

Table I: Basic tube characteristics

1.	Display type:	field sequential RGB
2.	Display size:	horizontal 2 inches, vertical 1.5 inches
3.	Focus:	electrostatic bipotential
4.	Resolution:	640 x 480
5.	Video rate:	57 MHz
6.	Horizontal scan rate:	94 KHz
7.	Vertical scan rate:	180 Hz
8.	Shadow mask translation frequency:	60 Hz
9.	Actuator voltage:	max 150 V
10.	Screen voltage:	12 kV
11.	Screen current:	40 μA nominal
12:	G2 voltage:	500 V
13.	G1 cut-off voltage:	38 V
14.	Power consumption:	< 3 W @ 300 fl screen luminance
15.	Horizontal deflection:	20 degrees @ 1.2 A
16.	Phosphor types:	P43 (green), P22R (red), P11 (blue)

bi-potential focusing. To do this, an enlarged machined cup was mounted on the G5 electrode, which then formed one half of the bi-potential lens. A gold band on the inside of the neck tubulation then formed the other half. This produced a longer throw between cross-over and focus plane and, consequently, lowered the magnification of the gun. Also, an astigmatism element was added to compensate for misalignments. Figure 6 shows a plot of spot size vs screen current at the center of the screen for 10 kV operation. At low drives of 10-20 μA, 3.0-mil spot sizes were observed, which grew to 7 mils at 55-60 μA. Our requirement was for a three-mil spot size at high drive. A carefully designed gun with more precise alignments should remedy the situation. Also just raising screen potential to 12 kV by itself will give a smaller spot size -- about four mils. Another method that will also reduce the spot size, which is now incorporated and being tested in present versions of the miniature CRT, is the use of a lenthened bipotential focus barrel on the electron gun. The estimated reduction in spot size is shown by the dashed line in Figure 6.

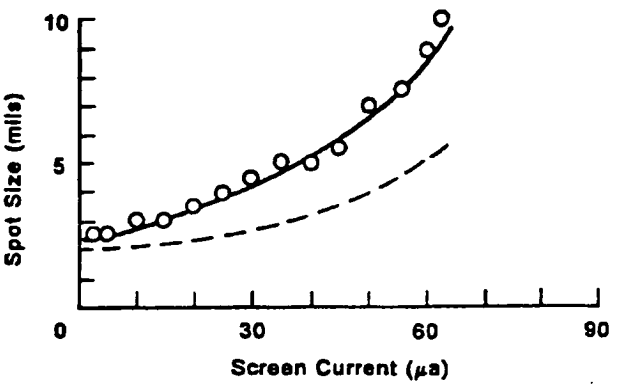


Figure 6: Spot size at center of screen vs screen current at 10 kV. Dashed line shows estimated reduced spot size using lengthened bipotential focus barrel.

Luminous Efficiency

In the present tube configuration, the shadow mask intercepts two-thirds of the beam current. The other third impinges on the screen. Figure 7 shows plots of the luminance vs screen current for the green, red, and blue phosphors. These measurements were taken using a Minolta Auto-Spot digital luminance meter. The green phosphor is P43 *(Gd_2O_2S)*, the red phosphor is P22R *(Y_2O_3)*, and the blue phosphor is P11 *(ZnS)*. The screen voltage for these tests was 10 kV. At 40 μa of screen current, we observed 660 fl for green. Since there are 144 fl in one lumen/inch^2, the green is providing 4.576 lumens/inch^2. Power flux is 0.148 watts/inch^2. Therefore, the efficiency for the green is 31 lumens/watt. Published figures under optimum conditions run as high as 40. Lumens/inch^2 for the red phosphor is 1.61 based on 232 fl, which translates to 10.9 lumens/watt for 40 μA of screen current. Published figures go up to 18. The blue phosphor emits 218 fl at 40 μa, which is 1.5 lumens/inch^2. Its efficiency is then 8.11 lumens/watt and published figures are about 11. Therefore, the average measured efficiency for red-green-blue is 16.0 lumens/watt. The power at 40 μA is 0.148 watts/inch^2 or 2.467 lumens/inch^2, which equals 355 foot-lamberts. Another way to look at the data is as follows: We can achieve our bench mark of 300 fl for green with only 18 μA of beam current. To achieve it for red will require 50 μA and for blue it will require 60 μA. However, to produce white light, blue and red are required in smaller proportions than the green. Even less blue intensity can be used if we switch to P22 blue phosphor.

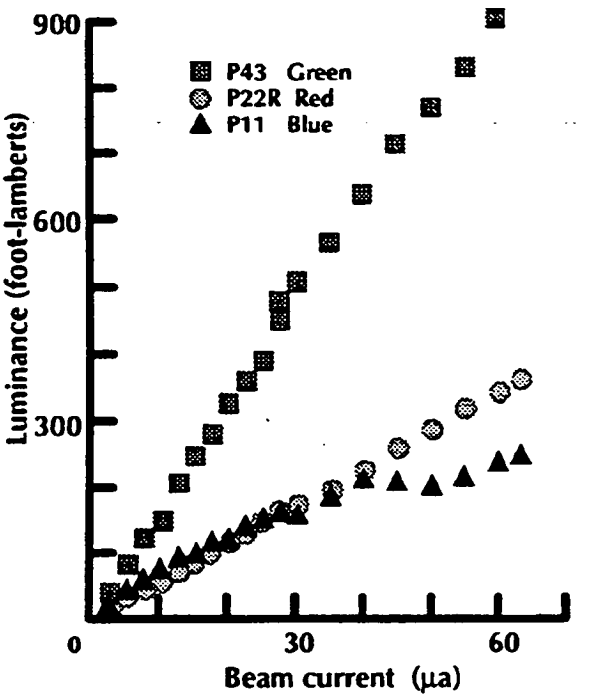


Figure 7: Luminance vs beam current

Power Consumption

The screen current required to achieve at least 300 fl is 30 - 40 μA. This means that the beam current (mask current + screen current) is about 120 μA. The power dissipation in the mask at 10 kV is then 0.8 watts for a totally white full brightness screen. This is seldom required and so a derating factor of 0.5 can be applied to this. This reduces the mask the beam is about 0.6 watts after derating. The cathode heater power is 0.4 watts at this power level. The horizontal yoke current is about 1.29 amps at 17° deflection and 10 kV. Since yoke current is a linear ramp, this current averages about 0.65 amps. A yoke resistance of 0.81 ohms gives a power dissipation of 0.53 watts. If this is increased by 15% for high frequency effects, we get 0.61 watts dissipation for the horizontal yoke. If we add to this a vertical dissipation of one watt, then the total power dissipation for the tube is 2.61 watts.

PLANNED IMPROVEMENTS

The "proof-of-concept" miniature CRT described here was fabricated using mainly techniques and materials commonly used in CRT manufacturing. Some of the planned improvements, on the other hand, will require alternative and sometimes more advanced methods. One of the main goals of this effort is to achieve a higher resolution of 1000 lines/inch in the horizontal direction, which in a 2" x 1.5" display would require 2000 x 1500 lines. Making the screen at present uses techniques that are readily available in television tube manufacture -- namely, photo deposition of phosphor stripes by passing light through a shadow mask. The ultra-high resolution of 1000 lines per inch will require that the phosphor lines be only eight microns wide with an extremely small size of allowable defects. This means a change from the conventional method of phosphor deposition. Either a cathephoretic deposition method or physical vapor deposition may be used.

Another requirement imposed by the ultra-high resolution is that the shadow mask will need to have slots that are only seven microns wide. This is close to the limit of what can be done by electro-forming nickel and is beyond what can be done using chem-milling techniques. There is also the possibility that nickel having such small dimensions will warp when subjected to high energy beam interception. In that case molybdenum is a better choice for the mask material. Some other method, such as plasma etching, will then be required to make the slots in the shadow mask.

To maintain high resolution and luminance, a smaller beam spot size will also be necessary. Since the periodicity of the color stripes is on the order of 25 μm, the optimum spot size diameter would be slightly larger than one mil. In addition using the lengthened bipotential focus barrel on the electron gun mentioned earlier, further reduction in spot size can be achieved by raising the cathode loading. This will be enabled without shortening the cathode life by the use of a new type of oxide coated cathode developed under an SBIR contract with NASA Lewis Research Center. Figure 8 shows emission current vs time test results over a 6000 hour period for both the improved oxide cathode and a conventional oxide cathode tested in the same close spaced diode configuration simultaneously in the same life test vehicle. Except for some chemical additives to the coating in the case of the improved oxide cathode, the same triple carbonate mix is used for both. The cathodes were operated

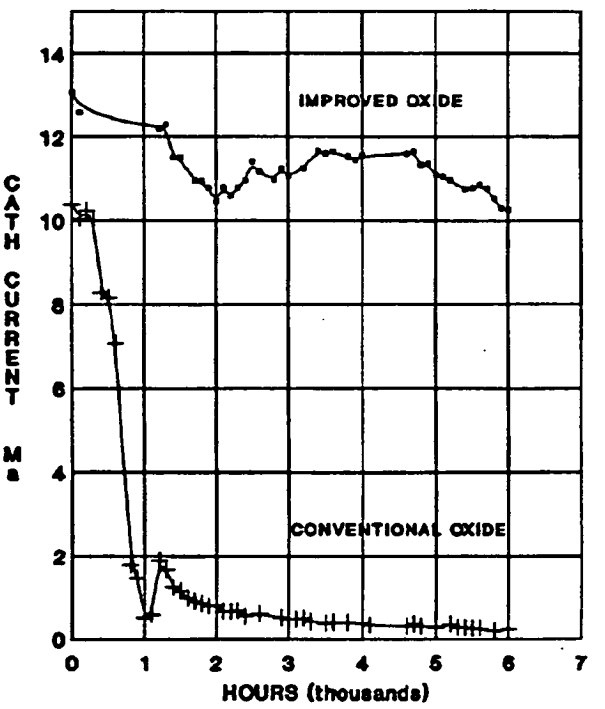


Figure 8: Diode life test results for improved and conventional oxide cathodes - DC operation

dc at a loading of 1 A/cm^2 at the center of the cathode and a peak loading calculated to be about 2 A/cm^2 at the edges.

The additional need for more precise alignment of the electron gun electrodes and alignment of the gun relative to the neck of the tube can be readily accomplished during electron gun/tube fabrication and assembly using established practices.

Table II: Projected tube characteristics

1.	Display type:	field sequential RGB
2.	Display size:	horizontal 2 inches, vertical 1.5 inches
3.	Focus:	electrostatic bipotential
4.	Resolution:	2000 x 1500
5.	Video rate:	563 MHz
6.	Horizontal scan rate:	281 KHz
7.	Vertical scan rate:	180 Hz
8.	Shadow mask translation frequency:	30 Hz
9.	Actuator voltage:	max 150 V
10.	Screen voltage:	12 kV
11.	Screen current:	40 μA nominal
12:	G2 voltage:	500 V
13.	G1 cut-off voltage:	60 V
14.	Power consumption:	< 1.5 W @ 300 fl screen luminance
15.	Horizontal deflection:	20 degrees (electrostatic deflection)
16.	Phosphor types:	P43 (green), P22R (red), P22 (blue)

The greater density of lines will require significantly higher horizontal deflection rates (281 KHz) and video rates (563 MHz). The faster electronic circuitry required is available and, in fact, such high video rates have been demonstrated in the CRT industry. A benefit of the narrower lines is a smaller displacement of the shadow mask and hence a reduced demand on the piezo-electric actuators.

Another planned improvement is the use of electrostatic beam deflection, which is very feasible because of the low scan angle. Elimination of the yoke means a substantial reduction in electrical power requirements (< 1.5 W) and also a significant reduction in size and weight of the tube. The much lower tube input power also means a big gain in basic efficiency of converting power input to light output, which is projected to at least double to about 4 lumens/watt for a 2" x 1.5" CRT.

A very valuable and quite significant planned improvement is the novel combination of beam indexing with the moving shadow mask approach. A major advantage offered by beam indexing is the elimination of moire fringe patterns, which commonly occur with shadow mask tubes [10] and result from the random phase between the video on-off signals and the periodicity of the mask. In conventional beam index tubes, the video on-off signals are timed so that the beam turn-on is always centered over the desired color element. Beam indexing heretofore has been incompatible with shadow mask tubes. This is not the case for the moving shadow mask tube described in this paper, in which video modulation of the beam can be indexed to the periodicity of the mask. This arrangement uses an ultraviolet light producing phosphor deposited on the back of the shadow mask and a photo-detector to

transmit a signal that triggers the video amplifier. Consequently, the video on-off signals are phase locked to the periodicity of the mask. Therefore, a much smaller beam size, which without beam indexing would need to be slightly larger than the mask pitch, can be used without introducing moire fringe patterns.

The beam index approach is wholly compatible with the goal to achieve the ultra-high resolution described above. In addition to the elimination of moire fringe patterns, a significant contribution to improving the resolution comes from the much smaller beam spot size permitted by beam indexing. One obvious benefit is the lesser amount of beam interception by the mask, which leads to less heating of the mask and consequent potential to warp, thereby improving color purity. This may even allow the continued use of nickel for the mask material. One scheme is to make the beam about as wide as the mask slot, turn it on as the leading edge enters the slot and turn it off as the trailing edge exits the slot. Then about the same amount of charge impinges on the mask as on the screen during the writing of a given element. Consequently, only 50% of the beam energy is transferred to the mask. This is a significant improvement over the 70% or more going to the mask in the case of the present moving shadow mask tubes and the 85% or more going to the shadow mask in conventional three gun tubes. Some reduction in electrical power consumption is also expected.

The projected characteristics for the tube embodying the above planned improvements are summarized in Table II.

ACKNOWLEDGMENTS

This work was supported under a Phase II SBIR contract, NAS3-26395, with the Electron Beam Technology Branch at the NASA Lewis Research Center, Cleveland, Ohio.

REFERENCES

1. Vancil, B., "A Moving Shadow Mask Color CRT for Miniature Displays," *SID Digest of Technical Papers*, Vol XXV, 1994, pp 393-396.
2. Vancil, B., "Color Display Tube", *U S Patent No. 5,198,730*, March 30, 1990.
3. Doyeux H. and House W.R., "Beam Index Cathode Ray Tubes", *Information Display*, January 1990, pp 12-15.
4. Bos P., Buzok T. and Vatue R., "A Full-Color Field-Sequential Color Display", *Proceeding of the SID*, Vol 26/2, 1985, pp 157-161.
5. Hijikigawa, M., "Future Prospects for TFT-LCDs," *SID Digest of Technical Papers*, Vol XXV, 1994, pp 165-166.
6. Werner, K., *Information Display*, September 1994, p 29.
7. Sheridan, T.B. and Zeltzer, D., "Virtual Reality Check," *Technology Review*, October 1993, pp 20-28.
8. Rheingold, H., *Virtual Reality*, Summit Books, New York, 1991.
9. Okano, Y., "Viewcam: A Flat-Panel Display for Viewfinder Applications," *SID Digest of Technical Papers*, Vol XXV, 1994 pp 177-178.
10. Shiramatsu N. and Inoue, A., "Removing Moire Patterns from Shadow-Mask CRTs," *Information Display*, June 1994, pp 12-16.

INTEGRATION OF SIMULATION WITH FIELDED EQUIPMENT USING DIS

Phil Landweer
BDM Federal
Albuquerque, NM 87106

ABSTRACT

A unique simulation was conducted in December of 1993 at the Depth and Simultaneous Attack Battle Lab in Ft. Sill, Oklahoma. A simulated fire support scenario with tanks, infantry fighting vehicles, artillery units, counter-battery radars, and associated command and control elements was provided by a Computer Generated Force, or CGF. A tactical situation display showed the locations of all combatants, as well as activities of interest such as detections, weapon firings, detonations, and communications as the simulated battle progressed in real-time. A DIS-compliant interface allowed the CGF to interact with actual and virtual fire support equipment, both sending and receiving Protocol Data Units (PDUs) to a variety of systems. These systems included a Digital Message Device at the simulated Fire Support Element, Forward Entry Devices at the forward observer and fire support team, low-cost computer units serving as a Fire Direction System or Fire Direction Data Manager interface for the Fire Direction Center and Multiple Launch Rocket System (MLRS) Battalion, respectively, and an MLRS Fire Control Panel Trainer. Thus, a seamless simulation was provided between constructive, virtual, and live simulations. The U.S. Army is advertising this as the first successful use of DIS protocols to integrate fielded equipment with each other, a simulated battle, and a training device without modifying the fielded equipment.

APPLICATION OVERVIEW

This project was conducted at the Depth and Simultaneous Attack Battle Lab (D&SA BL). The U.S. Army uses its Battle Labs to quickly investigate the utility of candidate systems, architectures, and tactics. The D&SA BL focuses on those systems which can be used for attacking the enemy from long distances or in a coordinated fashion, such as fire support systems.

The purpose of this project was two-fold. First, the Army Research Laboratory (ARL) Ft. Sill Field Element, which sponsored this effort, wanted to investigate how mission performance improved for beginning Artillery School students. Each student used a Multiple Launch Rocket System (MLRS) Fire Control Panel Trainer (FCPT) to execute Call for Fire missions within a simulated battle. Each student participated in the battle three times, and the timeliness with which the fire mission was executed wasmeasured. The other purpose of the project was to determine how well Distributed Interactive Simulation (DIS) Protocol Data Units (PDUs) could be used to integrate fielded equipment with each other as well as with constructive and virtual simulations.

CGF DESCRIPTION

The CGF system used was CIMUL8™/SPECT8™/DISIP8, a commercial off-the-shelf (COTS) software product. CIMUL8 is the simulation engine, and models tactically representative behaviors based upon user inputs. CIMUL8 also has a self-contained pre-processor for building up units and scenarios, as well as a post-processor for analyzing battle outcomes. SPECT8 is a graphical display system, and may be used to preview, replay, or watch a CIMUL8 run as it progresses. Finally, DISIP8 is a DIS-compliant interface used to both send and receive PDUs between CIMUL8 and other assets. SPECT8 can also be used to display the occurrence and effects of received PDUs.

SCENARIO DESCRIPTIONS

A ground combat scenario was simulated for this project. The "Blue" Forces (BLUFOR) consisted of an M1 tank platoon accompanied by an M2 Infantry Fighting Vehicle and two Improved TOW Vehicles (ITVs). Fire Support units including a Forward Observer (FO), Fire Support Team (FIST), Fire Direction Center (FDC), Fire Support Element(FSE) with a TPQ-36 fire-finder radar, MLRS Battalion, and MLRS Self-Propelled Loader Launcher (SPLL) were in direct support. The Opposing Force (OPFOR) had a Motorized Rifle Company with BMPs and T-80 tanks, a Self-Propelled Howitzer (SPH) Battery, and associated FDC.

Depending on the particular configuration used, these combatants were simulated using some combination of CIMUL8 (a constructive simulation), the FCPT (a virtual simulation), and fielded equipment (live simulations). The fielded equipment included Field Entry Devices (FEDs), a Digital Message Device (DMD), Low-cost Computer Units (LCUs), and a Fire Direction Data Manager (FDDM).

One interesting facet of this project was discovered upon arriving at the D&SA BL three days before the first demonstration was to occur. Initially, the scenario within CIMUL8 was located within northern Europe. Unfortunately, the FCPT could only execute fire missions using coordinates from the Ft. Sill firing range. So, the entire simulated battle was "moved" from a European battlefield to the Ft. Sill area. This translation was accomplished in a single afternoon, and included repositioning all combatants into realistic positions and ensuring that the required interactions would occur for the project.

DIS PROTOCOLS USED

DIS 2.0.3 PDUs were used to interface the simulations with each other. Specifically, Entity State, Firing, Detonation, Transmit, and Signal PDUs were used. Transmit and Signal PDUs were used to send TACFIRE messages between the various assets. A PC-based DIS interface developed by CAE-Link was used to transform tactical communications into DIS PDUs and vice versa. In this manner, the fielded equipment performed just as it would in a real combat environment.

ASSET CONFIGURATIONS

Four different configurations of the basic ground combat scenario have been used so far at the D&SA BL. Each of these will be described in turn. First, the aspects which are common to each scenario will be described.

Common Aspects

CIMUL8 simulated the movement, signatures, sensing, command and control, communications, engagements, firing, and lethality effects of all OPFOR units as well as the BLUFOR tanks, IFV, and ITVs. Whenever Transmit and Signal PDUs were generated by the live or virtual simulations, SPECT8 would display these events as green starbursts around the transmitting unit. This allowed everyone to watch the information flows as the battle progressed.

First Situation: FO, FIST, FDC, and FCPT

The FO's viewer was modeled within CIMUL8. As the OPFOR targets came into view, the targets were visually acquiredwithin the simulation. SPECT8 was used to display this event by placing a large diamond around the acquired targets on the tactical situation display. This served as the triggering event for an operator to enter the target coordinates into the FO FED and send an FR Grid TACFIRE message to the FIST FED. This message was actually transmitted as a set of

Transmit and Signal PDUs via Ethernet. Upon receipt of the message, the FIST operator would then pass the target coordinates on to the FDC. Another TACFIRE message was sent from the FIST FED to the FDC LCU, which was configured as a Fire Direction System (FDS). The FDC operator then generated a fire mission for the MLRS, and sent a Call for Fire message to the FCPT. The MLRS operator would then execute the mission, and fire upon the OPFOR within the simulated environment. Fire and Detonation PDUs were generated by the FCPT, which would damage or destroy the targets within CIMUL8.

Second Situation: DMD, FDDM, FDC, and Constructive MLRS

In this set-up, CIMUL8 was again used to start things off. A TPQ-36 fire finder radar was modeled, and could detect the OPFOR howitzers firing. As this happened, a large diamond was displayed around the targets on the SPECT8 screen. The target coordinates were then entered into the DMD by an operator, and a TACFIRE message sent via Transmit and Signal PDUs to the MLRS Battalion's FDDM. Upon receipt of the target by it's LCU, the Battalion would then task the FDC with a firemission. TACFIRE messages were sent between the LCUs. Upon the FDC sending a Call for Fire message to the MLRS, SPECT8 would display this message transmission. This cued an operator to use SPECT8's "Personal Control" capability to execute the fire mission. Thus, a constructive simulation was triggered by fielded equipment to complete the mission. Upon firing, the simulated rockets went to the targets and damaged or destroyed some of the OPFOR howitzers.

Third Situation: FO, FIST, FDDM, FDC, and FCPT

This case was essentially a combination of the previous two. CIMUL8 modeled the acquisition of the OPFOR by the FO, with TACFIRE then being sent from the FO FED to the FIST FED. Then, the FIST would pass on the target to the FDDM. The FDDM would task the FDC, which then passed on the Call for Fire message to the MLRS FCPT. The operator would then execute the mission, which would damage or destroy the OPFOR vehicles.

Fourth Situation: Constructive FR Grid, FDC, and FCPT

This was the set-up used for the actual ARL experiments. The FO and FIST were wholly modeled within CIMUL8, including their acquisition and communication activities. After the FIST received the target, an FR Grid TACFIRE message was generated by CIMUL8, with DISIP8 sending out Transmit and Signal PDUs. The FDC would receive these, with the operator then sending a Call for Fire to the MLRS FCPT. The student would then execute the mission, with theresults of the firing effecting the overall battle outcome. All relevant data was captured within CIMUL8. This allowed CIMUL8's post-processor to measure the timeliness of the students' actions for quantitative analysis. SPECT8 could also be used immediately after each trial to replay the battle, thereby providing immediate feedback to the student.

FUTURE APPLICATIONS

This seamless simulation technology may now be used for a variety of purposes. With additional FEDs, LCUs, and FCPTs, a large training exercise could be conducted within the U.S. Army Field Artillery School. Another potential application would be to integrate National Guard and Army Reserve MLRS units with an FDDM. The Guardsmen and Reservists could train at their home locations, with the overall simulation and FDDM located at Ft. Sill. This is particularly important for realistic training, since MLRS units would receive fire missions from an FDDM during combat, but Guard and Reserve MLRS units don't have FDDMs as part of their equipment. Finally, T-NET could be used as an enabling technology to link geographically distant equipment with each other, with DIS PDUs being passed from one location to the other via commercial satellite communications. Such technology could directly support Distance Learning projects.

LESSONS LEARNED

Probably the most challenging part of this project was getting all of the TACFIRE messages right so that the fielded equipment would operate properly. The Signal PDU is sufficiently flexible to contain any type of TACFIRE message. However, certain bytes of the FR Grid and Call for Fire formats are very critical! Also, the timing of TACFIRE message transmissions and acknowledgements must be closely adhered to if live simulations are to be integrated with each other and constructive and virtual simulations.

BIOGRAPHY

Phil Landweer works for BDM Federal, Inc. as the manager for Advanced Simulation. He has used digital modeling and distributed simulation for concept exploration, requirements analysis, pre-test analysis, test planning, test and evaluation, and tactics development to support a variety of government and industry organizations. His specific areas of interest include object-oriented simulation, functional modeling, and integrated computer graphics. Prior to joining BDM, he was an analyst at the Air Force Operational Test and Evaluation Center, where he used modeling and simulation to support a variety of Operational Test and Evaluation programs.

Environmental Technology

RAPID OPTICAL SCREENING TOOL - COMMERCIALIZATION OF AIR FORCE DEVELOPED TUNABLE LASER SPECTROMETER FOR ENVIRONMENTAL CHARACTERIZATION AND MONITORING

Bruce J. Nielsen
Armstrong Laboratory, Environics Directorate
Tyndall AFB, FL 32403-5323

Dr Greg Gillispie
Dakota Technologies, Inc.
Fargo, ND 58102-1809

David A. Bohne
Unisys Corp., Environmental Systems
St. Paul, MN 55164-0525

ABSTRACT

The cost of characterizing and monitoring U.S. government hazardous waste sites could exceed $100 billion utilizing traditional methods and technology. New sensor technologies are being developed to meet the nation's environmental remediation and compliance programs. In 1993, Armstrong Laboratory and Unisys Corporation signed a Cooperative Research and Development Agreement (CRDA) to commercialize fiber optic laser-induced fluorescence technology that had been developed with Air Force at North Dakota State University (NDSU). A consortia consisting of the CRDA partners, Dakota Technologies Inc., and NDSU submitted a proposal to the Advanced Research Projects Agency, Technology Reinvestment Project and won an award funding the commercialization. The result, Rapid Optical Screening Tool or ROST[1] is a state-of-the-art laser spectroscopy system for analysis of aromatic hydrocarbon-contaminated soil and groundwater. With ROST, environmental investigators are able to find, classify, and map the distribution of many hazardous chemicals in the field instead of waiting for reports to come back from the analytical laboratory. The Tri-Service research and development program leading to prototype laser spectrometers is summarized along with the technology transition. Results from laboratory and field demonstrations will illustrate the current system performance.

INTRODUCTION

The vision of U.S. Department of Defense (DOD), Tri-Service (Air Force, Army and Navy) scientists is about to become reality as a partnership between DOD, academia, and private industry evolves into a combined technology that can save millions of dollars in long-term hazardous waste site cleanup costs. The DOD has about 20,000 contaminated sites, most of which will require further characterization and many may require monitoring for 20 or more years, costing millions of dollars per year. The cost of site characterization and monitoring has traditionally been one-third or more of the total remediation costs.

Traditional methods employed during environmental site characterizations are time-consuming , yet often lead to insufficient or inadequate soil and groundwater data. The typical or

phased approach involves many steps, often widely separated in time, including: investigation design; grid layout; geophysics; soil boring; sampling (soil, soil gas, and groundwater); off-site analysis; and data evaluation. Due to the expense and time involved, sampling programs are usually conservative, typically providing the minimum amount of data necessary to complete the investigation. The understanding of hydrogeology and contaminant distribution is often poor and remediation designs fail. Too many times, steps must be repeated until the extent of contamination is satisfactorily defined. A fix is needed--one that could give a "snapshot" of the type and amount of contamination, help determine method of remediation, monitor the effectiveness of treatment, and indicate when the site is sufficiently "clean."

As a partial solution to the problem, the Tri-Services integrated laser spectroscopy with a cone penetrometer. The combination provides an opportunity to significantly expedite the characterization process by providing in situ, real-time data of both petroleum contaminant distribution and soil hydrogeology. This technology has been field-tested at numerous sites and is now being commercially deployed as a service by Unisys. Ongoing research will extend sensitivities, expand capabilities to detect other contaminants such as solvents, metals, and explosives; and make system operation more user-friendly for operating technicians. The end result is a technology that can significantly reduce the cost of site characterization and monitoring.

BACKGROUND

Development

The Site Characterization and Analysis Penetrometer System (SCAPS), developed jointly by the Tri-Services, has proven to be an effective technology for characterizing contaminated sites. The Tri-Services are cooperating on the development and implementation of cone penetrometers and associated technologies. The Army has provided leadership on developing SCAPS; the Waterways Experiment Station conceived the idea of combining optical measurements with cone penetrometers to determine chemical information about the soil. A patent, entitled "Device for Measuring Reflectance and Fluorescence of In Situ Soil," is now being licensed. SCAPS includes the truck-mounted cone penetrometer; physical and chemical sensors; environmental samplers; data acquisition, analysis, and graphical presentation hardware and software; and probe hole grouting. Unisys holds a sublicense for utilizing the patented technology.

Cone Penetrometry

The typical cone penetrometer is mounted on a 20-ton truck and driven to the site requiring characterization (Figure 1). Using the truck as a reaction mass, the penetrometer hydraulically pushes an instrumented conical rod into the ground to be characterized. The cone is pushed into the subsurface continuously at a rate of 2 centimeters per second. Signals from the cone are conveyed to the surface through cables located within the center of the push rods. The signals are processed by a computer located in the cone penetrometer truck. The cone penetrometer may characterize several aspects of the subsurface, depending on the types of sensors integrated into the penetrometer. Strain gauges measure the forces against the tip and sleeve of the cone tool allowing determination of soil type; i.e., sand, silt, clay, etc.; and stratification. Electrodes on the rod allow measurement of the electrical conductivity of the soil which are indicative of changes in

soil type or moisture and can often indicate the presence of contamination. Other sensors provide additional hydrogeological and chemical information regarding soil and contamination.

Soil, soil gas, and groundwater sampling can be performed using the cone penetrometer. To collect samples, the instrumented cone is removed from the push rods and specially designed sampling tools are attached. The sampling devices are also hydraulically pushed to the desired depth and a sample is collected. The sample is brought to the surface for subsequent analysis in the field or at an off-site laboratory. The cone penetrometer sensors provide information on hydrogeology and contamination; the samplers verify it. The real-time ability to receive and assess monitoring data on-site, without laboratory analysis, is critical. It facilitates decision-making during site investigation projects, while ensuring accurate and efficient completion of site investigations and optimization site remediation. Cone penetrometer technology can also be used to provide baseline data for intrinsic bioremediation modeling studies, to define excavation limits, and to monitor the progress of site remediation. Sampling, monitoring point installation, and many other capabilities exist when deploying a cone penetrometer for environmental investigations.

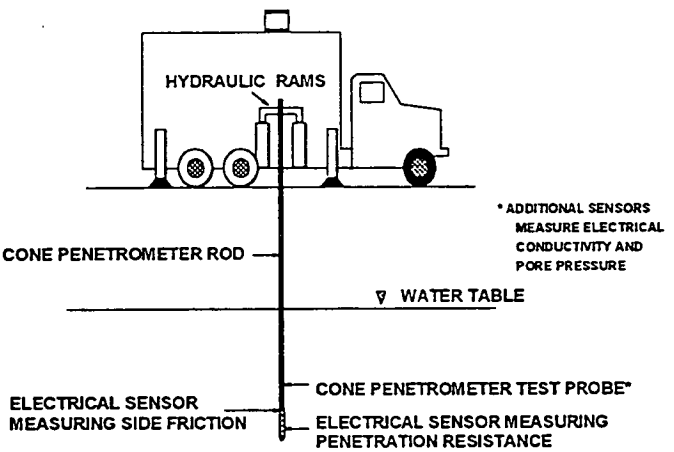


Figure 1. Cone Penetrometer

Laser Spectrometer Systems

One of the key sensors deployed for use with the cone penetrometer involves the use of laser systems to induce fluorescence of fuel products as the cone penetrometer probe is advanced into soils. Laser-Induced Fluorescence (LIF) has been shown to be useful in identifying petroleum contamination such as gasoline and JP-4 jet fuel. The first cone penetrometers fielded by the Army and the Navy made use of a fixed-frequency nitrogen laser developed by the Navy, but are now transitioning to tunable or multiple-wavelength laser systems. Armstrong Laboratory's Environics Directorate, working with North Dakota State University (NDSU), developed a transportable, laser spectrometer system using a Nd:YAG (neodymium:yttrium aluminum garnet) laser to pump a dye laser. The tunable, laser-generated ultraviolet light is transmitted through optical fibers for hazardous waste site characterization and monitoring.

Optical fibers are used to transmit ultraviolet light to monitoring points and return resulting light for the spectroscopic analysis. The detection system consists of either a monochromator, photomultiplier tube, and digital oscilloscope or a gated optical multichannel analyzer. A personal computer is used for system control, automated data collection, and

analysis. The system can identify aromatic hydrocarbons such as benzene, toluene, ethylbenzene, and xylene (BTEX), naphthalene, and polycyclic aromatic hydrocarbons (PAHs) by their fluorescent spectra. Jet fuel containing these components, is a common Air Force contaminant.

The basic detection approach takes advantage of the fact that certain substances fluoresce when particular wavelengths of light are absorbed. The transportable laser system is unique because its output may be tuned to select the optimum wavelength to stimulate fluorescence of the pollutants while minimizing potential interferences. The spectral emission including fluorescent lifetime, is somewhat like a fingerprint, useful for identifying the contaminant. The fluorescent intensity indicates concentration of the contaminant. This technology provides semiqualitative and semiquantitative information, on site, in minutes. The LIF response can be correlated to the total petroleum hydrocarbon (TPH) concentration within the soil. The system has been tested in the field with TPH detection limits as low as parts-per-million levels on soil when used with a cone penetrometer and in the laboratory at parts-per-billion levels for naphthalene in water using fiber optic probes.

Combined Technologies

The combined cone penetrometer and transportable laser spectrometer has been used at a variety of sites having aromatic hydrocarbon contamination. Sites characterized include fuels (jet, gasoline, kerosene, diesel, etc.), naphthalene, benzene, and coal tars. The tunable laser system is optimized for stimulating contaminants and detecting the fluorescence. Laboratory fluorescence spectra from fuels suggest that naphthalene produces the maximum fluorescence; consequently, a laser excitation wavelength appropriate for naphthalene is utilized during field investigations.

The system is designed to collect data in two different modes: "push" or "static." In the push mode, laser excitation frequency is fixed and LIF signal is monitored as the cone penetrometer probe is advanced, acquiring a fluorescence intensity-versus-depth (FVD) profile. Operation in the static mode, or with the probe stopped, allows collection of LIF multidimensional data sets, typically the fluorescence emission wavelength, intensity, and time of decay matrices (WTM). WTMs have proven to be useful in identifying various fuel types. The commercial product of this technology, known as ROST, is now providing state-of-the-art fuel-contaminated site characterization (Fig. 2).

Technology Transition

Armstrong Laboratory and Unisys Corporation signed a Cooperative Research and Development Agreement (CRDA) in 1993 to commercialize the Air Force-developed laser spectrometer system. A consortium consisting of the CRDA partners, Dakota Technologies Inc., and North Dakota State University submitted a proposal to the Advanced Research Projects Agency (ARPA), Technology Reinvestment Project (TRP). In December 1993, ARPA selected the proposal to receive a two-year $1,600,000 grant. The industry partners provide in-kind contributions and matching funds. The competition was very intense with only about 200 grants awarded from almost 3,000 proposals.

The ROST commercialization program will automate the collection and mapping of data, make equipment components smaller and more rugged, and develop user-friendly interface to allow easy use by environmental technicians. This instrument is also adaptable for monitoring well applications. For example, numerous probes can be installed at monitoring points and networked

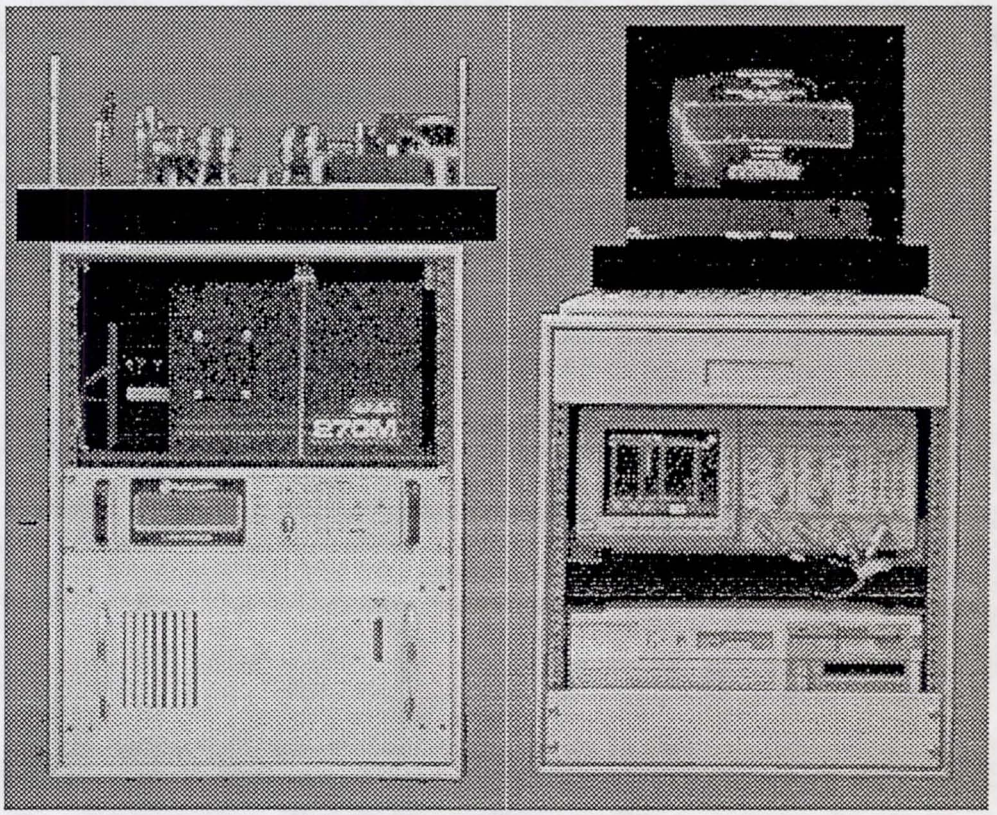


Figure 2. ROST Equipment

to a central location for continuous monitoring. ROST also has potential for industrial control monitoring and even medical diagnostics.

Use of the ROST system should result in substantial savings in costs associated with characterization, monitoring, and remediation of hazardous waste sites. Unisys is now offering site characterization services using the ROST system. *As a result of this technology transfer, DOD will benefit from application of technology and knowledge gained; the private sector will receive a highly transferable and profitable technology; the American economy will be helped; and all will benefit from a cleaner environment.*

RAPID OPTICAL SCREENING TOOL

Description

ROST employs laser-induced fluorescence spectroscopy for in situ analysis of petroleum hydrocarbons (Figure 3). Ultraviolet light is required to excite the fluorescence of most of the aromatic compounds in petroleum hydrocarbons. Pulsed light in this wavelength region is obtained in ROST by frequency doubling the output of a dye laser pumped by a Nd:YAG laser. Either the 2nd or 3rd harmonic of the Nd:YAG can be used as the dye laser pump. The laser source and detection system (spectrometer) are located in a cone penetrometer truck. The pulsed laser light travels via fiber optic cable to and from an optical module located near the cone rod. There light is directed through a sapphire window onto the surface of soil pressing up against the

window. Aromatic petroleum hydrocarbon molecules present will absorb the excitation light and emit fluorescence at longer wavelengths. The wavelength of light selected for excitation is in a range that is absorbed by aromatic petroleum hydrocarbons. A portion of the emitted fluorescence passes back through the window, returned to the surface, and imaged through a monochromator. The wavelength-dispersed radiation is converted to an electrical signal by a photomultiplier tube and the electrical signal is analyzed by a digital oscilloscope and computer. The incoming data are continuously processed and displayed in FVD profile for the entire cone penetrometer push. The fluorescence signal from 50 successive laser shots (taking a total time of 1 second) is averaged for each data point acquired and displayed. Since the cone is pushed at 2 centimeters per second, the spatial resolution of the FVD data is also 2 centimeters.

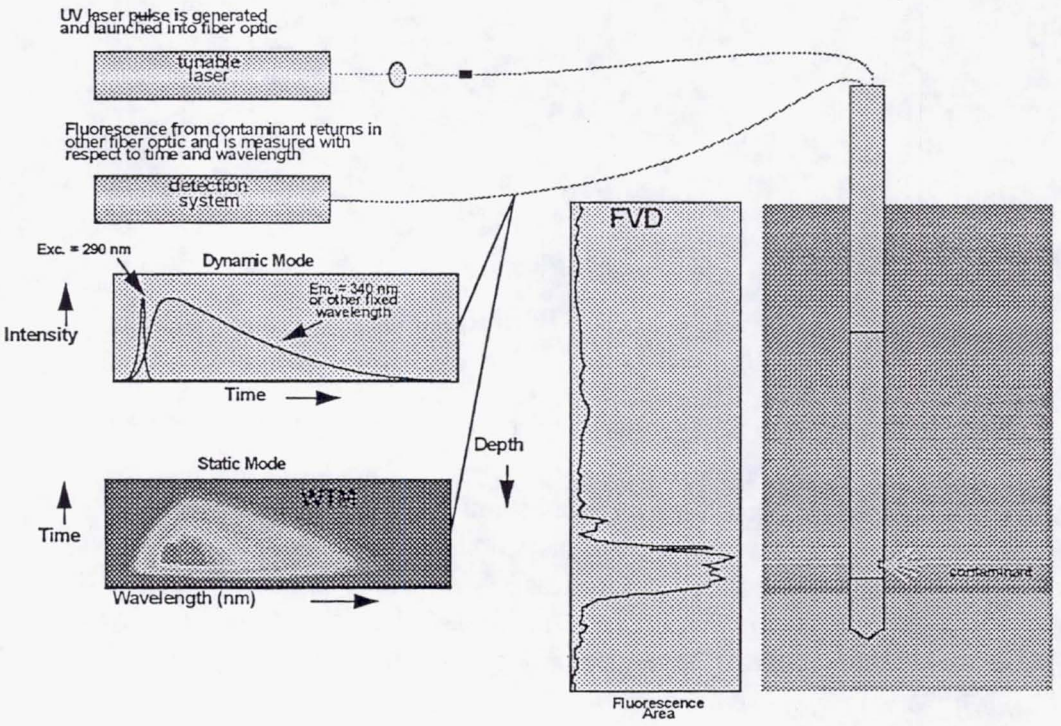


Figure 3 ROST Optical System Concept

In addition to FVD profiles, ROST can differentiate petroleum fuel types. This is accomplished by acquiring WTMs during a short pause (approximately 1 minute) in the cone penetrometer push. A WTM is a three-dimensional graph of fluorescence wavelength, fluorescence lifetime (i.e., time scale over which the fluorescence signals are emitted), and fluorescence intensity. Petroleum products have a distinctive fluorescence signature which allows the field operator to identify the approximate nature of the contaminant.. Emissions in the 260-300 nm range indicate single-ring aromatics like BTEX compounds. Emissions in the 300-350 nm range indicate two-ring aromatics such as naphthalene. Larger polycyclic aromatic hydrocarbons

fluoresce at wavelengths longer than 350 nm. WTMs are especially useful for determining if multiple sources of contamination are present. ROST can detect and characterize hydrocarbons such as gasoline, jet fuel, and diesel fuels (Figure 4). Chemometric algorithms for automatic recognition of fuel type are under development.

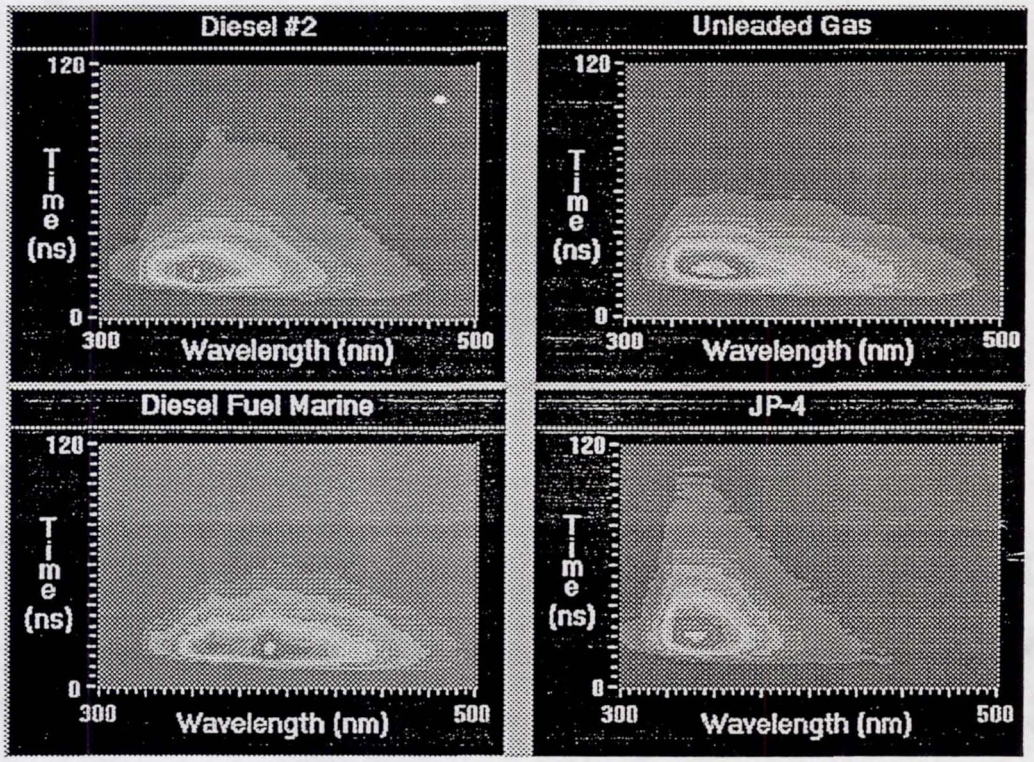


Figure 4. Fluorescence Emission Wavelength-Intensity-Lifetime Matrices (WTMs)

Benefits

ROST is extremely useful for soil and groundwater hydrocarbon contamination analysis when used with a cone penetrometer. ROST with variable wavelength (tunable), pulsed-laser source means that the excitation wavelength can be optimized for the contaminant of interest. Information regarding hydrocarbon type, depth, and distribution is available on-site at the conclusion of each push. In addition, geotechnical data are also collected. Typically, the vertical hydrocarbon profile (FVD) of a 30-foot push can be determined in less than 20 minutes. ROST, a self-contained, ruggedized system, can be permanently or temporarily installed on most new or existing cone penetrometer trucks.

RESULTS AND DISCUSSION

The ROST system is field proven and tested and is ready for wide-scale field screening. These technologies are being further refined and demonstrated within numerous DOD, DOE, and EPA programs. The results discussed are from laboratory and field demonstrations by the Tri-Service and ROST consortium. Two demonstrations were recently completed under the EPA

Superfund Innovative Technology Evaluation (SITE) Program and under an environmental Data Exchange Agreement (DEA) with the German Ministry of Defense. The purpose of the EPA SITE Program demonstrations is to evaluate innovative technology and report the results. The German DEA provides for environmental technology transfer including demonstrations to help both countries deal effectively with environmental problems.

Studies were also conducted at a number of Air Force Bases to demonstrate site amenability towards intrinsic bioremediation using the laser spectroscopy/cone penetrometer system to characterize the sites. To determine site amenability, acquired data were input to BIOPLUME® II, a computer model for in-situ contaminant biodegradation. The technology proved that it can be used to provide timely and accurate data for bioremediation.

The Tri-Services conducted a series of laboratory tests resulting in calibration curves with different fuels on various soil matrices. The calibration curve obtained in the laboratory for diesel fuel marine on a sand matrix indicates a detection limit that is lower than 30 mg/kg (ppm). The collection of WTMs for diesel #2, JP-4, unleaded gasoline, and diesel fuel marine show how each one has a characteristic pattern for fuel type identification (Figure 4).

A graphical representation of tip resistance, sleeve friction, and conductivity data collected by the cone penetrometer is shown (Figure 5). The ratio of the tip resistance to sleeve friction is

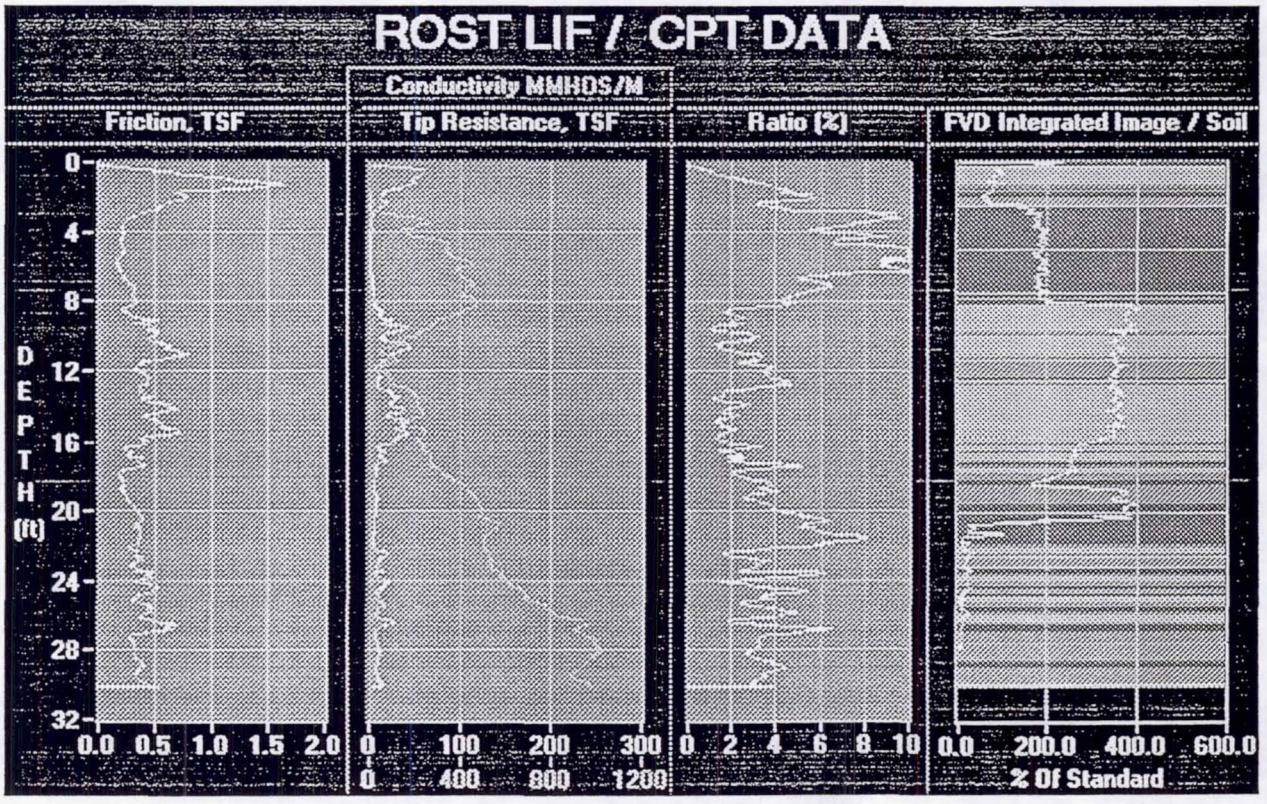


Figure 5. Real-Time Output

shown along with stratigraphy. Prior to each push, the ROST instrument performance is verified by measurement of fluorescence of a reference solution, which is contained in a cuvette cell that is pressed up against the sapphire window. FVD profiles reported as "percent of standard" are overlaid on the stratigraphy data. The fluorescence standard chosen to calibrate the ROST

instrument for this investigation was a equivalent to 10,000 ppm unleaded gasoline on sand. The vertical axis represents depth below ground surface. The horizontal displacement to the right is fluorescence intensity, which is a relative measure of petroleum hydrocarbon concentration.

Figure 6 shows FVD profiles in their relative positions along a transect, indicating how multiple ROST pushes can be used to characterize a site. In this example average fluorescence intensities between 30% and 150% of standard indicates that petroleum hydrocarbons were observed at most push locations.

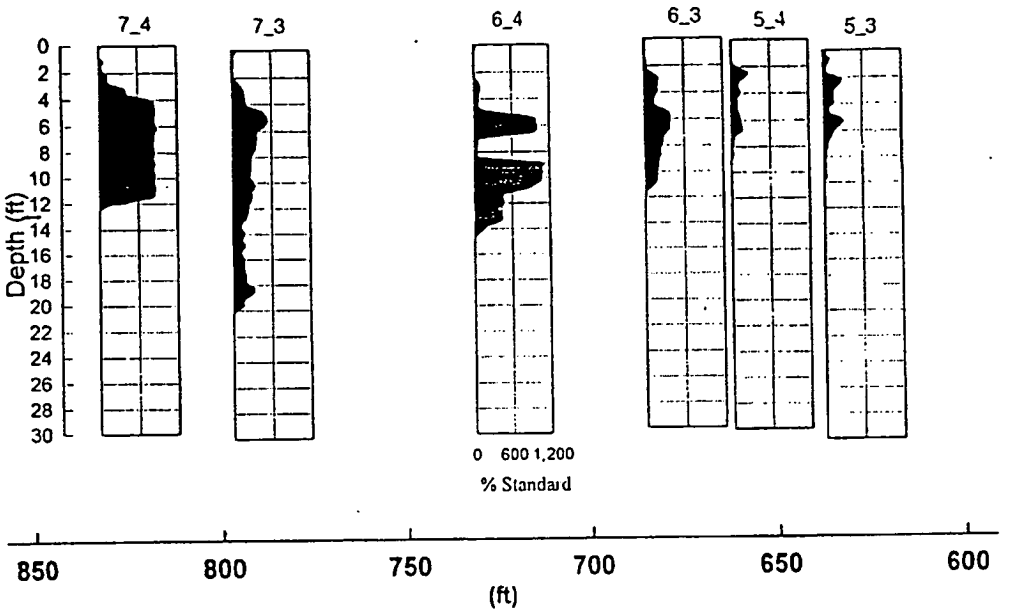


Figure 6. Fluorescence versus depth profiles

The analytical laboratory results collaborate the characterization of petroleum hydrocarbons by ROST. Areas having relatively high concentrations of petroleum hydrocarbons in soil samples correspond to areas where relatively high fluorescence intensities were observed. Furthermore, areas showing no fluorescence response are found to contain relatively low concentrations of petroleum hydrocarbons in laboratory samples. Figure 7 compares fluorescence intensity to the results of laboratory analysis for total semi-volatile organic compounds by EPA Method 8310 showing the relative fluorescence intensity data is very similar to that of the laboratory results. An excellent correlation was obtained between ROST and results using total semivolatile organic compound analysis as well as total recoverable petroleum hydrocarbons analysis. Data indicates that at this site, material containing more than 1 ppm total semi-volatile organic compounds by Method 8310, is readily observed by ROST. Data accumulated verify that ROST can reliable map subsurface petroleum contamination in situ, in real-time, and in continuous vertical fashion.

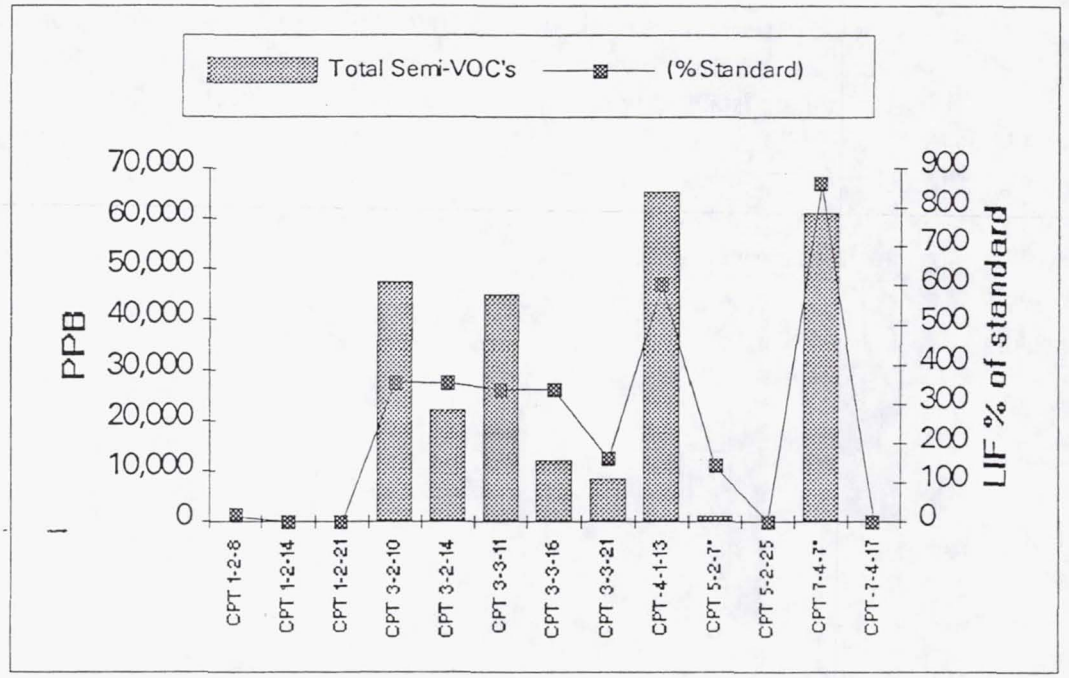


Figure 7. Fluorescence response versus analytical results

CONCLUSIONS

Screening for hazardous waste normally involves drilling of bore holes and monitoring wells. The process is slow, expensive, and results are often inconclusive. Remediation costs for U.S. government sites alone run as high as $500 billion. About 15 percent of that figure, or $75 billion, represents the price of screening and characterization. Laser induced fluorescence technology to detect aromatic hydrocarbons in situ is now a viable, field proven, commercially available technology. ROST and related technologies represent a landmark development in site characterization. Environmental investigators will be able to find, classify and map the distribution of many contaminants in days. Ongoing research will be developing techniques to detect and monitor contaminants such as chlorinated solvents, metals, and explosives which do not naturally fluoresce. Refining and demonstrating these technologies is at the heart of characterization, remediation and monitoring. These technologies may determine if remediation is needed, what remediation technology should be applied, whether the remediation is working, and when the cleanup effort is complete, all with the minimum of risk, time, labor, and cost.

Unisys Corp. and it's partners are now making the ROST system available for hydrocarbon contaminated site characterization at both government and commercial sites in North America and Europe. If ROST can lower the cost of testing from thousands of dollars per hole to just a few hundred dollars, the savings for government and industry could be enormous.

---

[1] ROST™ is a Unisys/DTI registered trademark, hereafter referred to as ROST.

PASSIVE MICROWAVE REMOTE SENSING OF ATMOSPHERIC WATER VAPOR, CLOUD LIQUID, AND TEMPERATURE

Steven J. Walter
Member of Technical Staff
Ground Based Microwave Application Group
Jet Propulsion Laboratory, California Institute of Technology
Mail Stop T-1182/3, 4800 Oak Grove Dr., Pasadena, CA 91109

ABSTRACT

Passive microwave techniques for sensing the earth's atmosphere provide powerful tools for understanding and monitoring its dynamics. These techniques exploit the microwave and millimeter-wave emissions of atmospheric constituents. Tuning a sensitive radio receiver, termed a radiometer, to these emissions allows the atmosphere to be probed remotely from satellites, airplanes, and the earth's surface. Microwave radiometers can sense water vapor, atmospheric temperature profile, cloud liquid, and trace concentrations of many atmospheric gases.

Microwave radiometers have applications in many areas: weather prediction, pollution forecasting, aviation, climatology, spacecraft tracking, geodesy, and atmospheric chemistry. Four specific applications for low-cost commercial radiometer systems are explored in this paper. Microwave temperature profilers (MTPs) can provide real-time detection of temperature inversions that trap automobile emissions and industrial pollution near the ground. Timely, accurate prediction of weather conditions conducive to creating high pollution levels has significant value in mitigating pollution hazards. Water vapor radiometers (WVRs) can measure the total atmospheric water burden (*e.g.*,water vapor and cloud liquid) which could enhance weather predictions and rain forecasts. Both MTPs and WVRs could support federal efforts to validate and improve the performance of global climate models. Critical climatic measurement needs include measurement of horizontal distribution of both cloud liquid and water vapor. Finally, airport-based radiometers could be used to predict aircraft icing conditions by detecting clouds bearing supercooled liquid. Ground-based radiometer determinations of the temperature profile and amount of cloud liquid coupled with ceilometer or radar determinations of cloud height allow identification of sub-freezing cloud liquid.

Widespread application of radiometric remote sensing systems has been restricted by costs. Plumbed-waveguide radiometer systems for monitoring water vapor, cloud liquid, or temperature profiles cost between $120K and $500K depending on the specific application. It is believed that significant price reductions can be achieved through mass production, application-specific designs, and use of monolithic or hybrid microwave integrated circuit (MMIC and MIC) technology. Recent advances in MIC technologies promise to reduce production costs by allowing radiometers to be etched on an integrated circuit. Conversion to MIC will also reduce the size and power requirements thereby lowering costs of the associated power, antenna pointing, and temperature regulation subsystems. It is also anticipated that the compact size and reduced power consumption will translate into improved temperature control which will increase radiometer stability. Preliminary studies indicate that system cost reductions ranging from a factor of three to eight appear feasible with minimal development costs. We believe these cost reductions will stimulate widespread commercialization of radiometric atmospheric remote sensing.

INTRODUCTION

Atmospheric radiometers have been primarily developed as research tools. Radiometers currently fly on satellites measuring atmospheric temperature, polar ozone, water vapor, stratospheric chlorine, and cloud liquid. Ground-based radiometers are being employed for research on climate, weather, satellite communications, astrophysics, and aviation safety. Due to these activities, radiometric techniques are well developed and the instrumentation is mature. Many current uses have commercial value. However, the widespread application of radiometric technology has been restricted by cost. This could change; with development of a commercial market for microwave devices and integration of microwave circuit technology, there is potential for significant cost reduction.

This paper intends to provide an overview of atmospheric radiometry and potential commercial markets. First, is a top-level analysis of how atmospheric radiometry is used to measure the atmosphere. This is followed by a general description of radiometer design and strategies for achieving future cost reduction. Finally, discussion will center on four applications with commercial value.

RADIOMETRIC REMOTE SENSING OF THE ATMOSPHERE

The atmosphere emits (and absorbs) a continuous spectrum of microwave and millimeter-wave radiation [1]. As illustrated by Figure 1, tropospheric absorption spectra are dominated by broad, discreet, water vapor and oxygen spectral lines superimposed upon a continuum. Although the 22 GHz water vapor line and 60 GHz oxygen band are the result of different processes, they are both primarily broadened by molecular collisions [1]. Continuum emissions are produced by water vapor, liquid water (clouds), and to a lesser extent, molecular oxygen [1].

The 22.2 GHz emissions of atmospheric water vapor are routinely used for remote sensing. Weakness of this spectral line and minor contributions from other radiative sources makes it a popular choice for many remote sensing applications. Measurement of water vapor emissions can be used to determine the columnar abundance of water vapor. Continuum emissions from cloud liquid are broadband and are therefore usually sensed in either the 30 to 38 GHz or 85 to 100 GHz spectral windows. Measurement in these spectral windows minimizes contamination from other sources of emissions. Radiometric measurements of cloud liquid allows the columnar liquid content of a cloud to be determined. In fact, radiometry is the only technique that

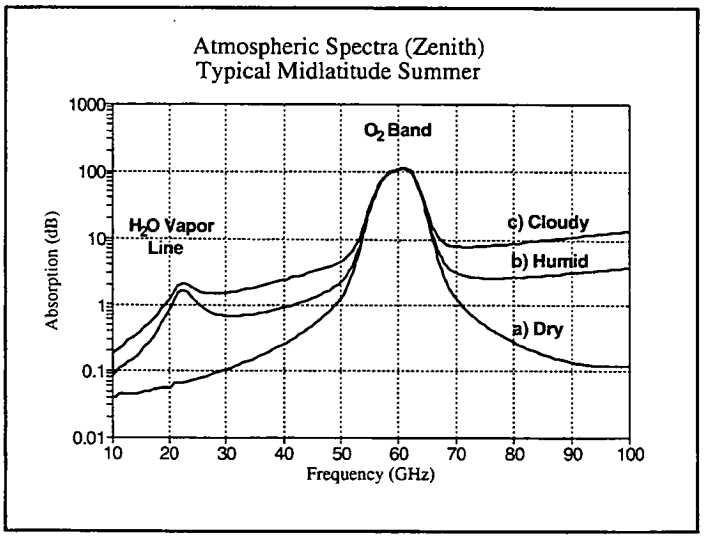


Figure 1. *Typical midlatitude summer atmospheric spectra for a) no water vapor, b) a humid day, and c)a humid and cloudy day.*

can determine a cloud's total liquid content. (Radar echoes are better correlated with drop size and shape than total liquid content.) Additionally, radiometry is the only technique able to accurately sense vapor along an arbitrary line-of-sight during cloudy weather. However radiometry can not be used to retrieve either cloud altitude or high-resolution vapor profiles. In contrast to the weak emissions of water, the 60 GHz oxygen band is strong or optically thick. As will be discussed, this makes this spectral band useful for measuring atmospheric temperature profiles.

The intensity of microwave or millimeter-wave energy emitted by an atmospheric gas can be quantified in terms of a brightness temperature [2]. A physical interpretation of the brightness temperature is that it is the temperature of a blackbody that emits the same intensity of radiation per unit bandwidth as the atmospheric gas being observed. This can be quantified using the Rayleigh-Jeans approximation

$$T_b(\nu) = \frac{\lambda^2}{2k} I(\nu) \qquad (1)$$

where $T_b(\nu)$ is the brightness temperature, k is Boltzmann's constant (1.381×10^{-23} J/K), λ is the radiation wavelength, and $I(\nu)$ is the radiation intensity [2]. The intensity of radiation received by a radiometer is the radiation being emitted minus the radiation being absorbed integrated along the atmospheric path being viewed. To calculate the radiation received from an emission region located a distance s_o from the observation point, the total radiative transfer needs to be calculatedusing

$$T_b(\nu) = T_{b0} e^{-\tau(s_o)} + \int_0^{s_0} T(s)\, e^{-\tau(s)} \alpha \, ds \quad , \qquad (2)$$

where the T_{b0} is the brightness temperature of the source or background, $T(s)$ is the temperature of intervening media at distance s, and $\tau(s)$ is the optical depth at distance s [2]. The first term represents the source and the second term represents the intervening media. Insight can be gained by examining this integral expression at the two limits of opacity. First we simplify the expression by assuming that the atmosphere is at a constant temperature, $T(s) = T$. This allows temperature, T, to be a multiplicative factor for the integral expression. For an optically thin atmosphere [2],

$$\int_0^{s_0} T(s)\, e^{-\tau(s)}\, ds \approx T\tau_{s_0} \qquad \text{for } \tau_{s_o} \ll 1 \text{ and } T(s) = T. \qquad (3)$$

In this case the brightness temperature is just the physical temperature times the opacity or optical depth. Since the optical depth is proportional to the concentration of the absorbing/emitting gas, then the brightness temperature is proportional to the line-of-sight gas content weighted by its physical temperature. In contrast, an optically thick atmosphere yields [2]

$$\int_0^{s_0} T(s)\, e^{-\tau(s)}\, ds \approx T \qquad \text{for } \tau_{s_o} \gg 1 \text{ and } T(s) = T. \qquad (4)$$

Thus, the brightness temperature of an optically thick gas is simply the media's physical temperature.

Given a sufficiently accurate description of the meteorological conditions, the atmospheric opacity or emission spectrum can be calculated. In contrast, the solution to the inverse problem is not unique. Emission measurements at a limited number of frequencies will not allow the various

atmospheric parameters to be uniquely determined. Interpretation of the radiometric measurements usually requires that some atmospheric parameters be estimated. For example, cloud and vapor sensing using a dual-frequency water vapor radiometer requires that the atmospheric temperature be estimated. The effect of an estimated parameter on the retrieval accuracy can be reduced with the judicious choice of radiometric frequencies. A detailed description of various techniques used to retrieve atmospheric water vapor and cloud liquid from measurements of brightness temperatures is beyond the scope of this article; however it is well documented in the technical literature [3, 4].

Optically thin regions of the atmospheric spectrum are used to sense atmospheric water. Thus radiometric measurements furnish a signal proportional to the atmospheric opacity times the physical temperature. The first step in retrieving water vapor and cloud liquid from emission measurements is to estimate a mean atmospheric temperature [3]. This is the average value of the physical temperature of the atmospheric water which ranges from 260 K to 280 K. It can usually be estimated to within 2% using surface temperature measurements. The mean temperature allows the opacity or atmospheric brightness temperature to be calculated. The atmospheric brightness is a sum of emissions from water vapor, liquid water (clouds), oxygen, and astronomical sources. To retrieve either water vapor or cloud liquid, corrections need to be applied for these other emission sources. Corrections for oxygen emissions can be computed from ground-based pressure measurements [3]. Beam-averaging causes most astronomical emission sources to be insignificant when measurements are made with a moderate beamwidth antenna ($\theta > 2^{\circ}$). The sun is an exception that usually frustrates attempts to make measurements in its direction. The other source needing correction is the microwave cosmic background which will appear as a constant 2.74 K offset in the radiative transfer equation [3]. Once these corrections are made, cloud liquid can potentially be determined from measurement at a single frequency. In contrast, water vapor determinations usually require two frequencies: one at the 22 GHz water vapor spectral line and the other at a cloud sensing frequency (to provide a correction for liquid emissions). Measurements at additional frequencies are used to improve the retrieval accuracy for both water vapor and cloud liquid [3, 4].

Temperature profiling exploits the optically thick emissions of the 60 GHz oxygen band. In this case, the brightness temperature equals the physical temperature of the gas [3, 5]. For the atmosphere which has a varying temperature structure, radiative transfer through the medium needs to be treated. This is realized with a weighting function, $W(s)$, that describes the distribution of received thermal emission as a function of distance between the emission source and the radiometer. For ground-based temperature profiling, frequencies are chosen with optical depths corresponding to a distance of a hundred meter to several kilometers. Assuming for simplicity that atmospheric opacity is constant with altitude, the weighting function is exponential $W(s) \sim e^{-s}$ [5]. Using the weighting function formulation, the measured brightness temperature reduces to

$$T_b = \frac{\int_0^\infty W(s)T(s)ds}{\int_0^\infty W(s)ds} \qquad (5)$$

where $T(s)$ is the temperature structure of the atmosphere with respect to altitude, s [5]. For a medium with a linear variation in temperature, *i.e.* $T(s) \sim s$, the measured brightness temperature equals the atmospheric physical temperature at a distance of one optical depth, s_a. That is where the

weighting function equals $1/e$. Therefore a temperature profile can be derived by measuring atmospheric emissions at a series of frequencies each corresponding to different optical depths. Once the average temperature is known for each optical depth then the data can be inverted to generate a temperature profile.

For ground-based and airborne applications the altitude resolution can be enhanced using a technique developed by Bruce Gary at JPL [5]. He exploits the fact that horizontal atmospheric temperature gradients are more than an order of magnitude smaller than vertical gradients. Making radiometric measurements at non-zenith elevation angles allows the effective altitude corresponding to a given optical depth to be varied [5]. The effective altitude for temperature retrieval at an elevation angle θ is just the optical depth at that elevation angle projected onto the vertical axis. Thus for optical depth s_a and an elevation angle of θ, the effective altitude h_a for the temperature retrieval is simply $h_a = s_a \sin\theta$. For a horizontally stratified atmosphere, the brightness temperature at elevation angle θ can be associated with an effective altitude h_a using [5]

$$T_b(\theta) = T(h_a) = T(s_a \sin(\theta)) \quad . \tag{6}$$

The atmospheric temperature profile can then be constructed using multi-frequency brightness temperature measurements at a series of elevation angles.

RADIOMETERS

Radiometers are sensitive radio receivers that generate an output voltage that is proportional to the power incident at the antenna. A generic block diagram for a total-power radiometer is illustrated in Figure 2. The actual implementation can employ either a heterodyne or direct detect receiver. Although the basic design parameters are determined by specific applications, there are some basic principles underlying radiometer design.

The power emitted by an atmospheric gas is simply

$$P_i = kTB \quad , \tag{7}$$

where k is Boltzmann's's constant, T is the brightness temperature of the gas in Kelvin, and B is the measurement bandwidth in hertz. The total power measured by the radiometer is the system noise which is the sum of the atmospheric and radiometer noise. The atmospheric noise ranges from 7 K to 300 K depending on the atmospheric opacity and radiometer noise ranges from 100 to 600 K depending primarily on the noise figure of the front end amplifier. Consequently, the total power incident at the radiometer measured with a 200 MHz bandwidth will range from 0.3×10^{-12} and 2×10^{-12} watts (or -95 dBm to -87 dBm). To measure the brightness temperature of a gas with 0.5 K resolution and the same bandwidth requires detecting power differences of 1.4×10^{-15} watts.

The input power levels also determine the system gain. The RF gain and IF gain (if applicable) are chosen to match the linear region of the detector with a typical value being 50 dB. DC gain is selected to generate a voltage that is optimized for the chosen analog to digital conversion technique. The system gain can then be expressed as a proportionality relationship between the incident atmospheric temperature, T_a, and the radiometer output voltage, V,

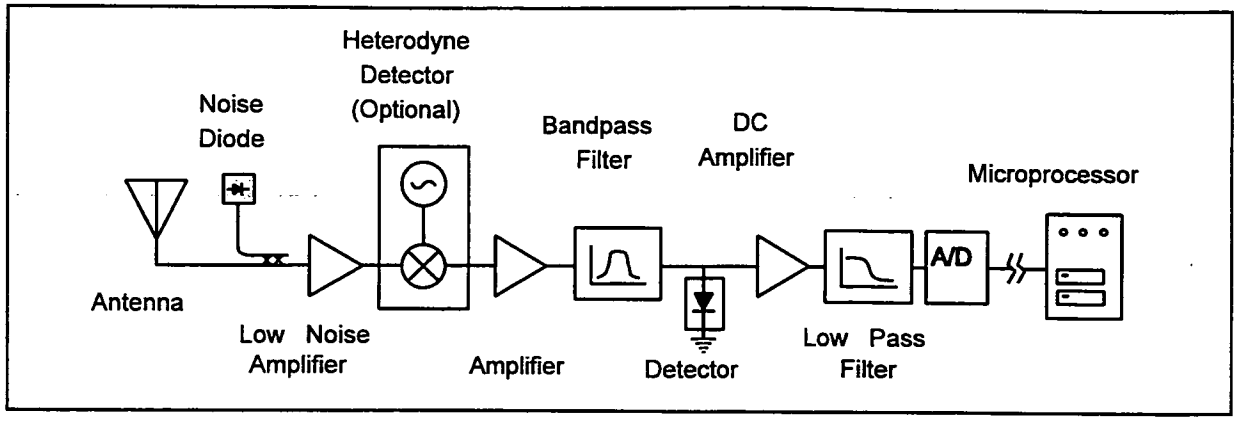


Figure 2. *Generic Block Diagram for a radiometer receiver.*

$$T_a = c\ (V - V_o) \quad , \tag{8}$$

where V_o is proportional to the receiver noise power and c is the system gain [4]. Reasonable values of system gain span from 1 V per 10 K to 1 V per 300 K. The requisite gain stability is simply the measurement accuracy divided by the total system temperature. For example, a 0.5 K noise temperature accuracy measured with 450 K of radiometer system noise (50 K atmosphere plus 400 K radiometer noise) requires 0.5/450 or 0.1% gain stability. In other words a 0.1% gain fluctuation will masquerade as a 0.5 K change in the incident atmospheric power.

The radiometer measurement precision is limited by fluctuations in measured thermal noise. The magnitude of these fluctuations is a function of the total system noise and can be reduced by integrating over time. To make a measurement with ΔT_B precision requires an integration time, τ,

$$\tau = \frac{T_s^2}{\Delta T_B^2 B} \tag{9}$$

where T_s is the radiometer system noise [2]. For measurement of a 260 K atmospheric temperature with 0.2 K precision using 300 K receiver with a 200 MHz bandwidth will require a 40 ms integration time. A wider bandwidth and lower receiver noise will reduce the integration time.

The instrumental frequency stability requirements are determined by the structure of the atmospheric spectrum and requisite measurement accuracy. If measurements are being made on the wing of a spectral line, then a frequency shift towards the spectral peak will cause an increase in the measured emissions. Requirements for stability are usually in the range of 100 kHz to 100 MHz depending on the application.

The primary technical challenge in radiometry is to develop a high gain (> 70 dB) receiver that is stable to a tenth of a percent. Temperature-induced gain and frequency changes tend to limit radiometer stability. Therefore, good thermal design and temperature control is essential for accurate radiometric measurements. Other sources of gain variations include power supply drifts, mechanical vibration, and RFI; however these can be minimized with good design practices.

There are several techniques for calibrating radiometers. Gain, c, can be calibrated with a measurement of two calibration targets each at a different temperature, T_{hot} and T_{cold}. The output voltage associated with each of these measurements, V_{hot} and V_{cold}, can then be used to calculate c [2]

$$c = \frac{T_{hot} - T_{cold}}{V_{hot} - V_{cold}} \quad . \tag{10}$$

Similarly, the receiver noise power, V_o, can also be calculated,

$$V_o = \frac{1}{2}\left((V_{hot} + V_{cold}) - \frac{1}{c}(T_{hot} + T_{cold}) \right) \quad . \tag{11}$$

Targets can be built from microwave absorbers. Measurements at ambient temperatures are usually used for one calibration point. Absorbers placed in liquid nitrogen (77 K) are another popular choice for a second calibration point. Construction of targets for high accuracy applications is a bit of an art and is discussed in detail in reference [6].

The gain can also be calibrated with one or more noise diodes. Injecting noise into the antenna signal path is an inexpensive method for monitoring radiometer gain and receiver noise. By using two diodes coupled with different strengths a real time calibration is possible. Other techniques for calibration include "tipping" the radiometer elevation angle to use the atmosphere as a calibration target and using internal hot and cold loads [4].

Antennas for radiometry do not need high efficiency; however they require low sidelobes and spillover. The need for reducing sidelobes and spillover is best illustrated with an example. Consider a sidelobe that contributes 1% to the total received energy: when directed toward the ground(T_{bg} ~ 300 K) it will introduce an error equal to 1% of T_{bg} or 3 K. Sidelobes and spillover are usually specified not to exceed 30 to 40 dB. To minimize sidelobes and eliminate spillover, radiometers tend to be designed with corrugated horn antennas.

In addition to the radiometer RF electronics there are a variety of application-specific design issues such as antenna beamwidth and pointing needs, power requirements, communications, instrument control, and data acquisition. An example of a general purpose water vapor radiometer is the JPL J-series radiometer which is shown in figure 3 [7]. It was developed to be compact, portable with self-diagnostic capability. It senses water vapor and cloud liquid at 20.7, 22.2, and 31.4 GHz. The antenna and RF electronics are housed in the top box and power supplies and instrument controller are housed in the bottom. The antenna is mounted horizontally and is pointed at a 45° mirror. This mirror can be rotated to vary the antenna elevation angle and the whole top unit swivels to vary azimuth pointing. This

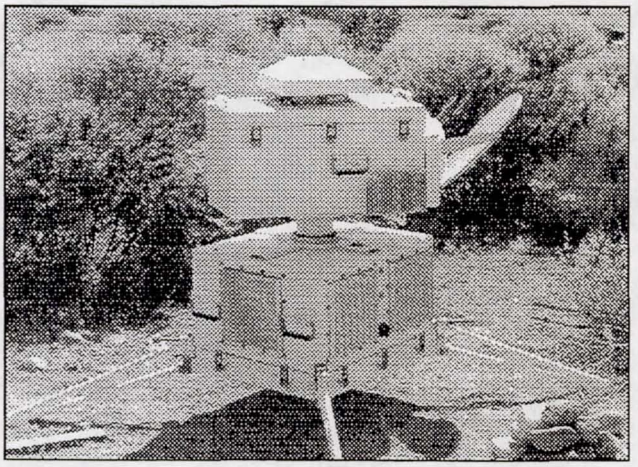


Figure 3. *The JPL J-series water vapor radiometer. Ref [7].*

radiometer has been a workhorse for JPL providing ground-based verification of satellite radiometer performance, collecting Ka-band propagation statistics for modeling earth-to-space telecommunication links, determining water vapor-induced atmospheric propagation delays, performing basic research in atmospheric dynamics, and participating in climate measurement campaigns.

A new generation of radiometers is being developed using monolithic and hybrid microwave integrated circuit (MMIC and MIC) technology. An initial proof-of-concept hybrid MIC radiometer channel was developed at JPL in 1992 see figure 4 [8, 9]. Since that time, several satellite instrument programs have started development of MIC and MMIC radiometers. These radiometers will be used to make astrophysical measurements, measure atmospheric water vapor, and monitor atmospheric temperature. For satellite applications, MIC and MMIC technology has been selected to achieve significant size, weight, and power reductions.

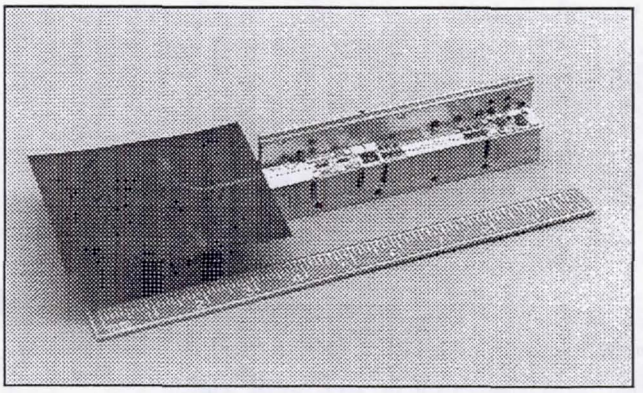


Figure 4. *A proof-of-concept, single-channel, MIC radiometer developed at JPL. Refs. [8, 9].*

We believe significant cost reductions can be achieved through application-specific design and use of MMIC and MIC technology. Cost drivers for radiometer design include antenna beamwidth, number of measurement frequencies, antenna pointing capability, size, power requirements, and thermal control. By performing a thorough review of a radiometer application, only required features will be included in the design. For example, it may be decided that only zenith pointing is necessary for weather observations. That design decision will eliminate the cost of mirrors and motors needed to steer the radiometer pointing angle. Additionally, when radiometer production increases from the current rate of several units per year to dozens per year then there will be a benefit to implement the electronics using MIC technology. Being able to etch a radiometer on an integrated circuit will reduce the labor costs associated with assembling the RF section. Additionally, the smaller size and lower power requirements will reduce the capacity of power supplies, enclosure size, temperature controller power consumption, etc. Each of these reductions will further drive down costs.

POTENTIAL COMMERCIAL APPLICATION OF RADIOMETERS

Air Pollution Forecasting

In urban areas, temperature inversions can trap auto emissions and industrial pollution near the ground creating significant health hazards [10]. Timely and accurate prediction of weather conditions conducive to high pollution levels have economic value. For example, forecasts can be used to regulate discharges from large emission sources. Several cities require power companies to switch to cleaner-burning fuels during periods when strong inversions are present. Since cleaner fuels are more expensive than the heavier, high-sulfur fuels, savings could be realized from improving forecast accuracy. Improved air quality can also be achieved in cities that have authority

to impose wood-burning (fireplace) bans. In addition, predictions allow alerts to be issued to groups that are most sensitive to air pollution such as young children, asthmatics, and the elderly.

Currently, local pollution forecasts are developed from general weather service predictions supplemented with local launches of radiosondes (balloons instrumented with meteorological sensors). Forecasts are restricted by the limited spatial and temporal resolution of these available data sources. A microwave temperature profiler would enhance existing observation systems since they can determine the atmospheric temperature profile as frequently as several times per minute. They can operate autonomously and can be accessed via telephone or radio link as required. This technology could significantly improve the temporal and spatial resolution of current balloon-based monitoring systems. The current cost of expendables such as radiosondes would make a low cost temperature profiler attractive. At current spending levels, savings realized over several years could be used to set up a profiler network allowing ground-level weather changes to be tracked.

It is worth mentioning that for this application there are several advantages of radiometric temperature profilers over a competitive temperature profiler technology, radio-acoustic sounding [11]. Profiling radiometers can be compact, reliable, low power, and inexpensive to operate. Additionally, since radiometers do not transmit acoustic or radio energy, they are "good neighbors" and can be deployed in urban environments. Although radiometric profilers do not provide the altitude resolution of a RASS, the aforementioned advantages still make them desirable for this application.

Rain and Weather Prediction

Currently the National Weather Service (NWS) has undertaken a massive program to update and modernize its data systems and forecast services. Resources are being earmarked to deploy 116 NEXRAD weather radars, launch additional GOES satellites, upgrade computer hardware, improve forecast models, and enhance surface weather monitoring. As the spatial resolution of weather forecasting models increase, there has been a growing interest in low-cost, autonomous sensors which improve spatial coverage. Low-cost radiometers could be a effective adjunct to the existing sensor complement.

Cloud sensing is an application where radiometers could clearly enhance existing and planned climate and weather observing networks. As stated earlier, radiometric sensing of clouds is the only technique capable of determining the total liquid content of clouds. Therefore, radiometers are able to monitor the total atmospheric water burden (liquid and vapor) in real time. It seems intuitive that a radiometric network that maps overhead water burden would improve weather forecasts.

Water vapor radiometers could enhance the NEXRAD radar product. As stated earlier, radar echoes are better correlated with size and shape of cloud droplets than with their liquid content. Concurrent WVR measurements can be used to relate liquid content of clouds to the radar returns. The NOAA Wave Propagation Laboratory has conducted preliminary studies showing value in radiometer support for radar measurements [12].

Finally, it should be noted that weather forecasting on the west coast of the United States suffers from the lack of data available over the ocean. Radiometer technology is also well suited for

deployment on ocean buoys. Several years ago, JPL developed a combined MTP/WVR data buoy for NOAA. Although NOAA continues to support radiometer development and demonstrations, NWS has not incorporated radiometers into their system upgrade partially due to their high cost.

Aircraft Icing Detection

Aircraft icing poses a serious winter hazard at many U.S. airports [3]. In freezing conditions, clouds can form with appreciable amounts of supercooled liquid water. Aircraft traversing these clouds accumulate ice as liquid freezes to the wings and fuselage This adds weight and increases drag. This hazard is most acute during descent prior to landing due to the airframe's sub-freezing temperature. Fortunately, commercial jets are equipped with anti-icing systems such as wing heaters. However, light and moderate-sized aircraft are usually unprepared to deal with severe winter icing. These aircraft would benefit from a ground-based airport surveillance system that can detect the presence of supercooled cloud liquid. Air traffic controllers who have been alerted to icing hazards can then appropriately advise and reroute vulnerable aircraft.

FAA-sponsored studies conducted by the NOAA Wave Propagation Laboratory demonstrate that passive microwave remote sensing techniques in conjunction with ceilometers or weather radars can identify winter icing hazards [3]. The ground-based microwave radiometers are sensitive to cloud liquid while being insensitive to ice. (Below 40 GHz, liquid water radiates more than two order of magnitude more energy per unit mass than ice [13].) Once liquid is detected, its temperature must be assessed. Combining the temperature profile with a ceilometer or radar measurement of cloud height furnishes an estimate of cloud temperature. When cloud liquid is present in sub-freezing clouds an alert can be issued to air traffic control. The NOAA group participated in a multi-year icing detection demonstration project, Winter Icing and Storm Program (WISP) at Denver's Stapleton airport that clearly established the value of radiometric measurements for icing hazard forecasting [14].

Climate Monitoring and Global Change

Recent measurements of increases in the concentration of greenhouse gases such as carbon dioxide and methane has sparked concern that we are changing our climate. In the debate over anthropogenic climate change, the accuracy of climate predictions has come under scrutiny. There are significant model uncertainties associated with the effects of water vapor and clouds on climate. Water vapor, the earth's principal greenhouse gas, represents a significant source of modeling error. Clouds can cool the earth by reflecting solar radiation back towards space or can warm the earth by reducing surface heat (IR) radiated to space. The total radiative effects of atmospheric water depends on its altitude, phase (vapor, liquid, or ice), and concentration.

Given the policy implications of global warming, several national and international programs are charged with improving the reliability of climate models. The Department of Energy's Atmospheric Radiation Measurement (ARM) program is an example of a program tasked with validating and improving the performance of global climate models [15]. ARM's charter is to improve the performance of general circulation and related atmospheric models as tools for predicting global and regional change. ARM recently began to instrument a cloud and radiation testbed (CART) in Oklahoma to provide data on the Earth's radiation balance spanning 90,000 km^2,

roughly the size of a climate model cell. They have plans to eventually instrument five more sites in regions with diverse climatic characteristics. The Oklahoma CART site is instrumented with a variety of remote sensing instrumentation including RASS temperature profilers, a LIDAR, wind radars, a WVR, a whole sky imager, ceilometers, etc. To realize their goals, ARM researchers have identified critical measurement needs including measurement of horizontal distribution of total-column cloud liquid and water vapor as well as, cloud mapping/imaging capability. ARM is also interested in airborne and ground based temperature profiles. Some of these needs can be met with radiometers, however their use has been limited by cost.

The Global Energy and Water Cycle Experiment (GEWEX) is an example of international effort to improve the understanding of the earth's energy fluxes and hydrological cycle. A significant objective of this program is to understand the impact of atmospheric water on weather and climate [16]. The Continental-Scale International Project (GCIP) is the focus of the GEWEX buildup phase which is funded through NOAA, DOE, and NASA. [16]. The goal of GCIP is to observe and model hydrological processes in the Mississippi River basin. This experiment will take advantage of a dense network of existing and planned atmospheric sensing systems. As in the case of ARM, the use of radiometers has again been limited by cost.

Finally, radiometric sensing could remedy inadequacies with surface cloud observations. Existing records of cloud cover are unreliable because of their subjective nature and dependence on weather observer training. Establishment of a 30 or 90 GHz standard data type for cloud liquid would go a long way toward generating a cloud record that could document climate change.

CONCLUSIONS

There is potential for widespread application of water vapor radiometers and microwave temperature profilers to monitor the atmosphere. Applications discussed in this paper includes weather prediction, climate monitoring, pollution forecasting, and aircraft icing detection. This market can be stimulated by a reduction in the cost of radiometers. It is believed that significant price reductions can be achieved through mass production, application-specific designs, and use of MIC and MMIC technology.

ADDITIONAL INFORMATION

There is a vast technical literature on all aspects of radiometry including radiometer design, retrievals, and applications. Good starting points are two general references on radiometry: *Atmospheric remote sensing by microwave radiometry* [3] which is a collection of review articles written by the experts in the field, and *Microwave remote sensing, active and passive* [13] which is a three volume text discussing all aspects of microwave remote sensing. Also, having developed satellite, airborne, and ground-based radiometers for a variety of remote sensing applications, JPL can be used as a resource for further information on radiometry. Finally, both NASA and JPL have a strong commitment to technology transfer and would be potentially interested in teaming with companies interested in commercializing radiometers.

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

BIBLIOGRAPHY

[1] Walter, S. J., "Tropospheric Water Vapor Microwave Spectrum: Uncertainties for Remote Sensing," *Proceedings of the Specialist meeting on Microwave Radiometry and Remote Sensing Applications,* E.R. Westwater, Ed., pp 203-210, U.S. Dept of Commerce, Boulder, CO 1992.

[2] Janssen, M.A., "An Introduction to the passive microwave remote sensing of atmospheres," *Atmospheric remote sensing by microwave radiometry*, M.A. Janssen, Ed., John Wiley & Sons Inc., pp 1-35, 1993.

[3] Westwater, E.R., "Ground-based microwave remote sensing of meteorological variables," *Atmospheric remote sensing by microwave radiometry*, M.A. Janssen, Ed., John Wiley & Sons Inc., pp 145-207, 1993.

[4] Elgered, G., "Tropospheric Radio-path delay from ground-based microwave radiometry," *Atmospheric remote sensing by microwave radiometry*, M.A. Janssen, Ed., John Wiley & Sons Inc., pp 215-250, 1993.

[5] Gary, B., "Observational results using the microwave temperature profiler during the airborne Antarctic ozone experiment," *Jour. Geophys. Res.*, **94**, pp 11223-11231 Nov., 1989.

[6] Stacy, J. M., Microwave Blackbodies for Spaceborne Receivers, *JPL Publication*, **85-10**, 1985.

[7] Janssen, M. A., "A new instrument for the determination of radio path delay due to atmospheric water vapor," *IEEE Trans Geosci. Remote Sensing*, **GE-23**, pp 485-490, 1985.

[8] Sukamto, L., M Janssen, and G. Parks, "MMIC receiver for water vapor radiometry," *NASA Tech Briefs*, **17**, Item 101, pg. 34, 1993.

[9] Sukamto, L., M. Janssen, and G. Parks, "Monolithic microwave integrated circuit water vapor radiometer," *Proceedings of the Specialist meeting on Microwave Radiometry and Remote Sensing Applications,* E.R. Westwater, Ed., pp 158-162, U.S. Dept of Commerce, Boulder, CO 1992.

[10] Graedel, T.E. and P. J. Crutzen, *Atmospheric change: an Earth system perspective*, W.H. Freeman and Co., New York, 1993.

[11] May, P.T., R.G. Strauch, K.P. Moran, and W.L. Eckland, "Temperature sounding by RASS with wind profiler radars: A preliminary study," *IEEE Trans. Geosci. Remote Sensing*, **28**, 19-28, 1990.

[12] Sassen, K., A.W. Huggins, A.B. Long, J. B. Snider, and R. J. Meitin, "Investigations of a winter mountain storm in Utah. Part II: Mesoscale structure, supercooled liquid water development, and precipitation processes," *Jour. Atmos. Sci.*, **47**, 1323-1350, 1990

[13] Ulaby, F.T., R.K. Moore, and A.K. Fung, *Microwave remote sensing, active and passive, Vol 1*, Addison Wesley Pub. Co., Reading, MA 1981.

[14] Stankov, B.B., E.R. Westwater, J.B. Snider, and R.L. Weber(1992), "Remote Measurements of supercooled integrated liquid water during WISP/FAA aircraft icing program," *J. Aircraft*, **29**, 604-611, 1992.

[15] DOE Global Change Research Program, ARM Outreach, **1**, No. 1, DOE, 1992.

[16] WMO (World Meteorological Organization), *Scientific Plan for the GEWEX Continental Scale International Project (GCIP)*, **WCRP-67**, *WMO/TD No. 461*, WMO, Geneva Switzerland, 1992.

A NOVEL FIXED-POINT MONITOR FOR RESPIRABLE COAL DUST

Charles D. Litton
U.S. Department of the Interior
Bureau of Mines, Pittsburgh Research Center
P.O. Box 18070
Pittsburgh, PA 15236

ABSTRACT

A novel detector which can distinguish fire smoke from diesel particulates was tested to determine its response to respirable dusts in the concentration range of 0 to 7 mg/m^3. The test results indicated that the detector response was linear over this dust concentration range for respirable coal dusts and non-linear for respirable rock dust. The response to respirable coal dust was also found to vary directly with the volatile fraction of the coal dust. Based upon these results, the use of this detector as a continuous monitor of respirable dust in underground coal mines has significant potential.

INTRODUCTION

Measurement of respirable dust in underground coal mines is of significant importance in maintaining a healthful environment for the mine worker. Current practice employs the acquisition of respirable dust on a filter cassette during an 8-hour working shift. The filter cassette is then dried and the mass of dust collected, then measured. The mass of dust measured divided by the total sample volume yields a time-weighted 8-hour average dust concentration. By regulations (30 CFR, Part 70.100) (*1*) this average respirable dust concentration cannot exceed 2.0 mg/m^3. Generally, the filter cassette is contained within a Coal Mine Dust Personal Sample Unit that is worn by mine personnel and continually samples the mine air at a nominal flow rate of 2.0 liters per minute. At the inlet to this unit is a small cyclone that allows only dust that is respirable to flow to the filter cassette. Respirable dust is generally considered to be dust particles with diameters less than 10 μm.

While this method of measurement does provide a valid representation of the miner's exposure to respirable dust levels during a typical work day, it does not provide sufficient information relative to areas in which high dust levels may exist, or the duration of excessive dust levels in these areas. In addition, if areas of high dust levels can be located, then there also exists a need to determine the effectiveness of various techniques to reduce these levels. The current practice is not designed to provide the continuous, or quasi-continuous, information necessary to make these determinations. As a result, there exists a need for a device which can be deployed as a continuous monitor of respirable dust levels at fixed locations or located on mine equipment. The device can be expected to encounter excessive levels of dust, droplets from continuous water sprays, and elevated levels of methane.

Such a device, then, needs to be permissible for use in flammable atmospheres; it must be capable of tolerating excessive levels of dust while remaining operational for periods of days or weeks; and it must be impervious to the presence of water droplets or excessive moisture that can produce significant measurement error. In addition, the device should be insensitive to types of coal dust so that its use is not severely limited.

Certain techniques to develop such a device are the subject of intensive research efforts by the Bureau of Mines. Optical techniques suffer from contamination by moisture, the presence of water droplets, type of coal, and size distribution of the respirable dust. Combined, these effects can produce intolerable measurement errors for an optical device. Continuous gravimetric sampling, such as the use of a device called the Tapered Element

Oscillating Microbalance (TEOM),[1] or the measurement of pressure drop across a filter as it collects respirable dust, offer greater potential, but still suffer from the presence of water droplets, excessive moisture, and, to a lesser degree, elevated levels of dust of prolonged duration.

During the past few years, a novel sensor has been developed by the USBM to distinguish between fire smoke and diesel smoke so that early-warning fire detection in diesel-operated mines is not compromised by the combustion products exhausting from diesel engines. A description of this detector and its principle of operation can be found in Reference 2. Briefly, the detector capitalizes upon the fact that smoke produced from fires contain a significant volatile fraction while smoke produced from diesel engines contain only a minute (if any) volatile fraction. When volatile smoke particles pass through a small heated chamber called the pyrolysis tube (air temperature ~300 °C), they devolatilize even further producing smaller particles but in much greater number concentrations. Smoke particles from diesel engines are unaffected as they flow through the pyrolysis tube.

Coal dust, like fire smoke, contains a significant volatile fraction, ranging from about 16% to about 40% depending on the type of coal. Even though coal dust is much larger in particle size than fire smoke, it is also known to devolatilize when subjected to temperatures in excess of about 450 °C. However, owing to the larger particle size, the number concentration of respirable dust particles is extremely low. At a mass concentration of respirable dust equal to 2.0 mg/m^3, and a volume mean diameter of 3.0 μm, the concentration on a number basis is only ~110 particles/cm^3. A hypothesis was presented that, upon devolatilization of this dust, the resultant number concentration of smoke particles could possibly increase dramatically to the point where the smoke particle level could be easily measured by the ionization chamber of the detector. If the hypothesis proved correct, then two additional questions needed to be addressed.

First, how does the detector's response vary as a function of respirable dust concentration, and second, how does the detector respond as a function of the volatile fraction of the respirable dust? If the hypothesis proved correct, and if these two questions could be answered, then the potential of this detector for use as a continuous monitor of respirable dust could be determined. One benefit is immediately obvious — namely, that operating the pyrolysis tube at temperatures of ~450 °C removes any uncertainty due to water droplets and excessive moisture, a major impediment to other approaches under investigation. Also, previous field tests of the detector indicated that the detector can survive extended periods of continuous operation (3-6 months) before dust contamination within the detector begins to degrade its performance. Consequently, even though prolonged exposure to excessive respirable dust levels could shorten this period, continuous operation for periods of 2 to 6 weeks should be readily obtainable before maintenance and cleaning are warranted.

The operation of the pyrolysis tube at elevated temperatures does pose a problem for operation in flammable atmospheres. However, it should be noted that the component of the detector containing the pyrolysis tube is a separate component that can be made permissible, and that the ionization chamber detectors and associated electronics are intrinsically safe. Further, the ionization chambers use a single radioactive source of Americium 241 with a total activity of 5.0 microcuries, which is the exempt level for this radionuclide.

In order to test the hypothesis discussed above and to answer the two fundamental questions about the detector and its response to respirable dust, a series of tests were devised. These tests and their results are presented in the sections that follow.

EXPERIMENTAL

Figure 1 shows the system used to determine the response of the diesel discriminating detector (DDD) to various concentrations of respirable coal dust. Prior to each series of tests, the dust generating system was adjusted to yield a stable concentration of respirable coal dust within the test chamber (*3*). The instrument used

---

[1]Reference to specific instruments does not imply endorsement by the U.S. Bureau of Mines.

to measure this concentration was a Tapered Element Oscillating Microbalance (TEOM) with a sampling location located 180° from the sampling location of the DDD and at the same height and distance from the wall as the DDD.

Continuous recordings of the respirable dust concentration, as measured by the TEOM, were obtained for each series of tests. In general, these data indicate remarkably constant levels of dust within the test chamber during the duration of each experiment.

Experiments were conducted using samples of Pittsburgh seam coal and Pocahontas seam coal with volatilities of 36.5% and 17.1%, respectively. Because respirable dust may also include rock dust, tests were also conducted using respirable rock dust. For Pittsburgh seam coal dust, measurements were made over a concentration range from 0 to 7 mg/m^3; for Pocahontas seam coal dust, over the range 0 to 3.6 mg/m^3; and for rock dust, over the range 0 to 3.0 mg/m^3.

During each series of experiments, the valve of Figure 1 was adjusted to dilute the dust entering the pyrolysis tube. With the valve completely closed, the dust flow was shut off and the air entering the pyrolysis tube contained no particles or dust. With the valve completely open, it was found that a small dilution flow was always present, reducing the actual dust concentration entering the pyrolysis tube by about 7%. By adjusting the valve, the dilution factor could be varied continuously from 0 to 0.93, its maximum value. If Q_D is the dilution flow and Q_T is the total flow, then the sample flow is $Q_S = Q_T - Q_D$. The dilution factor, D_f, is defined simply as

$$D_f = \frac{Q_s}{Q_T}, \tag{1}$$

such that the mass of respirable dust entering the pyrolysis tube, M_{PT}, is related to the mass of dust in the test chamber M_{TC}, by the expression

$$M_{PT} = D_f \cdot M_{TC}. \tag{2}$$

THEORY

The DDD (see Figure 1) consists of two components. The first component is a small housing that contains a "TEE" connector and the pyrolysis tube. The second component is a larger housing that contains the two ionization chamber detectors, small pump, and electronics. Dust from the test chamber enters the first component where the total flow is split into two separate and equal flows. One flow path goes directly through a short tube to ionization chamber No. 2, while the second flow path goes through the pyrolysis tube to ionization chamber No. 1. The two chambers are identical and share a common source of Americium 241 with an activity level of 5.0 microcuries. By applying a constant voltage to the source electrode, equal ion currents are established in the two ionization chambers which the electronics converts to voltages, V_p, and V_μ, corresponding to chambers 1 and 2, respectively.

The response of each measuring ionization chamber is related to the particle diameter, d_o, and number concentration, n_o, of particles within the chamber, via the expression

$$\frac{\Delta V}{V_o} = \frac{V_o - V}{V_o} = 1 - \frac{1}{K_o\, d_o n_o}\,(1 - e^{-K_o d_o n_o}) \tag{3}$$

where K_o is a chamber constant $\cong 0.0025\ cm^2/p$,
d_o is the number mean average particle diameter (in cm),
n_o is the average particle concentration (in p/cm^3),
ΔV represents the voltage decrease, and
V_o represents the starting voltage when no dust or smoke is present.

When no dust or smoke is present, $V_p = V_\mu = V_o = 10.0$ volts.

For respirable dust that has not thermally decomposed, the voltage reduction in chamber 2 is negligible, and the unpyrolyzed voltage, V_μ, serves as a dynamic reference voltage. The voltage difference, ΔV_T, can then be written as

$$\frac{\Delta V_T}{V_\mu} = \frac{V_\mu - V_p}{V_\mu} = 1 - \frac{1}{K_o d_o n_o}(1 - e^{-K_o d_o n_o}) \tag{4}$$

where d_o and n_o now represent the number mean average diameter and number concentration of the smoke particles produced from the thermal decomposition of the respirable dust.

For values of $d_o n_o \leq 120$, the product $K_o d_o n_o$ is less than 0.3, and Equation 4 can be expanded in a Taylor series to yield

$$\frac{\Delta V_T}{V_\mu} \cong \frac{1}{2} K_o d_o n_o. \tag{5}$$

In the actual device, the response of the detector, V_{DDD}, equals $\frac{1}{5}\Delta V_T$, and since $V_\mu = 10.0$ volts,

$$V_{DDD} = K_o d_o n_o. \tag{6}$$

RESULTS AND ANALYSIS

DDD Response

For Pittsburgh seam coal, experiments were conducted at two levels of respirable dust within the test chamber ($M_{TC} = 3.8$ and 7.1 mg/m^3). The M_{TC} values measured by the TEOM are shown in Figures 2 and 3, for these average dust levels. By varying the dilution factor, D_f, it was possible to span the dust concentration range of $0.87 \leq M_{PT} \leq 6.7$ mg/m^3.

For the Pocahontas seam coal, measurements were made at values of M_{PT} between 0.82 and 3.6 mg/m^3. And for rock dust, measurements were made at values of M_{PT} between 0.8 and 3.0 mg/m^3. The response of the DDD to each of these dusts is shown in Figure 4.

For Pittsburgh seam coal, the measured response was found to vary linearly with mass concentration according to

$$(V_{DDD})_{PGH} = 0.032\, M_{PT}. \tag{7}$$

For Pocahontas seam coal, the measured response also varied linearly, but with a reduced sensitivity due to its lower volatility, according to

$$(V_{DDD})_{POCA} = 0.019\, M_{PT}. \tag{8}$$

For rock dust, the measured response was found to vary in a non-linear fashion according to

$$(V_{DDD})_{RD} = 0.018\left(1 - e^{-1.37 M_{PT}}\right). \tag{9}$$

Even though rock dust is considered to be an inert dust, it does devolatilize but at a much slower rate than coal dusts (*4*). Because of this, the detector response to respirable rock dust is not surprising.

The relative responses of the DDD to Pittsburgh and Pocahontas coal dusts do not scale directly with the volatilities from proximate analyses of the dusts. However, data obtained by Hertzberg (*5*), et al. for these two coals under conditions of very rapid heating by a CO_2 laser, yield volatilities of 57% and 34% for Pittsburgh and Pocahontas dusts respectively. If these two values are used, then the relative responses scale directly with these values. Assuming such a relationship to be valid over a range of volatilities, then the following general expression results:

$$V_{DDD} = 0.056\, f_v \cdot M, \tag{10}$$

where f_v is the laser volatile fraction, and M is the respirable dust concentration, in mg/m^3.

Reproducibility and Noise

It should be noted that the data of Figure 4 actually represent the average of at least 4 measurements at each dust concentration, and, in some cases, as many as 16 individual measurements. For Pittsburgh seam coal dust, the maximum deviation of any individual measurement from the response given by Equation 7 was $\pm 9.2\%$. For Pocahontas seam coal dust, the maximum deviation of any individual measurement from the response given by Equation 8 was $\pm 5.2\%$. For rock dust, the maximum deviation of any individual measurement from the response given from Equation 9 was $\pm 11.6\%$. These values represent the maximum deviations, and the average deviations were found to be less than one-half these values. These data would indicate that the measurements for a particular dust are quite uniform and reproducible.

During the experiments, the inherent noise of the detector was measured to be ± 0.003 volts. Using Equations 7, 8, and 9, the equivalent dust noise levels are found to be ± 0.094, ± 0.158, and ± 0.133 mg/m^3, respectively.

Impact of Rock Dust

In any real application, the detector will encounter mixtures of coal dust and rock dust, and as a result, the question naturally arises as to the meaning of the detector's response under these conditions. Clearly, it is not possible to separate the two dusts using this detector. Now, since the detector is less sensitive to rock dust than to coal dust, the presence of significant levels of rock dust would not contribute very much to the detector's response. If the detector response is assumed to always be that for pure coal dust, then the question becomes one of what fraction of rock dust would begin to introduce significant error in the measurement.

The Mine Safety and Health Administration has indicated that any continuous monitor of respirable mine dusts shall be capable of indicating the actual level to within $\pm 25\%$. If the detector response is based upon the assumption of pure coal dust, then at what level of rock dust does the indicated level read less than 75% of the actual level present?

If f_i represents the mass fraction of the total dust that is rock dust, then the detector response to mixtures of Pittsburgh seam coal dust/rock dust can be estimated by combining Equations 7 and 9. Similarly, the detector response to mixtures of Pocahontas seam coal dust/rock dust can be determined by combing Equations 8 and 9. The mass fraction of rock dust at which the indicated respirable dust level is less than 75% of the actual value can then be determined as functions of the total respirable dust. The results are shown in Figure 5.

For total respirable dust levels up to 10 mg/m^3, the mass fraction of rock dust that could be present and still allow for a maximum uncertainty of 25% exceeds 0.30. For total respirable dust equal to 2 mg/m^3, rock dust could account for 45% of the total mass in a mixture with Pocahontas seam coal dust and 64% of the total mass in a mixture of Pittsburgh seam coal dust. Typical fractions of rock dust found in respirable dust samples rarely

exceed 0.30, and are more typically in the range of 0.15 to 0.20. Although these estimates need verification, there is the clear indication that the detector response is not seriously degraded by the presence of rock dust.

Comparison With Theory

Now, Equation 6 indicates that, theoretically, the response of the detector should vary linearly with the product of the diameter and concentration of the smoke particles produced during the devolatilization of the coal dust. Experimentally, it is found that the response varies linearly with dust mass concentration and laser volatility of the coal dust according to Equation 10. Setting the two expressions equal yields the effective smoke diameter-concentration product produced during the devolatilization of coal dust, or

$$(d_o n_o)_{SMOKE} = 22.4\, f_v M. \qquad (11)$$

Equation 11 represents the effective smoke yield in terms of the diameter-concentration product of the smoke.

Pyrolysis Tube

All of the data were obtained at a total pyrolysis tube power level of 33.5 watts (20 volts and 1.675 amperes). This operating point was chosen arbitrarily and does not necessarily reflect the minimum power levels required for the pyrolysis of respirable coal dust. The pyrolysis tube consisted of 0.008 inch Nichrome 60 wire wound onto a ceramic rod and sealed in a hollow Pyrex glass rod. This coil/rod element was then inserted into the air space of a tube made by joining two brass Swagelok elbow fittings.

The inside diameter for this tube was 20 mm, its length, 60 mm, and total volume 1.885×10^4 mm^3. The volume of the air space is this total volume less the volume of the coil/rod element (424 mm^3), or 1.843×10^4 mm^3 (18.43 cm^3). The flow through the tube is one-half the total flow, or 16.67 cm^3/s. Dividing the air volume by the flow yields the residence time of the dust within the pyrolysis tube, or $t_{RES} = 1.105$ s.

Assuming that the heat flow is radial from the surface of the Nichrome coil, and that the surface temperature of this coil may be calculated from the known current and voltage applied to the coil, then the average radiant flux and air temperature are calculated to be 4.8 watts/cm^2, and 490 °C, respectively, within this air space.

Hertzberg (*6*) has shown that the characteristic time for a coal dust particle to devolatilize decreases as the coal dust particle diameter decreases when the particle is subjected to a constant level of radiant flux. If it is assumed that the maximum respirable coal dust particle diameter is 10 μm, then at an average radiant flux of 4.8 watts/cm^2, the characteristic time to devolatilize is $\sim$0.55 s. For smaller coal dust particles, the time is less. Consequently, for the pyrolysis tube, the residence time of dust particles within the tube must be greater than 0.55 s. The value obtained above is twice this minimum value. Coal is also found to undergo rapid devolatilization when subjected to air temperatures in excess of 450 °C. The average value obtained above is 490 °C. In reality, the devolatilization is probably some complex function of both the radiative flux heating and the convective flux heating at an elevated temperature. For the current experiments, the conditions within the pyrolysis tube are sufficient to satisfy constraints of either type of rapid devolatilization.

However, additional research needs to be done to determine the minimum temperature and radiant fluxes within the pyrolysis tube in order to effect the rapid devolatilization process. Minimizing the level of power consumption is important in order to address the problems of electrical power necessary for the detector to function and of permissibility for use of the detector in flammable gaseous environments.

It is worth noting that for diesel particulate matter, the temperature at which devolatilization occurs is 550 °C, considerably greater than the average temperature of 490 °C of the pyrolysis tube. Although experiments were not conducted using this pyrolysis tube to determine its effect on diesel particulate matter, it is

expected that the diesel particulates would not significantly devolatilize. Additional research needs to be done to verify this aspect of the detector operation.

CONCLUSIONS

The data obtained for the response of a diesel discriminating fire sensor to levels of respirable coal dust up to 6.7 mg/m^3 indicate a linear response with mass concentration and laser volatility fractions of the dust. For respirable rock dust, the response was found to be non-linear due to the fact that rock dust is much slower to devolatilize then coal dust. Using the estimated combined response of the detector to mixtures of coal dust/rock dust, mass fractions of rock dust in excess of 0.3 would not seriously degrade the performance of the detector. But additional testing needs to be done to verify these estimates. Although there is clearly a need for further research to refine the detector and its practical application, the data presented in this report indicate that the detector has potential for use as a continuous monitor of respirable dust in underground coal mines.

REFERENCES

1. Title 30, Code of Federal Regulations, Part 70.100, p. 487, 1993.
2. Litton, Charles D. Diesel Discriminating Fire Sensor. USBM IC 9206, 1988, pp. 28-32.
3. Marple, V. A., and K. L. Rubow. An Aerosol Chamber for Instrument Evaluation and Calibration. Am. Ind. Hyg. Assoc. J., v. 44, 1983, pp. 361-367.
4. Hertzberg, M., K. L. Cashdollar, I. Zlochower, and D. L. Ng. Inhibition and Extinction of Explosions in Heterogenous Mixtures. Proceedings of Twentieth Symposium (International) on Combustion. The Combustion Institute, Pgh., PA., 1984, pp. 1691-1700.
5. Hertzberg, M. and I. Zlochower. Devolatilization Rates and Intraparticle Wave Structures During the Combustion of Pulverized Coals and Polymethylmethacrylate. Proceedings of 23rd Symposium (International) on Combustion, The Combustion Institute, Pgh., PA, 1990, pp. 1247-1255.
6. Hertzberg, M., I. Zlochower, and J. C. Edwards. Coal Particles Pyrolysis Mechanisms and Temperatures. USBM RI 9169, 1988, 39 pp.

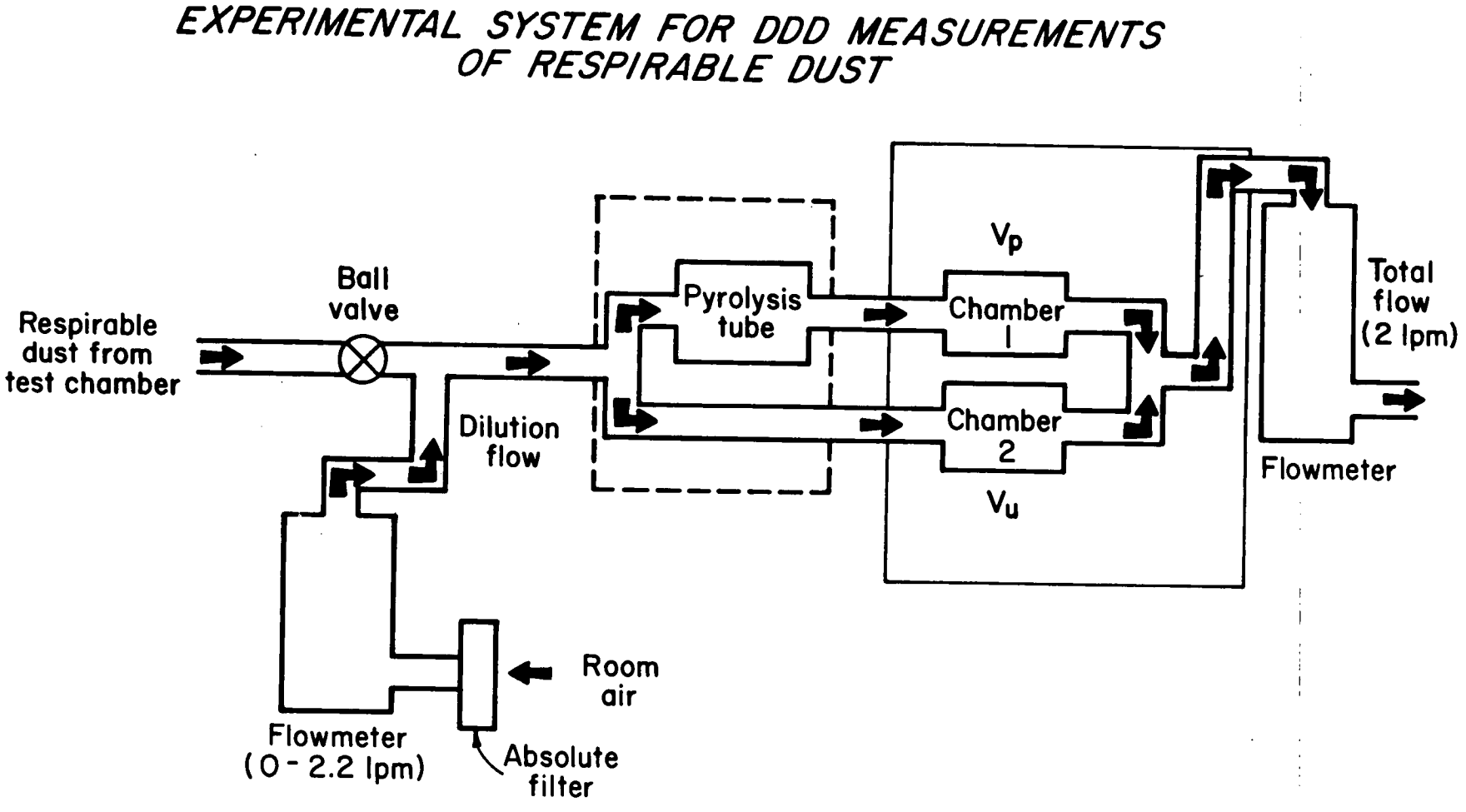


Figure 1. Schematic of the test configuration used to sample and measure the response of the DDD to levels of respirable dust produced in the aerosol test chamber.

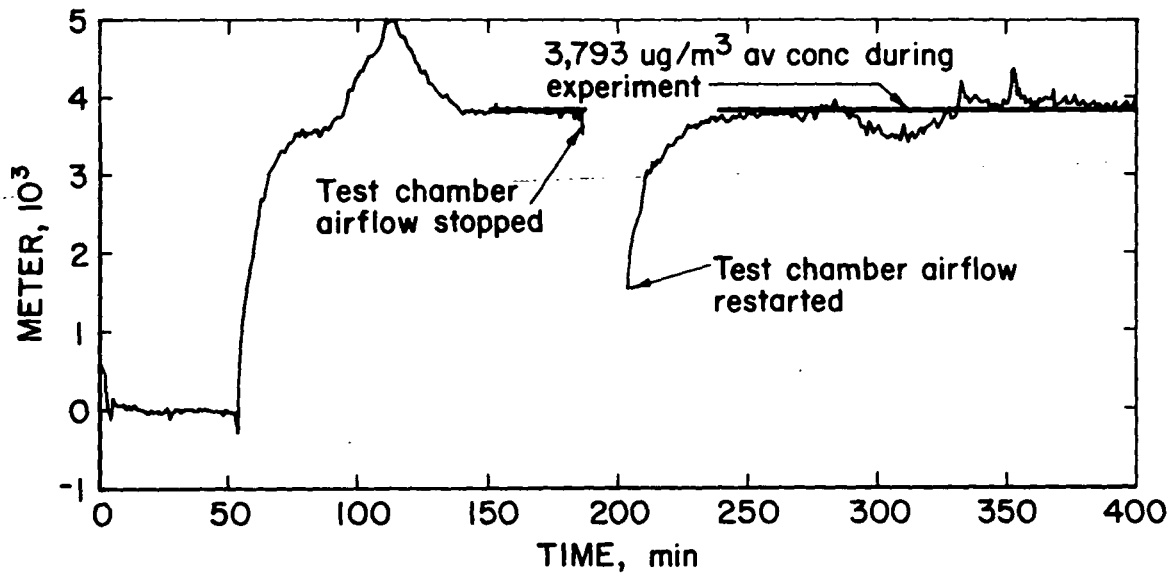


Figure 2. Respirable dust concentration in the test chamber as measured by the TEOM for Pittsburgh seam coal dust at a nominal concentration of 3.8 mg/m^3.

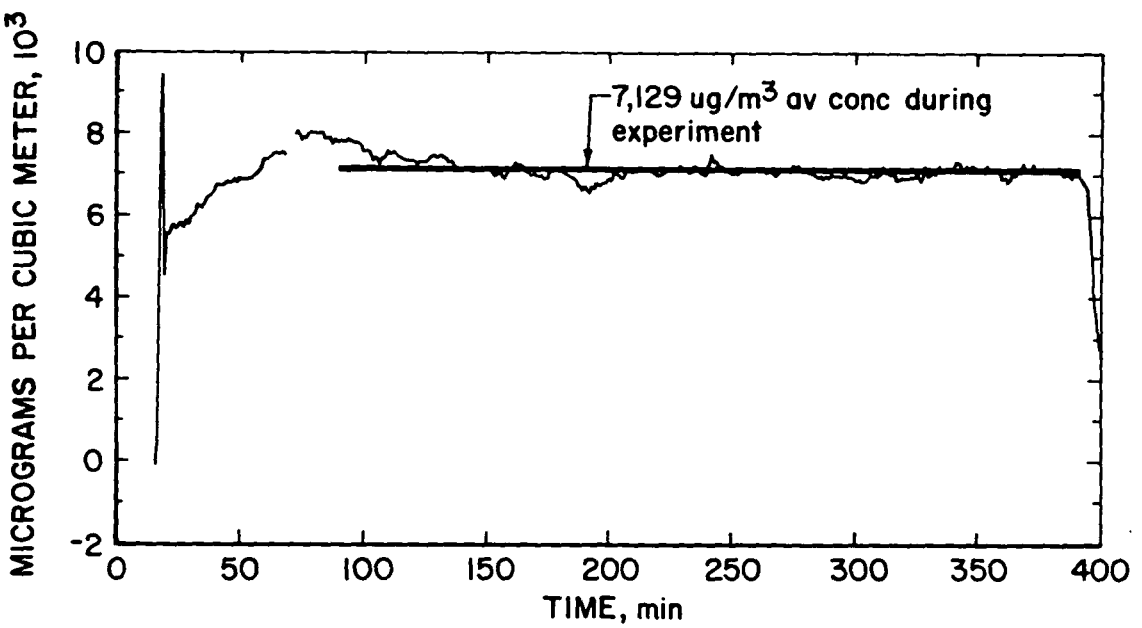


Figure 3. Respirable dust concentration in the test chamber as measured by the TEOM for Pittsburgh seam coal dust at a nominal concentration of 7.1 mg/m^3.

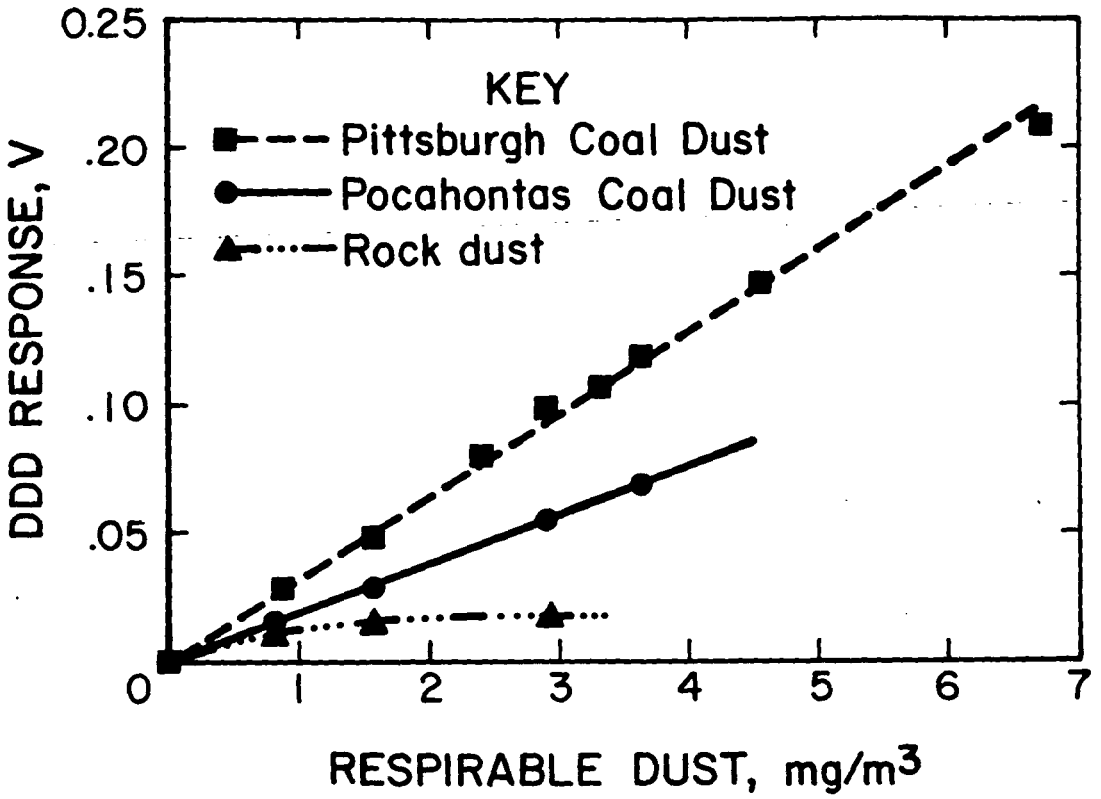
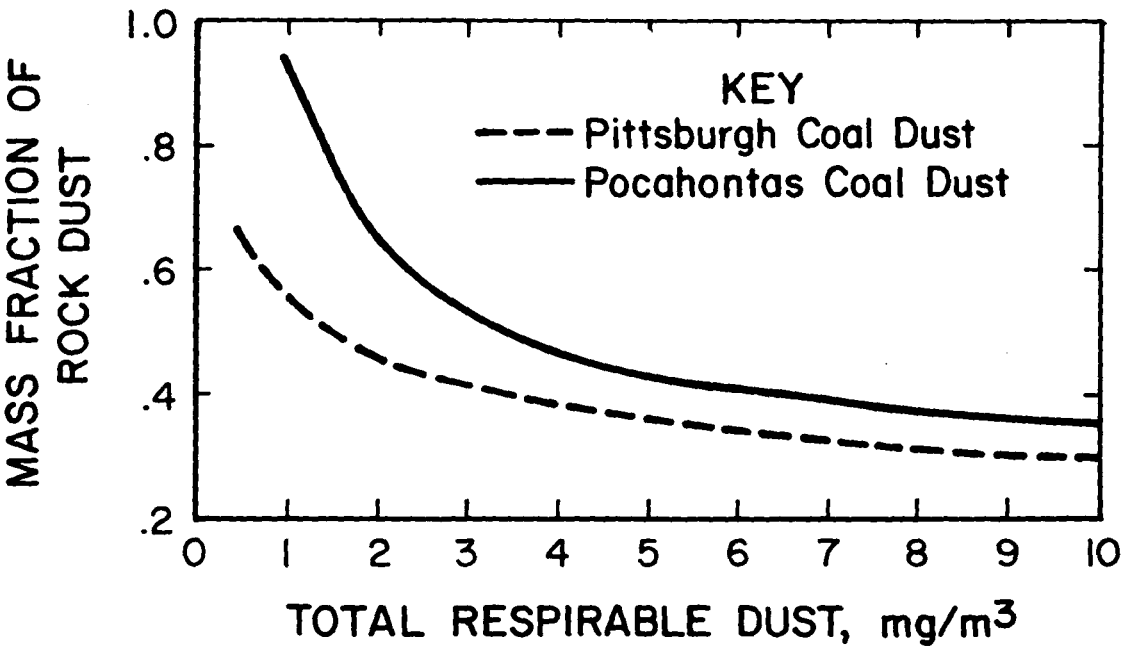


Figure 4. Measured response of the DDD to concentrations of Pittsburgh seam coal dust, Pocahontas seam coal dust, and rock dust.



gure 5. Estimated mass fractions of rock dust in coal dust/rock dust mixtures necessary to introduce a measurement error greater than 25%.

WORKSTATION-BASED NUMERICAL WEATHER PREDICTION SYSTEMS FOR OPERATIONAL USE AT THE KENNEDY SPACE CENTER

John Manobianco and Gregory E. Taylor
NASA KSC / Applied Meteorology Unit / ENSCO, Inc.
445 Pineda Court
Melbourne, FL 32940

John W. Zack
MESO, Inc.
185 Jordan Road
Troy, NY 12180

Walter A. Lyons and Craig J. Tremback
Mission Research Corporation / ASTER Division
P.O. Box 466
Ft. Collins, CO 80522

ABSTRACT

Weather support of ground and spaceflight operations at the National Aeronautics and Space Administration's (NASA) Kennedy Space Center (KSC) and the Air Force's Eastern Range at Cape Canaveral Air Station (CCAS) requires short range localized, accurate forecasts of winds, clouds, ceilings, fog, heavy rain, lightning, and low visibility. NASA and the Air Force have funded private corporations to develop and deliver numerical weather prediction systems to meet the operational forecasting needs at KSC/CCAS. The main component of the systems are the Mesoscale Atmospheric Simulation System (MASS) and Regional Atmospheric Modeling System (RAMS) models. These state-of-the-art mesoscale weather simulation models have been tailored specifically for short-range forecasting in the vicinity of KSC/CCAS and have been designed to run on high performance workstations. This paper provides an overview of MASS and RAMS as they have been designed for application to the forecasting problem at KSC/CCAS and highlights their existing and potential commercial applications.

GENERAL BACKGROUND

In a famous 1922 paper, eccentric British scientist L.F. Richardson envisioned the weather forecasting office of the future [1]. Since the atmosphere was a fluid, and thus could be predicted using the same equations of fluid dynamics developed by 19-century physicists and applied with great success in many fields, there seemed no reason why numerical weather prediction should not work. But the number of calculations required was staggering. Richardson's solution was to have a stadium-like arena filled with accountants, their slide rules and logarithm tables - each assigned a certain part of the computations, passing information back and forth under the command of a wand-waving conductor who would stand amidst the multitude.

While the endeavor promised long term employment for some ten thousand accountants, the Royal Meteorological Office declined to pursue the idea. But if one realizes that the Richardson's concepts actually describe the main components of today's modern computer, the idea was sound indeed. Sound enough that when the world's first digital computers became operational in the 1940's, John von Neumann and Jule G. Charney demonstrated the practicality of numerical weather prediction at the Institute for Advanced Study in Princeton, NJ [2]. Operational numerical weather forecasts commenced in the early 1950s, and since then, national meteorological centers have used the fastest and most powerful mainframe computers to predict hemispheric and global weather patterns on a daily basis.

The history of numerical weather prediction has been one of meteorology's success stories. The position and strength of larger scale weather features such as fronts, high and low pressure centers and jet streams can now be accurately predicted for a number of days in advance. But ironically, in the last two decades, it has been realized that predictions made by such computerized systems do not really forecast the 'weather' that is experienced at given point. Mesoscale, or intermediate scale weather systems, such as thunderstorm complexes, sea breezes, valley winds, and the like, are far too small to be resolved by today's operational numerical models. Researchers have intensively

studied such phenomena and began developing numerical weather simulation models suited for such 'mesoscale' phenomena.

SPECIFIC APPLICATION

The National Aeronautics and Space Administration (NASA) and United States Air Force have been conducting ground and spaceflight operations at the Kennedy Space Center (KSC) and Eastern Range at Cape Canaveral Air Station (CCAS) since the early 1960's. Weather support of operations at KSC/CCAS requires short range (< 24 h), detailed forecasts of winds, clouds, ceilings, fog and severe weather such as heavy rain and lightning. Forecasting these parameters for KSC/CCAS is a challenging task because the facilities in central Florida are located in a sub-tropical environment where there is an absence of significant large scale dynamical forcing during much of the year. Under these conditions, regional and local factors such as land/water boundaries, vegetation type and amount, and soil moisture play a dominant role in determining the short-term evolution of weather conditions.

The implementation of local, mesoscale modeling systems at KSC/CCAS is designed to provide accurate forecasts of specific thunderstorm-related phenomena such as precipitation and high winds thereby reducing downtime due to false weather advisories and alerts, hazardous weather events occurring without warning, and unnecessarily restrictive weather-based flight rules for manned and unmanned missions. In order to meet the forecasting needs at KSC/CCAS, NASA funded Mesoscale Environmental Simulations and Operations (MESO), Inc. to develop a version of the Mesoscale Atmospheric Simulation System (MASS). In addition, the Air Force funded Atmospheric Simulation Testing and Research (ASTER) division of Mission Research Corporation to develop the Emergency Response Dose Assessment System (ERDAS) which uses a version of the Colorado State University (CSU) Regional Atmospheric Modeling System (RAMS). NASA also funded ASTER to develop the Parallelized RAMS Operational Weather Simulation System (PROWESS).

The Applied Meteorology Unit (AMU) at KSC was formed in September 1991 by a tri-agency Memorandum of Understanding among NASA, the Air Force, and the National Weather Service. The AMU's mission is to evaluate and transfer research technology into the real-time weather support environment for the Range Weather Operations at CCAS and Spaceflight Meteorology Group at Johnson Space Center. The AMU is evaluating MASS and ERDAS and will be evaluating PROWESS, recommending and/or developing modifications to these systems as required, and transitioning them for operational use. This paper provides an overview of MASS, ERDAS, and PROWESS and discusses the existing and potential commercial application of these workstation-based mesoscale simulations systems.

MASS

System Overview

MASS is a limited area three-dimensional atmospheric modeling system. The system consists of: (1) a data preprocessor which can ingest a variety of data types and generate an initialization dataset for the model's prognostic variables, (2) a numerical-dynamical model which includes a detailed representation of atmospheric and surface processes, (3) a user interface which permits the user to configure the system parameters (e.g. the resolution of the grid) for a particular application, and (4) an output display package that permits the user to generate text and graphical displays of the model output in a variety of formats.

A detailed description of the current version of MASS can be found in the MASS Version 5.6 Reference Manual [3]. The atmospheric model is based upon the equations of conservation of mass, momentum and energy written in a finite difference form that can be solved on a Cartesian grid superimposed upon a map image plane. The equation set is hydrostatic (i.e. the vertical acceleration of air is assumed to be negligible) and is formulated in a normalized pressure vertical coordinate system which has the advantage of following the terrain. The model's prognostic variables are surface pressure, the two components of horizontal momentum, temperature, water vapor mixing ratio, cloud water/ice mixing ratio, rainwater/snow mixing ratio, surface interface (i.e. 'skin') temperature, subsurface temperature, surface cover layer water content, near-surface soil moisture content, root zone soil moisture content, and depth of snow on the ground. Each one of these variables must be assigned a value (i.e. initialized) for each grid point at the start of a simulation. This is the function of the data preprocessor.

The entire system (user interface, data preprocessor, atmospheric prognostic model and output display package) is designed to operate on a UNIX-based workstation. The system delivered to KSC in March 1993 was hosted on a Stardent 3000 vector processing workstation that at one time was the fastest floating point processor in its price class. However, the rapid pace of advancement in microprocessor technology has resulted in lower cost systems with

higher performance at the present time. As an example, a 24-h model simulation with a 45 km grid on a 55 x 50 horizontal matrix and 20 vertical layers (550,000 grid points) will execute in approximately 3.1 hours of CPU time on a 4-processor Stardent 3000 whereas the same simulation on a current state-of-the-art DEC Alpha Model 3000-600 workstation will require only 0.95 hours of CPU time.

MASS Initialization

Much of the real-time data required to initialize the MASS model is disseminated by the National Meteorological Center (NMC) in three separate data streams: (1) the Domestic Data Service (DDS), (2) the International Data Service (IDS), and (3) the Numerical Products Service (NPS). The DDS provides North American surface airways observations, ship and buoy reports, rawinsonde data and manually digitized radar reports. The IDS provides DDS-type information for the rest of the world. The NPS provides initialization analyses and forecast model output from the NMC and European Center for Medium-Range Weather Forecasts (ECMWF) operational models. The three data streams are transmitted to CCAS by the Zephyr Weather Information Service, Inc. These data are processed and stored in the McIDAS (Man-computer Interactive Data Analysis System) Interactive Data Display System (MIDDS) at CCAS. The MIDDS also processes data gathered by sensing systems such as the local meso-network of towers and the 50 MHz Doppler radar wind profiler operated in the vicinity of KSC. The real-time MASS modeling system developed for KSC/CCAS is depicted in Figure 1.

Real-Time MASS Configuration

The MASS data preprocessor and model have been run twice daily on the Stardent 3000 workstation since December 1993. The daily model forecast and data assimilation schedule consists of two 24-h coarse grid and two 12-h fine grid runs per day. The 24-h coarse grid run designated C00 is initialized with 0000 UTC data and assimilates hourly surface and manually digitized radar (MDR) data from 0000-0400 UTC. The 12-h fine grid run designated F12 is initialized with 1200 UTC data and assimilates 1300 UTC surface and MDR data. The 12-h forecast from C00 (valid at 1200 UTC) provides the first guess fields for the objective analysis of 1200 UTC data used for F12 initialization. Additionally, the 12-24 h forecast fields from C00 are used to specify boundary conditions for the F12 run. The cycle is repeated using 1200 UTC data to initialize the 24-h coarse grid run designated C12 and 0000 UTC data to initialize the 12-h fine grid run designated F00. The attributes and time table for the MASS configuration are summarized in Figure 2.

Commercial Applications of MASS

There are two potential modes of commercial application for MASS. They are summarized in Table 1. In the first mode, an integrated software/hardware system would be delivered to the end user. The user could then configure the system, execute the forecasts and display the results on the system. The supplier would then provide maintenance, support and possibly a feed of simulation system-ready data. In the second mode of commercial applications, the simulation system would be operated at a central site and the information generated by the system would be supplied to users as an information service. The simulation system would be configured to execute multiple simulation cycles per day and automatically disseminate localized weather information via transmissions of: (1) digital or graphical data over digital communications networks to the user's computer system or (2) images to the user's facsimile reception hardware.

ERDAS / PROWESS

ERDAS System Overview

The purpose of ERDAS is to provide emergency response guidance to operations at CCAS/KSC in case of a hazardous material release or an aborted vehicle launch. A major part of the system is the use of RAMS to provide local weather forecasts of the wind and turbulence fields for input to the dispersion models [4]. Although the system has been designed to run on local workstations at KSC/CCAS, it can and will be applied to a wide range of other emergency response and/or local weather forecasting applications.

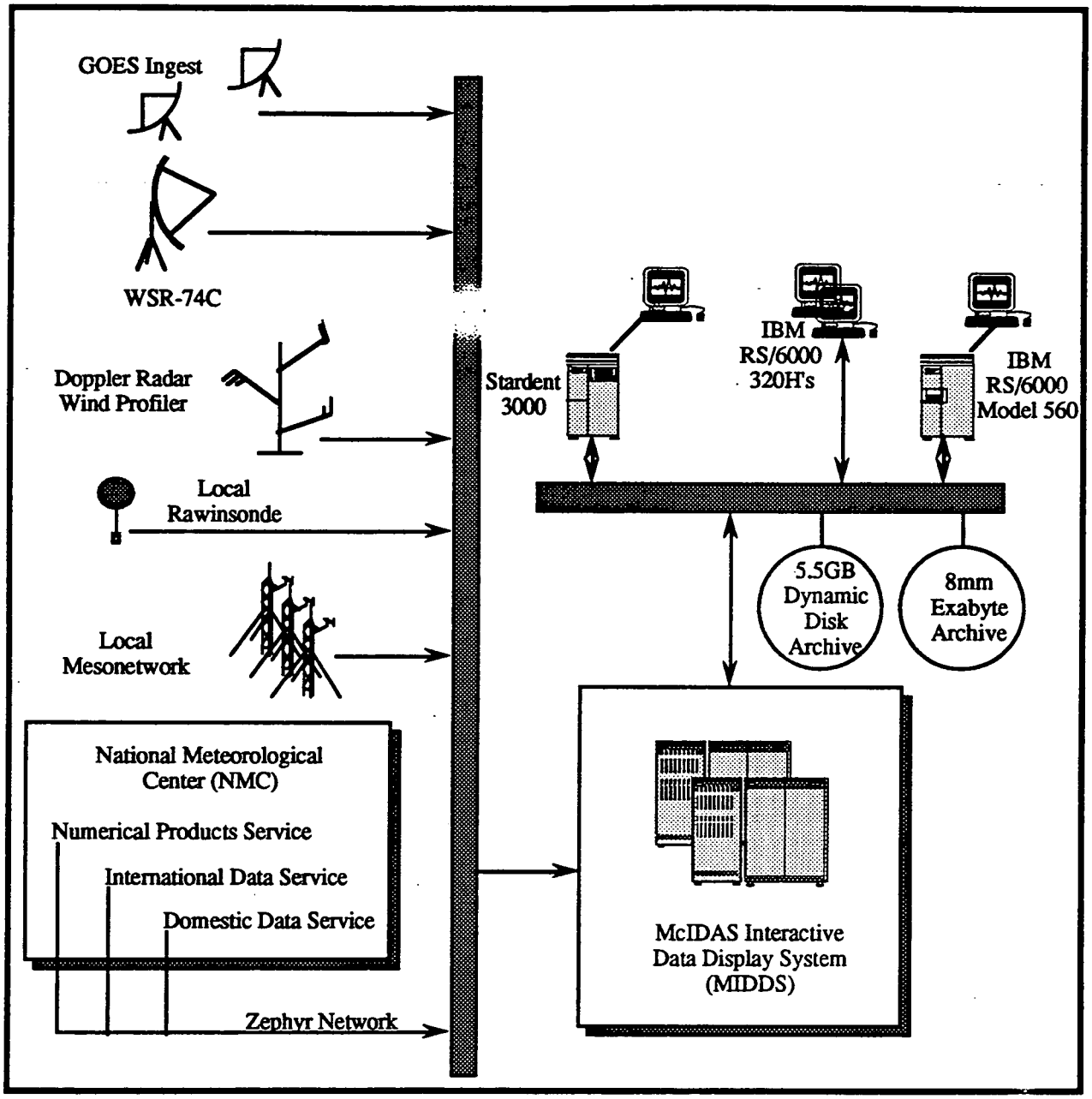


Figure 1. Schematic depiction of the real-time data ingestion and simulation system using MASS.

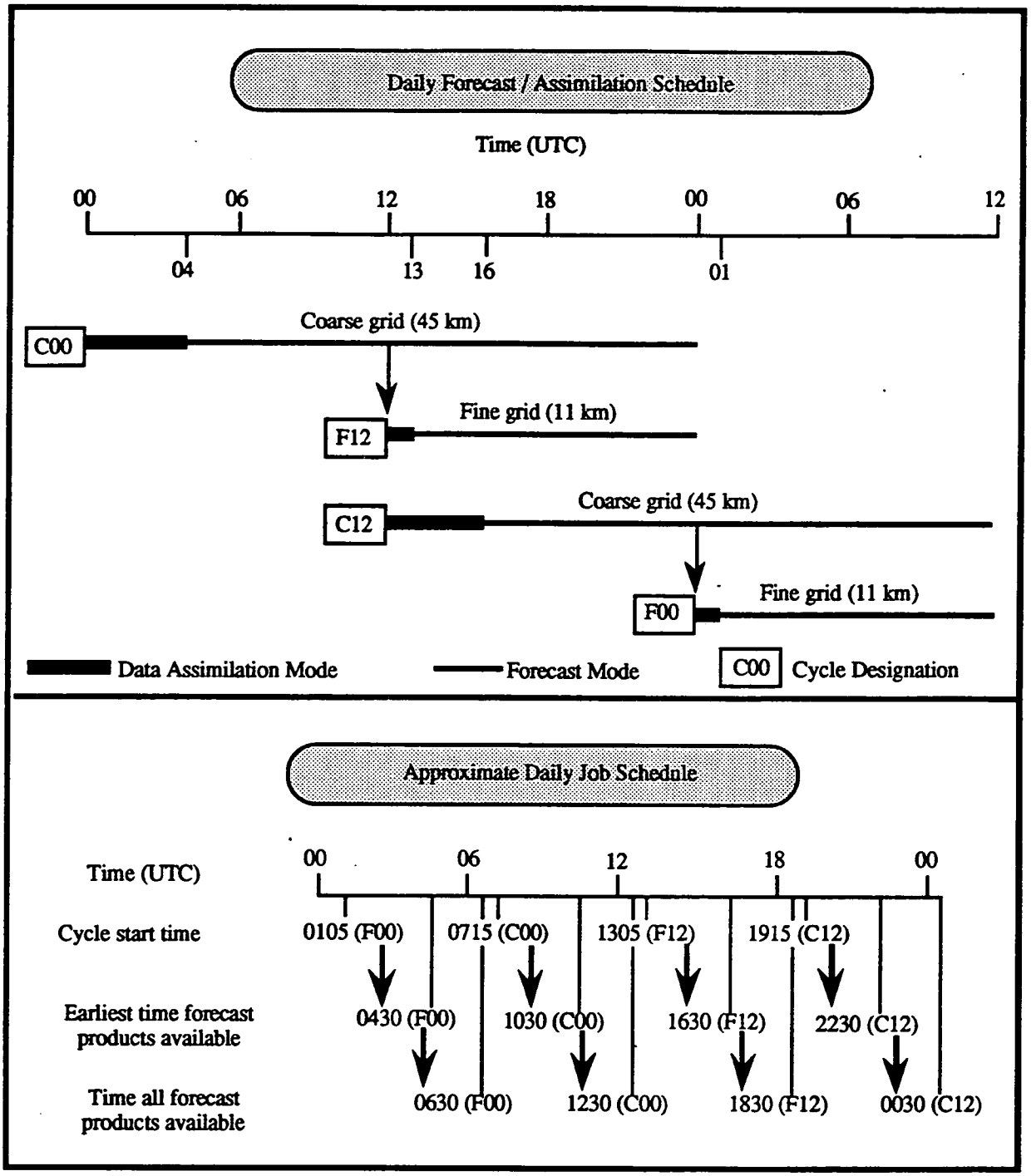


Figure 2. A schematic of the real-time daily forecast, data assimilation, and job schedule for the MASS pre-processor and model.

Table 1. Examples of commercial potential for a workstation-based high resolution atmospheric simulation system.

Mode 1: Integrated Software/Hardware System delivered to the end user and operated at the user's site to ingest atmospheric data, execute simulations and view simulation output. The cost of the system will decline as computational costs decrease during the next few years.	
Potential Users	*Applications*
University atmospheric or earth science departments	*Instructional aid:* simulations can be generated by instructors or students to illustrate or experiment with the principles of atmospheric physics. *Research tool:* simulations can be used to analyze specific cases or types of phenomena.
Private forecast companies	A forecast tool to provide additional information to customized local forecasts.
Television stations	High resolution forecasts over viewing area can be presented as a still or animated display during news broadcasts.
Commodities Dealers / Brokers	Simulation output can be used to provide updated forecasts over specific regions of interest for commodity trading decisions.
Electric & Gas Utilities	Use simulation output as input into dispersion models which can generate estimates of the areal extent and concentration of substances emitted from power plants
Mode 2: Integrated Software/Hardware System executed at an operations center to generate high resolution forecasts. Information from this system is then automatically delivered to the users in formats and quantities appropriate for a particular application.	
Potential Users	*Applications*
Resorts (i.e. ski, boating etc.)	A single page fax of local weather conditions for the planning of snowmaking operations at ski resorts; local wind forecasts for boaters.
Construction companies	A single page fax of local forecast data to permit the planning of weather-sensitive components of a construction project.
Individual User / Weather Hobbyists	Automated transmission (fax page or digital dataset) of fields of selected meteorological variables over local areas for general information.
Land-based transportation (i.e. trucking companies, government entities responsible for highway maintenance, etc.)	Provide information about local weather conditions along travel routes of trucking companies; high resolution local information transmitted to GIS systems operated by local governments and highway maintenance department for use in planning of snow removal, debris removal after severe storms and other operations.

ERDAS uses several key software components to accomplish its tasks: 1) RAMS to provide local meteorological forecasting and analysis, 2) HYPACT (HYbrid Particle And Concentration Transport package) to simulate the dispersion of the hazardous materials using the RAMS meteorological output, 3) visualization software to display and analyze the results from RAMS and HYPACT, and 4) a customized Graphical User Interface (GUI) to provide a consistent and easily-used interface for all of the software components. These components provide the control and visual depiction of the results of the local weather forecasts and resultant dispersion estimates.

Even with their relatively impressive performance, workstations have limitations. As the mesh size decreases, run times increase dramatically. Certain RAMS options, such as explicit cloud microphysics, also require substantial computational resources. For these reasons, the RAMS configuration in the initial ERDAS system at KSC/CCAS is limited to a 3 km inner mesh size and does not explicitly treat convective cloud formation. These restrictions will disappear as faster processors and clustered workstation systems become available at KSC/CCAS. The initial goal is to provide forecasters with new 24 hour forecast model output within six hours of initialization. The grid configuration for RAMS, therefore, has been chosen to be a 60 km mesh covering the southeastern United States, a 15 km grid covering most of the Florida peninsula, and a 3 km mesh covering a 110 km x 110 km region around KSC/CCAS. The 3 km grid represents the coarsest mesh that can resolve adequately both the sea breeze and island/estuary perturbations. As workstation processors become even more powerful, spawning a finer mesh grid (approximately 1000 meters) directly over KSC/CCAS can be readily accomplished.

For the local meteorological forecasts, RAMS is initialized and run twice daily from the standard synoptic data times at 0000 UTC and 1200 UTC. Thus, there is always a forecast available from which to run the dispersion estimates in case of an accidental or planned contaminant release. Fixed data inputs into RAMS include topography, land use, and climatological sea surface temperature patterns. RAMS has a non-homogeneous initialization with non-stationary boundary conditions since temporal variability in the outer boundary conditions is needed. Data sets such as NMC gridded analyses and forecasts, rawinsonde and surface data, as well as local meso-network and 50 MHz Doppler radar wind profiler data are used by the RAMS isentropic analysis package to initialize the model.

PROWESS System Overview

Previous research has shown that small inhomogeneities in surface heating such as those produced by the islands and estuaries surrounding KSC frequently affects the local weather by producing solenoidal circulations which can force cloud lines, rain showers, and even thunderstorms. Termed the Merritt Island Effect (MIE), these circulations occur on a significant fraction of the summer days with the Merritt Island Thunderstorms (MIT) occurring about once per week. The MIT is difficult to forecast as it is a local effect that is separate from the main Atlantic sea breeze front and frequently occurs earlier in the day. Since the MIT is often marginally electrified, it represents a recurring decision making problem for safety personnel in charge of outdoor work. False alarms due to the uncertainties in the forecasts creates a substantial loss of productivity in the summer months at KSC.

ASTER is currently developing a three-dimensional, turn-key thunderstorm forecasting system for KSC [5]. Called PROWESS for Parallel RAMS Operational Weather Simulation System, it will be driven by a similar graphical user interface as the ERDAS system. The RAMS grid configuration will be four nested grids with grid spacings of 60, 15, 5, and 1.25 km. The finest grid will cover a 120 km x 80 km region over the KSC area. The initial focus of the forecasting effort will be to simulate the development of the Merritt Island Thunderstorm.

An improved cloud microphysics module featuring substantially faster run times than previous RAMS modules will be used, along with other code efficiencies. Even so, with the finer mesh sizes than the ERDAS configuration, the computational requirements will be much greater. Current testing indicates that a computer system of between 250 and 325 megaflops will be required to produce forecast meteorology in about one-quarter of real time. This throughput will be attained using a RISC workstation cluster as a parallel computation platform networked with standard Ethernet connections. As a development system, eight IBM RISC/6000 workstations consisting of one Model 370 and seven Model 250 Power PCs are being used.

Commercialization of RAMS

Given that the RAMS model as embodied within the ERDAS and PROWESS, along with other numerical weather forecasting codes, would appear to be powerful tools, the question arises why such systems have not, until recently, been widely used, especially in the air quality and emergency response areas?

There are a variety of reasons, some based on true technological limitations, others based on regulatory and bureaucratic inertia. In the later regard, many federal regulatory agencies within the previous decade promulgated rules and guidelines mandating the use of air pollution dispersion models by a wide variety of organizations. Ten years ago computational choices were limited to a large mainframe computer (too expensive for all but a very few facilities), a minicomputer or the (then) new personal computer (PC). Many nuclear power plants opted for minicomputer systems, but even with these relatively powerful machines (by the standards of the early 1980s), real-time local weather forecasting was not feasible. Dispersion estimates were relegated to using local weather observations, often from a given on-site tower only 60 or 100 m in height. Given the great simplifications required to run dispersion models in near-real time, model design often employed assumptions viewed as conservative - predicting the very worst case impacts. This may help define the maximum possible impact of a radiological or toxic incident, but at the same time could result in the ordering of evacuations (itself an exercise with a predicted number of casualties in larger communities) which may in fact might not be warranted.

Obtaining permits for new pollution sources (a new factory or power plant) required estimation of the potential impacts of the several major emitted pollutants. The United States Environmental Protection Agency (USEPA) orchestrated the development of a suite of highly simplified dispersion models which could be run using only local weather observations as input. These model codes had the virtue of running on affordable computer platforms such as PC's and being relatively easy to use. Unfortunately, they also employed, in many cases, grossly oversimplified meteorology and highly conservative assumptions. By erring on the side of caution, they could produce estimates of

pollution impacts quite a bit larger than likely to occur. While perhaps desirable from the viewpoint of assuring absolute protection of public healthy and safety, these may well impose unnecessary costs and restrictions upon industry.

Thus for both emergency response and routine pollution permitting, a set of methodologies using simplistic models scaled to the affordable technologies of more than a decade ago became the norm. There has been little regulatory pressure aimed at changing these modeling methodologies even though much more sophisticated approaches (as exemplified by ERDAS) are available at increasingly lower cost.

It would appear the air quality modeling researchers must continue to demonstrate the utility of such codes in order to raise the consciousness of the regulatory branches of government in order to employ what are believed to be superior approaches. Aside from the problem of regulatory inertia, such models have not been used in the past for various other (real or perceived) reasons.

- The codes are very complicated and require considerable expertise to use, expertise only generally found within government and university research labs.
- Few practitioners in the field knew how to obtain such codes, or the training for running them, and where to seek guidance on interpreting model output.
- Little work had been done to evaluate the performance of the models.
- The data required to initialize the models was often hard to obtain and difficult to incorporate into the models.
- The models produce such voluminous output that the end user was overwhelmed by the shear volume of the results and had few tools to visualize and synthesize the results.
- The computing environments had no "user friendly" features such as Graphical User Interfaces (GUIs) which facilitate the initialization and running of the model and interpretations of the results.

However, as demonstrated by MASS, ERDAS and PROWESS, the situation is changing rapidly. A number of local-area weather forecasting codes are available from both government and commercial sources. Workstation technology is advancing at such a rapid pace that 'desk top' weather forecasting is a reality. The advent of networks of workstations using parallelized versions of the code now allow for even greater computational resources to be applied to a problem. By using workstation clusters, the model can be used at a number of levels of sophistication. If a user's requirements change so that faster turn around times, more sophisticated treatment of physical processes or finer spatial resolutions are desired, than either additional and/or faster processor nodes can be added to the system at minimal incremental cost.

Other relevant developments include the increasingly easy access to meteorological information required to initialize local weather forecasting systems. A number of commercial services now exist which make these data streams available through direct satellite broadcast or modem. Thus the same data resources available to national meteorological services are now available to a wide variety of users worldwide.

Most importantly, commercial firms (such as MRC and MESO, Inc.) are now providing to military, government and industrial end users all of the components required to effectively run a local weather forecasting system. Complete turn-key systems are available, including training and ongoing support.

Present and Future Commercial Applications of RAMS

Within the next year, the acquisition of a workstation cluster producing more than 100 megaflops throughput for a price of under $100,000 (less than a mid-range minicomputer of a decade ago) will be commonplace. This will allow for a wide variety of weather forecasting, emergency response and air quality modeling applications at the field office level. Codes such as MASS and RAMS can be purchased commercially, along with manuals, technical documentation, maintenance contracts and hot-line user support. New releases are routinely made available which take advantage of advances in both computational and atmospheric sciences.

At the research level there are over 60 users of the RAMS code worldwide. The advent of affordable high performance computing also allows for routine air quality modeling and operational forecasting deployment. In addition to ERDAS and the PROWESS systems, other current systems based on RAMS include:

- Weather News, Inc. Tokyo, Japan: the world's largest private weather forecasting company runs RAMS on a 10 km mesh over the Kanto Plain and at coarser mesh over Japan and surrounding waters. This provides highly localized forecasts for their utility, transportation and industrial clients.

- Theater Forecasting System: the ability to have detailed forecasts on the scale of military operations would have obvious benefit to military commanders. Codes such as RAMS are currently be evaluated as to their suitability to provide real-time forecasting support to improve tactical decision making in a variety of environments.

- Lake Michigan Air Directors Consortium: As a result of a Federal Court order, the states of Indiana, Michigan, Illinois and Wisconsin have agreed to developed a scientifically defensible plan to reduce emissions of smog forming pollutants in the Lake Michigan basin. In order to accomplish this, a complete system for modeling the pollution emissions, meteorology and smog chemistry of the Lake Michigan basin has been developed. RAMS has provided the meteorological component. RAMS is used to drive a state-of-the-art regional photochemical grid model aimed at replicating ozone levels in recent severe episodes. The alternative to meteorological modeling would have been to mount a prohibitively costly observational program in an 80,000 km^2 region comprised of more than 50% water.

- Iberdola, S.A.: This Spanish electric utility desired to estimate the impacts of a fossil fuel power plant located within complex mountainous terrain in northern Spain. The RAMS code was used to predict the highly localized airflows of the region and then to simulate the dispersion of pollutants released from the plant.

- Litigation Support: A major chemical spill in a mountain valley resulted in the filing of hundreds of law suits against a large transportation company. Many of the alleged victims of the dispersing toxic gas cloud resided in locales that did not appear to have been impacted by the gas. Since there were virtually no meteorological or pollutant measurements available in this remote region, RAMS was used to simulate the local weather patterns over a multi-day period. In addition, the model output was available to simulate the dispersal of the toxic gas cloud. While the law suit settled before going to court, such simulations are likely to play an increasing role in environmentally-related litigation.

The above represent selected examples of ongoing uses of models such as RAMS. There are many more potential applications, especially within the realm of industrial and commercial applications. It should be remembered that codes such as RAMS produce very detailed local forecasts of the fundamental meteorological variables of temperature, wind, moisture and pressure on a scale of kilometers with updates on the order of minutes. As such, the model can drive a variety of other modules of relevance to pollutant dispersion, energy production, transportation, agriculture and natural resources management.

Some additional examples of RAMS applications which could be implemented using existing workstations and user interfaces include:

- Utility Load Forecasting: Energy use is critically dependent upon temperature. Even a several degree error in forecasting daily maximum or minimum temperatures can result in a serious under or over estimate of the required generating capacity for an electrical utility. To meet such demands, many utilities will purchase power from other utilities. However, last minute purchases can often be at a premium price. Inability to meet demand can result in even worse scenarios: brownouts, rolling blackouts, or as occurred in Tokyo in the mid-1980s, a nearly citywide blackout.

- Water Resource Management: RAMS ability to provide very detailed forecasts of thunderstorm rainfall or mountain snowfalls can be useful in the management of irrigation and flood control systems.

- Regional Snowfall Prediction: Snowfall can be highly localized, especially in mountainous regions and around the Great Lakes. Models such as RAMS can be used to predict variations in snow on a county by county basis or even finer scales. Transportation and highway maintenance operations can be properly forewarned of impending local heavy snows and blizzards. Utilities can be greatly impaired by heavy snows, and deploying maintenance crews before hand can often assist in the rapid restoration of service.

- Emergency Response: Realistic simulations of dispersion of radionuclides or toxics from design accident scenarios in nuclear and chemical plants can result in more cost-effective design or sheltering and evacuation plans and implementing appropriate safety systems. The quality of actual forecasts in the case of an emergency can be significantly improved.

REFERENCES

[1] Richardson, L.F., 1922: 1922: Weather Prediction by Numerical Process. Cambridge University Press, 236 pp. (Reprint with a new introduction by Sidney Chapman, Dover Publications, 1965, 236 pp.).

[2] Charney, J.G., R. Fjortoft and J. von Neumann, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, 6, 309-318.

[3] MESO, 1994: *MASS Version 5.5 Reference Manual*, 118 pp. [Available from MESO, Inc., 185 Jordan Road, Troy, NY 12180].

[4] Lyons, W.L., and C.J. Tremback, 1993: A prototype operational mesoscale air dispersion forecasting system using RAMS and HYPACT. *Preprints 86th Annual Meeting and Exhibition*, Air and Waste Management Association, Denver, CO, 16 pp.

[5] Lyons, W.L., C.J. Tremback, R.L. Walko, R.A. Pielke, and W.R. Cotton, 1994: Design of an operational forecasting system for localized and sea breeze thunderstorms at the Kennedy Space Center. *Preprints 10th Intl. Conference on Interactive Information Processing for Meteor., Oceanography, and Hydrology*, Amer. Met. Soc., Nashville, TN, 213-218.

THE USE OF MICROWAVE INCINERATION TO PROCESS BIOLOGICAL WASTES

Sidney C. Sun
Regenerative Life Support Branch
NASA Ames Research Center
Moffett Field, CA 94035-1000
(415) 604-4835

Venkatesh Srinivasan
The Bionetics Corporation
NASA Ames Research Center
Moffett Field, CA 94035-1000
(415) 604-1417

Mark E. Flynn
The Bionetics Corporation
NASA Ames Research Center
Moffett Field, CA 94035-1000
(415) 604-1335

ABSTRACT

The handling and disposing of biological or biohazardous solid waste matter is a difficult problem for hospitals, research laboratories, and industry. The National Aeronautics and Space Administration (NASA) faces the same challenge in space. To meet this challenge NASA is developing regenerative life support systems that will process and recycle waste materials into resources that can be used to sustain humans living in space for extended durations. Plants provide critical functions in a regenerative life support system in that they photosynthesize carbon dioxide and water into complex organic molecules and oxygen. The edible portions of plants can provide food for the crew and the inedible portions can be processed into recyclable materials. The Regenerative Life Support Branch at NASA Ames Research Center has been evaluating the microwave incineration process as a means of oxidizing inedible plant matter into carbon dioxide and water that can be used by plants in plant growth chambers. Microwave incineration is a technology that is simple, safe, and compact enough for home use. It also has potential applications for institutions that produce biological or biohazardous waste. Researchers at NASA Ames Research Center have run several sets of tests with a commercially available Japanese microwave incinerator to establish its viability in processing biological material. Plant matter was incinerated to produce ash, reducing the waste to 13% of its original weight. One goal of the tests was to determine whether or not the incinerator generates toxic compounds as a byproduct of the combustion process. This paper will describe the results of the tests and the preliminary analyses of the resulting ash and exhaust gas, as well as the significance of the results and the implications for commercial applications of the technology.

INTRODUCTION

Processing of solid wastes is becoming more important as the availability of landfill space dwindles. Incineration of solid waste is one way to reduce the amount of waste that is discarded. Biological wastes form a special category of solid wastes. These wastes can come from hospitals, research centers, universities, veterinary clinics, and veterinary schools. Wastes generated in medical institutions and biological research laboratories are notable because much of the waste is

hazardous. Hospitals generate 4.54 to 9.08 kg/bed/day of waste, 10% of which is infectious [1]. Efforts to control costs, the danger of AIDS, and an increased concern for occupational safety are driving medical institutions towards using more disposable items. Consequently the volume of infectious medical wastes has increased [2]. To solve their refuse problems, hospitals are turning increasingly to incinerators to handle pathological, infectious, and administrative wastes.

There are many different types of incineration systems in use right now [3]. These systems are usually sized to serve municipalities or large buildings and are designed to process large amounts of material at a time. Smaller units suitable for individual households, laboratories, or medical practices are less common. The use of microwaves to support incineration is even more novel in the United States.

Researchers at NASA have been looking at solid waste processing technologies that could be applied to human space exploration. Long duration missions on space stations, planetary/lunar exploration, and settlement missions will require large amounts of air, food, and water to sustain humans in space. The launch and onboard storage costs of these resources is prohibitive. These factors drive the need to minimize the amount of materials that need to be launched into space and maximize the recycling of waste products into usable resources.

The Laboratory-Scale Controlled Ecological Life Support System (Lab-Scale CELSS) is being developed at NASA Ames Research Center to study the issues associated with recycling gaseous, liquid, and solid wastes within a closed system. Phase I of Lab-Scale CELSS is focused on recycling the carbon in a closed system containing plants and a solid waste processor. One of the areas currently under investigation is determining if plants can be sustained in an atmosphere of exhaust air from a solid waste processor. The solid waste processor being used is a microwave incinerator. The incinerator exhaust air needs to have sufficient levels of carbon dioxide to support photosynthesis. In addition, harmful compounds such as carbon monoxide, sulfur oxides, and nitrogen oxides must be held below levels at which they are toxic to plants.

Incineration is not without its own set of problems. One problem is the emission of toxic pollutants as byproducts of the process. It is clear that if new incineration technologies are to be used in the public sector, then the issue of toxic pollutants must be addressed.

TECHNOLOGY

The solid waste processor that is being used in the Lab-Scale CELSS Project at NASA Ames Research Center is a microwave incinerator manufactured in Japan. This type of incinerator is being used in Japanese households to incinerate garbage, but its use in the United States has been very limited. The microwave incinerator is 50 cm wide, 38 cm deep, and 86 cm high [4]. The maximum power consumption of this unit is about 2000 W. The incinerator can process a batch of material up to 2 kg in mass and 5 ℓ in volume.

Waste matter is placed into a processing container that slides out of the incinerator on a set of rails, as shown in figure 1. The container, when slid back into the incinerator, serves as the primary combustion chamber. There is a microwave susceptor in the bottom of the chamber. Waste matter in the primary combustion chamber is heated using microwaves at 2,450 MHz. When the waste is heated, it begins to lose moisture in the regions near the top surface. As drying on the surface of the waste progresses, microwaves penetrate further into the waste until almost all of the moisture is removed. After most of the solid waste has dried, the microwave susceptor within the chamber absorbs the microwaves, heating the material around it, causing decomposition and partial oxidation.

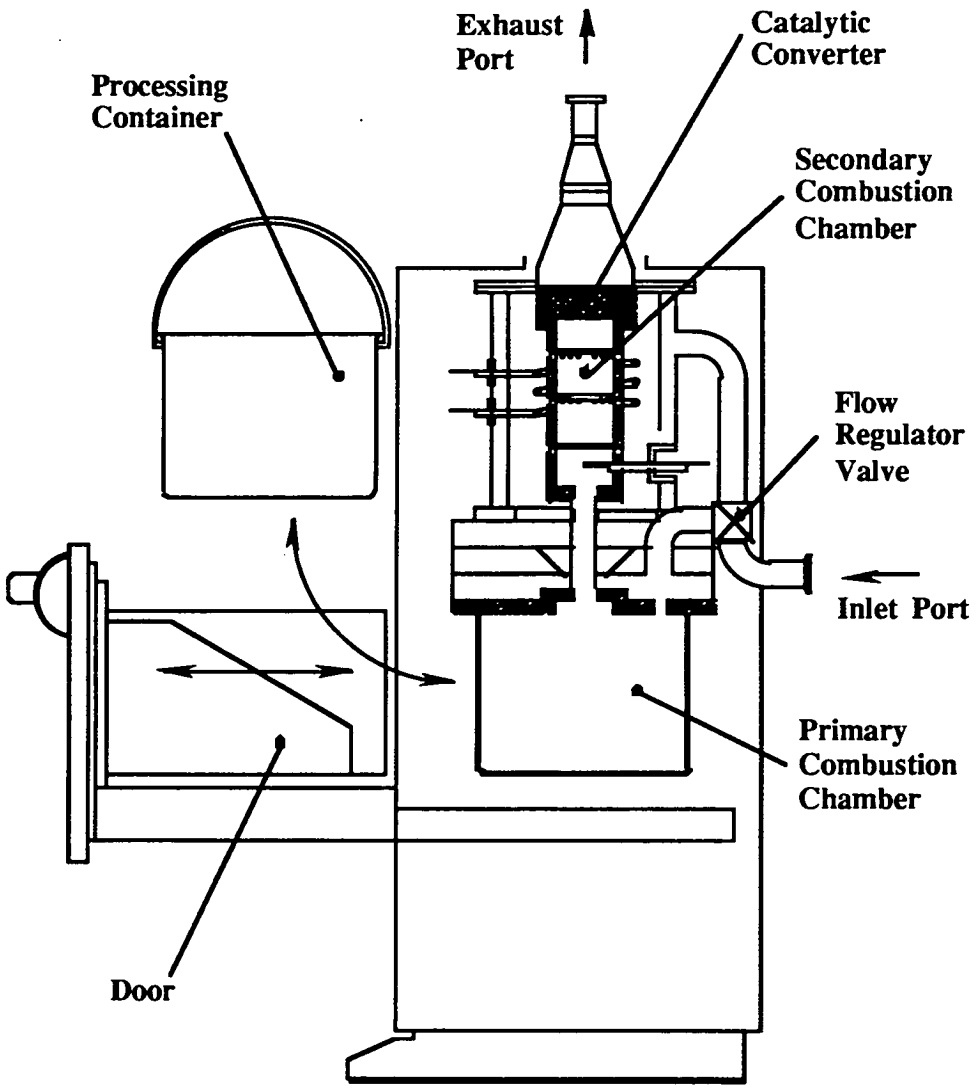


Figure 1. Microwave Incinerator.

A secondary chamber is mounted above the primary chamber, and a gas passage tube separates the two chambers. The gases resulting from the decomposition of waste in the primary chamber pass to the secondary combustion chamber where they are ignited. The temperature of the flame is 1000°C to 1200°C. The secondary chamber contains a catalyst bed which consists of platinum group metals supported in a honeycomb. This catalyst is preheated using an electrical heater. Gases resulting from the decomposition of waste in the primary chamber and products of the flame pass through the catalyst honeycomb where they are oxidized. After a period of approximately two hours, the incineration is complete, and the unit enters a cool down mode. At the end of the processing, a fine ash remains.

This commercial unit was designed to draw in ambient air to support the combustion process and to cool the supporting electronics. A blower draws the air in through the bottom of the unit. A portion of this air is directed into the combustion chamber. The remaining air stream serves two purposes: (1) to cool the electronics and (2) to keep the shell of the incinerator cool.

Both air streams are mixed together, before leaving the unit. Remixing both air streams serves to cool the otherwise hot exhaust from 250°C to just over 100°C.

TEST METHODOLOGY AND RESULTS

In the initial phases of the Lab-Scale CELSS Project, wheat will be grown inside a crop growth chamber. This chamber will serve as a biological processor to recycle carbon dioxide from the oxidation of solid waste. When the wheat reaches full maturity, it will be harvested, and the inedible portions (stems, leaves, husk, and roots) will be incinerated during the next cycle of crop production. For the purposes of the tests reported in this paper, only wheat straw (stems and leaves) was incinerated.

Waste and Ash Composition

For the initial testing phase, the microwave incinerator was operated within a laminar flow hood in order to capture and duct away the exhaust. A test sample of 25.7 gm of wheat straw was loaded into the incinerator. Incineration of the wheat straw resulted in an 87% reduction in the weight, leaving 3.4 gm of ash. The wheat and the ash were analyzed to determine composition. The results of the analysis are summarized in table 1.

Table 1. Weight Percents of Major Elements in Wheat Straw (Feed) and Ash (Residue)

Sample	Carbon gm (wt %)	Hydrogen gm (wt %)	Oxygen gm (wt %)	Nitrogen gm (wt %)	Other gm (wt %)	Total gm (wt %)
Wheat Straw	10.3 (40.1)	1.5 (5.8)	11.3 (44.0)	0.5 (1.9)	2.1 (8.2)	25.7 (100.0)
Ash	0.4 (11.7)	0.1 (2.9)	0.7 (20.5)	0.02 (0.6)	2.2 (64.3)	3.4 (100.0)
Percentage Reduction	96.1	93.3	93.8	96.0	0	86.7

Approximately 92% of the weight of the wheat straw (feed to incinerator) consisted of the major elements listed in table 1. The remaining 8% consisted of small quantities of inorganic elements, including potassium, calcium, phosphorus, magnesium, and sodium [5]. Table 1 shows that over 90% of the main elements are volatilized and oxidized.

Combustion Products

Carbon dioxide and water were expected to be the major products of the incineration process and small amounts of combustion byproducts were expected as well. In an effort to characterize the exhaust from the combustion process, the microwave incinerator was modified at NASA Ames Research Center to prevent the mixing of combustion air and cooling air. A one-inch diameter pipe with several sampling ports was attached to the incinerator. From these ports, temperature and carbon dioxide and oxygen concentrations were measured throughout incineration. Figure 2 shows the temperature and carbon dioxide profile of the exhaust gas.

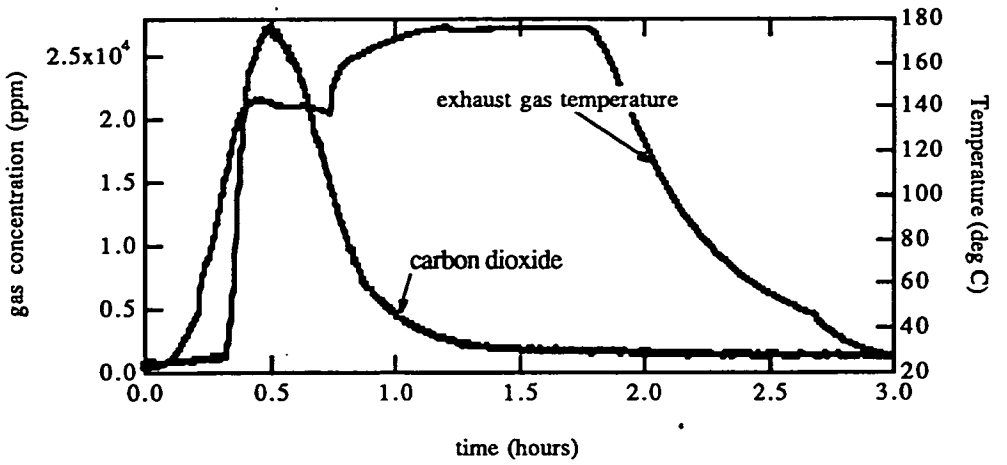


Figure 2. Carbon dioxide and exhaust gas temperature profile.

In addition, the gas from the exhaust was sampled at different points in time. The gas was first trapped using thermal desorption tubes. The trapped compounds were desorbed and focused onto the end of a capillary gas chromatography column. The compounds were chromatographically separated and monitored on a mass spectral detector over a mass range of 50 to 300. Trace amounts of numerous volatile organics and products of combustion were observed. Benzene and benzaldehyde ranked highest. Their concentrations and the exhaust temperature at which they were observed are listed in table 2. Also listed in table 2 are the spacecraft maximum allowable concentrations (SMAC's) [6]. SMAC is the maximum allowable concentration of a chemical that is permissible aboard manned spacecraft. In addition, the exhaust gas was sampled at different times during the incineration cycle using carbon monoxide detecting Draeger tubes. The maximum carbon monoxide concentration was less than 5 parts per million (ppm).

Table 2. Peak Concentrations and SMAC for Some of the Major Gaseous Byproducts

Components of Exhaust	ng/l	ppm (mole/mole)	SMAC (ppm)	Temperature (°C)
Benzaldehyde	44	0.013	0.091	90
Benzene	28	0.006	36	135
Carbon monoxide	6250	< 5	23	135

The microwave incinerator is currently undergoing a second set of characterization tests before being integrated into the Lab-Scale CELSS testbed. The goal of these tests is to quantify the total amount of different compounds produced in one incineration cycle. The test setup shown in figure 3 uses air of known composition (from gas bottles) as feed air. The products of incineration will be cooled, compressed, and stored in 240 gal storage tanks. Once the preliminary tests are completed, the air from these storage tanks will be sampled for a more complete analysis. The wheat and the ash will also be analyzed more thoroughly to account for the major inorganic compounds. Preliminary tests are currently in progress.

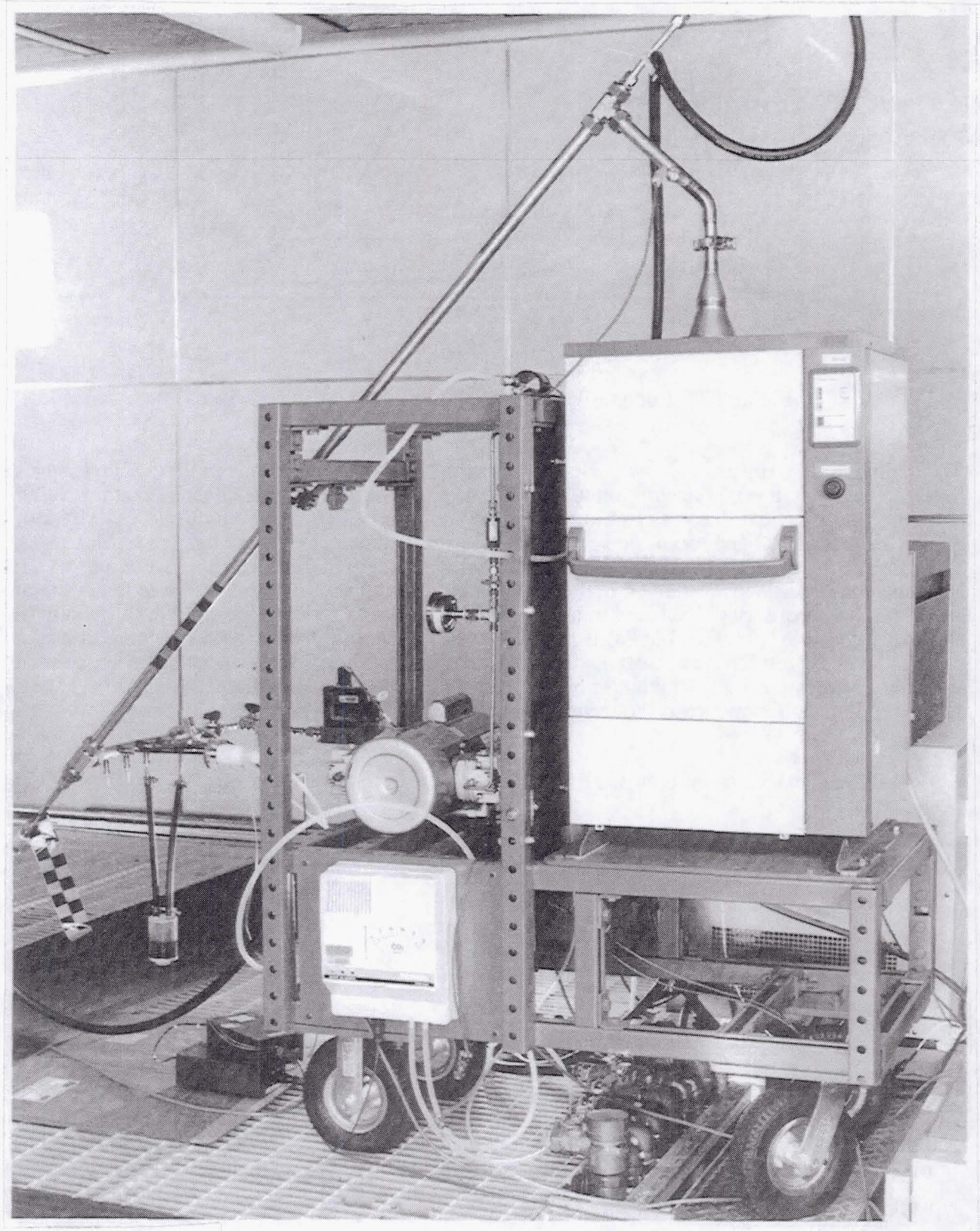


Figure 3. Microwave incinerator subsystem testbed.

TECHNOLOGY APPLICATIONS

Researchers at NASA Ames Research Center have been evaluating a microwave incinerator for use in recycling solid biological waste. Testing to date has shown that the microwave incinerator as a means of solid waste processing has numerous positive characteristics.

1. It is easy to use.

   Operation entails turning the power on, opening the drawer, loading the waste into the processing container, closing the drawer, and hitting the start button. When the incinerator is done, the drawer can be opened, and the processing container can be removed and emptied.

2. It is compact.

   Similar in size to a household trash compactor, the incinerator occupies very little space. It is small enough to fit in a kitchen or a laboratory.

3. It is safe.

   The exterior of the chamber remains cool-to-the-touch during operation. The drawer has an interlock that prevents the drawer from being opened when incineration is in progress.

4. The emissions are clean and environmentally safe.[1]

   When incinerating wheat straw, the combustion process produces mainly carbon dioxide and water vapor. Tests on wheat straw have indicated the presence of other compounds at very low levels.

5. It is versatile.

   A microwave incinerator can process a wide variety of biological wastes.

6. It reduces garbage disposal.

   When incinerating wheat straw, the incineration process reduces the mass of the waste to approximately 13% of its original mass.

7. It is simple to hook up.

   No special fuel or gas is needed to operate the unit. It runs on standard 110 VAC electrical power. If operated outdoors, the unit can exhaust to the atmosphere directly. If operated indoors, a vent or duct to carry away the exhaust would need to be added.

The results described in this paper show that the incinerator is effective at processing inedible plant matter, one form of biological waste. The authors hypothesize that other forms of

---

[1] Some types of waste (polyvinylchloride (PVC) products, for example) that produce refractory and toxic byproducts in any incineration process will be unsuitable for use in the microwave incinerator.

biological waste could also be processed. The high temperatures generated in the unit are sufficient to kill infectious agents. Medical institutions and biology research laboratories could therefore use microwave incinerators to process their infectious wastes. Their small size, ease of use, and relative low cost (approximately $6500) make them feasible for use in small laboratories and individual medical or veterinary practices. One issue that still needs to be addressed is the incineration of materials such as polyvinylchloride (PVC), which is found in items such as syringes [7]. Some form of additional processing or filtering would be required to remove the toxic compounds resulting from the combustion of these materials.

The rising costs of garbage collection and disposal provide economic incentives for the increased use of incinerators. Besides decreasing the burden on landfills and other sites of solid waste disposal, microwave incinerators could provide a potential energy source for heating homes and water by capturing the thermal energy produced by the incineration process.

Microwave incinerators could also be used on boats and recreational vehicles. A simple electrical hookup is all that would be required. Usage of microwave incinerators on boats could reduce the amount of garbage that is disposed of overboard or bagged, stored, and returned to port. Users of recreational vehicles could use their incinerators while on the road, minimizing the amount of garbage that is disposed of at campsites or RV parks.

The safety and versatility of microwave ovens has already been established. Microwave ovens are commonplace in homes and offices. Given the merits of the technology, microwave incinerators could eventually become just as useful and commonplace.

REFERENCES

1. *Centralized Incinerator Study for the South Florida Hospital Association*, Cross/Tessitore & Associates, P.A., Orlando, Florida, December 1985, pp. 17-18.

2. Issues in Medial Waste Management - Background Paper, U.S. Congress, Office of Technology Assessment, OTA-BP-0-49, U.S. Government Printing Office, October 1988, p. 1.

3. Freeman, H.: Incineration: What Role? In *Environment and Solid Wastes* edited by Framers, C. W. and Auerbach, S. I., Butterworth Publishers, 1983, pp. 425-433.

4. Suzuki, J., Hosaka, M., Kawasaki, Y., Nishino, A., and Sogen, K.: A Microwave Burning Processor for Waste Disposal. *Journal of Microwave Power and Electromagnetic Energy*, Vol. 25, No. 3, 1990.

5. Bubenheim, D. and Wignarajah, K.: Incineration as a Method for Resource Recovery from Inedible Biomass in a Controlled Ecological Life Support System. *Life Support and Biosphere Science: The International Journal of EarthSpace* 1(3): In press. (1994)

6. Leban, M. and Wagner, P.: Space Station Freedom Gaseous Trace Contaminant Load Model Development, SAE Paper No. 891513, 1989.

7. Frietas, J., Panasonic Industrial Company, telephone conversation, October 4, 1994.

Pressurized Spray Stripping and Cleaning Techniques

NOT AVAILABLE AT THIS TIME

TURBINE/ BRUSH RECYCLING PIPE CLEANING SYSTEM

Rudy J. Werlink
NASA DM-ASD
Kennedy Space Center, Fl 32899

ABSTRACT

The Turbine Brush Cleaner (patent applied for) provides a improved method of cleaning the inside of piping and tubing. The mechanical action of the brush aids the cleaning action of the cleaner(s) which are much more environmental acceptable than CFC's (Freon). It offers advantageous over existing rotating brush systems which require flexible drive shafts and a facility air source for power. The elimination of the drive shaft allows a much more compact design which can transverse smaller diameters and up to 90 degree bends over longer distances. Another advantage is reduced contamination of the cleaned sections without a trailing drive shaft. The cleaning system consists of an electrical motor and pump which provide liquid flow to spin the turbine and deliver the cleaning solution. The outflow of the pipe section being cleaned is recycled back to the pump inlet after filtration. . Test results on 3 foot lengths of sample tubes using several water based environmentally friendly cleaners will be discussed. The sample tubes were contaminated with 4 types of difficult to remove greases. The solution is recycled through filters, minimizing the waste stream.

Introduction:

The Turbine/Brush was created in response to the need for new pipe and tube cleaning methods which will be required as the production of Chloro-florocarbons are phased out at the end of 1995. At Kennedy Space Center tubes must be cleaned and verified to high levels for oxygen service as stated in KSC-123-C "Surface Cleanliness of Fluid Systems". Established cleaner has been CFC-113 which contains Chlorine in a stable form and has high ozone destruction potential. The approved methods used CFC-113 flushes to clean with a final CFC-113 flush for verification. The effluent is collected and evaporated with the remaining Non-Volatile Residual weighted. The maximum allowed is 1 mg non volatile residual per 1 ft2 of lateral inside area.

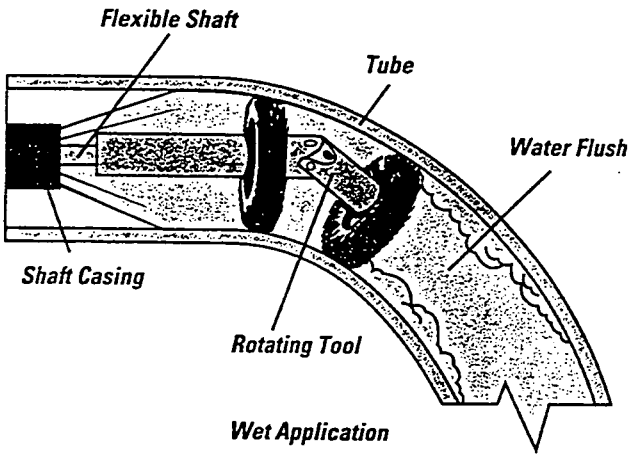
Existing Technology

With the use of new non-CFC cleaning fluids, mechanical action such as a commercially available rotating circular brush greatly enhances the removal process. The use of one product was demonstrated such as the Goodway AT-100[1] which uses a air motor to rotate a flexible shaft which spins a circular brush. a plastic casing surrounds the flexible shaft and guides the flowing water and cleaner solution to the brush. This system was demonstrated by cleaning of 3 foot tubes pre-contaminated with grease.

---

[1]Manufactured by:
Goodway Technologies Corporation
420 West Ave.
Stamford, CT 06902-6384

Figure 1
Cross Sectional View of the Goodway Cleaning Brushes



While the system was fairly effective in removal of the grease when used with a aqueous cleaner such as Reentry , The following aspects for improvement were noted:

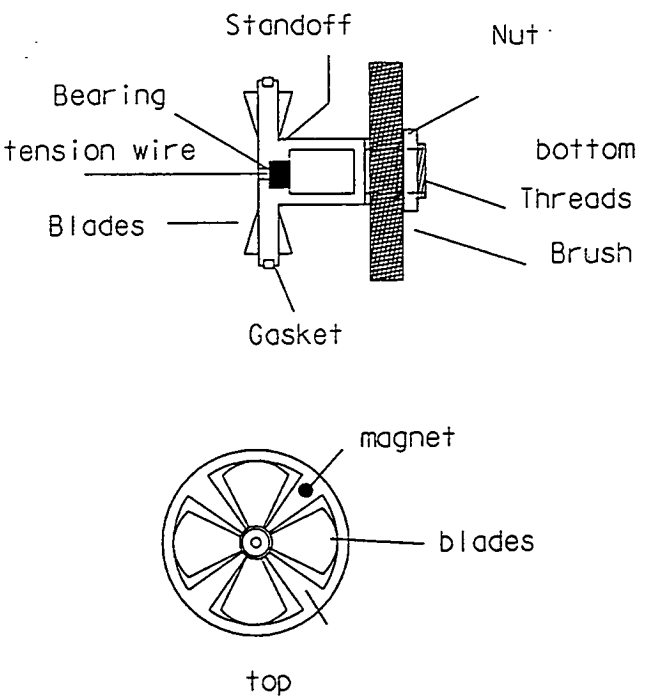
- The drive shaft flexible cable and plastic casing were big sources of contamination.
- The dual brush assembly could not travel through elbows-
- The maximum length to clean tubes was limited by friction between the inside of the pipes and the plastic casing.
- The smallest diameter tubing with bends that could be cleaned is 1.25 inches

Turbine/Brush Invention

The Author of this paper designed the following device to greatly improve on the short comings of the commercial system a Turbine and standoff with a removable circular brush was designed and fabricated out of stainless steel. A patent application is in filed in the United States Patent Office under serial number 8/207,313. Upstream flow pushes the assembly and rotates the brush to provide cleaning action. The rotation takes place on a bearing inside the turbine with the longitudinal speed controlled by a thin stainless cable held in tension by the operator. The pressurized cleaning solution spins and pushes the brush assembly with the tension on the wire controlled by the operator to provide location control in the pipe. A small permanent magnet provides spin rate and longitudinal location information using a coil pickup and voltage display with the turbine in the pipe. The brush assembly is compact enough to travel through 90 degree bends and can transverse through long sections since friction between the Teflon coated tension

wire and the inside is much lower than the drive shaft of commercial systems. The Turbine/brush can be fabricated for varies pipe sizes with a 1 inch diameter and a 2 inch diameter unit already produced. Figure 1 shows the Brush assembly detail and in the pipe.

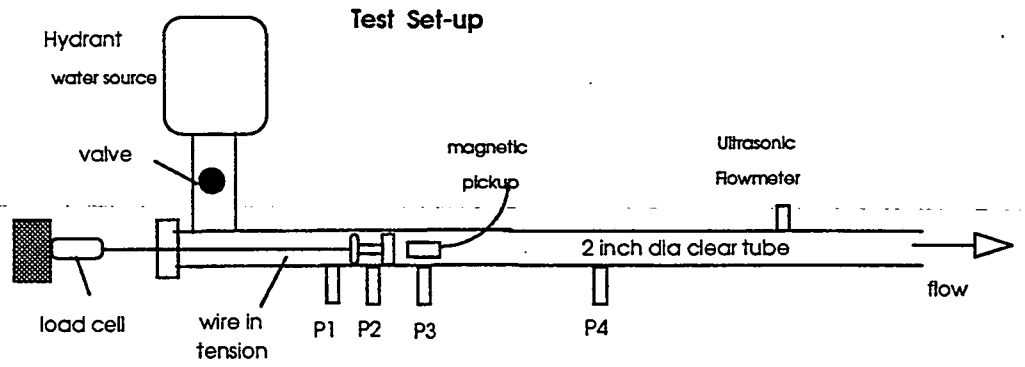
Figure 2 Turbine Brush cut-away views



Operational parameters

The Operational characteristics were measured experimentally for the 2 inch diameter turbine the experimental setup used a 486 laptop PC and Lab View software version 3.0 to display real time data and write to disk in a ASCII file. Sampling rate = 58 per sec.

transducer	**range**
load cell	0-50 lbs
P1	14.7-60 psia
P2	14.7-50 psia
P3	12-30 psia
P4	12-30 psia
coil	pulse counter--generated by rotating turbine magnetic -coil pickup on tube.



The test generated the following data which represents about the minimum flow-rate required to rotate the 2 inch turbine brush:

TABLE 1

Operational Data for 2 Inch Turbine

load cell	tension on wire	8.94 lbs
P1	upstream before turbine	3.10 psig
P2	turbine to brush standoff	1.55 psig
P3	back of brush	.16 psig
P4	downstream	.38 psig
pick-up coil volts	rotational rate	350-500 rpm
Flow-rate ultrasonic	reading (noisy)	55.19 gpm
Flow-rate calculated	calculated [2]	72.41 gpm

values averaged using 50 sample points with a sampling rate of 53.3 samples per second

Cleaning System Design

A complete system to generate flow using cleaning solutions and provide filters to reuse the cleaning solution has been designed and fabricated by NASA Special Projects and the Prototype shop at Kennedy Space Center . The system is designed to clean varies pipes and tubes using the turbine/brush and flowing aqueous cleaning solutions buy electrically driven pump . The system is transportable by Pickup Truck, the only facility requirement is 220 volt AC for the pump motor. The Prototype system is designed for cleaning up to 2 inch diameter pipe and tubing up to 100 feet in length.

The effluent is filtered through particulate filters which are replaceable with micron ratings from 1 to 50 microns and returns to the supply tank. A secondary filter loop is to remove traces of cleaner from the rinse tank or provide a higher level of cleaning to the cleaning solution tank.

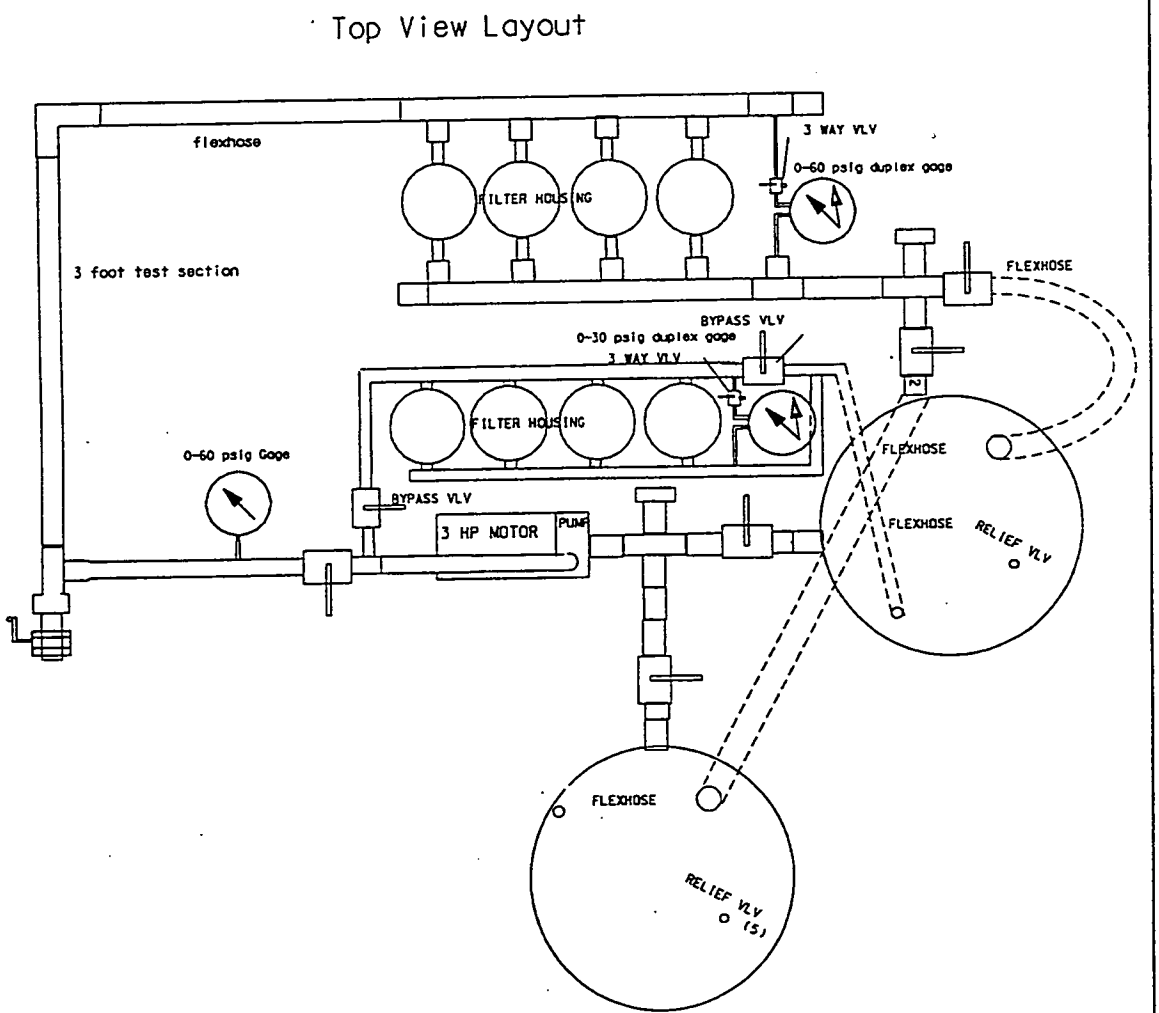
---

[2] Calculated based on the formula $q = a \bullet \sqrt{(2 \bullet p \bullet g)/\rho}$

a= cross section areas of pipe

p= P4 pipe outlet pressure

Figure 3 - 2 inch Turbine/Brush Reuse Cleaning System



Turbine/Brush cleaning effectiveness procedure

To evaluate cleaning effectiveness, a smaller recycling electric/pump system was used with the 1 inch diameter cleaning system. This system uses a 5 micron rated particulate filter. Two cleaning agents were evaluated at ambient temperature:

- Brulin 815GD-10% mixture with water
- Amberclean L12-50% mixture with water

Pre-cleaning

A set of 8 3 foot long 1 inch diameter stainless steel tubes were used as sample sets .
The tubes were pre-cleaned using standard CFC 113 flushing per KSC-123-C Cleaning Specification level 300.

Sample tube contamination

The samples were then contaminated with 4 greases which were hard to remove and common at Kennedy Space Center. The mixture was prepared per the following:
dissolve a mix of .5 grams each of the following in 500 ml of Freon-113:

- Dow Corning 55
- Dupont Krytoz 240 AC
- Amoco Rycon #2
- Chevron Moly Grease

This resulted in a grease mixture concentration of 4 mg/ml.
A 10 ml quantity of this mixture was pipetted into the 8 sample tubes with stainless steel capped ends each tube was horizontally attached to a rotary drive and vacuum pump. As the pipe was rotated under low pressure, heat was applied to the pipe exterior with a heat lamp. The Freon-113 evaporated, leaving 40 mg of grease evenly distributed over the inside of the pipe.

Sample tube cleaning:

Three sample tubes were cleaned by allowing the rotating Turbine/brush to move up and down the pipe. Three tubes were cleaned without the turbine using the cleaning solution only. One tube was not contaminated and cleaned and one tube was not processed from the pre-clean stage to the Non-Volute Residual (NVR) analysis. This set of 8 tubes consisted of one run.
All sample tubes were rinsed with distilled water and dried after cleaning and before verification.

NVR analyzes of sample tubes:

Each tube was rinsed by capping the ends with stainless steel caps and inverting. with about 400 grams of CFC-113 solvent in two rinses of 200 grams each. The effluent was evaporated down to 10 milliliters using a rotary evaporator in order to recycle the solvent. The 10 ml's of remaining solvent was evaporated to dryness in a 100 deg C oven in a pre-cleaned and pre-tarred aluminum dish. The dish was then cooled in a desiccator and weighted. The results were reported as mg's of Non-Volute Residual. Blank tubes were run before and after each sample tube to verify the lab processes.

Results

The evaluation of the Turbine/Brush showed that the system was very effective in removing the contamination from the tubes. Blanks were run to validate the sampling and lab technique. The following factors were varied to show effects:

- Two types of cleaners were used for comparison Brulin and Amberclean. The Brulin was considerably less effective based on two sets of data. see Graph 1
- The Turbine/brush aids in effectiveness in cleaning vs the cleaner alone. see Graph 2
- Doubling the cleaning time form 3 to 6 minutes does not produce significantly better results indicating that 3 minutes is sufficient time. see Graph 3.
- Data scatter is fairly high indicating the need for more testing for statistical significance.
- The recycling of the cleaning fluid through 5 micron filters is effective in maintaining cleaning fluid quality. Lab tests show the contamination is not accumulating in the fluid or re depositing on the cleaned tube. Table 2 summarizes the testing conducted showing values for mg's of NVR, % removed, and mg/ft2 based on verification using NVR and CFC-113.

Recommendation:

There are many factors affecting the effectiveness of the cleaning process which could be investigated. Several factors are: concentration of water to cleaner, Temperature of cleaning solution, and flow-rate of the cleaning solution.

- The 2 inch Turbine/brush Prototype system will be evaluated using Amberclean and other cleaners for compliance to the KSC standard of 1 mg per ft and placed in operation at KSC.
- The Turbine/brush design will be improved to achieve more torque and lower flow requirements.

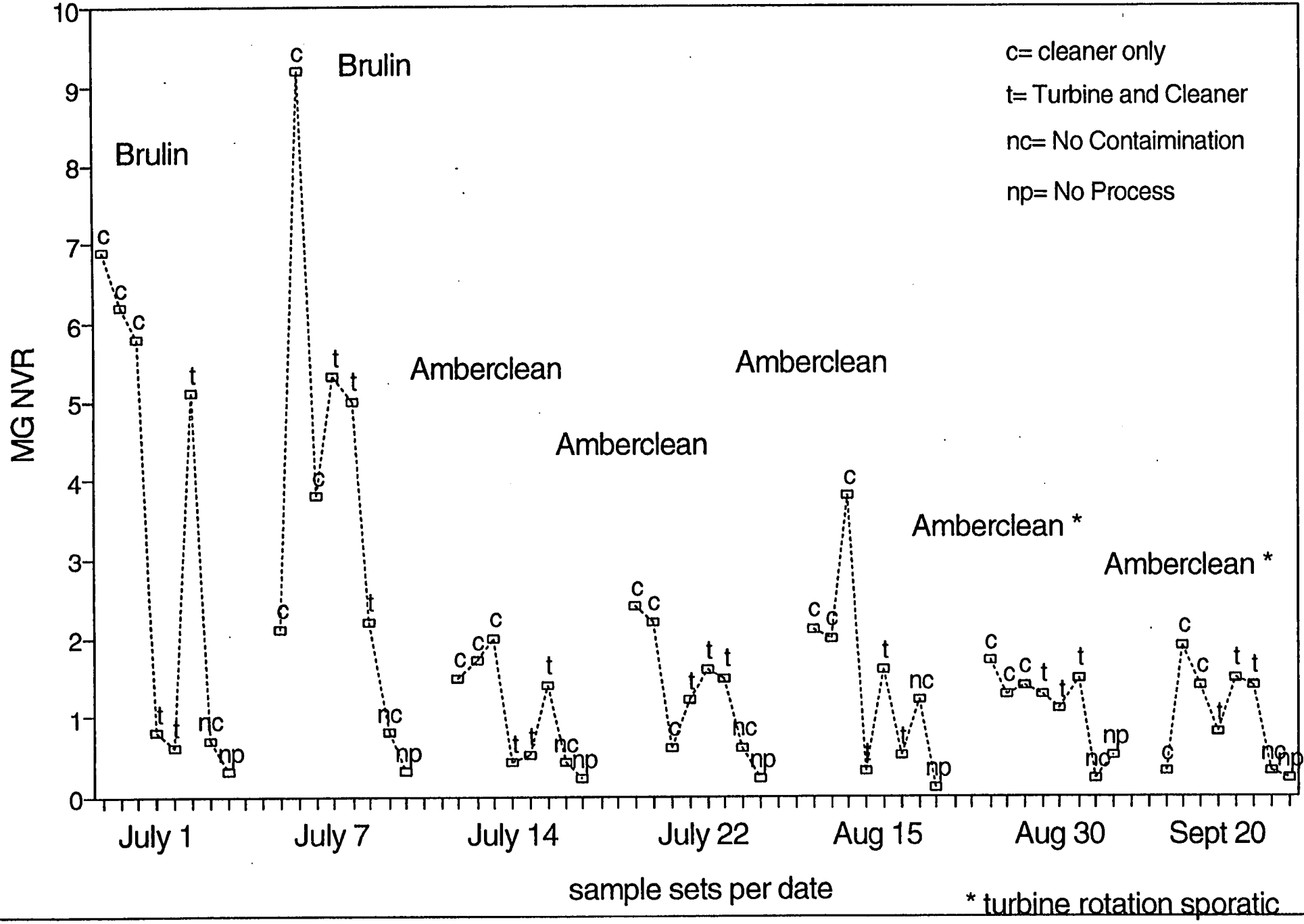
TABLE 2
Summary of test results

Sample Set	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
Brulin	2 Cleaner	6.9	82.75	8.67
July 1	3 Cleaner	6.2	84.5	7.79
3 minutes	8 Cleaner	5.8	85.5	7.29
	4 Turbine	0.8	98	1.01
	6 Turbine	0.6	98.5	0.75
	7 Turbine	5.1	87.25	6.41
	1 Cleaner -no	0.7	98.25	0.88
	BLK no process	0.3	99.25	0.38
Brulin	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
July 7	6 Cleaner	2.1	94.75	2.64
3 minutes	2 Cleaner	9.2	77	11.56
	BLK Cleaner	3.8	90.5	4.77
	3 Turbine	5.3	86.75	6.66
	7 Turbine	5	87.5	6.28
	4 Turbine	2.2	94.5	2.76
	1 Cleaner-no	0.8	98	1.01
	8 no process	0.3	99.25	0.38
Amberclean	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2

July 14	8 Cleaner	1.5	96.25	1.88
3 minutes	1 Cleaner	1.7	95.75	2.14
	BLK Cleaner	2	95	2.51
	6 Turbine	0.4	99	0.50
	4 Turbine	0.5	98.75	0.63
	2 Turbine	1.4	96.5	1.76
	7 Cleaner-no	0.4	99	0.50
	3 no process	0.2	99.5	0.25
Amberclean	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
July 22	Blk Cleaner	2.4	94	3.02
3 minutes	3 Cleaner	2.2	94.5	2.76
	4 Cleaner	0.6	98.5	0.75
	1 Turbine	1.2	97	1.51
	2 Turbine	1.6	96	2.01
	7 Turbine	1.5	96.25	1.88
	4 Cleaner-no	0.6	98.5	0.75
	8 no process	0.2	99.5	0.25
Amberclean	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
Aug. 15	7 Cleaner	2.1	94.75	2.64
6 minutes	3 Cleaner	2	95	2.51
	8 Cleaner	3.8	90.5	4.77
	Blk Turbine	1.6	96	2.01
	4 Turbine	0.5	98.75	0.63
	2 Cleaner-no	1.2	97	1.51
	1 no process	0.1	99.75	0.13
Amberclean	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
Aug. 30	2 Cleaner	1.7	95.75	2.14
6 minutes	1 Cleaner	1.3	96.75	1.63
	8 Cleaner	1.4	96.5	1.76
	7 Turbine	1.3	96.75	1.63
	Blk Turbine	1.1	97.25	1.38
	3 Turbine	1.5	96.25	1.88
	4 Cleaner-no	0.2	99.5	0.25
	5 no process	0.5	98.75	0.63
Amberclean	DESCRIPTION	NVR (MG)	% REMOVE	MG/FT^2
Sept. 20	6 Cleaner	0.3	99.25	0.38
6 minutes	7 Cleaner	1.9	95.25	2.39
	2 Cleaner	1.4	96.5	1.76
	1 Turbine	0.8	98	1.01
	8 Turbine	1.5	96.25	1.88
	3 Turbine	1.4	96.5	1.76
	4 Cleaner-no	0.3	99.25	0.38
	5 no process	0.2	99.5	0.25

Graph 1

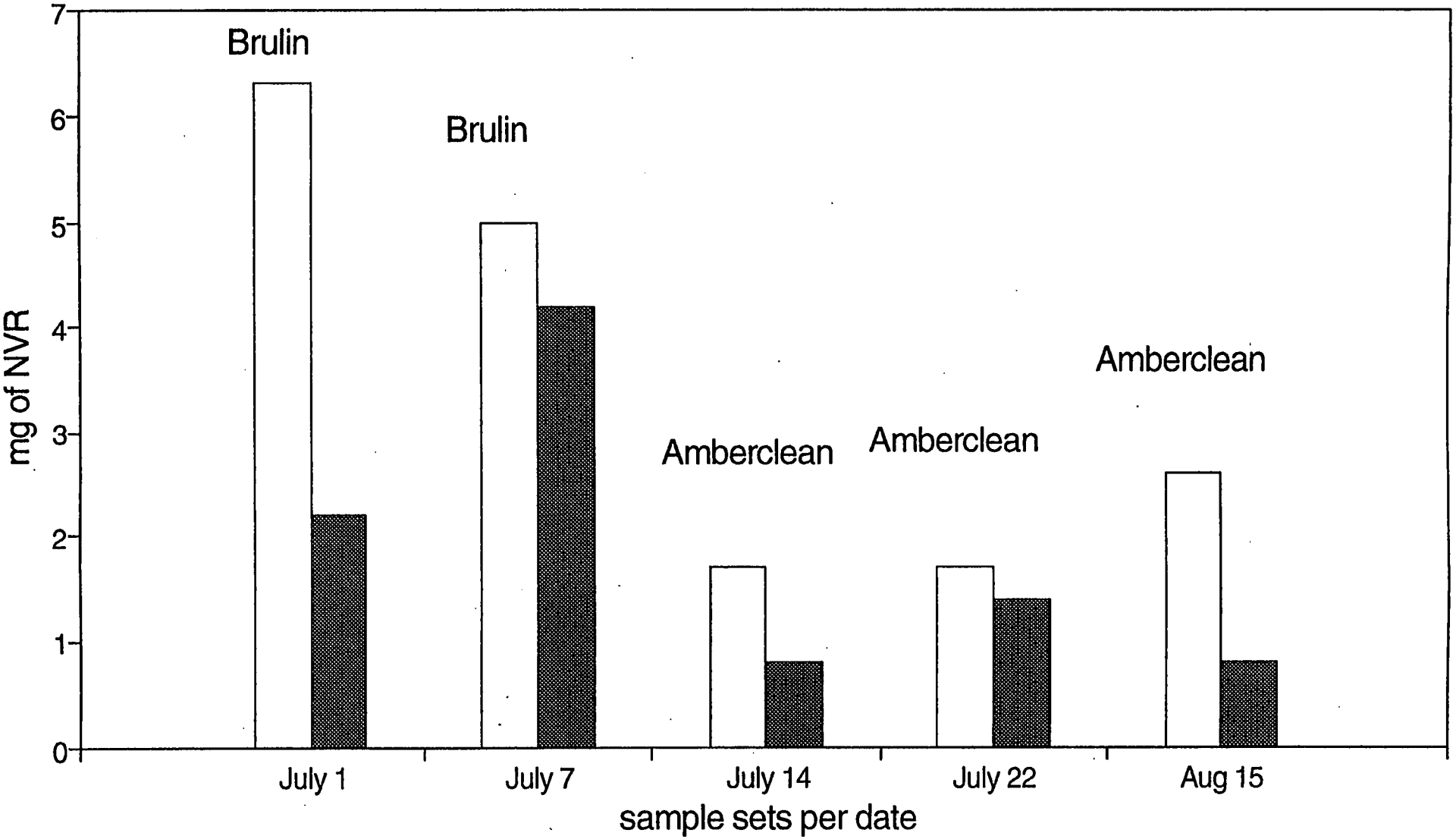
comparison of 3 foot contaiminated tubes using Brulin and Amberclean



* turbine rotation sporatic

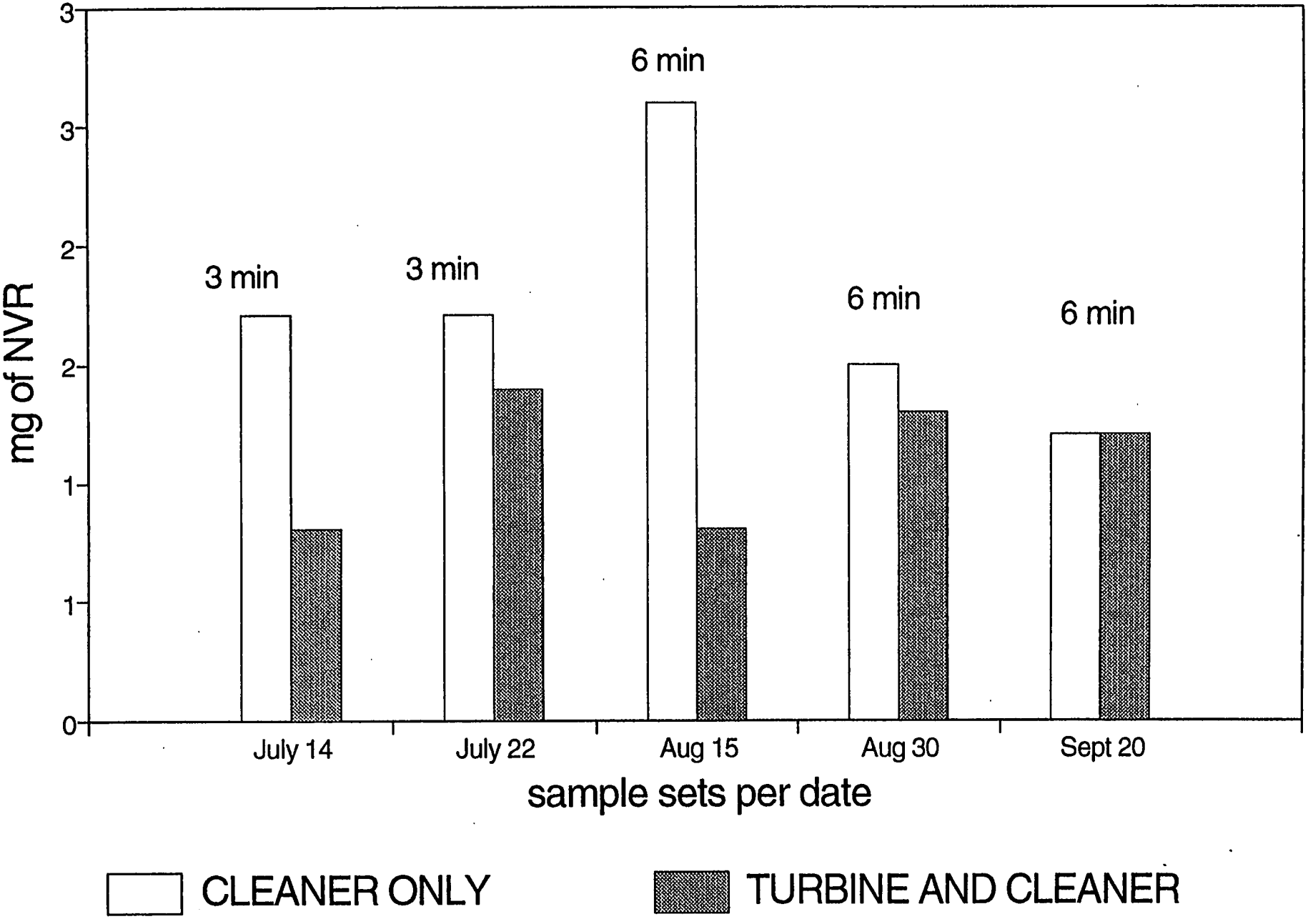
Graph 2

Turbine vs Cleaner only
3 sample tubes averaged per run



Graph 3

Amberclean Turbine vs Cleaner Doubling Cleaning Time



PHOTOVOLTAIC POWER WITHOUT BATTERIES FOR CONTINUOUS CATHODIC PROTECTION AND AN ALTERNATE PHOTOVOLTAIC/ULTRACAPACITOR COMBINED POWER SOURCE

Wallace W. Muehl, Sr.
Department of the Navy, Coastal Systems Station
Dahlgren Division, Panama City, FL 32407-7001

ABSTRACT

The Coastal Systems Station (COASTSYSTA) designed, installed and started up, on 20 January 1990, a state-of-the-art stand alone photovoltaic powered impressed current cathodic protection system (PVCPSYS) not requiring any backup power for steel and iron submerged structures. The PVCPSYS installed on a 775 foot steel sheet piling of a Navy dock bulkhead provides complete, continuous corrosion protection. This PVCPSYS has been in operation for more than 4 years, not requiring any repair or maintenance and is environmentally clean. Initial cost savings of the PVCPSYS versus conventional cathodic protection system was $46,000. A PVCPSYS has been installed on another 800 foot bulkhead, 21 May 1993, and is providing complete corrosion protection without backup power. Performance is well documented. Other potential applications are mothballed ships, locks, dams, bridges, pipelines and similar structures. These systems are considered a major advance by Sandia and the Department of Defense (DOD) Photovoltaic Review Committee. An ultracapacitor, a recent hi-tech development, that is environmentally clean, will be incorporated in the PVCPSYS, when required, to enhance the system's capability. A photovoltaic/ultracapacitor combined power source operating under adverse conditions, and/or to satisfy or meet regulations will assure cathodic protection, including pipelines carrying combustibles or other products that could otherwise create environmental problems. Patents are pending on this PVCPSYS and the photovoltaic/ultracapacitor powered systems.

The objective of the initial project was to successfully demonstrate that renewable energy can efficiently and economically replace or be used instead of continuous non-renewable power sources. An opportunity to clearly show that photovoltaic power is practical and reliable was the result of a recommendation to provide cathodic protection to the Naval Diving and Salvage Training Center bulkhead.

The COASTSYSTA in Panama City, Florida, has broken new ground in the application of solar energy for cathodic protection. Photovoltaic arrays without battery backup have been connected to the 775 foot-long steel sheet piling of a dock bulkhead via a cathodic protection system, to prevent corrosion on that steel structure in a salt water environment.

Cathodic protection, as the name signifies, is the process by which, in the COASTSYSTA impressed current type application, the entire steel sheet piling is transformed into a cathode via a series of anodes mounted in PVC standoff racks, in the water, next to the piling. When direct current (DC) energy is applied to the anodes and sufficient electrical potential is attained by current flow from the anodes via an electrolyte (seawater) to the piling, the corrosion is transferred to the anodes, preventing piling corrosion.

Mr. Wally Muehl, electrical/mechanical engineer at COASTSYSTA, was evaluating power sources to protect the Naval Diving and Salvage Training Center bulkhead when he focused on photovoltaics. Although there were 10 other impressed current cathodic protection systems installed on the docks, all were powered by a continuous power source with the current rectified to DC. Of these 10 systems, 5 were down from 1 to 1 1/2 years and 3 were down more

than 2 years due to rectifier failures and/or the power source secured due to construction and as a result, no corrosion protection was provided. PVCPSYS's would have continued to provide power and corrosion protection and would not have been affected by these type power outages.

The Naval Diving and Salvage Training Center is in a separate location from these docks, and it was determined that power was not readily available and would be expensive to provide rectifiers on the dock due to the dock configuration. Rectifiers would also pose a safety hazard on the dock that is regularly used for diver and salvage training. This bulkhead was 12-years old and other than the initial coating, received no corrosion protection.

Mr. Muehl developed a state-of-the-art solar powered impressed current cathodic protection system for submerged steel and iron type structures without requiring any battery backup power. Innovations in design and method of operation permits the photovoltaic arrays to easily provide and maintain complete continuous corrosion protection without the necessity of DC power backup such as batteries. Battery backup power is considered costly and an environmental problem. To date, all impressed current systems require a continuous DC power supply in order to provide cathodic protection.

The COASTSYSTA photovoltaic power system is a fixed-axis system which is suitable for the Panama City latitude of 30°10'N, 85°22'W. The tilt of the adjustable arrays were set at latitude instead of +15° in January 1990, and have not been changed. This is a good indication that other areas with good distribution, but lower insolation levels, would be excellent prospects for a similar type of photovoltaic powered system. For higher latitudes, there are several other options to improve system performance without battery backup. These include one-axis east/west tracking, two-axis north/south, east/west tracking, or simple adding a module or two to meet the additional current requirements.

The engineer in charge, Mr. Muehl, designed, prepared the specifications, and monitored the installation, also had two other problems that had to be considered and resolved in order to install an impressed current cathodic protection system. The first problem was ensuring that the steel piling had electrical continuity. Another problem was providing sufficient impression of current "carry over" to overcome a 155-foot section of piling that had to be bypassed, and provide cathodic protection, without anode placement in the area having a water depth of 27 feet, where diving takes place. Both problems were overcome in the design.

To facilitate the use of a photovoltaic powered cathodic protection system without battery backup, the steel sheet pilings were provided an initial one-time only preconditioning polarization for a predetermined continuous time period to the extent that these pilings were initially polarized to a relatively high negative potential by a temporary DC power source. The photovoltaic power system was provided with blocking diodes to prevent any possibility of current reversal. It is to be noted that evolution of a protective hydrogen film is merely a by-product of the preconditioning polarization at the higher negative potentials. Additionally, depending upon the environment and if higher (more negative) polarized potentials could be maintained other than required to provide basic complete cathodic protection, formation of thicker calcareous deposits having protective value over a period of time could occur. The initial DC power for polarization can be provided by a DC power source such as a portable motor driven DC generator or a portable motor driven DC welder.

COASTSYSTA PVCPSYS tests performed, along with other data obtained, provides a further explanation that the anode/seawater/cathode piling structure acts like a battery and when allowed to rest, the polarity level recovers and is electrochemical in nature. An electrochemical lead/acid battery, for example, can recover charge if allowed to rest after serving a load. The electrochemical reaction reverses slightly when the load is disconnected, however, a capacitor without an external current source cannot recover by simply removing the load. It is believed

that the one-time only initial preconditioning polarization (controlled conditions) of the structure embeds single hydrogen atoms in the steel sheet piling that can also migrate and diffuse in the structure. This system delays the decay of the negative potential and permits the photovoltaic arrays to supply sufficient power allowing the system to easily provide complete continuous cathodic corrosion protection including cloudy, overcast, rainy and nighttime conditions without the necessity for DC power backup such as batteries.

In summary, the foregoing novel method and system of a one-time only preconditioning or prepolarizing the structure prior to energizing the photovoltaic solar array on-line with the system, provides a relatively higher negative potential that has a slow rate of decay. This permits the use of regulated photovoltaic solar energy with excess available power, and without any backup power, to easily provide complete continuous corrosion protection, including cloudy, overcast, rainy and nighttime conditions, with excellent polarization levels and improving with time. An analogy may be that the steel structure becomes very effectively polarized, and will remain so by the variable DC charge affect provided by the simple solar array system, much like a piece of steel or iron can become magnetized by the application of a DC electrical current.

The installation, start up, and continuing operation, including underwater inspections, are well documented to date by COASTSYSTA and verified on site, during the day and at night by the U.S. Army Corps of Engineers, Construction Engineering Research Laboratory, Naval Energy Program Office and members of DOD Photovoltaic Review Committee. The average amount of available sunshine for the three weeks prior to these organizations visit, per data provided by the National Weather Service, averaged 24 percent.

This system has been in operation almost 4 3/4 years without requiring any maintenance or adjustment. A patent is pending on the new technology. Other possible applications are mothballed ships, docks, dams, locks, bridges, marinas, offshore structures and pipelines in various environments.

The estimated cost in 1985 of a conventional cathodic protection system requiring continuous DC power was $75,000.00 and the estimated cost in 1990 was $108,000.00. The PVCPSYS cost at contract completion was $61,316.00, complete and ready to use. The initial cost savings by installing a PVCPSYS was in excess of $46,000.00.

The DOD Photovoltaic Review Committee and Sandia National Laboratories consider this successful and cost effective system a major advance for the application of photovoltaics.

A photovoltaic power system without any backup power has been installed on another 800 foot bulkhead. The two previous 400 foot conventional rectifier powered impressed current cathodic protection systems were modified to allow this conversion. The PVCPSYS successfully started operation on 21 May 1993, without any backup power and is providing complete continuous corrosion protection. A state-of-the-art data collection system is provided, that along with other capabilities, will monitor, report, analyze and record simultaneously the solar energy output DC volts, DC amps and the DC negative potential voltage of the steel sheet piling on a personal computer that is MS-DOS compatible located about 1/2 mile away from the site.

An ultracapacitor, that is also environmentally clean, will be incorporated into the PVCPSYS when required to enhance the system's capability. A photovoltaic/ultracapacitor combined power source operating under adverse conditions and/or to satisfy or meet Government regulations will assure cathodic protection for submerged or buried steel or iron type structures, including pipelines or the like carrying combustible or similar products that

could otherwise create environmental problems. Westinghouse Electric Corporation acquired the rights to manufacture the ultracapacitor and has expressed an interest in the combined PVCP/ultracapacitor system.

The ultracapacitor technology developed over the past 12 years by Pinnacle Research Institute, Inc. (PRI), provides both high energy and power. The capacitor breaks all paradigms associated with conventional batteries and capacitors. The ultracapacitor has no conventional dielectric. Unlike batteries there is no chemical reaction; therefore, the ultracapacitor can be discharged and rapidly recharged by photovoltaic power for an indefinite number of times. The ultracapacitor has 100 times more power per ounce and is significantly smaller than conventional capacitors. Most importantly, a second generation of ultracapacitors has been recently developed that are capable of releasing energy gradually. In the electric powered automobile industry, for example, some initial applications will be for hill climbing and acceleration. According to PRI, ultracapacitor applications are being discussed with automobile manufacturers. Also, PRI is a subcontractor of the Department of Energy Highbred Electric Vehicle Program. An ultracapacitor, if required for the COASTSYSTA PVCPSYS installed in 1993 would weigh about 18 pounds.

With reference to this manuscript, it is with pleasure that I acknowledge the "helpful cooperation and information" received from the following personnel:

- Dr. Michael G. Thomas and Dr. Hal Post, Senior Members, Technical Staff, Photovoltaic Research Dept., Sandia National Laboratories, Albuquerque, New Mexico.

- Mr. James F. Jenkins, P.E., Corrosion & Metallurgical Engineer, Naval Civil Engineering Laboratory, Port Hueneme, California.

* Mr. L.E. Humble, Photovoltaic Programs, Energy Program Office, Naval Weapons Center, China Lake, California.

* Mr. Roch A. Ducey, Principal Investigator and Ms. Jearaldine I. Northrup, Research Engineer, U.S. Army Construction Engineering Research Laboratory, Champaign, Illinois.

- Mr. Thomas F. Lewicki, P.E., Facilities Corrosion Program Manager, HQ Air Force Civil Engineering Support Agency, Tyndall Air Force Base, Florida.

- Dr. Thomas M. Cawthon, Hydrogen Program Manager and H. Dana Moran, Manager, Research and Technology Applications, National Renewable Energy Laboratory, Golden, Colorado.

- Dr. K.C. Tsai, President, Pinnacle Research Institute, Inc., Los Gatos, California.

- Navy Divers and Dive Locker, Coastal Systems Station, Panama City, Florida.

* Members of the DOD Photovoltaic Review Committee.

Video and Imaging

FLAT PANEL PLANAR OPTIC DISPLAY

James T. Veligdan
Brookhaven National Laboratory
Upton, NY 11973

ABSTRACT

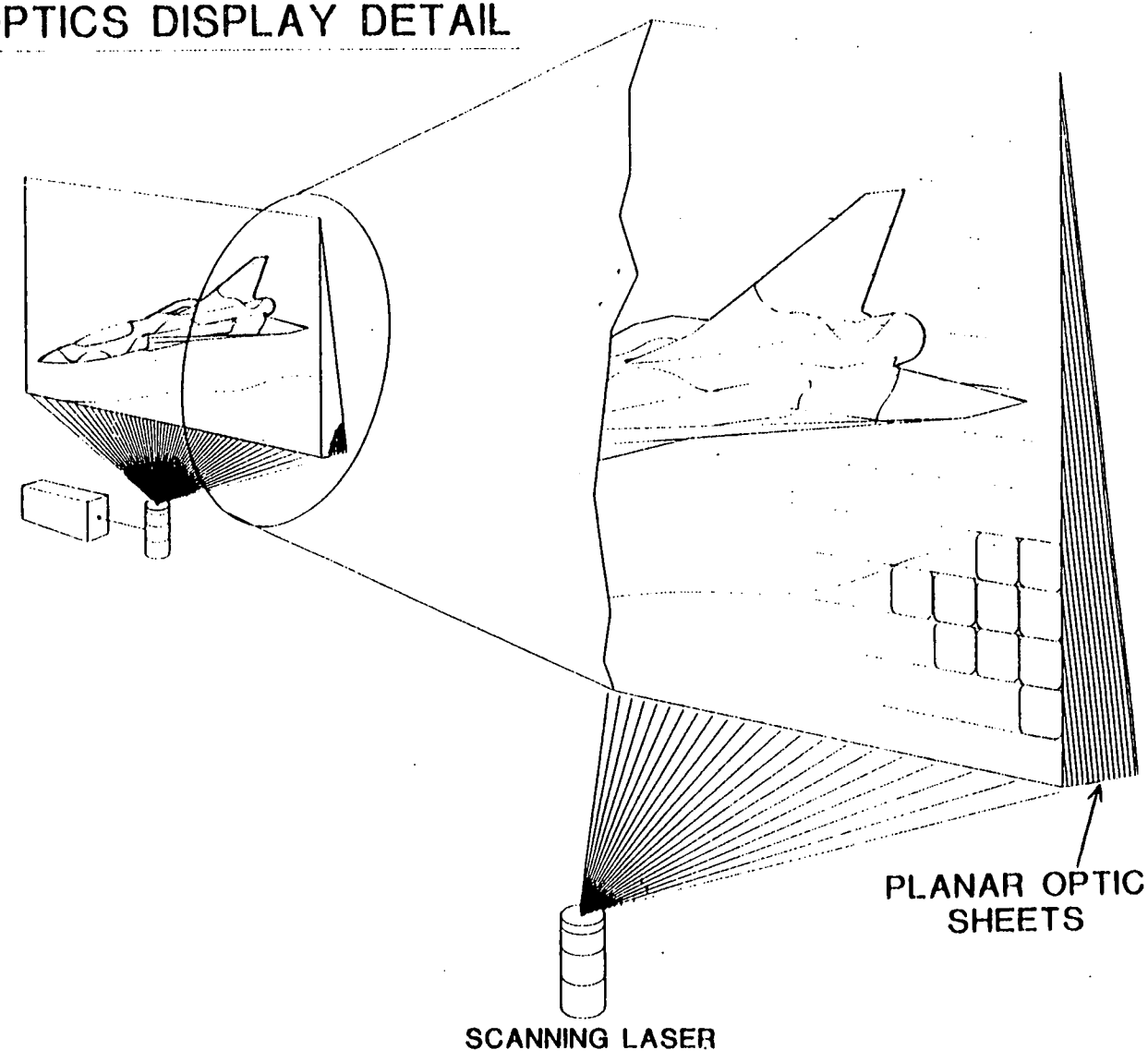
A prototype 10 inch flat panel Planar Optic Display, (POD), screen has been constructed and tested. This display screen is comprised of hundreds of planar optic glass sheets bonded together with a cladding layer between each sheet where each glass sheet represents a vertical line of resolution. The display is 9 inches wide by 5 inches high and approximately 1 inch thick. A 3 milliwatt HeNe laser is used as the illumination source and a vector scanning technique is employed.

BACKGROUND

For approximately 40 years, researchers have been searching for a flat panel TV screen. The president of the United States has identified Flat Panel Displays as a technology he sees as vital to the economic future of the Unites States.[1] From data communications in business, to entertainment on the family room TV, the Cathode Ray Tube (CRT) has proven itself to be indispensable as a means of displaying information. It ranks high in brightness, reliability, manufacturability and cost. However, its volume, weight, and power consumption have severely limited its portability which is required for the development of many new technologies. Future advancements of video display technologies will be dependent, to a large extent, on advances in flat panel displays.

Work on flat panel displays has been advancing on many fronts simultaneously with the largest fraction of research being conducted in the area of liquid crystal displays (active and passive.) Other popular technologies have been plasma displays, electro-luminescent devices, flat CRTs, and spatial light modulators.

FIGURE 1
PLANAR OPTICS DISPLAY DETAIL



Although it has been known that lasers could provide an inherently high brightness and high resolution display, there has never been a method to accomplish this safely or compactly. Conventional laser projection, like the type employed at laser light shows, can provide a bright image on a flat screen in a tightly controlled environment where no one is able to intercept the laser beam with their eyes. Such a system would be unsafe to the eye and illegal in an uncontrolled home environment. A rear projection laser system could be used which could contain the dangerous levels of light, however, the physical size of such a device is no smaller than that of a conventional CRT. The Planar Optic Display, (POD), being described here uses neither front nor rear projection optics. It is an internal projection system which has the brightness of rear laser projection and the compactness of LCD screens.

OPERATION THEORY

[At the outset, it must be emphasized that the new technology presented here is the flat panel Planar Optic Display screen and not the use of a scanning laser beam to produce a video image. Therefore the lasers themselves and the laser scanner will receive only a basic discussion.]

Fiber optic wave guides have been well understood and used for decades. An internal fiber known as the core (refractive index n) is surrounded by a cladding (refractive index <n) so that light which enters the fiber within a known acceptance angle is confined within the fiber. This confinement occurs because of a process known as total internal reflection. The same process occurs if the internal core is a sheet of glass rather than a fiber of glass. It is understood that each internal core sheet must be adjacent to a sheet of cladding to ensure total internal reflection. Such a device, when constructed with many sheets or planes of glass is called a planar optic display. See Fig. 1. A planar optic device is analogous to a fiber optic device, however, there are a few very important distinctions.

In a fiber optic, the angular information of the incident light beam is lost as the light exits the fiber. In a planar optic sheet, the angle of the incident light (in the plane of the sheet) is preserved at the exit of the sheet. This is a very important characteristic because one now has the capability to direct a laser beam into the entrance of a planar optic sheet and have the same scanning laser beam exit the sheet at a predetermined location. This is crucial to understanding the operation of this Planar Optic Display, (POD), flat panel screen.

Another difference between fiber optics and planar optics is the effect of diffraction on the propagating laser beam within the device. In a fiber optic the beam diameter is confined by the fiber and can never be larger than the fiber diameter. However, in the planar optic sheet, a propagating laser beam will continue to expand in one direction due to the effect of diffraction.

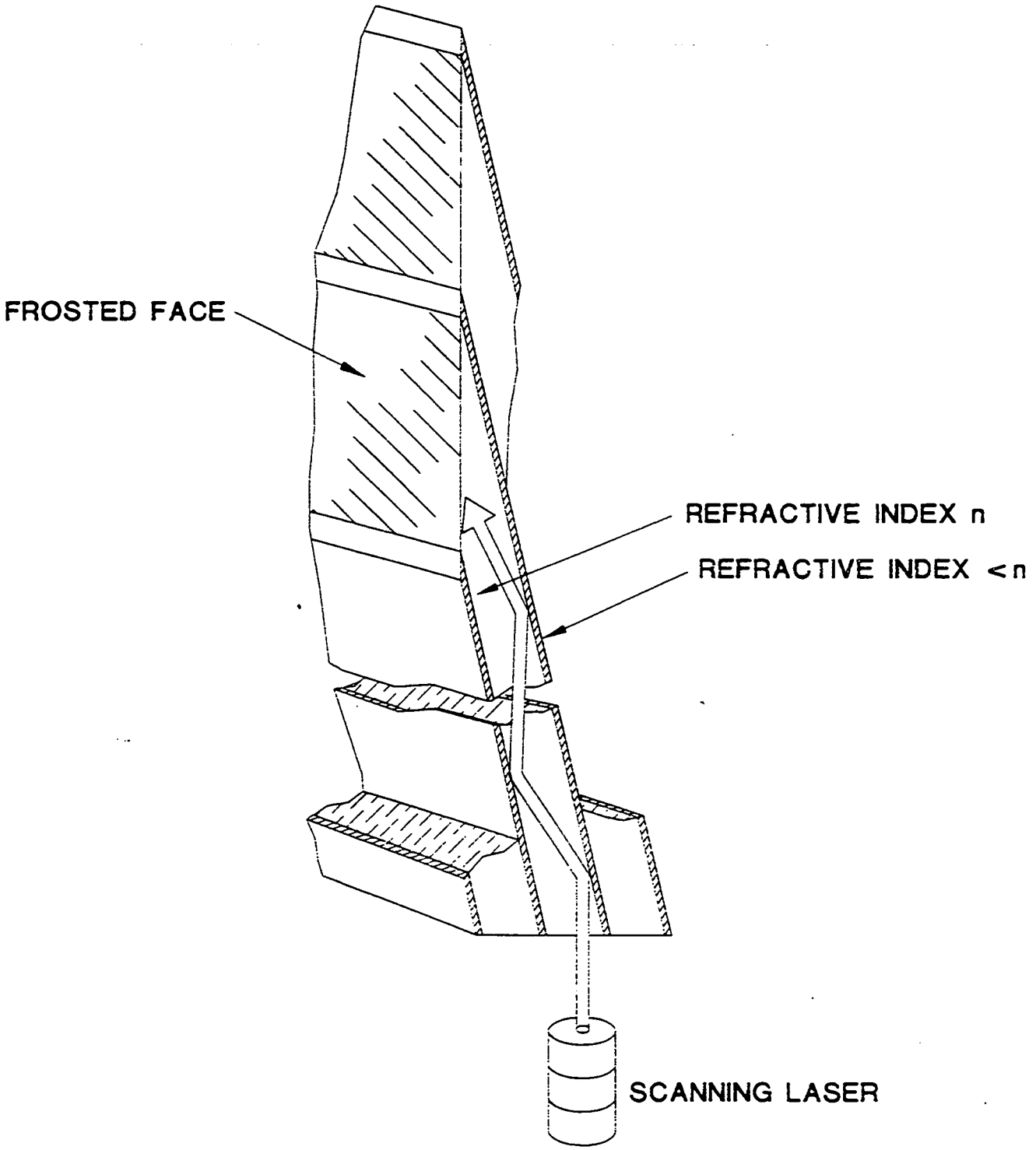
A single planar optic sheet is not very useful by itself, but if you were to stack 1000 different length sheets together, a new type of flat panel screen becomes evident as shown in Figure 1. For the purpose of this discussion, we consider a flat panel POD with a 2 meter diagonal display; 1 meter high by 1.33 meters wide. The panel contains 1,000 planar sheets, each 50 microns thick, for a total thickness of approximately 5cm. (A 6 ft. diagonal display, 2 inches thick)

Vertical Resolution

Figure 2 shows a detail of a section of the planar optic screen. The laser beam enters the planar optic display from the bottom. [For clarity, the laser beam diameter is shown to be much smaller than the planar core, but in reality the laser beam diameter should match the core diameter.] The cladding confines the laser within the core and the scanner rasters the laser beam along the length of the planar sheet. Each planar sheet corresponds to exactly one vertical line of resolution. Therefore, to attain a HDTV format with 1,000 lines of vertical resolution, the screen must contain 1,000 planar sheets. The laser light exits each planar sheet at the front frosted face which

FIGURE 2

PRINCIPLE OF PLANAR OPTICS



diffuses the beam to provide an extremely wide viewing angle, like conventional CRTs. This diffusive face is also makes the screen totally eye safe even if the scanner failed and the laser beam were fixed at a single spot.

Horizontal Resolution

Although the vertical resolution is determined by the number of planar sheets, the horizontal resolution can be limited by 2 factors; the modulation rate of the laser and the spreading effect of diffraction on the laser beam. As the diode laser beam is being scanned from left to right, its current must be modulated fast enough to produce 1,000 discrete pixels. Therefore to achieve a horizontal resolution of 1,000 for a high definition display, the laser diode must be turned on and off (or varied) 1,000 times during each horizontal scan. This high resolution would require the diode laser to be modulated at 1MHz (One million times per second). A conventional frame rate (30 frames per second) would require that the diode be modulated at 30 MHz. Therefore, the diode laser does not appear to be a limiting factor in attaining optimum horizontal resolution.

The effect of diffraction, however, will ultimately limit the attainable horizontal resolution. As the laser beam travels within the planar sheet, a 50 micron diameter laser beam would expand far too much to provide high resolution at the face of the screen 1 meter away. By expanding the laser beam to a diameter of 500 microns, the diffraction effect is now reduced so that the spot size at the screen surface is only 1 mm. Since the screen is 1.33 meters wide, the horizontal resolution limit is 1330. In order to obtain the optimum horizontal and vertical resolution simultaneously, a rectangular, rather than circular, beam profile is needed. Fortuitously, the diode laser, by its very nature, provides just such a beam profile. Thus, the ideal laser beam size for a 2 meter flat panel POD is 50 microns by 500 microns.

A planar optic screen as shown in Figure 1 can be made as small or as large as desired, within the capability of manufacturers to produce the core material for the planar optics. For larger screens, one must use diodes with higher output powers or couple the output from several diodes. The laser scanner for a POD screen can be either mechanical, electro-optic, or acousto-optic depending on the need. For extremely high rate scanning applications, an active matrix liquid crystal display (AMLCD) can be placed on the input face for individual pixel control. Then all 1,000 horizontal lines could be scanned at once. For color applications, each of the three primary colors can be scanned sequentially to reduce the number of required LCD pixels.

Three dimensional TV (using light polarization) becomes a possibility with this Planar Optic Display since each planar optic sheet retains the polarization information of the incident laser beam. A second set of diode lasers with opposite polarization would be integrated with the scanner such that the two polarizations are time interlaced on the surface of the screen.

The question of manufacturing the planar optic screen should now be addressed. Although every POD made to date has used glass, a production screen should be plastic due to weight and cost considerations. After researching this subject it was discovered that the screen manufacturing problem has already been addressed by DOW Chemical Company. They have developed an inexpensive process to coextrude alternating layers of transparent PC/styrene-acrylonitrile with alternating indices of refraction. This is exactly what is needed for a POD screen. Furthermore, they have managed to extrude the material so that the layers are as thin as 1 micron. An excellent description of the process is contained in the literature[2].

EXPERIMENTAL RESULTS

The initial 'proof of principle' experiment consisted of a 2.5 cm (1 in.) planar optic display where the individual glass wave guides were 75 microns thick. The glass sheets were bonded together with a low refractive index and low viscosity epoxy (#328) manufactured by Epoxy Technology Inc. Although the resolution looked good (approximately 35 lines/cm), the contrast was poor because the display was quite opaque. We decided that future displays should be black to improve overall contrast. The next 15 displays were constructed using various epoxies and blackening agents in order to obtain maximum contrast. These displays were 10cm (4 inch), diagonal, and used 200 micron thick glass, as we did not have the thinner glass in large sizes. The optimal display used lampblack (from Fisher Scientific) mixed with VA-6 epoxy (from Epoxy Technology Inc) as the bonding agent for the sheets of glass.

The Prototype

In order to make the 25 cm (10 inch) display, 145 sheets of glass (200 microns thick) were prepared as follows. The material was Schott D 263 borosilicate glass with a refractive index of 1.52. The sheets of glass were cut so that the first sheet was a full size sheet. This becomes the rear of the screen. Each successive sheet was cut 1 mm shorter and placed on top of the previous sheet. The last, and front-most sheet, is now only a thin strip of glass. The sheets were bonded together with the VA-6 epoxy (refractive index = 1.50) and the front face was ground flat to act as the display surface. The bottom face of this stack of glass was ground and polished to act as the light input surface.

In order to test the Planar Optic Display, it was necessary to use a laser projection system. This system was built by Laser Images Inc. (Van Nuys, CA) and uses a He Ne laser which emits 3 milliwatts of light. Vector scanning is accomplished through the use of low mass mirrors on high speed galvanometers. The animation software is called "Choreographics" and runs on a Macintosh computer.

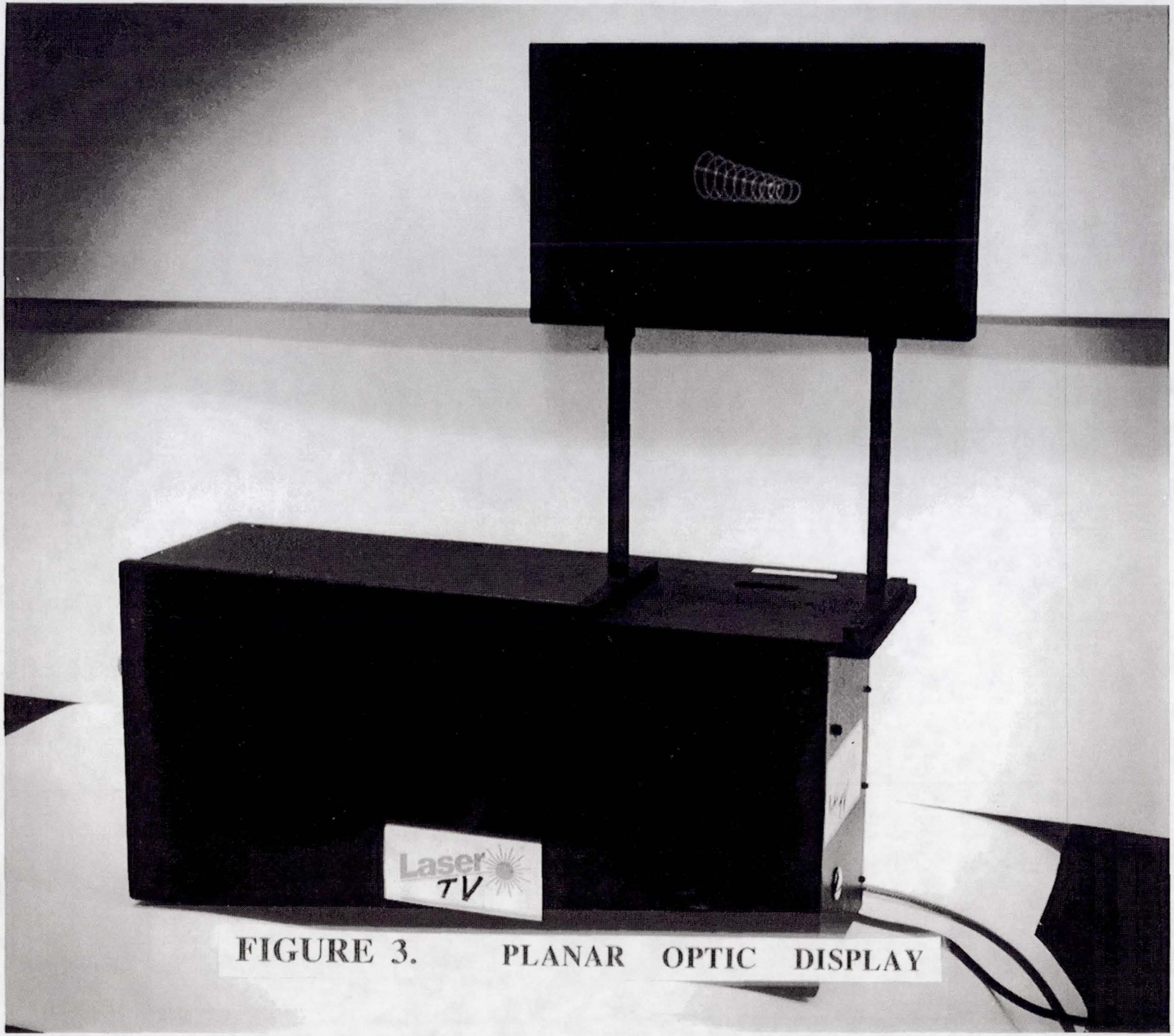


FIGURE 3. PLANAR OPTIC DISPLAY

Display Characteristics

The diffusive nature of the screen allows for a very wide viewing angle, approximately 120 degrees. The brightness of the display has no inherent limit since the source laser controls the ultimate brightness. When the laser is off, the screen appears flat black. Because of this contrast, the colors appear brilliant. The measured resolution of this prototype display is 142 lines vertical and 400 lines horizontal. A photograph of the Planar Optic Display and its scanner is shown in Figure 3.

3 Dimensional Viewing

This display can be used for 3 dimensional (3D) viewing if a polarization algorithm is used and the viewer wears polarized glasses. The POD was found to maintain the polarization information of the laser beam, however a small amount of scattering of the polarization occurs at the diffusive display face. This will not have a deleterious effect on 3D operation since the amount of scattering was measured and found to be small; a crossed polarizer extinction ratio of 100 was measured.

CONCLUSIONS

The Planar Optic Display has been successfully demonstrated using vector scanning techniques running animation sequences. The display has high brightness, high contrast and an exceptionally wide viewing angle.

ACKNOWLEDGMENTS

I would like to thank Mr. Dwayne Branch for his skill in the cutting, handling and bonding of the microthin glass. Mr. Barry Lafler deserves much thanks for his help in the grinding and polishing of the display. I also wish to thank Mitch Hartman at Laser Images for providing the laser scanning system.

REFERENCES

1. *Electronic Engineering Times*, March 22, 1993 Pg. 24
2. *Journal of Plastic Film & Sheeting*, Vol. 4, 1988, Pages 104-115.

DUAL-MODE NON-WASHOUT LIQUID CRYSTAL DISPLAY

William H. McKnight
Naval Command Control and Ocean Surveillance Center
R D T & E Division Code 573
San Diego CA 92152-5000
(619) 553-2485; (619) 553-6068 FAX; e-mail mcknight@nosc.mil

ABSTRACT

A common problem in optical displays is that of washout wherein bright ambient light conditions generate a tendency to diminish the contrast and relative brightness of a transmissive (e.g., CRT/television, etc.) display. In the case of a transmissive (lighted from behind) liquid crystal display, addition of a partially transmitting mirror to the polarizer/liquid crystal panel sandwich in a position adjacent to the light source (behind the panel and rear polarizer) permits the display to operate in both a transmissive and reflective mode simultaneously (as required by ambient light conditions) so that instead of washout under bright ambient light, the reflective mode will enhance the brightness and contrast of the display.

BACKGROUND

A common version of a twisted nematic liquid crystal display (LCD) is depicted in Figure 1 wherein the light source is behind the panel and thus this LCD is operated in a transmissive mode. The twisted nematic liquid crystal molecules are captured in a matrix of tiny cells, each functioning as a light valve or pixel. Each pixel is, of course, electronically addressable for the purpose of supplying the data signal specifying the light level or grey scale level for each pixel. When a pixel is in a quiescent state (no voltage applied between the two optically transparent but electronically conducting plates which contain the liquid crystal), linearly polarized light striking the pixel (from either direction as the panel functions symmetrically) will have its axis of polarization rotated ninety degrees by virtue of the optical properties of the helically shaped twisted nematic molecules. This situation is depicted in Figure 1 where the rear polarizer passes light linearly polarized perpendicular to the plane of the figure as is indicated by the dots in the light ray path and in the rear polarizer. Since, as is indicated in the upper half of the liquid crystal panel (LCP), no voltage is applied, the axis of polarization is rotated ninety degrees and is passed by the front polarizer which has its polarization axis oriented vertically as is indicated by the vertical arrows in the light ray path and the front polarizer. In this case light will pass out the front and be seen by the viewer and the light valve/pixel is considered open.

When a voltage is applied between the LCP surface plates as shown in the bottom half of Figure 1, the molecules are "stretched or straightened out" in response to the electric field

such that they are re-oriented and distorted and they no longer act to rotate the axis of polarization of light passing through them. In this case the crossed configuration of the front and rear polarizers act to block any light from emerging from the front polarizer and the light valve is considered closed. Thus the presence or absence of this controlling voltage determines whether or not the valve is open or closed (light or dark) with this particular configuration of front and rear polarizers.

For this (or any other transmissive) display to operate most effectively (good contrast and brightness), a relatively darkened ambient lighting situation is required. When this is not the case (such as when a bright light is shown on the face of the LCD) the relative brightness and contrast is seriously dimenished. This is the classic problem of washout.

An alternative mode for operating the LCD is in the so-called reflective mode wherein the light source is not located behind the LCP but rather in front such as is the case from ambient light in a lighted room. In this situation, depicted in Figure 2, ambient light (which is artificially supplied in the case of a dark environment) traveling from left to right is vertically polarized as indicated by the vertical arrows in the light ray path and the front polarizer. A fully reflecting mirror has now been inserted in place of the light source behind the panel which was used for illumination in the case of the transmissive mode of operation. As before, a voltage applied between the transparent conducting plates containing the liquid crystal (bottom half of Figure 2) serves to prevent the molecules from rotating the linear polarization axis and thus the light is blocked from striking the mirror which causes this pixel to appear dark. With no voltage applied however, as depicted in the upper portion of Figure 2, the twisted nematic molecules once again serve to rotate the polarization axis ninety degrees, thus allowing the light to pass through the second (back) polarizer and strike the mirror. The mirror then, in effect, serves as a behind-the-panel light source in the manner of the transmissive mode of operation as described above. That is, the light from the mirror passes back through the rear polarizer, then the panel itself where the polarization axis is once again rotated ninety degrees and then on through the front polarizer and out to the viewer. Thus, these pixels appear bright.

An alternate relative orientation of the two polarizers permits a reverse contrast operational mode for each of the transmissive and reflective modes described above. This is best visualized if one imagines that the rear polarizer is rotated by ninety degrees in each of the two modes (transmissive and reflective) described above. Figure 3 shows a transmissive version of this configuration where the arrows in the rear polarizer denote a vertical axis of polarization. Thus, following the ray diagram with the arrows showing the axis of polarization, one sees that no light emerges when no voltage is applied (top half of Figure 3) and the pixels do pass light when voltage is applied. This reverses the contrast relative to the previously described transmissive mode of operation. In a similar manner, the contrast is reversed relative to the original description of the reflective

mode by orienting the two polarizers in the same direction as is depicted in Figure 4. This configuration also provides bright pixels when voltage is applied and dark pixels with no voltage, which is, of course, opposite to the original configuration for the reflective mode. Of course there are other possible combinations of polarizer orientations, each providing one or the other of the two contrast modes available.

DESCRIPTION

To reduce or correct for the washout problem (which only occurs in the transmissive mode) one simply needs to place a partially transmitting mirror in front of the light source and behind the rear polarizer as shown in Figure 5 (this is otherwise the same configuration as Figure 1). In so doing, the ambient light which would otherwise serve to washout the transmissive mode display would be partially reflected by the mirror and thus function in a reflective mode. Figure 6 shows the effect of adding the partially transmitting mirror to the LCP configuration of Figure 3. Of course, the effect of adding the partially transmitting mirror would be the same for all possible polarizer configurations for the transmissive mode of LCP. That is, adding the partially transmitting mirror would serve to enhance the contrast of a transmissive LCP in all cases by using ambient light which would otherwise cause washout. The percentage of light permitted to pass through the mirror would be optimized for a given application in accordance with the extent of the ambient light anticipated to otherwise create washout and also in accordance with brightness, contrast and other operational requirements.

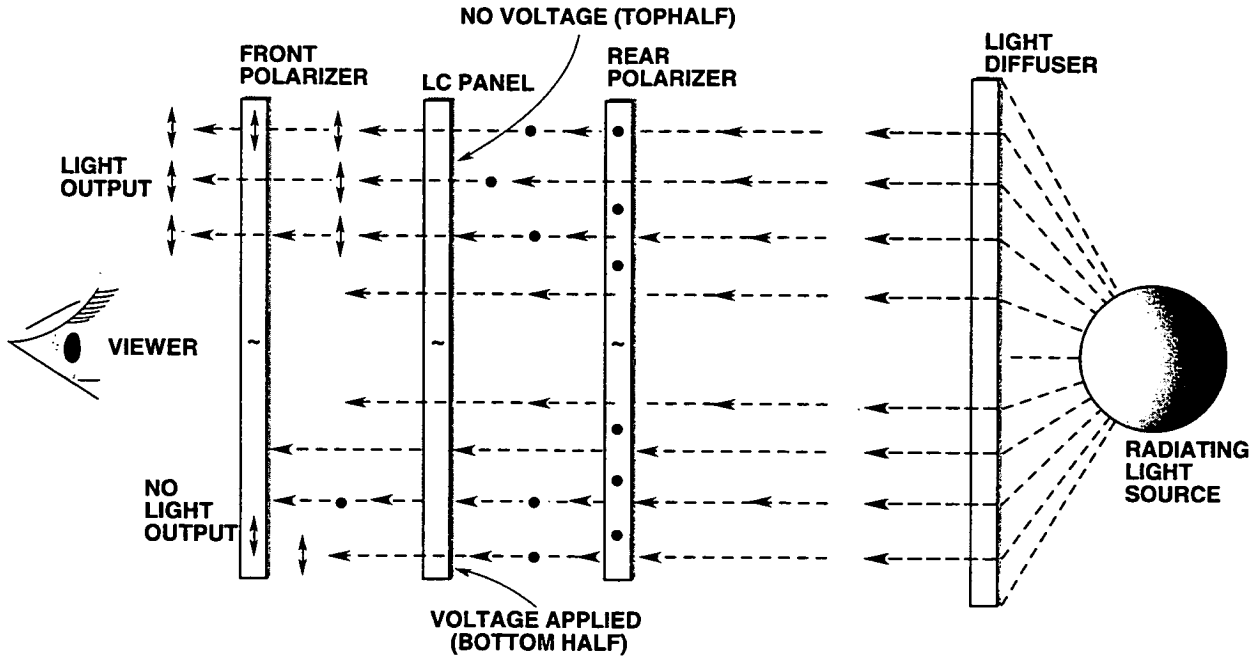

NO VOLTAGE (TOPHALF)
FRONT POLARIZER
LC PANEL
REAR POLARIZER
LIGHT DIFFUSER
LIGHT OUTPUT
VIEWER
NO LIGHT OUTPUT
RADIATING LIGHT SOURCE
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 1

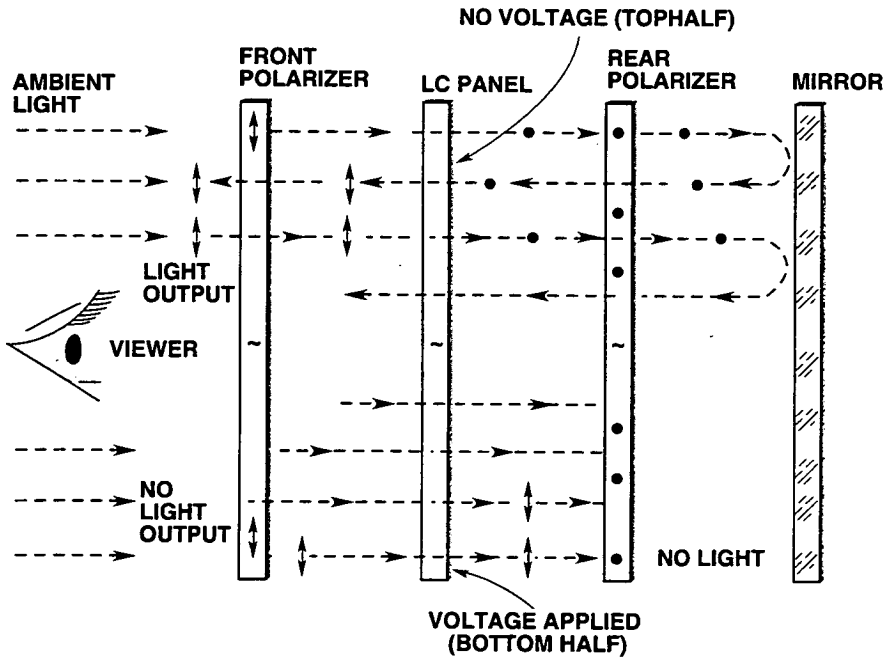

NO VOLTAGE (TOPHALF)
AMBIENT LIGHT
FRONT POLARIZER
LC PANEL
REAR POLARIZER
MIRROR
LIGHT OUTPUT
VIEWER
NO LIGHT OUTPUT
NO LIGHT
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 2

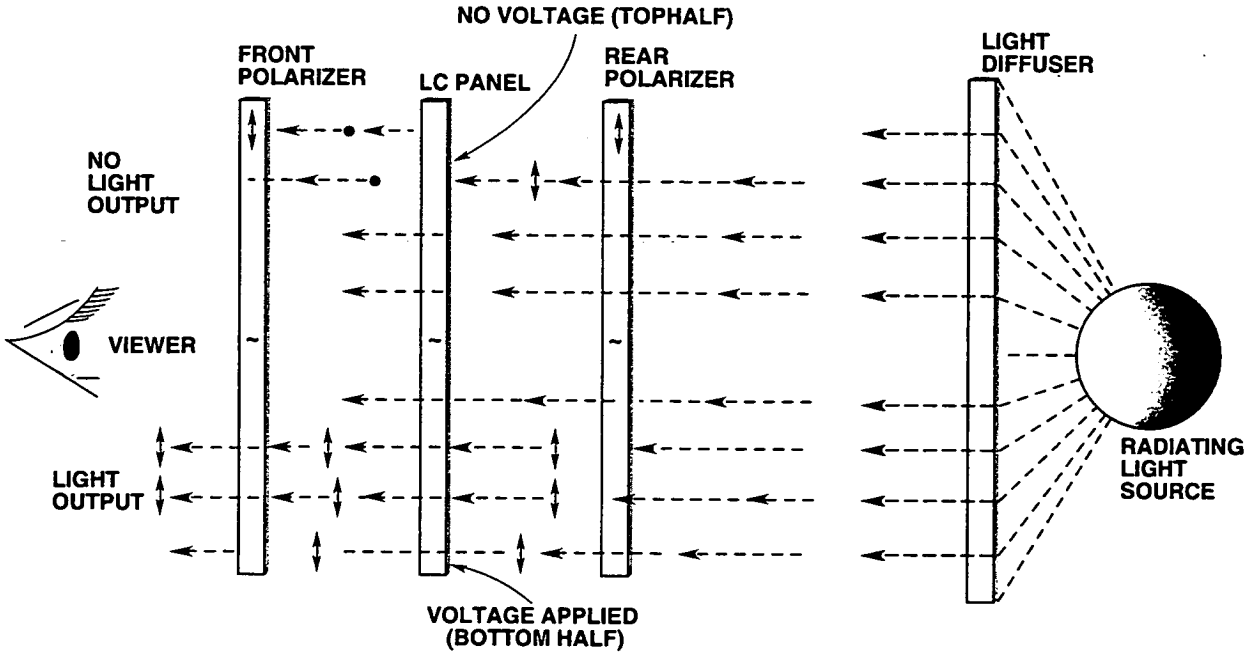

NO VOLTAGE (TOPHALF)
FRONT POLARIZER
LC PANEL
REAR POLARIZER
LIGHT DIFFUSER
NO LIGHT OUTPUT
VIEWER
LIGHT OUTPUT
RADIATING LIGHT SOURCE
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 3

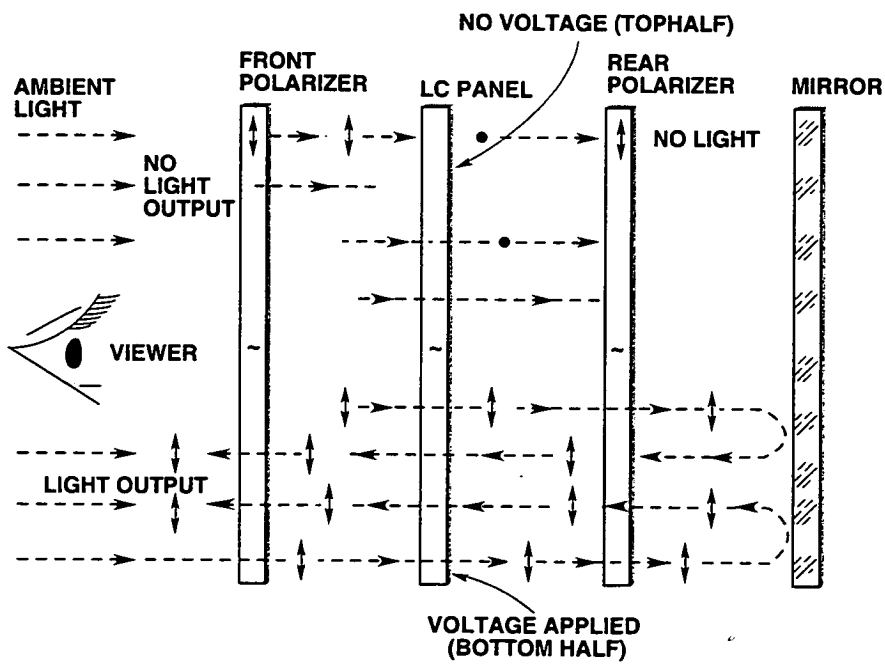

NO VOLTAGE (TOPHALF)
AMBIENT LIGHT
FRONT POLARIZER
LC PANEL
REAR POLARIZER
MIRROR
NO LIGHT
NO LIGHT OUTPUT
VIEWER
LIGHT OUTPUT
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 4

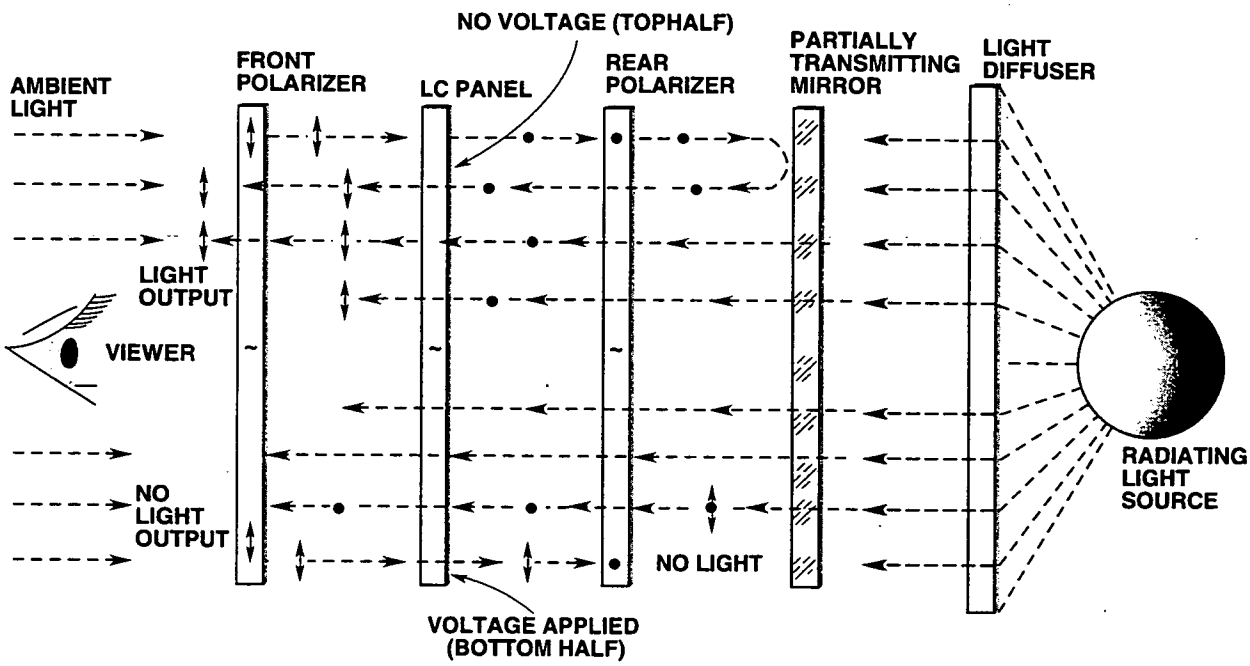

NO VOLTAGE (TOPHALF)
AMBIENT LIGHT
FRONT POLARIZER
LC PANEL
REAR POLARIZER
PARTIALLY TRANSMITTING MIRROR
LIGHT DIFFUSER
LIGHT OUTPUT
VIEWER
NO LIGHT OUTPUT
NO LIGHT
RADIATING LIGHT SOURCE
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 5

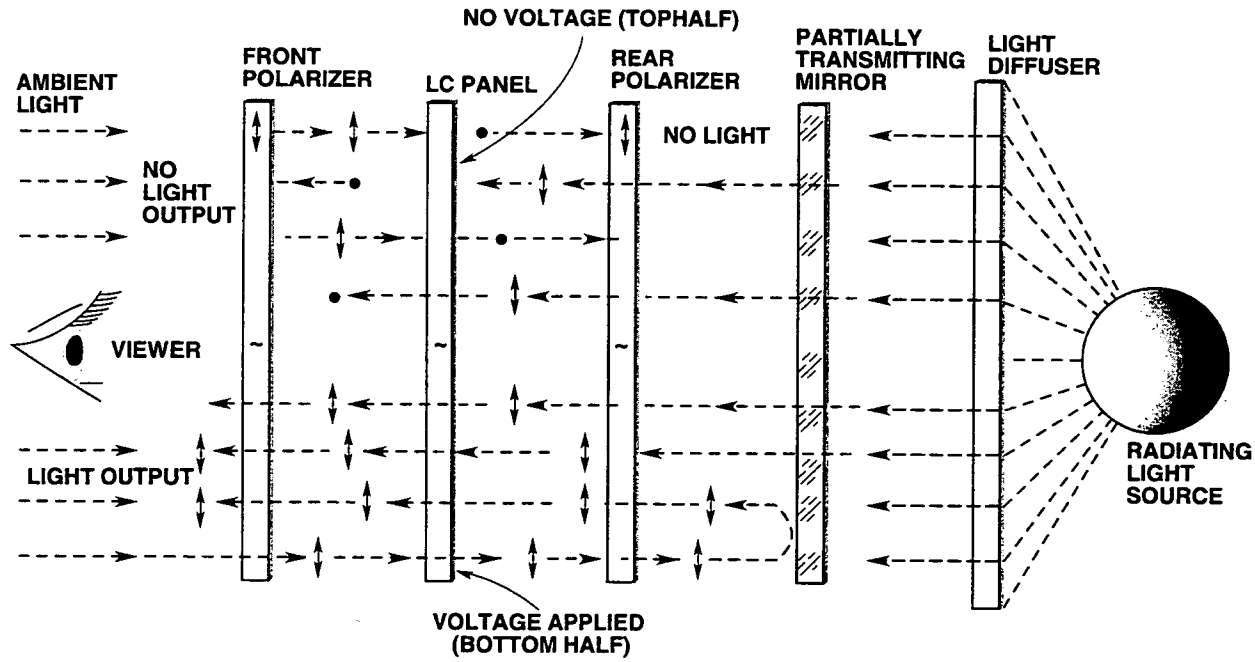

NO VOLTAGE (TOPHALF)
AMBIENT LIGHT
FRONT POLARIZER
LC PANEL
REAR POLARIZER
PARTIALLY TRANSMITTING MIRROR
LIGHT DIFFUSER
NO LIGHT
NO LIGHT OUTPUT
VIEWER
LIGHT OUTPUT
RADIATING LIGHT SOURCE
VOLTAGE APPLIED (BOTTOM HALF)

FIG. 6

DEVELOPMENT OF CMOS ACTIVE PIXEL IMAGE SENSORS FOR LOW COST COMMERCIAL APPLICATIONS

Russell C. Gee, Sabrina E. Kemeny, Quiesup Kim, Sunetra K. Mendis, Junichi Nakamura, Robert H. Nixon, Monico A. Ortiz, Bedabrata Pain, Craig Staller, Zhimin Zhou, and Eric R. Fossum
Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109

ABSTRACT

The Jet Propulsion Laboratory, under sponsorship from the NASA Office of Advanced Concepts and Technology, has been developing a second-generation solid-state image sensor technology. Charge-coupled devices (CCDs) are a well-established first generation image sensor technology. For both commercial and NASA applications, CCDs have numerous short comings. In response, the active pixel sensor (APS) technology has been under research. The major advantages of APS technology are the ability to integrate on-chip timing, control, signal-processing and analog-to-digital converter functions, reduced sensitivity to radiation effects, low power operation, and random access readout.

JPL has been exploring a complementary metal-oxide-semiconductor (CMOS) APS technology. CMOS is a widely used microelectronics technology for microprocessors, memory devices and application specific integrated circuits (ASICs). Use of CMOS for image sensors can reduce fabrication costs by over a factor of three compared to CCD image sensors. Thus, the JPL-developed technology has widespread appeal for low cost commercial applications such as video phones, computer input devices, surveillance and robotics.

The JPL CMOS APS technology is being successfully transferred to industry. The sensors feature TTL-compatible operation (5 volts or less), random accessibility, 75 dB dynamic range, fixed pattern noise less than 0.1 %, read noise in the 15-25 electron r.m.s. range, and low power.

This paper will discuss the development of CMOS APS technology, its performance characteristics, and its application to low cost commercial products.

1. INTRODUCTION

Imaging system technology has broad applications in commercial, consumer, industrial, medical, defense, and scientific markets. The development of the solid-state charge-coupled device (CCD) in the early 1970's led to relatively low cost, compact imaging systems compared to vidicons and other tube technology. The CCD has advanced as the microelectronics industry has improved silicon material quality and device fabrication technology. Today, in mass production, CCDs are made at the rate of over 10 million imager chips per year in Japan (Sony, Matsushita, and NEC dominate production) mostly for video camcorder applications. At this production rate, a CCD has a manufacturing cost of approximately $10-$15 per chip, or about $50/Mpixel (million pixels). Unfortunately, these large production runs are mostly used in vertically integrated products so that the cost for low volume external purchase of CCDs is typically much higher. Megapixel CCD sensors, desired for low volume applications, are typically made in the U.S. or Europe rather than Japan and cost in the neighborhood of $1,000/Mpixel. Scientific-grade defect-free sensors can cost as much as $10,000/Mpixel. (HDTV format sensors with 2M pixels, will enter production in Japan in a few years and will lower the cost of megapixel sensors significantly.)

The major reason why megapixel CCDs are so expensive is related to the high cost of fabrication equipment that must be amortized over low volume production runs. Furthermore, modern CCD technology is a significant departure from mainstream microelectronics fabrication technology -- complementary metal-oxide-semiconductor or CMOS, which is used for most microprocessor and ASICs. CMOS technology is backed by an enormous worldwide R&D workforce and infusion of capital. In contrast, advancement of CCD technology is limited by a lack of both investment capital and worldwide level of effort.

Technology 2004, Video/Imaging (1994).

CMOS technology advancement has been rapid. This advancement has been following the well-known trend that microelectronic device feature size decreases by approximately a factor of two every five years. In large volume production, six-inch CMOS wafer fabrication costs approximately $1,000 per wafer. Thus, a CMOS image sensor with a 10 micron pixel pitch might have a manufacturing cost of approximately $10/Mpixel, or about five times less than a CCD. For lower volume production, the cost of fabricating a six-inch CMOS wafer is about the same as fabricating a four-inch CCD wafer. A six inch wafer yields about three times the number of large-sized chips as a four-inch wafer so the manufacturing cost of a CMOS image sensor would be approximately three times less than a CCD image sensor.

The use of CMOS presents an additional opportunity for significantly reducing imaging system cost, power and mass as well as improving reliability. A CMOS-based image sensor can be readily integrated with on-chip timing, control, signal chain and analog-to-digital converter (ADC). Unlike a CCD system that requires a large number of power supply voltages, clock drivers that can drive large capacitances, a discrete component signal chain, and an ADC chip, the CMOS sensor can be a single-chip camera system with a full digital interface. The image sensor can communicate directly with a microprocessor or computer, significantly reducing system complexity and concomitant development time. Other advantages of the CMOS APS are: TTL-compatible operation (0-5V), only a single power supply is required, electronic shuttering, readout windowing, variable integration time and pixels in the array can be addressed randomly.

The major hurdle to realizing the economic benefit of utilizing CMOS-based image sensors has been the performance of the sensor. Until recently, CCD imager performance has been vastly better than its CMOS counterpart. In this paper, a high performance CMOS sensor technology competitive with CCDs and suitable for many scientific, commercial, consumer, industrial, medical and defense applications is described.

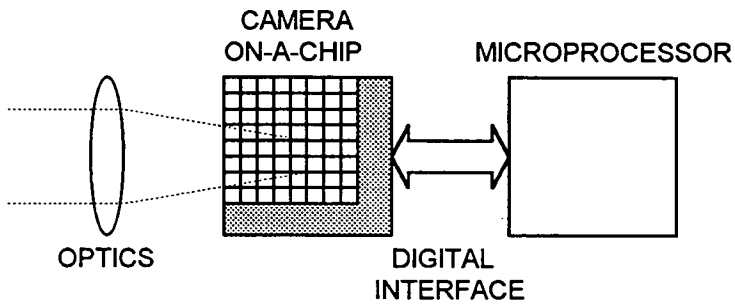


Fig. 1. Block diagram of highly integrated, low mass, low power, compact imaging system

2. CMOS ACTIVE PIXEL SENSOR

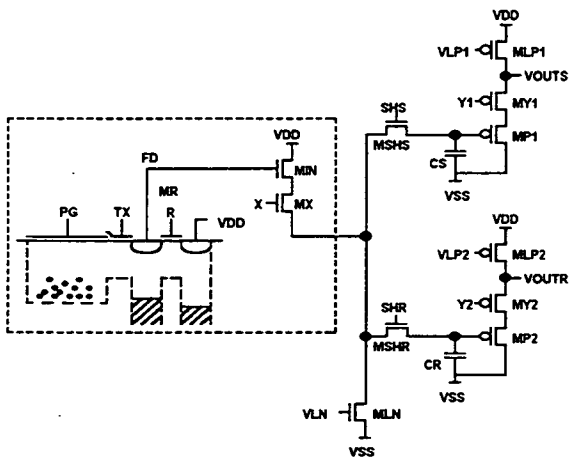


Fig. 2. Circuit diagram of CMOS APS

The Jet Propulsion Laboratory, California Institute of Technology, has recently developed a CMOS active pixel image sensor (APS) technology. This technology is a second generation solid state imager technology that greatly improves the performance of CMOS image sensors to a level comparable to CCDs [1-3]. Each pixel consists of a photoactive region that is a MOS photogate detector, similar to the structure employed in CCDs. The pixel also contains a transfer gate and a floating-diffusion source-follower output amplifier, similar to those employed in the output stage of a CCD. The output transistor is within the pixel, hence the name active pixel sensor. The in-pixel source-follower converts the photogenerated signal into a voltage. The pixel is addressed by a row select switch, and the output of the transistor is fed to a vertical wire running down the column. The voltage on this column bus is sensed by an amplifier located at the bottom of each column.

The signal is sampled onto a holding capacitor for readout. The per-column signal chain has two capacitors, one for sensing the output of the floating diffusion after reset, and the second for sensing the output following intra-pixel transfer of the signal charge. The two capacitors are buffered by a second source-follower stage that is scanned and selected for readout. The differential output permits correlated double sampling (CDS) of the pixel that suppresses pixel kTC noise, 1/f noise, and fixed pattern noise due to threshold voltage offset. The signal chain is shown in figure 2.

A layout of a photogate CMOS APS pixel is shown in figure 3. As seen in figure 3, the optical fill-factor (percentage of pixel area designed for photodetection) of the APS is approximately the same as an interline CCD (25-30%), but lower than for a full frame CCD. On-chip microlenses are used on interline CCDs to boost the effective optical fill-factor to over 60% and could also be used with the CMOS APS by requiring an additional backend processing step (as is the case for on-chip color filter arrays). JPL has performed intra-pixel laser scanning of its CMOS APS devices and has found that for n-well, n-channel implementations, significant response is obtained in regions not designed for photodetection. Photons generating carriers in these regions are not blocked by polysilicon, and the generated carriers diffuse laterally to the collecting potential well. While this introduces a small amount of crosstalk, the resultant per-pixel quantum efficiency is measured to be close to that of a full frame CCD.

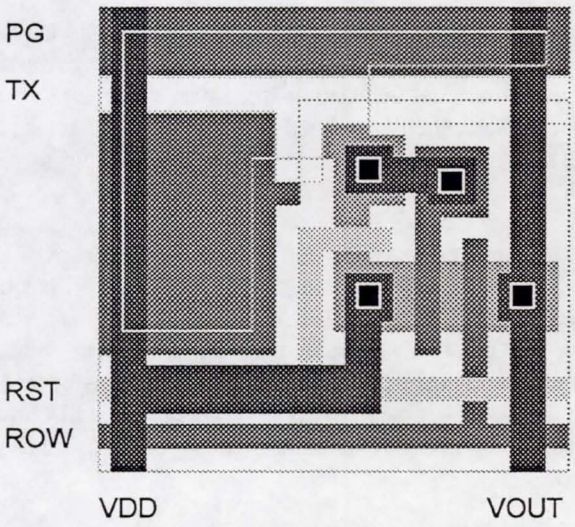


Fig. 3. Typical layout of CMOS APS photogate pixel

An absolute quantum efficiency curve for a CMOS APS implemented with approximately a 25% optical fill factor and no coatings is shown below in figure 4. Note the absence of fringe patterns in the measured quantum efficiency normally associated with CCD overlapping polysilicon gates. Also, the good near infrared (NIR) response of this n-well, n-channel device allows for scientific imaging in this spectral band (0.7-1.0 μm). Improvement in blue/UV response is desired and can be achieved using phosphors (e.g., lumogen) and/or anti-reflection coatings. Improved device design is also expected to boost blue response. The pixel can also be implemented using a photodiode detector structure. The photodiode has the advantage of increased blue response by eliminating the polysilicon overlayer, but has larger capacitance (lower conversion gain, μV/e⁻) and its kTC noise cannot be suppressed on-chip. Thus, the signal-to-noise ratio remains nearly the same as for the photogate implementation, though this structure is simpler to design and operate. A pinned photodiode structure, as that employed in interline CCDs, can be used to suppress kTC noise, but introduces a non-standard CMOS process variation.

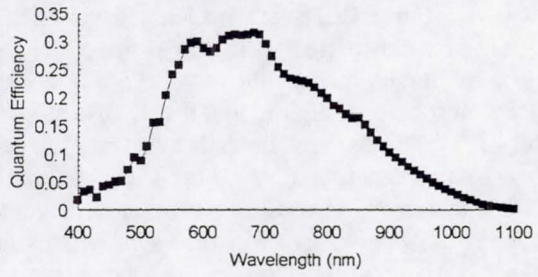


Fig. 4. Measured absolute quantum efficiency in 20 μm CMOS photogate APS pixel (no coatings).

Output-referred conversion gain in the CMOS APS depends on the capacitance of the floating diffusion output node. Typical values are 7 μV/e⁻ (n-well, n-channel), and 3 μV/e⁻ for a photodiode. So-called "full-well" of the APS sensor is determined by the saturation of the signal chain rather than the photogate potential well capacity, and is typically 1.2 V output-referred, or 170,000 electrons for the photogate device. Increasing or decreasing the supply rails results in a change in saturation level of about 0.5 V/V.

Readout noise in the CMOS APS is presently limited by excess noise from the pixel output transistor, though theoretically limited by kTC noise on the sampling capacitors at the bottom of the column. These capacitors are typically 4 pF yielding a theoretical output-referred kTC noise of 45 μV r.m.s. Typical output referred noise levels are 180 μV r.m.s. with 5 V power supply operation, and 100 μV r.m.s. at 3 V power supply operation. Thus, the experimental noise level is typically 25 electrons r.m.s. with 76 dB of dynamic range. Noise as low as 14 electrons r.m.s. has been obtained at 3 V operation. Improvement in these values is expected in the next year.

Room temperature dark current in the CMOS APS is typically 200 mV/sec, or 1 nA/cm^2 -- typical of MOS devices including most CCDs. In the photodiodes, average dark currents an order of magnitude less are typically observed, though the percent fluctuation is greater, as is the incidence of "hot" or white pixels. Cooling is expected to reduce dark current, which is utilized in scientific CCD sensors. The use of non-standard CMOS fabrication steps can be used to reduce dark current to levels comparable to scientific inverted-surface CCDs, but for most commercial applications this is not necessary.

Fixed pattern noise (FPN), often a concern in active pixel image sensors, has been reduced to negligible levels. The D.C. offset variations seen from pixel to pixel has been observed to be comparable to CCDs -- typically 1-2%. The conversion gain/quantum efficiency variations are dominated by column-to-column variations, since threshold offset per pixel is suppressed by the CDS operation. A double-delta sampling (DDS) technique for on-chip suppression of column-to-column variations has been developed at JPL that suppresses column-wise FPN to less than 0.1% sat. -- a nearly unobservable level. Off-chip FPN suppression for removing pixel-to-pixel variations is typically employed in scientific CCD applications (dark frame subtraction) and is readily applicable to CMOS APS sensors.

The voltage-mode, random-access readout of the CMOS APS allows functions not easily implemented using CCDs. The nominal CMOS APS architecture uses row and column decoders for selecting pixels for readout. Window-of-interest readout is easily implemented in the CMOS APS and is useful for star trackers and optical communications. Variable integration periods for different windows can also be achieved -- a function useful for tracking stars of greatly different magnitudes, or for scientific sensors for spectroscopy, where some spectral bands have weak signals. Windowed readout can also be used for electronic panning in large arrays, where a limited instantaneous field of view is desired. Such an approach is useful in surveillance applications.

The voltage mode readout has another large advantage over CCDs. Since CCDs are read out by physically transporting the signal charge to the output amplifier, charge must be transported with nearly perfect charge transfer efficiency (CTE), i.e., no charge can be lost due to traps or spilling en route to the amplifier. For a large number of transfers (e.g. 10,000) the transfer efficiency per transfer must be very high (e.g. 0.999999) so that the net transfer efficiency ($0.999999^{10000} = 0.99$) is reasonable. This high CTE requirement leads to several inherent problems with CCDs. First, CCDs require large clocking voltages (10-15 volts) to maintain high CTE. Secondly, CCD performance degrades with increasing array size unless CTE is increased. Also, CCD performance degrades with increasing readout rate since CTE drops at higher transfer speeds. CCD performance degrades in the presence of trap-inducing radiation (especially protons) and CCD performance degrades at low temperatures due to trapping of signal charge. On the other hand, the CMOS APS technology does not suffer from these limitations.

Power dissipation in the CMOS APS technology is very low. Although column-wise readout requires many source-follower circuits operating in parallel, these circuits are typically biased at 10 μA. These circuits are only activated to sample the data onto the holding capacitors so the duty cycle is low, perhaps 1% or less, depending on array size. Driving analog data off-chip requires a larger bias current to charge cable capacitance at the serial data rate. Only one source-follower is on at a time so that the situation is comparable to a CCD output amplifier. However, in the CMOS APS the amplifier supply voltage is only 3-5 volts (compared to a CCD biased at perhaps 20 volts) so that the CMOS APS dissipates a factor of 4 or more less power. Typical APS power dissipation at serial data rates of 100 kpixels/sec to 1 Mpixels/sec is under 10 mW.

3. ON-CHIP ELECTRONICS

Integration of on-chip electronics leads to an enormous decrease in system power dissipation as well as in system electronics volume and mass. Since radiation shielding of the electronics volume is often required for deep space missions, additional leverage for mass reduction is obtained. CMOS technology has been developed specifically for very large scale integration (VLSI) of microelectronic circuits. Implementation of the image sensor in CMOS enables massive on-chip electronics integration. These electronics include timing and control electronics, and output signal chain. For example, JPL has demonstrated a 128x128 element CMOS APS chip that requires only +5 V power and a master clock to continuously produce video output. The chip has additional digital control input lines for commanding the window of readout (by inputting the addresses of the window boundaries) and for digitally controlling the interframe integration time (by inputting a 32-bit word delay). The chip has integrated per-column CDS circuitry, and integrated DDS circuitry for suppressing column-wise FPN to below 2% sat. The chip was implemented in 1.2 μm CMOS technology through a commercial foundry and has a 19.2 μm photodiode-type pixel pitch.

On-chip analog-to-digital conversion (ADC) can permit a full digital interface, since output data is digital. There are many approaches to on-chip ADC and a full discussion is presented in reference [4]. JPL has demonstrated a small image sensor chip (32x32 elements) with a column-parallel single-slope ADC architecture. The on-chip ADC had 8 bit resolution and operated at 30 frames per second. The ADC was found to have excellent linearity and capable of at least 10 bit resolution (i.e., less than 500 μV noise). Most interesting was the incorporation on-chip ADC reduced on-chip power dissipation from 7 mW to 5 mW. This is ascribed to the power reduction obtained by using digital output amplifiers rather than analog output amplifiers that more than offset the power of the on-chip ADC.

4. FUTURE IMPROVEMENTS

Improved performance through continued R&D is expected in the next few years. Reduction in noise and increased digital resolution through oversampled ADC technology [5] is anticipated. Improved quantum efficiency through the use of optical coatings, backside thinning and smaller CMOS feature sizes is also expected. At the present time, several megapixel-class CMOS APS sensors are being designed for fabrication. Very high speed imaging for large format sensors is also expected to be demonstrated. "Smart sensors" are also expected to be demonstrated through the use of on-chip CMOS signal processing circuits.

5. ACKNOWLEDGMENTS

The work presented in this paper was carried out at the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the National Aeronautics and Space Administration, Office of Advanced Concepts and Technology. The authors appreciate the support of C. Kukkonen, S. Khanna, and V. Sarohia of JPL, and G. Johnston of NASA Headquarters. The authors also appreciate discussions with A. Dickinson, S. Eid and D. Inglis of AT&T Bell Laboratories.

6. REFERENCES

1. E.R. Fossum, *Active Pixel Sensors -- Are CCDs Dinosaurs?,* in CCD's and Optical Sensors III, Proc. SPIE vol. 1900, pp. 2-14, (1993).

2. S. Mendis, S.E. Kemeny, R. Gee, B. Pain, and E.R. Fossum, *Progress in CMOS active pixel image sensors,* in Charge-Coupled Devices and Solid State Optical Sensors IV, Proc. SPIE vol. 2172, pp. 19-29 (1994).

3. S.Mendis and E.R. Fossum, *CMOS active pixel image sensor,* IEEE Trans. Electron Devices, vol. 41(3), pp. 452-453 (1994).

4. B. Pain and E.R. Fossum, *Approaches and analysis for on-focal-plane analog-to-digital conversion,* in Infrared Readout Electronics II, Proc. SPIE vol. 2226, paper 24 (1994).

5. S. Mendis, B. Pain, R. Nixon, and E.R. Fossum, *Design of a low-light-level image sensor with an on-chip sigma-delta analog-to-digital conversion,* in CCD's and Optical Sensors III, Proc. SPIE vol. 1900, pp. 31-39 (1993).

ENHANCING THE GALILEO DATA RETURN USING ADVANCED SOURCE AND CHANNEL CODING

Kar-Ming Cheung
Jet Propulsion Laboratory
Pasadena, CA 91109

Kevin Tong
Jet Propulsion Laboratory
Pasadena, CA 91109

Todd Chauvin
Jet Propulsion Laboratory
Pasadena, CA 91109

ABSTRACT

Data compression and error correction are two indispensable building blocks in modern digital communication systems. Data compression conserves transmission and storage bandwidth by removing redundancy in the source data. Error correction introduces redundancy to the data in a controlled fashion to eliminate channel errors. A combination of both techniques ensures efficient and reliable transmission of information from one point to another. In this article, we describe a part of the advanced data compression and error correction coding schemes that we plan to use to enhance the data return from the Galileo spacecraft. The Galileo spacecraft is now on its way to Jupiter. In April 1991, the spacecraft's 4.8-m high-gain antenna failed to fully deploy when commanded. The only way to communicate between the spacecraft and Earth is through the use of one of the two low-gain antennas. A major effort is now being conducted to maximize the data return and to achieve an expected 70% of the original science objectives. This includes mission replanning, in-flight reprogramming, and major ground upgrades. In this article, we discuss the error-correction coding and the image compression schemes only.

INTRODUCTION

In this paper, we describe a portion of the advanced source and channel coding schemes that we plan to use to enhance the data return from the Galileo spacecraft. The Galileo spacecraft, which was launched in October 1989, is now on its way to Jupiter. Its mission includes releasing a probe into the Jovian atmosphere, Io flyby, probe data capture and relay, Jupiter orbital insertion, and 10 satellite encounters (with Jupiter's moons Ganymede, Callisto, and Europa). The Galileo Project involves many years of effort. In April 1991 the Galileo team commanded the spacecraft to open the 4.8-m high-gain antenna (HGA). The HGA failed to completely deploy. All indications are that 3 of the 18 ribs are stuck to the antenna's central tower. Several unsuccessful attempts have been made to free the stuck ribs. If the HGA fails to deploy, the only way to communicate between Earth and the spacecraft is through the use one of the two low-gain antennas (LGAs). If the current configuration (ground and spacecraft) remained unchanged, the telemetry data rate would be 10 bits per second at Jupiter arrival, compared to the originally expected data rate of 134 kbits per second in the planned HGA configuration. A study was conducted from December 1991 through March 1992 to evaluate various options for improving Galileo's telemetry downlink performance [1]. The recommendations from this study, which are now being implemented, include mission replanning, in-flight re-programming of the flight software, and major ground upgrades. Image and instrument data as well as spacecraft calibration and monitoring data are expected to be heavily edited and compressed using the

Galileo's onboard processors, which has severely limited computation and memory resources. It is expected that data compression will increase the effective data rate by about a factor of 10; other spacecraft and ground enhancements, including the advanced error correction coding, would increase the data rate by about another factor of 10.

We first describe the use of the integer cosine transform (ICT) scheme for lossy image compression. The ICT can be viewed as an integer approximation of the popular discrete cosine transform (DCT) scheme, which is regarded as one of the best transform techniques in image coding. Its independence from the source data and the availability of fast transform algorithms make the DCT an attractive candidate for many practical image processing applications. In fact, the ISO/CCITT standards for image processing in both still-image and video transmissions include the two-dimensional DCT as a standard processing component in many applications. Other than the plain ICT image compression, we also developed other image analysis and post processing schemes. We collaborated with Ames scientists and the Galileo Solid State Imaging (SSI) Team to conduct an extensive Principal Investigator (PI)-in-the-loop evaluation of the ICT compression scheme. The goal of this experiment was to find the appropriate compression parameters and the optimal compression/distortion trade-off. We introduced the idea of an addressable truth window, where a small window of the image is losslessly compressed and the rest is lossy compressed, to achieve multiple science objectives in the Galileo severe bandwidth-critical environment. We developed a compression ratio prediction scheme based on the generalized Gaussian function (GGF) to facilitate mission planning. We also developed a number of post-pass image restoration techniques to remove the undesirable blockiness and checkerboard effects due to ICT compression.

The heavily edited and compressed image and instrument data will be protected by an advanced error-correction coding scheme, which uses interleaved variable redundancy Reed-Solomon (RS) codes as outer codes, and a long constraint length convolutional code as the inner code. The ground decoder, which is implemented in software in a multi-processor workstation, employs a feedback mechanism that passes intermediate decoding information from the outer code to the inner code to facilitate multiple pass decoding. This decoder is known as the feedback concatenated decoder (FCD), and is able to achieve a final bit error rate (BER) of 10^{-7} at a 0.65 dB signal-to-noise ratio (SNR).

Finally we discuss the issue of interaction between data compression and error control (containment/detection/correction) processes in the Galileo communication system design. The ICT and most other data compression schemes have the undesirable effect of error propagation. That is, a small loss in compressed data can cause a big loss in reconstructed data. The nature of error propagation depends on the compression schemes being used. In the Galileo LGA mission operation scenario, the volume of data returned will be drastically reduced. To maximize the scientific objectives with the limited transmission bandwidth of the LGA, most of the data are expected to be heavily edited and compressed. These valuable data must be safeguarded against catastrophic error propagation caused by channel errors and other unforeseeable errors. Galileo's ICT scheme is equipped with a simple but effective error containment strategy. The basic idea is to insert synchronization markers and counters at regular intervals (every compressed 8 lines of data in the 800 line images) to delimit uncompressed data into independent blocks. If an anomaly occurs during the transmission of data, the decompressor can search for the synchronization marker and continue to decompress the rest of the data. The error containment strategy guarantees that error propagation will not go beyond 8 lines of data.

The data compression and error correction schemes described above have been implemented and tested. Other than deep space applications as described in the Galileo

scenario, these algorithms are also ideal for commercial applications. The ICT scheme can be used in low-cost high-speed image/video commercial applications. The FCD scheme can be used for power-constrained and bandwidth critical satellite communications.

Galileo's Image Compression Scheme

The Galileo image compression scheme is a block-based lossy image compression algorithm that uses an 8 x 8 ICT. The ICT was first proposed in [2], and was streamlined and generalized in [3][4]. The elements in an ICT matrix are small integers with sign and magnitude patterns that resemble those of the DCT matrix. Also the rows of the ICT matrix are orthogonal. The integer property eliminates real multiplication and real addition operations, thus greatly reducing the computational complexity. The orthogonality property ensures that the inverse ICT has the same transform structure as the ICT. Notice that the ICT matrix is only required to be orthogonal, but not orthonormal. However, any orthogonal matrix can be made orthonormal by multiplying it by an appropriate diagonal matrix. This operation can be incorporated in the quantization (dequantization) stage of the compression (decompression) scheme, thus sparing the ICT (inverse ICT) from floating-point operations, and at the same time preserving the same transform structure as in the floating-point DCT (inverse DCT). The relationship between the ICT and DCT guarantees efficient energy packing and allows the use of some fast DCT techniques for the ICT. The ICT matrix used in the Galileo mission is given as follows

1	1	1	1	1	1	1	1
5	3	2	1	-1	-2	-2	-5
3	1	-1	-3	-3	-1	1	3
3	-1	-5	-2	2	5	1	-3
1	-1	-1	1	1	-1	-1	1
2	-5	1	3	-3	-1	5	-2
1	-3	3	-1	-1	3	-3	1
1	-2	3	-5	5	-3	2	-1

Figure 1 shows the rate-distortion performance of the ICT scheme compared to the JPEG scheme. Simulation results indicate that the difference in performance between using the floating-point DCT and the ICT is unnoticeable.

In addition to the baseline ICT compression scheme, we have performed an analysis study of the compression scheme and have developed useful features and tools to support the ICT compression task. Features include the implementation of an addressable truth window and region of interest. Tools include the development of compression ratio prediction and post-pass-image restoration techniques. We have also gained invaluable experience in cooperating and working with the PIs, the main users of the restored images, in developing a compression scheme that will meet their science objectives, given the bandwidth-limited environment.

PI-in-the-Loop Evaluation of ICT Compression

The Galileo compression scheme was evaluated with the involvement of the SSI team scientists, the principal users of the images compressed by the compression algorithm. An extensive PI-in-the-loop visual evaluation of the ICT compression scheme was conducted with the Remote Payload Systems Research Group at NASA Ames to find the optimal compression/distortion trade-off [5]. We also collaborated with the Vision Group at NASA Ames to generate a number of customized quantization tables for the Galileo compression task [6][7].

In this experiment, we used a new human behavioral and perceptual test technique (known as Progressive Division), largely derived from the Space Station Program, to quantify the ultimate usefulness of compressed-transmitted-decompressed-displayed still and video images. Galileo spacecraft imagery (800 x 800 pixels x 8 bits of gray) were compressed using the ICT in conjunction with four different quantization tables. Different types of monochromatic astronomical images were studied by members of the SSI team and others using the present methodology. The Progressive Division method involved presenting the first observer with a very broad range of quantization levels and requiring that he or she select either the upper or lower half of the range presented (if possible) as having the more acceptable image quality. Again, two identical images were presented side by side but without any identifying information. The chosen half-range was then presented to the second subject, who again had to judge between the upper or lower-half of the now-bisected range. This was continued with subsequent image evaluators until an acceptably small quantization range was found. This process can be computer automated. This approach made it possible to converge on an acceptable solution rapidly and with remarkably good consistency between scientists.

Addressable Truth Window (TW)

The Galileo compression scheme is lossy . To facilitate the science objectives, the concept of a truth window (TW) is introduced, where a small window of the image is losslessly compressed and the rest is lossy compressed. The TW is a fixed 96 x 96 window that can be located anywhere within the image. The TW is losslessly compressed using the Huffman encoding module of the ICT compression algorithm to perform this compression; thus no new compression code was added to the compression program, conserving memory onboard the spacecraft.

The TW can be used by the PI to preserve important details in the image for a more thorough investigation, or as a statistical reference to the rest of the image by the developed image restoration techniques

Compression Ratio Prediction

The amount of compression achievable by any compression scheme varies considerably, being highly dependent on the data characteristics. To facilitate mission planning and operation, a compression ratio prediction scheme is being developed. Our research efforts produced a compression ratio prediction scheme that uses the known statistics of the imaging camera and the type of image expected to obtain the entropy of the image. Using the estimated entropy of the image, we in turn correlate it with data in a pre-generated lookup table to predict the compression ratio.

An interesting result is that for planetary images the probability distribution of the adjacent pixel differences can be modeled with a generalized Gaussian function (GGF) of the form

$$f_{\nu,\sigma}(x) = \frac{\alpha_\nu}{\sigma} e^{-\beta_\nu \left|\frac{x}{\sigma}\right|^\nu}$$

$$\alpha_\nu = \frac{\nu}{2} \left(\frac{\Gamma(\frac{3}{\nu})}{\Gamma^3(\frac{1}{\nu})}\right)^{\frac{1}{2}}$$

$$\beta_\nu = \left(\frac{\Gamma(\frac{3}{\nu})}{\Gamma(\frac{1}{\nu})}\right)^{\frac{\nu}{2}}$$

We found that, unlike other natural images, most planetary images have a shaping parameter ν much less than 1 ($0.4 \leq \nu \leq 0.8$). This is compared with the Normal distribution where the shaping parameter is 2 and the Laplacian distribution where the shaping parameter is 1. This result is important for modeling of images for further studies when the actual images are unavailable.

Post Processing Techniques

A number of post-pass image restoration techniques have been developed to remove the undesirable blockiness and checkerboard effects inherent in a block-based transform compression scheme. The main goal of these restoration techniques is to restore the image without increasing distortion. This is important because standard restoration techniques make the image visually better at the expense of reducing detail thus compromising scientific accuracy. For example, a median filter technique for reducing random noise makes the image more visually pleasing, but removes details and creates an image that is further from the original image.

The restoration techniques developed attack the problem in the frequency domain and the spatial domain during the decompression process. First, frequency coefficients are adjusted within the range of possible original values. Then linear filtering [8] is performed with the constraint that frequency coefficients stay within their range of possible original values. This creates a restored image that probably is closer to the original image.

Galileo's Error-Correction Coding Scheme

The Galileo error-correction coding scheme uses a (255,k) variable redundancy RS code as the outer code, and a (14,1/4) convolutional code as the inner code. The RS codewords are interleaved to depth 8 in a frame. The redundancy profile of the Reed-Solomon codes is (94, 10, 30, 10, 60, 10, 30, 10). The staggered redundancy profile was designed to facilitate the novel feedback concatenated decoding strategy [9][10]. This strategy allows multiple passes of channel symbols through the decoder. During each pass, the decoder uses the decoding information from the RS outer code to facilitate the Viterbi decoding of the inner code in a progressively refined manner. The FCD is implemented in software on a multiprocessor workstation. The code is expected to operate at bit signal-to-noise ratio of 0.65 dB at a bit error rate of 10^{-7}. Figure 2 shows the schematic of the FCD architecture. In this paper, we discuss the implementation and operation aspects of the FCD task only. The FCD novel node/frame synchronization scheme is discussed in [11]. The FCD code selection and performance analysis are

discussed in detail in [12]. The FCD gap-processing and error-recovery schemes will be discussed in a forthcoming paper.

The (255,k) Variable Redundancy Reed-Solomon Code

All RS codes for the Galileo mission use the same representation of the finite field GF(256). Precisely, GF(256) is the set of elements

$$GF(256) = \{0, \alpha^0, \alpha^1, \alpha^2, ..., \alpha^{254}\}$$

where α, by definition, is a root of the primitive polynomial

$$p(x) = x^8 + x^7 + x^2 + x + 1$$

(i.e. $p(\alpha) = 0$).

In the encoding/decoding process, each power of α is represented as a distinct non-zero 8-bit pattern. The zero byte is the zero element in GF(256). The basis for GF(256) is descending powers of α. Note that this is the conventional representation, not Berlekamp's dual basis [13]. The RS generator polynomial is defined as

$$g(x) = \prod_{i=0}^{n-k-1}(x - \alpha^{\beta(i+L)}) = \sum_{i=0}^{n-k} g_i x^i$$

where n denotes the codeword length in bytes and k denotes the number of information bytes, and α^β is a primitive element of GF(256). The parameter β is chosen in some applications to minimize the bit-serial encoding complexity. Since the Galileo RS encoders are implemented in software, there is little advantage in preferring a particular value of β. The parameter L is chosen such that the coefficients of g(x) are symmetrical. This reduces the number of Galois field multiplications in encoding by nearly a factor of 2.

The Galileo mission utilizes four distinct RS codes. We define RS(n,k) to be an RS code which accepts as input k data bytes and produces as a code word n bytes, where $n > k$. An RS(n,k) code can correct t errors and s erasures if $2t + s \leq n-k$. The codes are referred to as RS(255,161), RS(255,195), RS(255,225), RS(255,245). Specifically, the parameters β and L of the four codes are

- RS(255,161) $\beta = 1, L = 81$
- RS(255,195) $\beta = 1, L = 98$
- RS(255,225) $\beta = 1, L = 113$
- RS(255,245) $\beta = 1, L = 123$

The four RS codes, which are interleaved to depth 8, are arranged in a transfer frame as shown in Figure 2. The RS decoders use a time-domain Euclid algorithm to correct both errors and erasures. The details of the decoding algorithm is discussed in [14].

The (14,1/4) Convolutional Code and Its Parallel Viterbi Decoder

The (14,1/4) convolutional code used for the Galileo mission is the concatenation of a software (11,1/2) code and an existing hardware (7,1/2) code. The choice of convolution code is constrained by the existing (7,1/2) code which is hardwired in the Galileo Telemetry Modulation Unit (TMU), and by the processing speed of the ground FCD. The generator polynomials of the (11,1/2) code and the (7,1/2) code in octal are (3403, 2423) and (133, 171) respectively. The generator polynomials of the equivalent (14,1/4) code are (26042, 36575, 25715, 16723).

The Viterbi decoder for the (14,1/4) code is implemented in software in a multiprocessor workstation with shared memory architecture. The use of a software decoder is possible due to the slow downlink rate of the Galileo Mission. The advantages of a software-based decoder are that the development cost is low and it allows the flexibility to perform feedback concatenated decoding. We examined two different approaches to parallelize the Viterbi algorithm: 1) state-parallel decomposition in which each processor is equally loaded to compute the add-compare-select operations per bit, and 2) round-robin frame decoding that exploits the multiple processors by running several complete, independent decoders for several frames in parallel. Our early prototypes indicate that the first approach requires a substantial amount of inter-processor synchronization and communication, and this greatly reduces the decoding speed. The second approach requires minimum synchronization and communication, since each processor is now an entity independent of the other processors. The performance scaling is nearly perfect. We chose the round-robin approach for the FCD Viterbi decoder. The details of the FCD software Viterbi decoder implementation are described in [15].

Redecoding

Redecoding, as shown in Figure 2, involves using information fed back from code words successfully decoded by the RS decoder to improve the performance of Viterbi decoding. A correctly decoded RS bit forces the add-compare-select operation at each state to select the path that corresponds to the correct bit. The Viterbi decoder is thus constrained to follow only paths consistent with known symbols from previously decoded RS codewords. The Viterbi decoder is much less likely to choose a long erroneous path because any path under consideration is pinned to coincide with the correct path at the locations of the known symbols. Each RS frame is decoded with 4 feedback passes. In the first pass, only the first code word RS(255,161) is decoded. In the second pass, the fifth codeword RS(255,195) is decoded. In the third pass, the third and seventh codewords RS(255,225) are decoded, and finally in the fourth pass, the second, fourth, sixth, and eighth code words RS(255,245) are decoded. During each pass, the decoder uses the decoding information from the Reed-Solomon outer code to facilitate the Viterbi decoding of the inner code in a progressively refined manner. The details of the FCD redecoding analysis are described in [12].

Interaction Between Data Compression and Error Control Processes

Packet loss and other uncorrectable errors in a compressed data stream cause error propagation, and how the error propagates depends on the compression scheme being used. In the Galileo Low Gain Antenna Mission scenario, the volume of data returned will be drastically reduced. To maximize the scientific objectives with the limited transmission bandwidth of the low gain antenna, most of the data (image and non-image) are expected to be heavily edited and compressed. These valuable compressed data must be safeguarded against catastrophic error propagation caused by packet loss and other unforeseeable errors.

The ICT scheme for SSI data is equipped with a simple but effective error containment strategy. The basic idea is to insert sync markers and counters at regular intervals to delimit uncompressed data into independent blocks, so that in case of packet loss and other anomalies, the decompressor can search for the sync marker and continue to decompress the rest of the data. In this case, the interval is chosen to be 8 lines of uncompressed data. The error containment strategy guarantees that error propagation will not go beyond 8 lines of data. We have also considered other options to prevent error propagation, but these options usually result in great onboard implementation complexity or excessive downlink overhead. For example, a self-synchronizing Huffman code can be used to contain errors, but it is difficult to implement. A packetizing scheme with varying packet sizes can be used to contain errors (by matching packet boundaries and the compressed data block boundary), but the packet headers introduce excessive downlink overhead in the case of SSI data.

The SSI ICT error containment scheme works as follows. On the compression side, every 8 line block of data is compressed into a variable length compressed data block. The DC (steady-state bias) value is reset to zero at the start of every 8 lines, thus making every 8 lines of compressed data independent of other 8 lines. A 25-bit sync marker and a 7-bit modulo counter are inserted at the beginning of every 8 lines. The sync marker is chosen to minimize the probability of false acquisition in a bursty channel environment. The 25-bit sync marker pattern in hex is 024AAAB. Simulation results indicate that this sync marker gives a probability of false acquisition of less than 10^{-8}. The decompression scheme consists of two program modules: the SSI ICT decompression module and the error detection/sync module. The SSI ICT decompression module reconstructs the data from the compressed data stream, and the error detection/sync module checks the prefix condition of the Huffman codes to detect any anomaly. When an anomaly is detected, a sync marker search is initiated to detect the next available sync marker. Decompression resumes from there on and the reconstructed blocks are realigned using the modulo counter. The corrupted portion of the data is flagged and reported.

The downlink overhead of the SSI ICT error containment scheme is a function of compression ratio (CR) and image width (W). It is measured by the percentage of sync data (sync marker and counter) compared to the compressed data, and is given by the following equation:

$$\frac{4 \times CR}{8 \times W}$$

For example, an 800 x 800 SSI image has the following overhead as a function of the compression ratio:

Compression Ratio	Overhead
2	0.00125
4	0.00250
8	0.00500
16	0.01000

Potential Commercial Applications

The data compression and error containment and correction schemes described above have been implemented and tested. Other than deep space applications as described in the Galileo scenario, these algorithms are also ideal for commercial applications. The multiplication-free ICT scheme can be used as a substitute for the DCT scheme to reduce the implementation complexity and/or increase the speed in Joint Photographic Experts Group (JPEG) and Motion Picture Experts Group (MPEG) applications. Other image

analysis and processing techniques developed in this task, such as compression ratio prediction, image restoration, and image enhancement, can be used for commercial image processing. The Progressive Division methodology we used in the ICT subjective evaluation may find use in a variety of commercial applications where video imagery must be transmitted without incurring perceptible amounts of quality loss. The software FCD scheme demonstrates its superior error-correction capability, and can be used for power-constrained and bandwidth-critical mobile and satellite communications.

Acknowledgement

The research and development described in this paper was carried out by Jet Propulsion Laboratory, California Institute of Technology, under a contract with National Aeronautics and Space Administration.

References:

[1] L. Deutsch and J. Marr, "Galileo S-Band Study, Final Report," JPL Publication, March 1992.
[2] W. Cham, "Development of Integer Cosine Transform by the Principle of Dyadic Symmetry," IEE Proceedings, Vol. 136, Pt. I, No. 4, August 1989.
[3] K. Cheung, F. Pollara, and M. Shahshahani, Integer Cosine Transform for Image Compression, TDA Progress Report 42-105: January - March 1991, Jet Propulsion Laboratory, Pasadena, May 15, 1991.
[4] K. Cheung and K. Tong, "Proposed Data Compression Schemes for the Galileo S-Band Contingency Mission," Proceedings of the NASA Space and Earth Science Data Compression Workshop, April 1993, Snowbird, Utah.
[5] R. Haines, Y. Gold, T. Grant, S. Chuang, "Subjective Evaluations of Integer Cosine Transform Compressed Galileo Solid State Imagery," NASA Technical Paper 3482, July 1994.
[6] A. Watson, A. Ahumada, M. Young, "ICT Quantization Matrix Design for the Galileo S-Band Mission," NASA Technical Memorandum, 1993.
[7] A. Watson and A. Ahumada, "Preservation of Photometric Accuracy in ICT-Compressed Imagery," NASA Technical Memorandum, 1993.
[8] A. Borden and E. Majani, "Optimal Restoration Filters for A Posteriori Removal of Blocking Artifacts in Block Transform-Compressed Data," to be submitted.
[9] E. Paaske, "Improved Decoding for a Concatenated Coding System Recommended by CCSDS," IEEE Transaction of Communications, Vol. COM-38, August 1990.
[10] O. Collins and M. Hizlan, "Determinate-State Convolutional Codes," TDA Progress Report 42-107: July - September 1991, Jet Propulsion Laboratory, Pasadena, November 15, 1991.
[11] J. Statman, K. Cheung, T. Chauvin, J. Rabkin, and M. Belongie, "Decoder Synchronization for Deep Space Missions," TDA Progress Report 42-116: October - December 1993, Jet Propulsion Laboratory, Pasadena, February 15, 1994.
[12] S. Dolinar and M. Belongie, "Enhanced Decoding for the Galileo Low-Gain Antenna Mission," Proceedings 1994 IEEE International Symposium on Information Theory, Trondheim, Norway, June 27 - July 1, 1994.
[13] E. Berlekamp, "Bit-Serial Reed-Solomon Encoder," IEEE Transaction of Information Theory, Vol 28, 1982.
[14] R. McEliece, "The Decoding of Reed-Solomon Codes," TDA Progress Report 42-95: July - September 1988, Jet Propulsion Laboratory, Pasadena, November 15, 1988.
[15] T. Chauvin and K. Cheung, "A Parallel Viterbi Decoder for Shared Memory Architecture", presented in the SIAM Conference on Parallel Signal/Image Processing on Multiprocessor Systems, Seattle, Washington, August 1993.

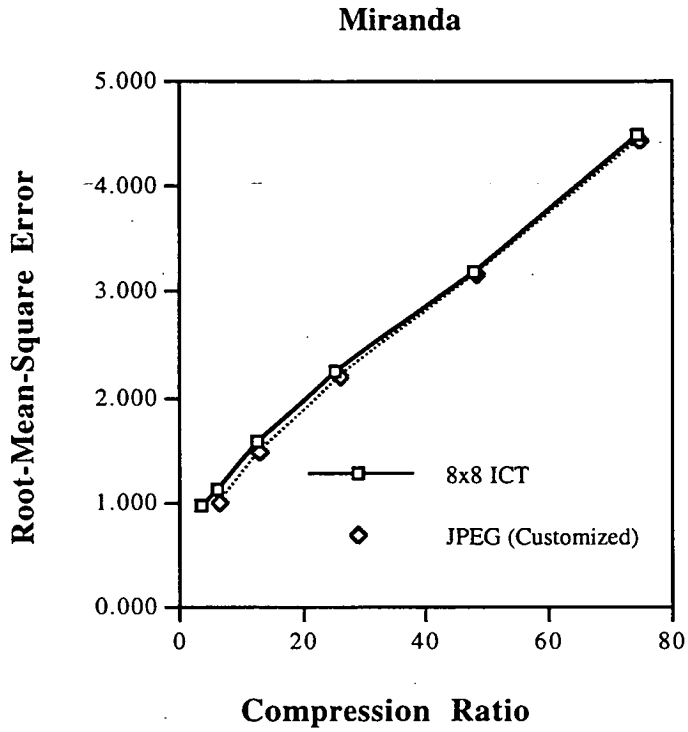


Figure 1. Rate-Distortion Performance of ICT

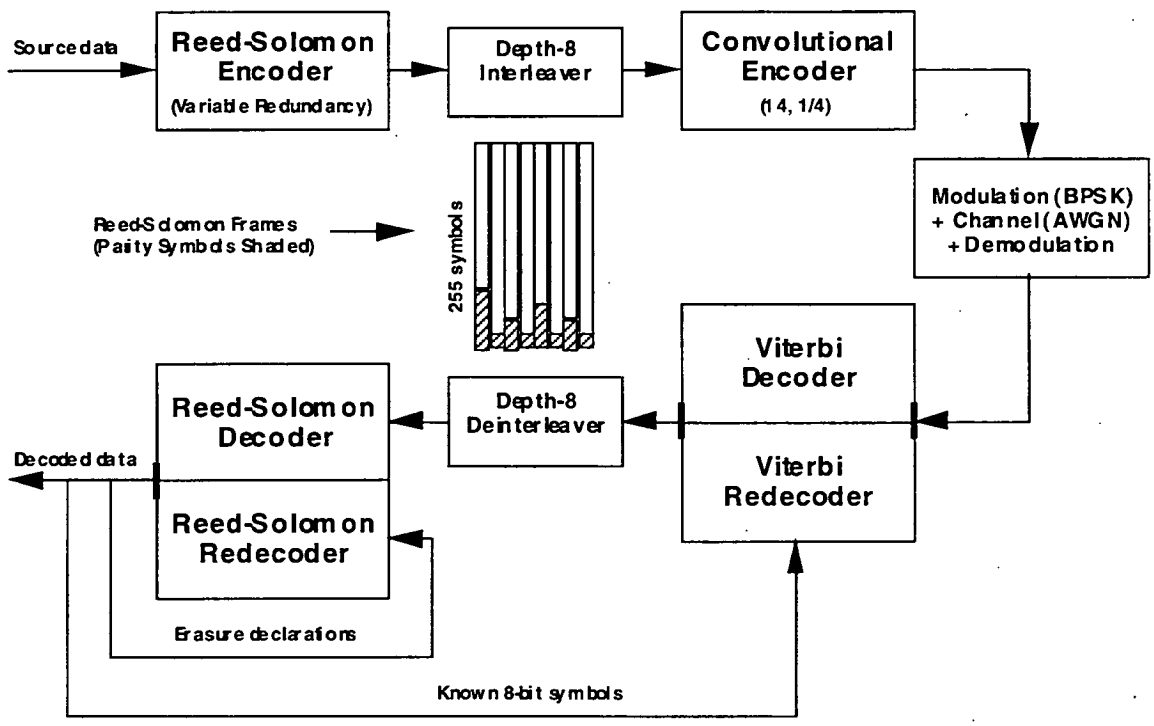


Figure 2. Schematic of the FCD

OPTICAL PRODUCT GRADE SENSOR FOR PROCESS CONTROL

John M. Oestreich
U.S. Bureau of Mines, Dept. of the Interior
729 Arapeen Drive
Salt Lake City, Utah 84108-1283

William K. Tolley
U.S. Bureau of Mines, Dept. of the Interior
729 Arapeen Drive
Salt Lake City, Utah 84108-1283

David A. Rice
U.S. Bureau of Mines, Dept. of the Interior
729 Arapeen Drive
Salt Lake City, Utah 84108-1283

ABSTRACT

The U.S. Bureau of Mines (USBM) is developing an optical sensor to instantaneously measure mineral concentrations in flotation froths and other mineral processing streams. Better process control sensors are needed for clean, efficient mineral processing. The optical sensor consists of a commercial color video camera, a video-capture board, a computer, and a USBM-developed computer program that evaluates the color information. One major innovation in this device is the use of the color vector (a calibration standard from commercial broadcasting) to estimate composition. Color vector angles have proven to be a reliable measure of the subtle color changes resulting from differences in mineral concentrations. The concentrations of chalcopyrite and molybdenite (common copper and molybdenum ores) in concentrates were successfully measured with the optical sensor. These products can be measured as dry mixtures, slurries, or in-situ in the flotation froths that are used to separate the minerals. The optical sensor was also adapted to measure the concentration of fossilized resin in coal with an innovative optical technique using fluorescent light. Image analysis in conjunction with color measurements successfully recognized different sizes of bubbles in a flotation froth. The USBM optical sensor has great potential as a process-control sensor for metallurgical processes.

INTRODUCTION

There are few fast, reliable methods to obtain on-line analyses of product streams in mineral processing operations. The minerals industry generally uses wet chemical analyses, X-ray fluorescence (XRF) or neutron-activation to determine compositions of mineral products. These analyses are often made at the end of processing after further improvement in product quality is impossible. Fast, accurate process-control sensors can significantly improve profitability and reduce waste generation by extracting minerals more completely and selectively. Good process control also minimizes reagent consumption and improves down-stream water quality. The need for better process controls has been noted in the literature [1].

One of the most common industrial mineral processing operations is froth flotation. In flotation, reagents are added to a mineral slurry to make selected minerals hydrophobic. Air is then bubbled through the slurry to concentrate the hydrophobic mineral into a froth that can be skimmed off the slurry. Many sequential stages, each employing a different suite of reagents, are often needed to effectively separate the minerals of interest from the waste rock. Closely monitoring every step of a multi-stage flotation operation is prohibitively slow and expensive with current on-line analyzers. However, monitoring every stage in the flotation circuit may be practical with inexpensive optical sensors. Since modern copper flotation plants process as much as $1000 worth of copper every minute, a delay of several minutes in detecting substandard product can be extremely costly.

Flotation plant operators routinely use visual observation of the color and consistency of flotation froths to estimate how well the flotation is performing [2,3]. Automating this visual analysis of froth compositions would be invaluable in flotation control. Earlier USBM research demonstrated that froth color can be used to quantify product grade in the separation of molybdenite from chalcopyrite, and of chalcopyrite from quartz and pyrite [3].

The U. S. Bureau of Mines (USBM) is developing an inexpensive optical sensor to measure mineral concentrations in flotation froths and other process streams. Commercial versions of the USBM sensor could cost as little as one to two thousand dollars. These sensors should be inexpensive enough that many stages in the flotation circuit could be routinely monitored. The result would be significantly improved control of these operations.

An innovation in the development of this sensor was the use of color vector angles to quantify the color of mineral mixtures. The color vector is used in commercial television broadcasting to calibrate color video cameras. In laboratory experiments, the color vector angle has proven to be a reliable measure of froth composition in flotation.

In addition to composition, the optical sensor is ideal for measuring bubble size to assess froth texture. Currently, operators visually assess flotation performance by noting both the color and the bubble size of the froth. Incorporating size analysis into the optical sensor could be a powerful tool to boost flotation performance.

Chalcopyrite and molybdenite were selected for the initial optical sensor research for several reasons. These two minerals often occur together in copper porphyry ores and are recovered as co-products from copper mines. The large difference in color between brassy yellow chalcopyrite and dark grey molybdenite is ideal for color measurements.

Application of the optical sensor to coal resins and other mineral systems is in progress. Non-mineral applications of this sensor are also under consideration. Opportunities to use or license this technology are available from the USBM.

COLOR MEASUREMENTS

Colorimetry, spectroscopy, and other techniques based on light absorbance have been used for many years in diverse analytical techniques. Recently, it has been demonstrated that modern video cameras can be used for automated colorimetric analyses [4, 5, 6]. However, analytical techniques that use transmitted light have limited applications with opaque minerals. The USBM is developing an optical sensor based on reflected light to analyze compositions in opaque mineral processing streams. A small video camera detects the reflected light. A telephoto lens on the camera allows the sensor to be placed well away from the process streams.

The optical sensor system consists of a color video camera, a light source, a video-capture board, a computer, and a USBM-developed computer program. A schematic of the entire optical sensor system is shown in Figure 1. The video camera is a Panasonic[1] WV-F250BH with separate high-resolution charge-coupled device (CCD) detectors for each of the three primary colors: red, green, and blue. A zoom lens allows the iris, focus, and field of view within the camera to be precisely adjusted to optimize the sensor response. The sensor's computer is an IBM-compatible 486DX2-66 personal computer (PC) with a fast VGA video card and 8 Mb of memory.

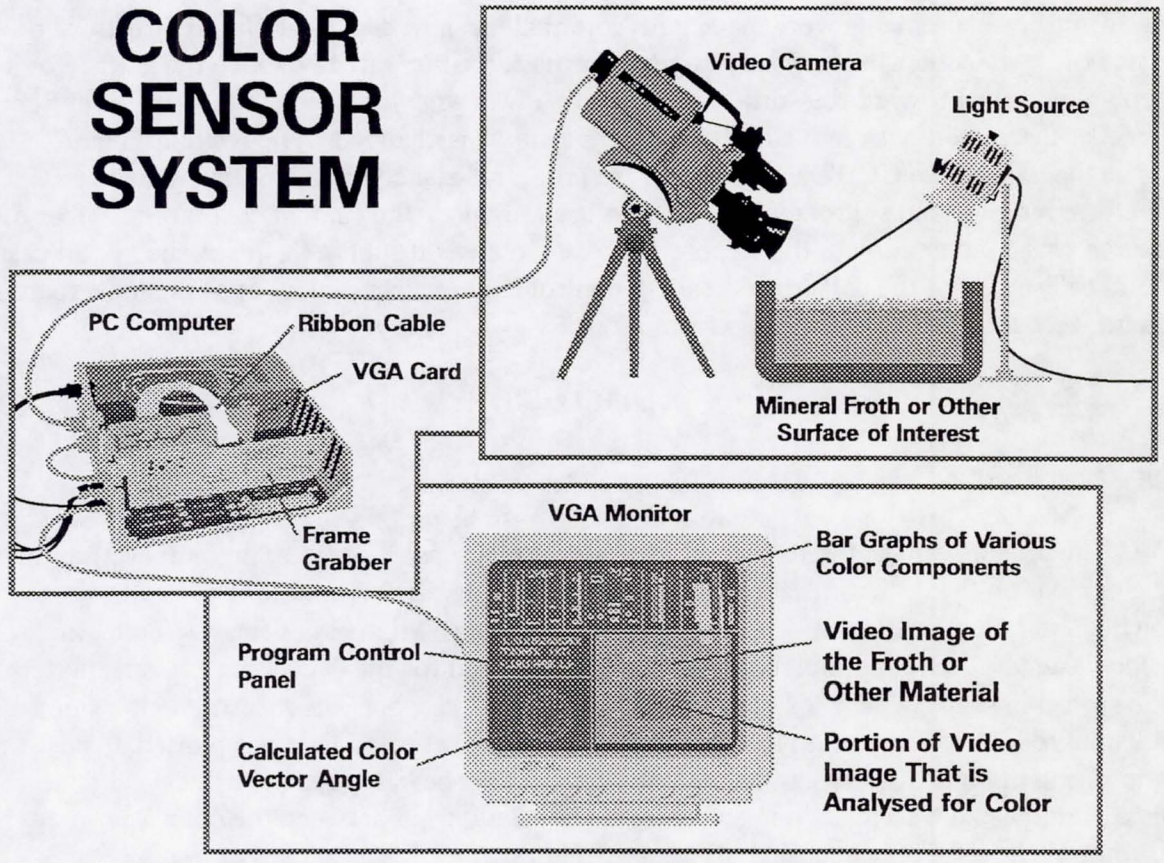


Figure 1. Schematic of color sensor for measuring composition of mineral mixtures.

---

[1]Reference to specific trade names does not imply endorsement by the USBM.

The camera transmits National Television Standards Committee (NTSC) composite video images through a cable to a Hauppauge Win/TV video-capture board in the optical sensor computer. A USBM-developed C-language computer program using LiveWindows video processing library functions displays and manipulates the color information on the computer. The program measures the red, green, and blue (RGB) color components for each pixel within a user-selected region of the image. The RGB data are then averaged over the region and used to calculate the "chrominance red" (C_R), "chrominance blue" (C_B), and the color vector angle using the equations [7, 8]:

$$C_R = 0.877 * (0.701 * \text{Red} - 0.587 * \text{Green} - 0.114 * \text{Blue}) \quad (1)$$

$$C_B = -0.493 * (-0.299 * \text{Red} - 0.587 * \text{Green} + 0.886 * \text{Blue}) \quad (2)$$

$$\text{Color vector angle} = \arctan (C_R/C_B) \quad (3)$$

Compositions are then evaluated by comparing the measured color vector angle to a previously stored calibration curve. The computer can relay the composition data to an automatic process control system or act directly as a process controller. Color vector angles are allowed to exceed 360 degrees to avoid discontinuities in calibration curves. Color vector angles provide a useful and easily interpreted measure of the concentration of chalcopyrite-molybdenite mixtures.

Color measurements were made with a normal incandescent light placed directly above the surface of the mineral samples. The camera was placed either directly above the surface or aimed steeply downward toward the surface. The camera was white-balanced with a white reference whenever the camera was moved or the camera setup was changed. The white balance adjusts the gain of the CCD's so the colors are properly balanced in the output signal. A waveform monitor and vectorscope were also used to check the camera calibration. Most of the research was conducted with the camera 30 to 50 cm away from the froth surface. In all cases, careful calibration of the camera and care in controlling the physical layout of the light source and camera were required to produce consistent results.

DISCUSSION

Color Analysis of Dry Chalcopyrite-Molybdenite Mixtures

The primary aim of the research was to develop a sensor for flotation; however, initial development on the optical sensor was conducted with dry, binary mixtures of chalcopyrite and molybdenite rather than flotation froths to simplify testing. Mixtures of finely ground, high-purity chalcopyrite ($CuFeS_2$) and molybdenite (MoS_2) were used for the dry mixtures. The mixtures were spread in shallow pans for analyses with the optical sensor. The measured intensities of the individual red, green, and blue (RGB) components in the reflected light are plotted against composition in Figure 2. As can be seen, the relationship between the intensities and mineral concentration is hard to interpret. In contrast, the relationship between the color vector angle and the composition is simple and obvious as shown in Figure 3. The color vector angle varies strongly and nearly linearly with changes in the composition. The graph shows color measurements taken on four different days to indicate the repeatability of these measurements. The correlation between color and composition was calculated to be:

$$\% \text{ Molybdenite} = -458.5 + 4.15*(\text{Vector Angle}) - 0.00762*(\text{Vector Angle})^2 \quad (4)$$

Over the four days of testing, the average error in the composition estimates was 3% of full scale.

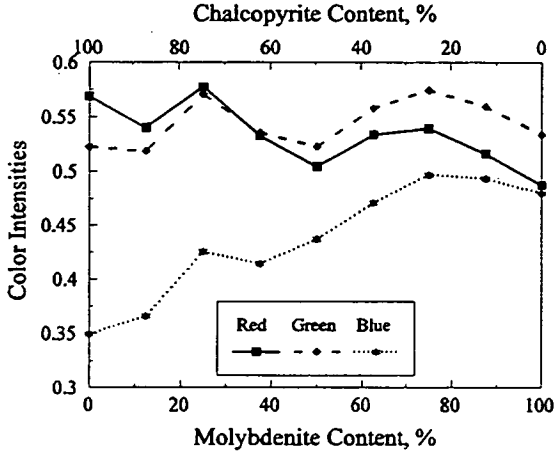


Figure 2. Intensities of red, green, and blue components of video signal for dry mixtures of chalcopyrite and molybdenite.

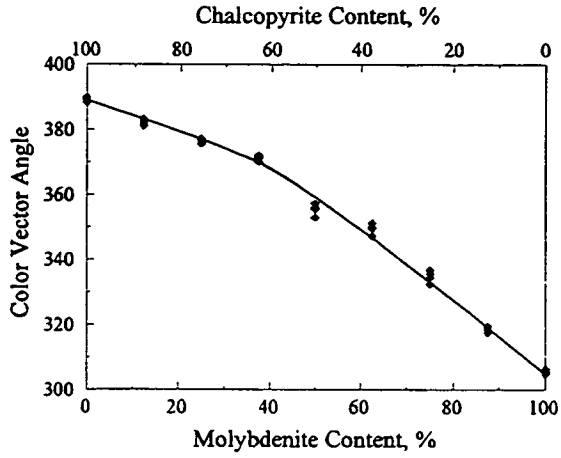


Figure 3. Color vector angle versus molybdenite content of dry mixtures of chalcopyrite and molybdenite.

Color Analysis of Chalcopyrite-Molybdenite Flotation Froths

Froth flotation is one of the most widely used methods of concentrating minerals. In flotation, reagents are added to a mineral slurry to make specific minerals hydrophobic. Air is then bubbled through the slurry to concentrate the hydrophobic mineral into a froth that can be skimmed off the slurry. The primary goal of this research was to develop an optical sensor to measure the composition of these flotation froths. Measuring froth compositions is challenging because the froth is a heterogenous mixture of minerals, water, and air. However, a reliable technique to analyze for froth composition would have a great impact on the minerals industry in improving process control and reducing reagent usage, wasted resources, and energy use.

An important flotation operation in minerals processing is the separation of molybdenite from chalcopyrite. Any molybdenum in the copper product is lost, and copper contamination can significantly reduce the value of the molybdenum product [9]. In current practice, the grade of the products is generally measured only at the end of the flotation process.

Chalcopyrite-molybdenite flotation froths were prepared using copper sulfide and molybdenum sulfide concentrates obtained from a commercial flotation plant. The copper concentrate contained roughly 80% chalcopyrite and 1% molybdenite. The molybdenum concentrate contained 1% chalcopyrite and 80% molybdenite. Both concentrates contained about 15% silicate minerals and minor amounts of other minerals. Slurries containing 5% mineral, 95% water, and a small amount of flotation reagents were used in this research. Flotation froths were formed by bubbling air though these slurries in a Denver flotation cell. Small samples of the froth were collected during the optical measurements and analyzed with flame atomic-absorption

spectrometry to determine the froth compositions. These analyses were used for the calibration curves to define the relationship between composition and color.

Good correlation was found between the composition of the molybdenite-bearing froth and the color vector angle. Figure 4 shows the color vector data plotted against mineral concentration in the froth. Separate curves for chalcopyrite and molybdenite are shown because silicates and other minerals are also present in the froth.

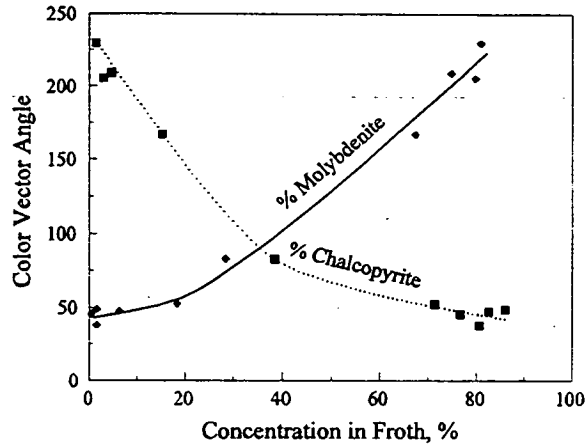


Figure 4. Color vector angle molybdenum and copper content of froths produced in laboratory Denver cell by floating commercial concentrates.

The brightness of the light reflected from flotation froths tends to vary extensively because the froth does not have a simple smooth surface. The top of a flotation froth bubble can reflect light directly into the optical sensor that is 2-3 times as bright as the shadowed edges of a bubble. One advantage of using the color vector angle is that it is

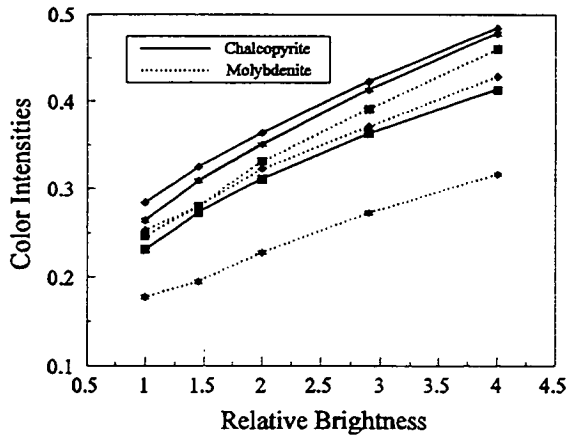


Figure 5. The effect of image brightness on individual RGB intensities measured for dry chalcopyrite and molybdenite samples.

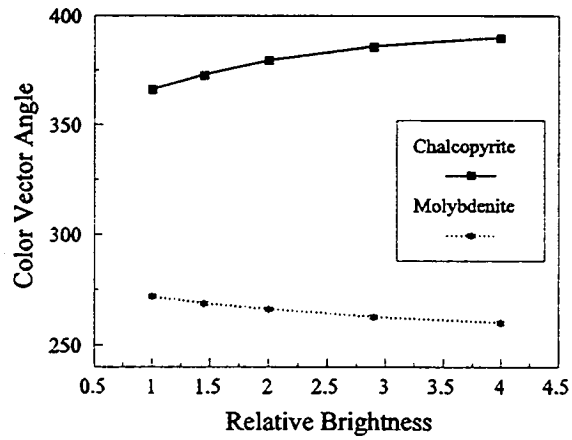


Figure 6. The effect of image brightness on color vector angle reading for dry samples of chalcopyrite and molybdenite.

relatively insensitive to changes in intensity of the light reaching the sensor. Figure 5 shows that quadrupling the image brightness almost doubles the measured RGB component intensities. The same change in image brightness has a much smaller effect on the color vector angle, as shown in Figure 6. This is because the color vector angles vary with the ratios of the RGB components in an image rather than their absolute brightness. When variations in image brightness occur, the individual RGB intensities are a noisier measure of color than color vector angles. The reduced noise in the color vector angle measurement is valuable because variations in image brightness

caused by irregular shapes, changes in lighting, or shadows are hard to avoid in an industrial environment.

Bubble-Size Analysis

The size of the bubbles in flotation froths is important. Controlling bubble size is quite difficult in part because it is very hard to measure reliably. Several high-resolution images collected with the optical sensor are shown in Figure 7. Images collected with the optical sensor have been successfully interpreted with an image analysis program. An image analysis program is now being adapted to allow the optical sensor to simultaneously measure bubble size and composition.

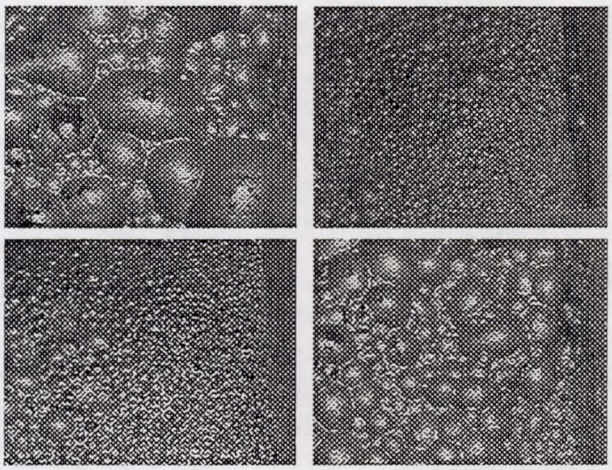


Figure 7. Images of flotation froths collected with optical sensor (actual images are 5.5 inches wide).

OTHER APPLICATIONS OF THE OPTICAL SENSOR

The primary objective of this work is to develop a sensor to analyze flotation froths. However, the optical sensor has been tested in several related applications. The only limit to potential applications of the optical sensor is that some measurable optical property (color, image shape, brightness, etc.) must vary in response to changes in the physical property that needs to be measured (concentration, temperature, pH, etc.).

Chalcopyrite-Molybdenite Slurries

Mineral slurries are produced in leaching, flotation, grinding, and many other metallurgical operations. Thus, measuring the concentration of minerals in slurries could become a very important application of optical sensors. To test the utility of the optical sensor in slurry applications, chalcopyrite and molybdenite were suspended in pure isopropanol. The use of isopropanol as the liquid medium rather than water was necessary because molybdenite is naturally too hydrophobic to disperse in water; isopropanol is transparent to visible light and does not interfere with the color measurements. The relationship between color vector angle and composition of the slurries is shown in Figure 8. The plot shows that color vector angle

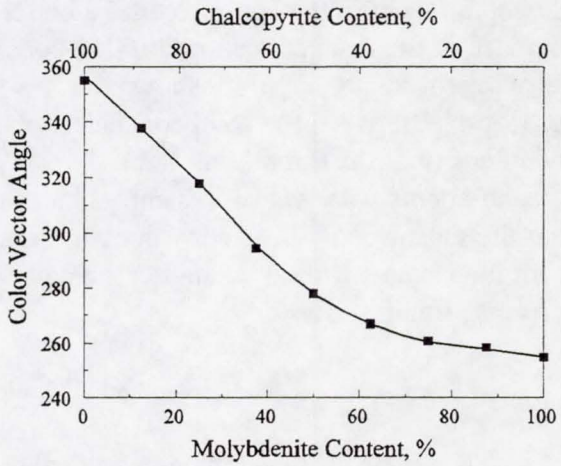


Figure 8. Color vector angle versus molybdenite and chalcopyrite content of isopropanol slurries at 20% solids.

correlates smoothly with composition. Thus, analyses of mineral slurries with optical sensor appear to be practical.

Coal Resin

In the results described to this point, product grades have been calculated from the color vector angles. However, another measure of color may be used if it is better suited to the particular application. One such application is determining the concentration of coal-resin in products recovered from coal by flotation. Fossilized tree resin is often found in the coal beds of central Utah. The resin is used in printing inks to speed drying and reduce smearing, and in varnishes, adhesives, and even chewing gum. A pound of resin is worth roughly 50 times as much as a pound of coal. Some coals contain as much as 10% resin.

Industrially, resin concentrates are assayed with time-consuming wet chemical methods which are difficult to automate. Fast optical resin assays could significantly improve the control of resin flotation operations and enhance the recovery of this valuable material.

The optical sensor system was used to determine the resin concentrations in the products obtained by flotation. Coal resin fluoresces when exposed to ultra-violet (UV) light, and the light fluorescing from the resin can be used to measure the resin concentration. In this case, the resin fluoresces a blue-colored light; the coal neither produces nor reflects any light so the color does not change as the composition of the mixture changes. As a result, the intensity of the blue component in the light provides a better measure of the resin concentration that does the color vector angle. Figure 9 shows the relationship between the resin concentration and the intensity of the fluorescent light. Measurements with wet resin samples indicated that the sensor could be used with damp resin from filter-cakes without changing the color response significantly.

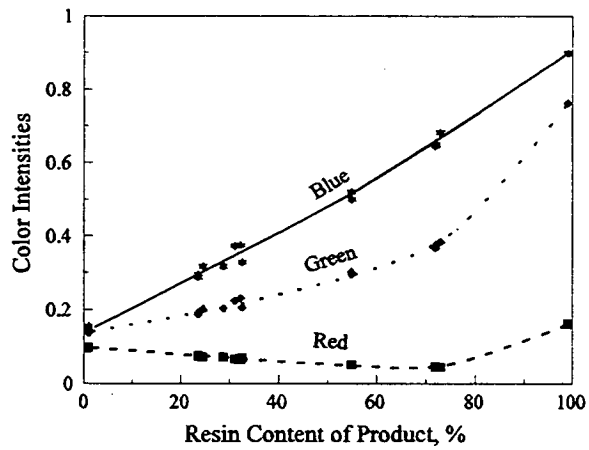


Figure 9. Intensity of blue component in light fluorescing from coal containing various amounts of resin The intensity is a nearly linear measure of resin concentration.

COMMERCIAL POTENTIAL

The USBM is developing an optical sensor that can instantaneously measure mineral concentrations in flotation froths and other mineral mixtures. Optical sensors are a valuable new tool that could augment analyses obtained from conventional XRF and neutron activation analyzers currently used in metallurgical plants. Inexpensive optical sensors could provide more complete data instantaneously from a large number of points along a metallurgical process. The optical sensor could be applied in many product streams where the color of the product varies with composition.

The cost of the optical sensor could be reduced by connecting several cameras to a single optical sensor computer or using a network of fiber-optic cables to allow a single camera to simultaneously observe several locations. A single light source could also provide identical lighting to many locations through several fiber-optic cables.

Optical sensors are ideal for process control applications because they are inexpensive, fast, and non-invasive. One obvious application of the color sensor technology would be to control the addition of a chalcopyrite depressant in a molybdenite flotation circuit. A laboratory demonstration of this particular process control application has already been conducted. A pump which added a chemical to suppress chalcopyrite flotation was activated when the optical sensor determined that the chalcopyrite concentration of the molybdenite product was too high. The optical sensor is compatible with current process controllers and would not require major modification of the current concentration plants.

The primary objective in the research was to develop a sensor to analyze flotation froth compositions. However, the sensor also has performed well in measuring the compositions of dry mineral mixtures and mineral slurries. Conceivably, the sensor could be used to analyze compositions of mineral mixtures on conveyor belts, mineral slurries in launders, or many other metallurgical process streams.

CONCLUSIONS

A practical optical sensor capable of on-line measurement of flotation product grades is being developed and tested at the USBM. The sensor consists of a video camera, a video capture computer board, a computer, and a USBM-developed computer program that interprets the color. One major innovation in this research is the use of the color vector (a calibration standard from commercial broadcasting) to characterize the color of mineral mixtures. Clear correlations were derived to relate color vector angles to mineral compositions. Color vector angles proved to be easier to interpret and more resistant to minor variations in lighting than other measures of color tested in the research program. These correlations between color vector angle and composition have been successfully used to measure the compositions of flotation froths, mineral slurries, and dry mineral mixtures containing chalcopyrite and molybdenite. In addition, the control capabilities of the sensor were demonstrated in the laboratory. The sensor was able to measure when the chalcopyrite content of a molybdenite product exceeded a specified limit and adjust reagent feed rates to the process to maintain the specified product grade.

The optical sensor also has been used to measure the composition coal resin flotation products. This system was unique in that light emitted by the resin under ultraviolet illumination was used to measure concentration. Intensity of the emitted light proved to be nearly linearly related to resin concentration.

A further use of the optical sensor is to measure physical size properties of the mineral system. Images collected with the optical sensor have also successfully been processed with image analysis to distinguish differing bubble size distributions in flotation froths.

Further refinement of the optical sensor is expected. The use of fiber optics promises greater capacity both for collecting color information and for providing standardized illumination. Multiplexing through fiber optics also may allow a single camera to monitor many points in a process simultaneously. The optical sensor is a valuable addition to sensor technology for process control in mineral processing plants.

REFERENCES

1. Cecile, J.L. and Villeneuve, J., "Control of Sulfide Flotation", *Internat. J. Mineral Processing*, **33**, 185-191 (1991).

2. Weiss, N. L., "Minerals and Their Properties", *SME Mineral Proc. Handbook*, SME, New York, NY, 2-2 to 2-17 (1985).

3. Gebhardt, J.E., Tolley, W.K., Ahn., J.H., , "Color Measurements of Minerals and Mineralized Froths," *Minerals and Metall. Process.*, May 1993, 96-99 (1993).

4. Sanford, R.L., Meredith, D.L., Spears, D.R., Computer Vision Application in Mineral Processing Research, *Proc. of 1992 IEEE Ind. Application Society Annual Meeting*, **2**, Houston, TX, 2013-2019 (1992).

5. Henkel, S., Video Photometric Analysis Enhances Copper Recovery During Refining, *Sensors*, **10**, No. 10, 8 (1993).

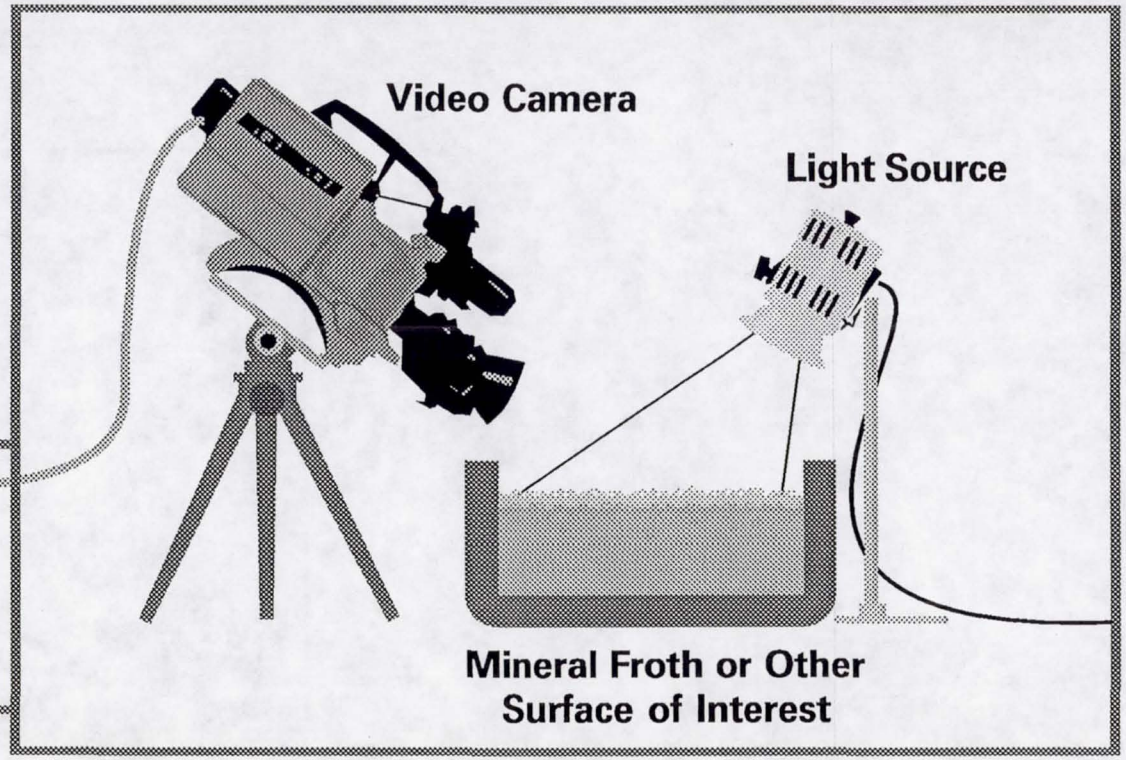
6. Cocanour, J.B., Harbuck, D.D., and Odekirk, S.B., Real Time Process Control Using Video Photometry, 1994 Annual TMS Meeting, San Francisco, CA. (1994).

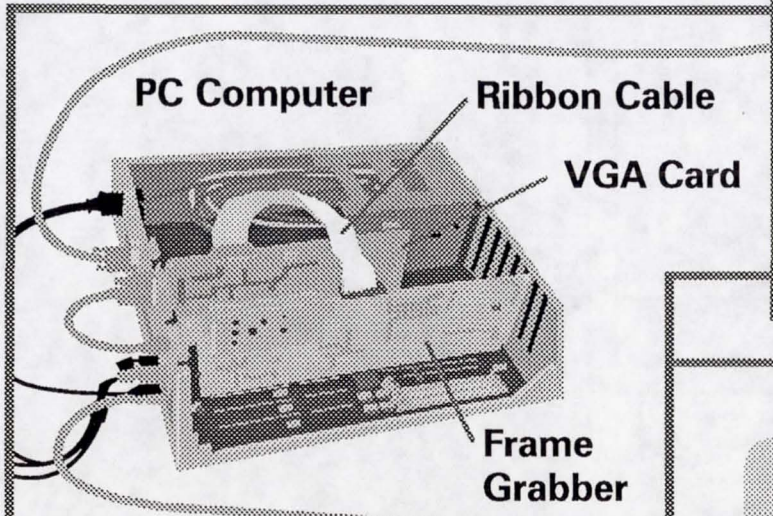
7. Lambert, D., *LiveWindow Users Guide*, Revision 1.00, Software Interphase, Inc., Foster, RI, 3-7 (1992).

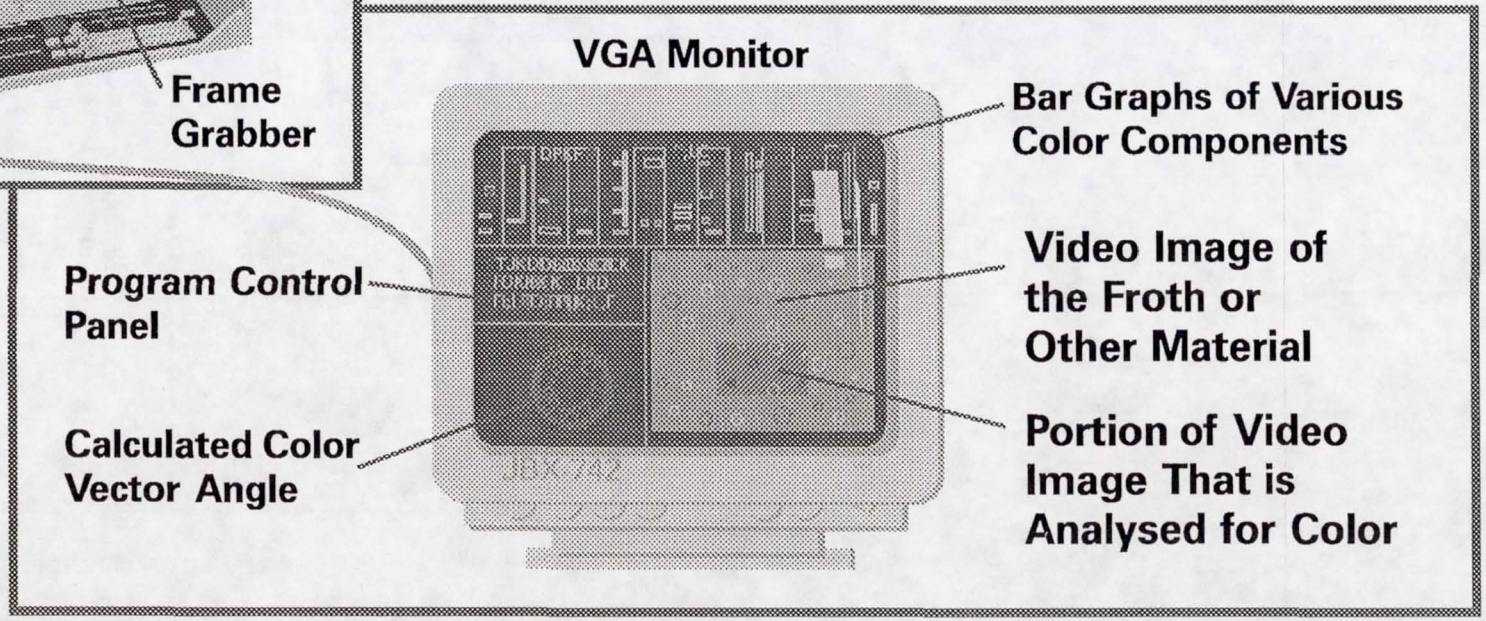
8. Nillesen, A.H., *Desktop Video Data Handbook*, North American Philips Corporation, 2-22 to 2-35 (1993).

9. Sutulov, A., Flotation Recovery of Molybdenite, Annual Volume, 1977, The Metallurgical Society of CIM, 37-47 (1977).

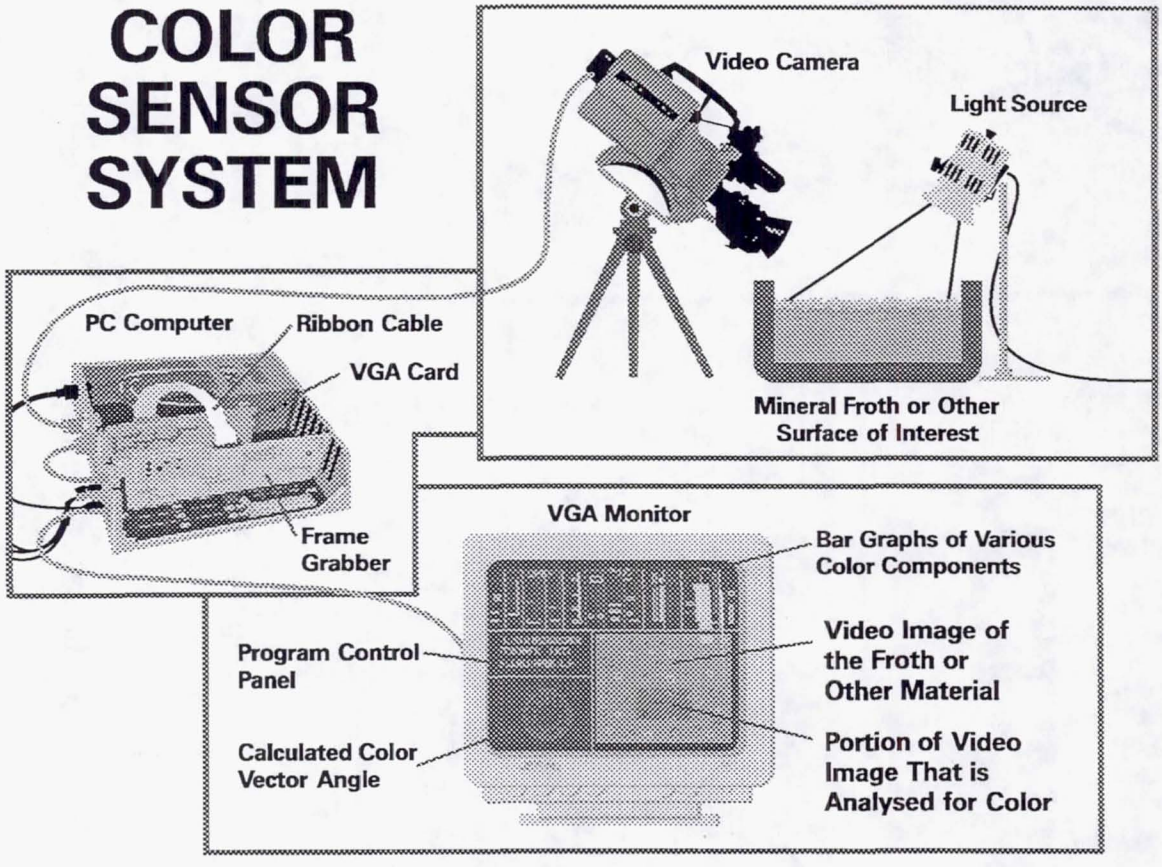
COLOR SENSOR SYSTEM







COLOR SENSOR SYSTEM



ON-LINE ANALYSIS AND PROCESS CONTROL USING LINE-SCAN VIDEO PHOTOMETRY

John B. Cocanour III
U.S. Bureau of Mines
Salt Lake City Research Center
Salt Lake City, UT 84108

Donna D. Harbuck
U.S. Bureau of Mines
Salt Lake City Research Center
Salt Lake City, UT 84108

H. Quinn Stevenson
U.S. Bureau of Mines
Salt Lake City Research Center
Salt Lake City, UT 84108

ABSTRACT

The line scan video photometer, developed by the U.S. Bureau of Mines (USBM) of the Department of the Interior, is an inexpensive, personal computer (PC)-based, software-driven instrument which is used for the real time, on-line, chemical analysis of liquid solutions. The novelty of the instrument lies in the integration of its various components: (1) a true Red/Green/Blue (R/G/B) color video camera; (2) a unique, line-scan color peak detection technique which digitizes the R/G/B video signals on a PC-Card with a potential color resolution of over 772 billion color hues; (3) software ratiometric color sampling which eliminates color drift, and (4) software-based control, which uses the digitized color signals as control variables for various process control methods. Because color hues are often representative of chemical concentrations, this system can be used for chemical analysis. It also has potential applications in the pharmaceutical industry, in the food and beverage industries, and in the petroleum industry. Commercial potential of the video photometer has already been industrially demonstrated for the on-line analysis of copper in decopperization electrolytic circuits.

NEED FOR ON-LINE SENSORS AND PROCESS CONTROL

To optimize recovery processes for improved efficiency and reduced costs, there is a great need for new low cost, rapid, on-line analytical and process control techniques. With traditional analytical methods, samples must often be removed from the process stream and sent to a nearby instrument or laboratory where they are mixed with reagents and diluted before analysis can be performed with typical equipment such as atomic absorption spectrophotometry (AAS), inductively-coupled plasma spectrophotometry (ICP), or colorimeters. Many analytical instruments use an open flame or x-rays which could pose hazards in industrial environments. After laboratory results are obtained, which

could be minutes, hours or days in some cases, then process parameters can be adjusted to optimize the operation. In the meantime, the process is operating inefficiently. Currently, several on-line sensors are available which are placed directly in solutions or slurries, such as pH or Eh probes. These sensors must be periodically removed to be cleaned. Because of contamination, the reliability and durability of such equipment is questionable.

USBM researchers have shown that color analysis using video easily lends itself to on-line procedures and has many advantages over standard analytical techniques. The line scan video photometer offers an innovative, cost-effective ($16,000), rapid technique for on-line, noninvasive analysis. Other recently developed video-based instruments are slow, experience color drift, have repeatability problems, and have limited color resolution. The video photometer overcomes these problems. It is able to discern color hue and digitize the video color signal into a distinct value which can be directly correlated to solution concentration. The digitized color value can then be used as the manipulated variable for software-based control. This paper describes some of the theory behind the video photometer and gives an example of a potential commercial application.

THEORY AND DESCRIPTION OF LINE SCAN VIDEO PHOTOMETER

Line scan video photometry, as defined by USBM researchers, is the digitizing of a true R/G/B line video signal to characterize the absorptive or reflective indicators in the visible region of the electromagnetic spectrum. The video photometric sensor and the interfacing equipment for on-line analysis and process control are shown in Figure 1. The set-up includes a Sony DXC-930[1] three chip R/G/B color camera, a line scan PC card, a calibration and process stream chamber, a 3200 Kelvin tungsten-halogen light source, and an OPTO22 Mistic controller. Figure 1 also shows a possible application where two circuits are analyzed and controlled simultaneously.

The Sony DXC-930 provides precise control over picture parameters so that optimal image color and contrast can be easily reproduced. This $6000 camera has all the gain adjustments of higher priced professional cameras. A very fine dichromatic prism in the camera separates light into its base colors, i.e., red, green, and blue. The video signal from the camera outputs color values as floating voltages.

The line scan PC card, Figure 2, is the heart of the photometer. It solves the drift problems which plague other video color sensors. It allows for the simultaneous monitoring of many streams. Using a patent-pending technique, it digitizes the R/G/B video signals for a potential color resolution of over 772 billion color hues. In the PC card, the floating R/G/B video signals from the camera are referenced to zero by a level-setting circuit. The resultant analog voltage value is then presented to a 16 bit analog-to-digital (A/D) converter, and through the video timing circuit, the color values of the sample stream are
captured.

---

[1] Reference to specific products does not imply an endorsement by USBM.

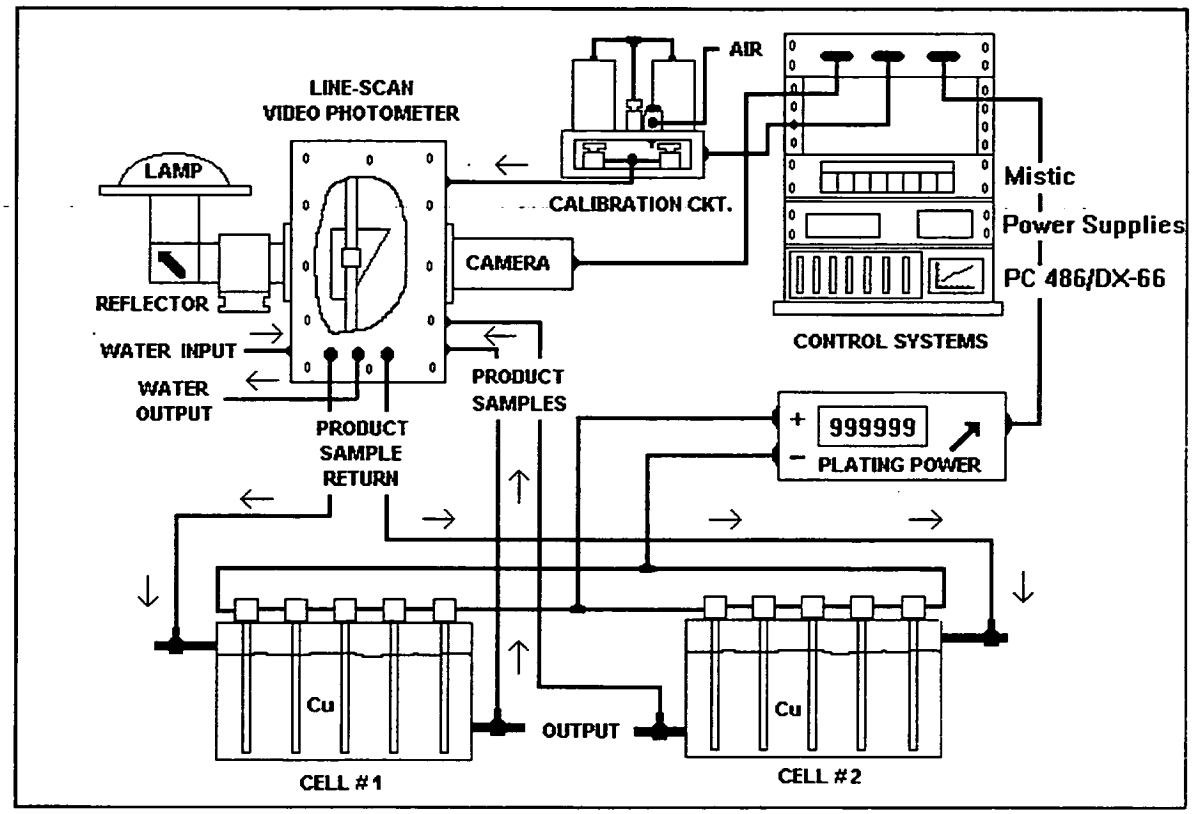


Figure 1. Line Scan Video Photometry.

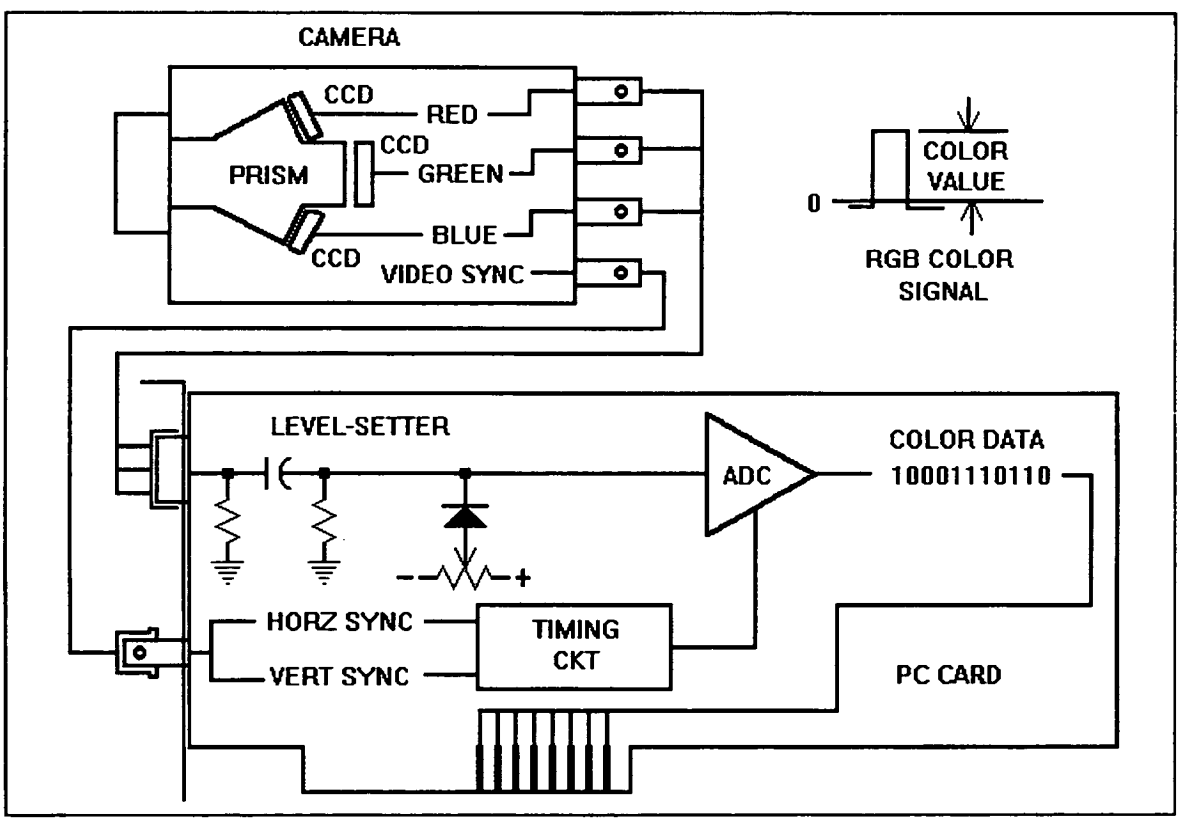


Figure 2. Line Scan PC Card.

Calibration of the photometer is two-fold: electronic and analytical. The camera is electronically calibrated using absorption through clear water as a light standard. Water runs through the center cell of the light absorption chamber as shown in the front view of Figure 3. All color readings are read as a ratio to this standard. If there is any light drift, it will not affect the color reading.

Analytical calibration requires two steps. First, because different solutions vary in their absorptive density, the light absorption chamber was designed to allow for a wide range of color densities. A side view, Figure 4, reveals that the chamber is wedged shape and can slide up and down through the light camera path to allow for placement at the optimum color absorptive density. This motion is accomplished with a 200 step per revolution stepper motor which is under software control. This up and down motion allows for the adjustment of optimum color absorptive densities for the line-scan video photometer. Second, known standards can be automatically fed to the chamber through the product cells to calibrate the photometer for each specific application, see front view in Figure 3. A computer program automatically converts absorption values to g/L, a more recognizable measure of concentration than absorption.

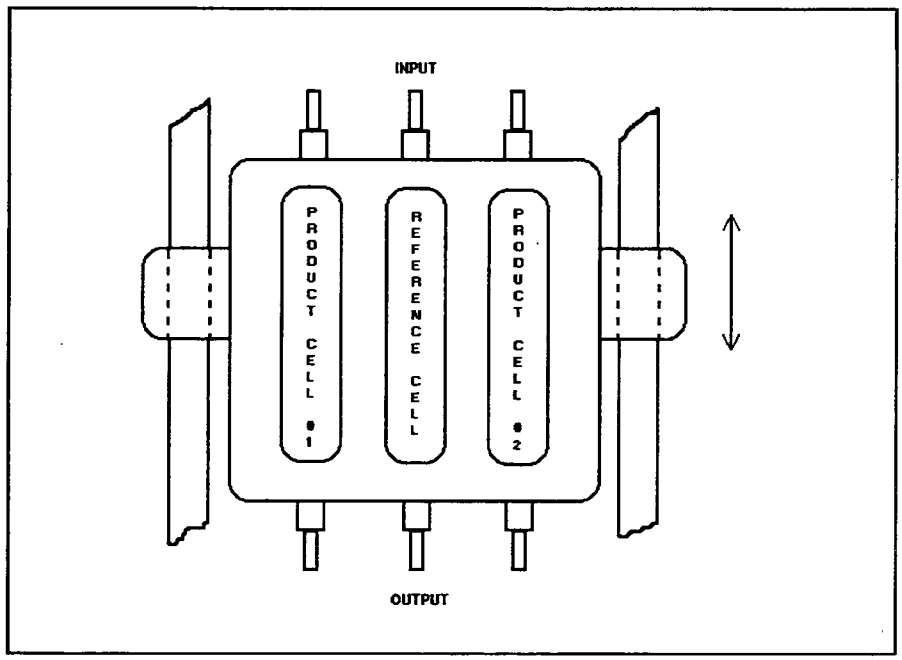


Figure 3. Front View of Light Absorption Chamber.

Because color is the essential factor for this sensor, an appropriate light source is important. In addition, a luminance level of 100 IRE (Institute of Radio Engineers) at 125 candle power is required for proper operation of the camera. For this work, a 3200 Kelvin tungsten-halogen light was used, and the camera was set to match this source. It is necessary that the light source not be exposed to any outside light interference which would pollute the R/G/B video indicators.

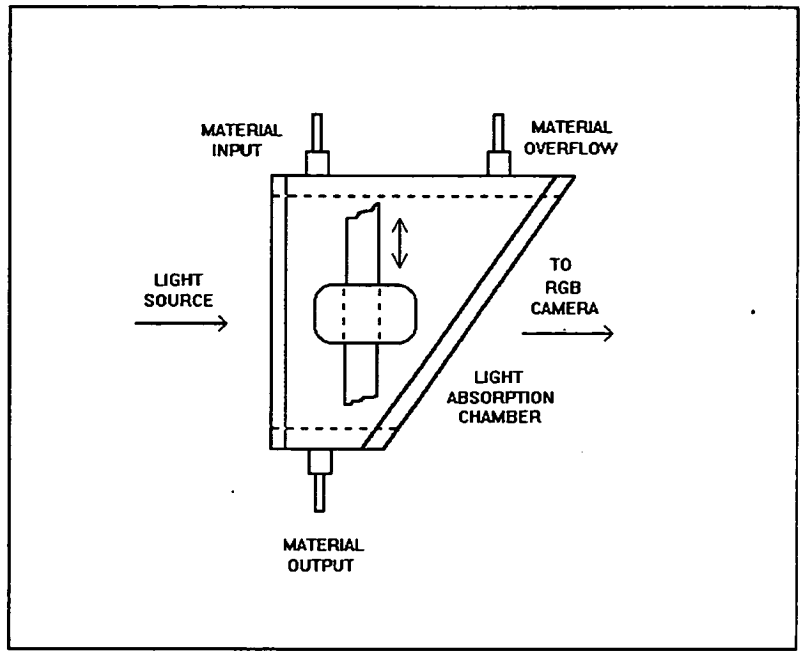


Figure 4. Side View of Light Absorption Chamber.

The OPTO22 Mistic controller acts as D/A and A/D convertors. The controller is used to read temperature and control outside pumps, valves, and/or power supplies. The Mistic product was chosen because it is an inexpensive, multiple sensor input/output (I/O) smart controller which accommodates RS-422 communications (the normal industrial computer communication link). The controller is supported by software drivers written in the C programming language. The actual control algorithms, including PID (Proportional, Integral, and Differential) control are executed by the computer through the developed software rather than through traditional Programmed Logic Controls (PLCs). Not only is this less expensive, it also allows for the integration of innovative control methods, i.e., fuzzy logic, neural networks, adaptive control, and the introduction of various color models. While traditional analytical equipment is typically priced from $60,000 to $130,000, the video photometer costs are only $16,000, (a price which includes a computer, video camera, and all control hardware and software).

APPLICATION OF LINE SCAN VIDEO PHOTOMETRY

Because copper sulfate has a distinct bluish-green color, on-line analysis and control of copper in metallurgical operations is a prime application for the video photometer. Two areas have preliminarily been identified as needing such analysis and control in the copper industry: decopperization electrolytic circuits of copper refineries, and electrowinning circuits in solvent extraction, electrowinning (SX/EW) copper plants. In the decopperization circuits, solutions containing copper sulfate along with impurities such as nickel, arsenic, and bismuth are cascaded through electroplating cells to recover copper. If the copper concentration drops too low, deadly arsine gas is formed. In one refinery, electrolyte is manually sampled approximately every 2 hours and analyzed. If the copper concentration is not in the appropriate range, the flow of electrolyte between the cells is manually adjusted. Because of the dead time needed for analysis and manual

control of the valves, copper values are often lost, and deadly arsine gas can be formed. Installation of on-line analysis and process control equipment could maximize the recovery of salable copper and improve safety by avoiding the formation of arsine gas.

In SX/EW operations, copper is efficiently plated when the concentration of copper in solution is approximately 25 to 35 g/L. Outside of this range, copper recovery diminishes. On-line analysis of copper and control of plating current would maintain the desired copper recovery setpoint.

To demonstrate copper recovery utilizing the line-scan video photometric system, a 1-liter laboratory-size copper electrolytic plating cell was fabricated. Lead plates were used as the anodes while a copper slab acted as the cathode. Solution entered the cell using a peristaltic pump with a 4 to 20 ma interface to the Mistic controller. A DC power supply provided the needed current density to deposit copper on the cathode. Solution in the cell was maintained at a constant level by having an overflow drain as an outlet. The goal of this demonstration was to use digitized color indicators to control the flow rate of solution pumped into the cell, thus controlling copper recovery.

Analytical calibration was performed by preparing solutions with varying concentrations of copper sulfate and H_2SO_4, and a constant nickel impurity. Care was taken to simulate actual electrolyte concentrations. However, because many of the impurities in the electrolytic solution contain insufficient color to interfere with the copper, the synthetic solutions contained only copper, nickel, and H_2SO_4. To test process control of the copper circuit, a solution containing 40 g/L Cu, 5 g/L Ni, and 185 g/L H_2SO_4 was placed in the electrolytic cell. The control point was set at 40 g/L Cu and plating was initiated. As plating proceeded and the copper concentration decreased, the pump automatically responded through PID control to pump in additional solution to maintain copper concentration at 40 g/L. The solution being pumped to the cell contained 45 g/L Cu, 5 g/L Ni and 185 g/L H_2SO_4. Figure 5 shows how quickly steady state was attained. Random samples were taken during the run to compare the copper reading from video photometer with AAS analysis. All results were within +/- 0.3 g/L Cu. To see how well the circuit would handle a disturbance, water was poured directly into the cell 30 minutes into the run to decrease the copper concentration. Figure 5 shows how PID control restored the system to steady state.

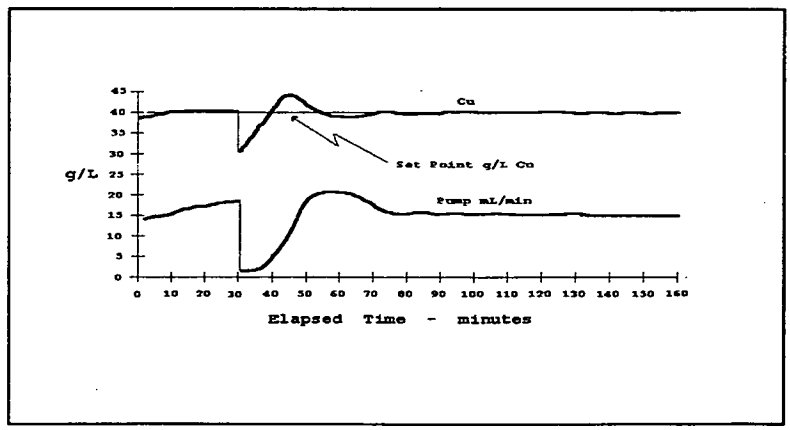


Figure 5. PID Control of Copper Electrolytic Cell Using Video Photometer.

COMMERCIAL USES

Because the line-scan video photometer can distinguish so many color hues, it fares especially well in metallurgical and chemical processing where many metals, chemicals, and their complexes have distinctive colors and the color intensity varies with concentration. The video photometer was tested on-site at a decopperization circuit in Texas. The instrument successfully analyzed copper concentration and was able to control the circuit. Other potential commercial uses for the video photometer would be the on-line analysis and process control of solids, slurries or froths of homogeneous color. The technology could be used in the pharmaceutical industry where color is indicative of medicinal chemical concentrations. It could also be used in the food and beverage industries to analyze and control homogeneously-colored solid and liquid streams. The technology could potentially be used in the petroleum industry to detect colored impurities in fuels, etc. Another potential application would be the on-line analysis and control of colored gaseous streams.

The line-scan video photometer is a complete system for on-line analysis with automatic standardization, and rapid response process control based on analytical results. It provides real time analysis, real time representation of R/G/B values, and real time calculation of PID control. Judging by the interest thus far shown, adaptation by industry should be rapid.

AUTOMATED DIE AND WIRE BOND INSPECTION USING MACHINE VISION FOR MULTI-CHIP MODULE MANUFACTURING

Jonathan E. Ludlow
Project Leader
Acuity Imaging Inc.
Nashua, NH.

ABSTRACT

This paper contains a description of progress to date on a project designed to develop an automated machine vision inspection system for die and wire bonds in MCMs and other microelectronic applications. The requirements of the applicable MIL specification for optical inspection are described and defined in terms of a functional specification for automated inspection. The balance of the paper contains a description of a breadboard prototype inspection system and the approaches to image acquisition, enhancement, segmentation and feature extraction that have been developed for wire and die bond inspection

INTRODUCTION

This paper is a report on the progress of a Small Business Innovative Research (SBIR) project aimed at automating visual inspection of die and wire bond as it applies to inspection of Multi-Chip Modules (MCMs). The work is being performed for the US Navy Naval Surface Warfare Center (NSWC), White Oak Laboratory in Silver Spring Maryland. as NSWC Contract N60921-93-C-0037 under the technical direction of Mr. Victor Newton and Mr. Kenneth Johnson of NSWC. The Contract Officer is Ms. Donna Jackson .

The contractor, Acuity Imaging (formerly Automatix Inc. and Itran Inc.), is a significant player in the field of applied machine vision with a particular emphasis on solutions for end-users and system integrators in electronics manufacturing and process control.

A SBIR project consists of a Phase 1 conceptual design phase and a Phase 2 feasibility demonstration and prototype development phase. The Phase 2 effort was spilt into two separate sub-phases. The first of these, which has been completed, was aimed at developing and demonstrating automation of the critical inspection processes on a breadboard prototype. The second which has recently commenced will consist of the development of a practical inspection system to be delivered to the Navy.

The objective of this paper is to report on the feasibility of wire and die bond inspection and to describe the approaches that have been identified for inspection and measurement. An equally important goal is solicitation of input from individuals and organizations that are active in MCM manufacturing and deployment, so that the balance of this program can focus on successful commercialization and provision of a useful tool for the user community.

SUMMARY

Acuity Imaging have completed the first part of a two year effort aimed at automating inspection of wire and die bond for multi-chip modules. The first year's effort was aimed at demonstrating critical inspection processes on a breadboard prototype which was implemented on a standard platform based machine vision system. Inspection processes have been developed for die location and orientation, die attach, wire bond shape, wire bond location and wire connectivity and separation.

The successful demonstration of these inspection processes required the development of several novel approaches to image segmentation and analysis. These include:

- The use of a programmable light source in combination with image arithmetic to improve discrimination of parts and background.
- The use of correlation in rotation and scale space to locate wire bonds.
- The development and use of tracking filters to follow wires over irregular background scenes.
- The use of connectivity analysis and mathematical morphology to check wire connectivity and separation.

The principal goals for the balance of Phase 2 of this project are the further development of the inspection algorithms and the integration of a prototype wire and die bond inspection system that can be delivered for evaluation and use in a Navy laboratory or production facility.

CONTEXT

Wire bonding is a commonly employed means of creation of electrical connection between a semiconductor die and a leadframe or substrate. Applications of wire bonding range from the humble 16 lead DIP which might contain elementary logic to advanced "mainframes on a chip" such as the DEC Alpha processor. Wire bonding is also widely used as the means of interconnect in MCM manufacture since it is a well-understood process and the equipment required is widely available.

In its essence the process of wire bonding consists of making an electrical connection using fine (typically 25 micron diameter) gold or aluminum wire between pads on the die and sites on the substrate or lead frame. The integrity of the bond and the wire is essential to the subsequent performance of the device.

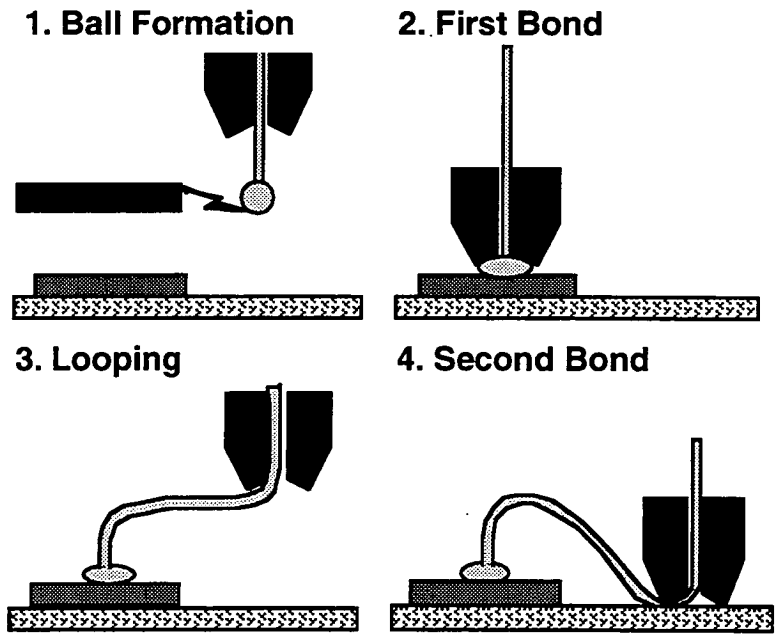


Figure 1. The Ball Bonding Process

The process of wire bonding is dependent on the formation of a fusion bond between the wire material and the pad (or substrate) metalization. The fusion of the wire material with the metalization of the pad or substrate is effected by various combinations of heat, ultrasonic energy and contact pressure. The two major classes of bonding process are *ball bonding* and *wedge bonding*. These processes vary in the details of how the wire is prepared and presented, how the fusion process is promoted, and the morphology of the resulting bonds at either end of the wire. The ball bonding process is illustrated in Figure 1. The wedge bond process differs in the shape of the tool, the manner in which the wire is guided and the shapes of the bond produced. Figure 2 illustrates the different shapes of bond produced.

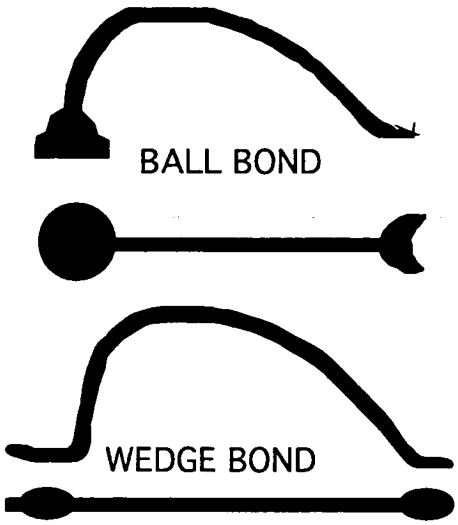


Figure 2. Ball and Wedge Bond Shapes

Precap Inspection

Third Optical Inspection or Precap Inspection is the term typically used to refer to the inspection step that is employed in the assembly process prior to encapsulation or sealing. The importance of this inspection stage derives from the fact that this is the last point at which a part can be reworked or scrapped before significant value is added. This inspection is typically performed by human inspectors using low and medium power microscopes (and very high powers of concentration). In the case of the devices destined for defense applications the form of this inspection is defined by MIL STD 883c "Test Methods and Procedures for Microelectronics". The method that is applicable to MCMs is Method 2017 -- Internal Visual Inspection (Hybrid).

Even though wire bonding is a relatively well understood and controlled process, the continued need to bond at high speeds and the trend to packages with ever higher numbers of interconnect has meant that inspection has remained an essential part of the assembly and packaging process in cases where high yield and reliability is required. This is particularly true when MCM integration is considered,because the value added during the assembly stage is significantly greater that in the case of single chip packages. It has become a commonplace that MCM's need what are referred to as Known Good Die (KDG)[1] for acceptable yield. This is due to the high value added during assembly and the fact that a part with 20 dice with an individual yield of 97 % will only have a module yield of 50% (0.97 to the power of 10). A similar effect will occur if the reliability of the assembly process is not near perfect. The high cost of scrapping or rework can be avoided if good assembly is confirmed prior to commitment of further dice. What is important for single chip modules (SCMs) is essential for MCMs.

WHY AUTOMATE INSPECTION ?

The short answer to this question is that the traditional manual inspection is labor intensive and is not a good fit with a state-of-the-art microelectronic manufacturing.

The Phase 1 report to the Navy summarized a survey of failure modes in hybrid devices [1,2] that reported that wire bond failures (at 23%) were the second most common failure modes in the population of hybrid devices surveyed. A further study [3] reported the experience of 19 microelectronics manufacturers with the cost and reliability of MIL-STD 883c Method 2017. The overall conclusion was that though inspection costs were growing as a percentage of manufacturing cost, the results obtained were still unreliable and subject to difficulties in interpretation.

The traditional motivations for automated inspection are improvements in reliability and manufacturing cost reduction. Reliability improvement is derived from the inherent consistency of an automated inspection process, which is not subject to fatigue or difference in interpretation between operators. Economy results when an automated inspection system is able to improve defect detection rates and reduce scrap, work in progress (WIP), or scrap costs. A further advantage of automated inspection is that it is by definition scalable.

---

[1]Known Good Die are die that have been electrically tested to confirm functionality prior to assembly. In the assembly of conventional single chip packages this step occurs after the chip has been integrated in the package.

There are also a number of indirect benefits of automated inspection that relate to the micro-electronic manufacturing and development process in general and the current trends in military micro-electronics in particular. These are:

Integration with Automation. As manufacturing processes become more integrated there is a need to integrate inspection with the overall process and to remove manual handling from the process flow. Automated inspection fits very well with this requirement.

Data for Statistical Process Control (SPC). It is becoming widely understood that it is better to implement well controlled manufacturing processes than to rely on inspection to detect defects introduced by uncontrolled processes. This argument is commonly used to suggest that inspection is not a value-adding step in manufacturing. However, SPC requires measurement to effect control and the systems required for automated inspection (and the measurements performed) are for the most part the same as those required for SPC. Technology developed for automated inspection will continue to be of use for process control and monitoring.

Reliable Source of Defect Data. The collection of defect data is an important part of inspection. An automated inspection system can be readily integrated with an automated defect data collection system.

Support Rapid Prototyping and Process Development. Successful prototype device and process development is strongly dependent on the ability to verify assembly and to quantify variations due to changes in process parameters.

Higher Yields From Short Runs. Defense electronics are typically manufactured in short runs and small lots. There is little opportunity to get the bugs out of the manufacturing process during an lengthy startup build. Automated inspection will allow the learning (and yield) curve to be steepened.

Even the Cleanest People are Too Dirty for Clean Rooms People are a significant source of contamination. Automating the inspection process will allow reduced handling and exposure to contamination prior to encapsulation

Approaches to Automatic Bond Verification

The author is aware that a wire bond that looks like a very good bond may fail to do the one thing that a bond is required to do. This is to conduct signals reliably during the life of the device. It is for this reason that the utility of optical inspection is not universally accepted and that there is a substantial body of opinion that holds that the only way to determine whether a reliable interconnect has been constructed is to perform a test that verifies the presence of the required area of diffusion between the wire and the pad metalization. The required pull or shear test can confirm conclusively the presence of a bond between the wire and pad. However, the relative complexity and invasive nature of such "objective" tests suggests that there is a useful function to be performed by an inspection method that can confirm that there is a bond of the right shape in the right place connected by a correctly formed wire loop to a second bond which is also correctly formed and in the right place.

OVERALL REQUIREMENTS

The goals incorporated in the last sentence of the above paragraph form the basis of the applicable parts of Method 2017 in MIL-STD 883c - Manual Visual Inspection (Hybrids) . The near term goals of the current project have, therefore, been defined in terms of automating the appropriate parts of Method 2017.

Method 2017 In MIL-STD 883c

Table 1 describes the measurements and inspection processes specified in Method 2017 that are applicable to MCMs. Method 2017 typically describes conditions that are to be considered

unacceptable. For clarity of reading these criteria have been inverted in the table. The figures in parentheses are the criteria for Class S applications.

Gold Ball Bonds	Ball bond diameter shall be greater than 2.0 and less than 5.0 times the wire diameter. Ball bond wire exit shall be within the perimeter of the ball. The wire center exit shall be within the area of the bonding pad. *Intermetallic formation shall not extend radially more that 0.0001 ins. completely around that portion of the gold ball bond located on the metal*
Wedge Bonds	Ultrasonic wedge bond width must be greater that 1.2 and less than 2.5 times the wire diameter. Ultrasonic wedge bond length must be greater that 1.5 and less than 5 times the wire diameter. Thermosonic and thermo-compression wedge bond width must be greater that 1.2 and less than 3.0 times the wire diameter. Ultrasonic wedge bond length must be greater that 1.5 and less than 5 times the wire diameter. [Other length to wire width ratios are specified for RF devices with aluminum and gold wire]
Crescent Bonds	The tail of a crescent bond shall be greater than 1.2 and less than 5 times wire diameter. Bond length shall be between 0.5 and 3 times wire diameter. The tool impression must cover the entire width of the wire
General (applicable to all bonds)	Bond must be at least 50% (75%) on pad on the die. Bond must be 100% on pad on the substrate. The wire exiting the bond must not pass over another bond. Bonds must not be closer that 0.001 ins. to unconnected unglassivated metalization. For glassivated metalization this distance is 0.0001 ins. Wire bond tail length shall not exceed two wire diameters at the die pad or four wire diameters at the package post. The degree to which the bond or the tail is allowed to cover the exit metalization of a pad varies according to the class of device and type of bond. In general, if metalization is visible that forms a conducting path at least 0.0001 ins. wide between the pad and the exit the joint is legal.
Wire	Wire separation must be at least two wire diameters away from the bond. This separation requirement is reduced to 0.001 ins. close to the bond. Wires path must have some "arc". Wires must not have sharp bends or nicks. *Wires must not come within 0.005 ins. of the lid of the package*

Table 1. Summary of Method 2017 Requirements Applicable to Wire Bonds
Items in italics are not currently being considered for automatic inspection

REQUIREMENTS FOR AUTOMATIC INSPECTION

In order to implement a practical automatic inspection system the inspection criteria defined above were transformed into a measurements specification. This specification is summarized below:

Die and Substrate Location The position of the die on the substrate must be precisely measured. This measurement must be in terms of X, Y and Theta and must be made using features on the die that do not vary from die to die. This last requirement rules out the use of the edge of the die as a location feature since the die edge location will vary from part to part.
Die and Substrate Orientation The orientation of the die and substrate must be determined unambiguously. This means that die that has been misplaced by 90 or 180 degrees must be detected.
Wire Bond Location The wire bond center and outline must be located. This information will be used to determine the percentage coverage of the bond pad and to detect excessive proximity to other die structures.
Wire Bond Morphology The width and length of the wire bond must be determined as must the length of the tail and the exit angle of the wire.
Wire Connectivity Wire connectivity between specified die and substrate pads must be confirmed.
Wire Trajectory and Separation The required separation between adjacent wires must be confirmed in 2 or 3D.

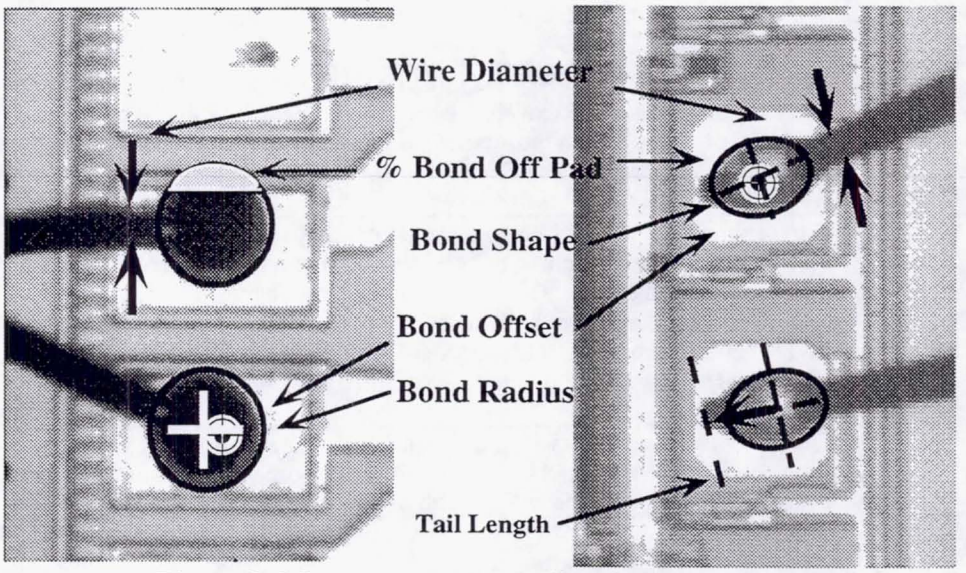


Figure 3. Measurements on Ball and Wedge Bonds

Figure 3 illustrates graphically the dimension and location measurements that are required for ball and wedge bonds. In both cases the "second" bond - at the other rend of the wire will be significantly different in shape.

PROTOTYPE IMPLEMENTATION

The balance of this paper describes aspects the breadboard prototype wire bond inspection system that has been implemented by Acuity to inspect the NSWC UDSP multi-chip module prototype.

Optics

Design studies suggest that a useful wire bond inspection system will require at least two levels of magnification. One of these will be high magnification with a field of view (FOV) in the 0.020 to 0.030 ins. range for locating die and inspecting bonds and another in the 0.2 to 0.25 ins. range for die attach and wire connectivity checking. This range can be provided by utilizing multiple lenses and cameras, or using a motor driven zoom or turret head lens. For the breadboard prototype this complexity was avoided by using a single lens with a FOV of approximately 0.060 ins. This setup allowed bonds to be located with reasonable accuracy, while allowing the shorter wire runs to fit into a single FOV. The lens employed was a Leica MonoZoom 7, a lens that is widely used in industrial and electronic inspection applications

Lighting Bonds and Wires

The provision of suitable lighting is critical to all machine vision applications. It was clear from the literature and from experience in Phase 1 that this is particularly true of die and wire bond inspection. The basic requirement for this application was that both coaxial and programmable low angle diffuse ring lighting be available[2] The coaxial illumination requirement can be met by specifying a lens system that incorporated "through the lens" illumination . The low angle ring lighting was provided by a custom made illuminator that consisted of three concentric rings of diffusers that mounted below the objective lens (Figure 4).

---

[2] In this context coaxial illumination is bright field and ring light is dark field illumination.

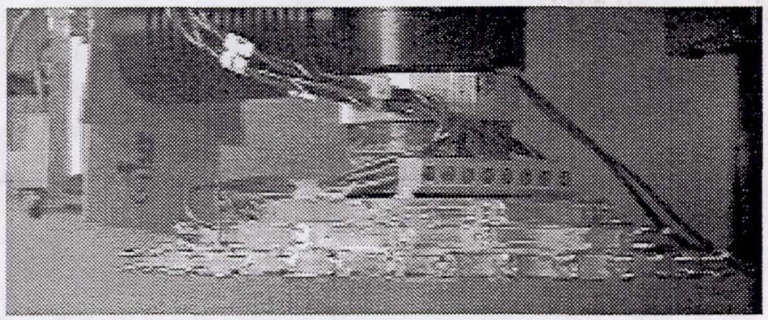


Figure 4. Prototype Lighting System for Wire Bond Inspection

Lighting for Enhanced Segmentation

The underlying reasoning behind the provision of programmable lighting is the requirement to separate the various components of the MCM in greyscale (i.e. brightness) in the images stored in the vision system. The important point to note is that when illumination is changed from coaxial to diffuse low angle the image of certain surfaces may change from dark to light or *vice versa* . Figure 5 shows two pairs of images of wire bonds and wire taken with different lighting conditions

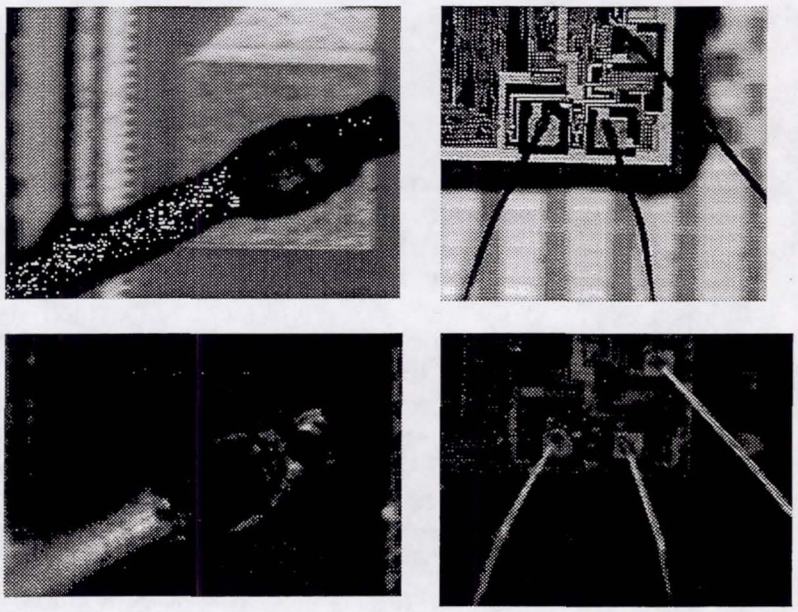


Figure 5. Bonds and Wire under Different Lighting

Image Enhancement

Wire bond inspection is one of the many cases where the "as acquired" image can not be reliably segmented without further enhancement. The two principal enhancement operations that are most appropriate for this application are *Mathematical Morphology* and *Image Arithmetic*. An example of the former is the use of a morphological filtering operation to remove probe marks that might otherwise confuse measurement of bond dimensions. Arithmetic combination of images provides the means to enhance the contrast between the features of interest and the background. Figure 6 shows the results of combination of the images from Figure 5. The contrast between the bond and the die and between the wire and the die and die attach material has been greatly enhances in both cases. Successful application of this techniques requires careful selection of parameters and control of image contrast.

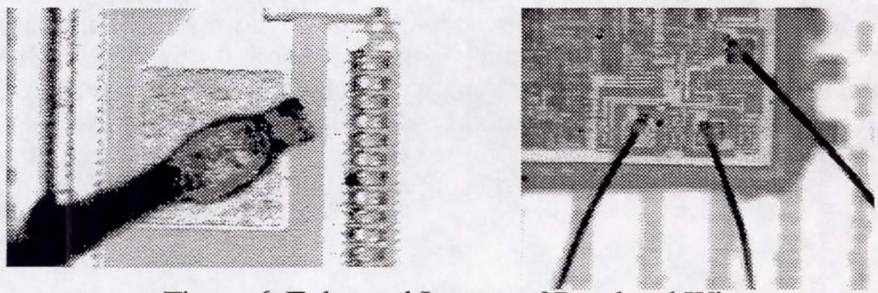


Figure 6. Enhanced Images of Bond and Wire

FEATURE EXTRACTION

Extraction of useful information from electronic images is the working definition of machine vision. The preceding section describes how images that can be successfully analyzed can be best obtained if the object is to inspect wire bonds. It is now time to cut to the chase and describe how useful measurements can be extracted from these images.

Die Location

The requirement is to locate the component - either die or substrate with respect to the component to which it is physically attached or to which its location is referenced.. This requirement derives both from the need to check component location and orientation as specified by Method 2017 and the need to provide the inspection system with precise data on component location for use by other inspection processes.

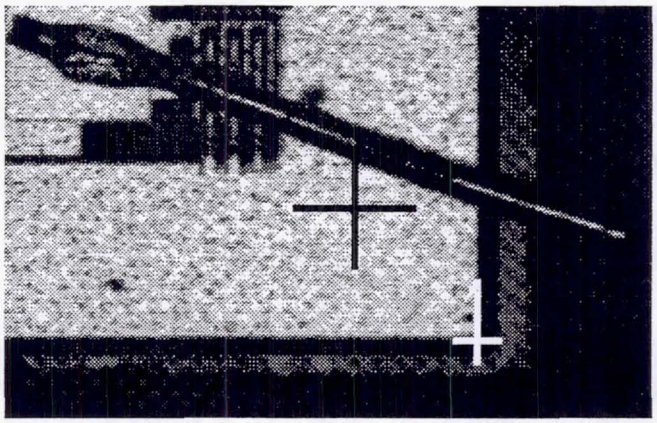
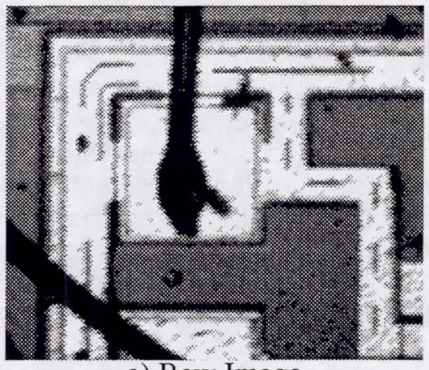


Figure 7. Feature Location For Die Location On a NSWC UDSP Die

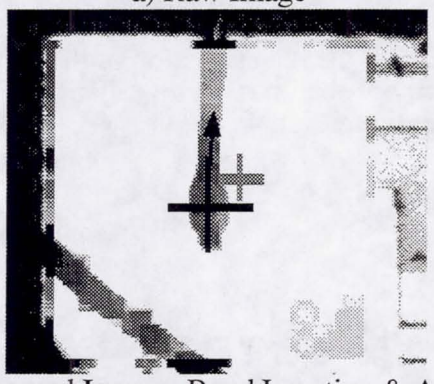
The most suitable approach is to locate two or more features on the die surface and to use knowledge of the location of these features in die space to calculate the current die offset. Under most circumstances normalized correlation pattern matching is the clear choice for location of die features Normalized correlation has the ability to locate features with high noise immunity and insensitivity to variations in illumination. In cases where a feature may be partially obscured the generalized Hough pattern match can be used instead of normalized correlation. Figure 7 shows an image in which normalized correlation has been used to locate the corner of a die. A white cross has been drawn on the location where the stored template that represented of the corner of a die best matched the image under examination. The black cross is placed on the expected nominal location. The distance between the crosses is the offset of this corner of this die.

Bond Location And Shape

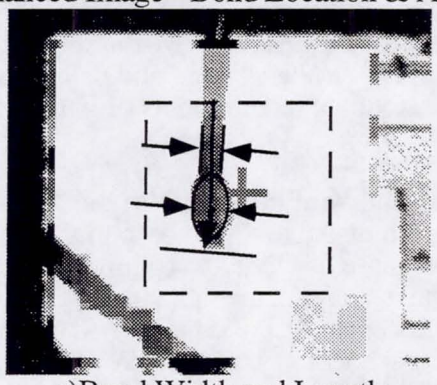
Each wire bond must be located with respect to the die in such a manner that its coverage of the die or substrate pad can be determined. Once the bond is located critical bond dimensions can be measured. The process currently implemented for wedge bonds is illustrated in Figure 8 below. The process starts with an enhanced image of the bond and surrounding die surface. The initial bond location is then performed using an adaptation of the normalized correlation pattern matching technique in which a synthetic template is used to search for the bond in X, Y, angle, and scale. Once the bond center is located the bond width and tail length are measured and the bond location and outline compared with the bond pad outline. This last step is used to determine whether the allowable percentage of bond off the pad has been exceeded. In the case of ball bonds the radius of the ball is measured using a variant of the Hough edge fitting algorithm



a) Raw Image



b) Enhanced Image - Bond Location & Angle



c)Bond Width and Length

Figure 8. Bond Location & Measurement

Wire Location

Measurement of wire bond location and morphology is challenging in terms of its requirements but is not outside the mainstream of machine vision applications. However, the need to follow wires and check their end points and separation is a much less constrained problem. In analyzing the requirements of this application two factors stood out:
1) Traditional threshold based feature extraction techniques are unlikely to be successful at determining wire path when the wire crosses a number of backgrounds and is not constrained in height or path.
2) Separation checking will be very computationally intensive if a large number of wires have to be checked.

The approach to wire location and separation checking that is being developed takes account of both of these factors. The basic element is a tracking filter which has high immunity to background noise and blur of the wire image. This filter is applied to suitably enhanced image to produce an ordered list that represents the path of a wire. The use of this filter is illustrated in Figure 9 which shows wire being successfully followed from bond to bond over mottled backgrounds and in images where the wire is significantly out of focus. This sort of tool is particularly applicable to MCM development and prototype applications such as the Navy UDSP where wire runs may have significant curvature.

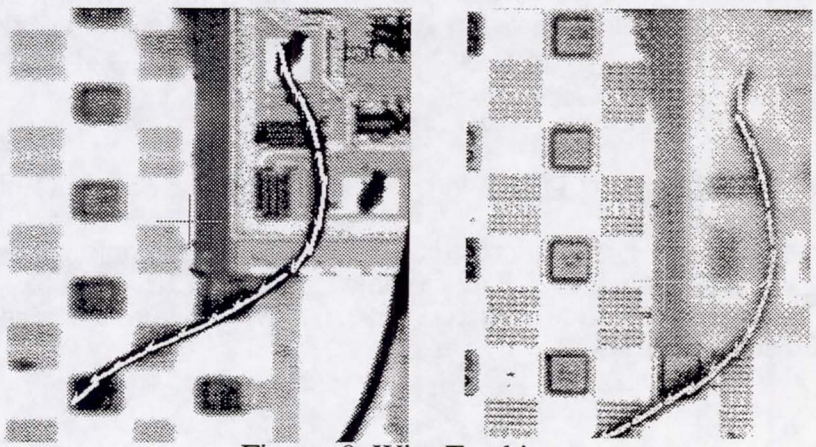


Figure. 9. Wire Tracking

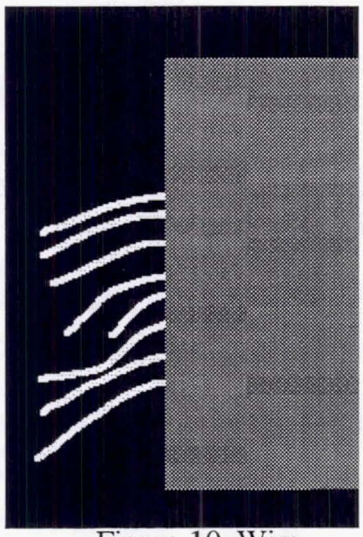


Figure 10. Wire Separation Checking

Wire Separation Checking

The approach taken to wire separation checking is to use existing capabilities of the machine vision system to determine whether any two wires are too close to each other. The tools used are binary connectivity and morphology. The procedure is to use the wire location information generated by the wire location process to make a synthetic image in which all the wires in a region are represented (Figure 10) . A connectivity analysis is then employed to determine how many connected white blobs (wires) are present. Once this is done a morphological dilation operator is applied to the image to increase the diameter of the representation of the wires. A subsequent connectivity analysis will reveal any pairs of wires that were insufficiently separated

CONCLUSIONS AND PLANS FOR DEVELOPMENT

Automatic machine vision based optical inspection has the potential to offer significant benefits in the area of verification of wire and die bonds in SCM's, MCMs, and hybrids. The work performed to date has established that there are image enhancement and feature extraction techniques available that are a good match with the functional requirements of wire and die bond inspection.

The breadboard prototype has established the feasibility of inspection of die and wire bonds on MCMs and has validated a number of approaches to machine vision based inspection. The objective of this paper is to present the results of this phase of work and to inform the user community of the state of development. During the next year a deliverable prototype will be built and tested that will incorporate the inspection processes described above. This prototype will be implemented on hardware that will allow the inspection process to take place at a rate that is a better match to production rates than was achieved with the prototype hardware.

Input from manufactures of MCM and users of wire bonding processes will be critical to successful further development in this are. It is Acuity Imaging's intention to actively solicit this input during the next phase of this project.

References

1. Blanton, J., How to Prepare for a 1772 Audit, Defense Electronic Supply Center, November 1985.
2. Licari, J.J. & Enlow, L.R., Hybrid Microcircuit Technology Handbook; Materials Processes, Design, Testing & Production, Noyes Publications, Park Ridge, NJ, 1988.
3. Siu, B., "Pre-Cap Visual Inspection", MicroCIM AHAP Meeting Proceedings, January 28-29, 1992, Rancho Cucamonga, CA.

3-D DIGITAL ROBOTIC OPTICAL INSPECTION DEVICE

Bruce R. Altschuler, COL, USAF, DC
Dental Research Detachment, Walter Reed Army Institute of Research
US Army Medical Research and Materials Command
Ft. George G. Meade, Maryland 20755

ABSTRACT

The DROID, or 3-D Digital Robotic Optical Inspection Device, is an advanced tool for extremely rapid 3-D medium range inspection and numerical mensuration, using triangulated and electro-optically structured laser light, and from at least one, to several, videocameras--some mounted on articulated robotic arms, to provide precise quantitative measurement of any objects within the inspected field of view. Unlike fixed-jig 3-D scanning devices, this tool uses electro-optic, not mechanical moving parts, for data acquisition, and can inspect singly or cooperatively with other DROIDS, providing mutual calibration of scene data.. Freed from any fixed mounting requirement, the DROID may be adapted for mobility, and can go to the object. 3-D measurement of parts while still on a machining lathe or mill are possible, as well as 3-D cooperative inspection for dynamic assembly by multiple DROIDS, dynamic inspection of warping or flexing parts such as plastics during cooling, or human motion envelopes for safety equipment or prosthetic customized design. The DROID is ideally suited for tele-medicine as a potentially quasi-autonomous tool for rapid accomplishment of tedious surgical sub-tasks--micro-suturing of nerves, blood vessels, etc., when the 3-D/4-D inspection is coupled with manipulative robotic end-effectors. Major economies in surgical costs, automated assembly, and closed loop automated manufacturing are possible with creative implementation of the DROID concept, especially where coupled with its self-generating 3-D archival data bases. Applications include: robot builders of space stations; robot assistant surgeons; automated hazardous waste cleanup robots; automated interactive machine tool inspectors; closed loop automated manufacturing of flexible materials (including hot metals, plastics and composites).

INTRODUCTION

In recent years industry has progressed slowly in the field of 2-D inspection and robotic manufacturing. Most robots are relegated to pre-programmed repetitive tasks involving assembly. A few feeble applications in 2-D have begun involving welding and verification of pre-positioned part assembly. Many of the futuristic promises of robotic "CYBORGS" have been compromised by the lack of effective machine vision sensory capability and the lack of effective "smart" recognition algorithms. Fast, accurate, and versatile 3-D mapping in a 4-D (time-dynamic) capacity is a requirement for mobile robots to achieve utility in next generation rapid prototyping and minimal run production on a flexible computerized assembly line. Only through 3-D sensory information input in the form of digital (numerical) coordinate data, appropriately filtered and processed, will the robot have the capability to progress in interpreting its environment in a contextual and "comprehending" way. Between the dumb and senseless robot of today, and the "CYBORG" brilliant, interpretive "Robocop" or "Robomedic" of science-fiction of tomorrow, lies a realizable and practical interim step of the "DROID", a tool

with a high 3-D sensing capability, high mobility potential, with proprioceptive articulated arms and end-effectors, and upgradable computer interpretive capabilities. This paper describes progress and the potential technology transfer capabilities of the DROID in its present and near-term upgradable configuration. With the DROID, new applications can be envisioned because the DROID has the ability to provide 3-D data in 4-D (moving and changing) environments, can be mobile and bring itself or a suitable probe (such as a 3-D 'camera' on a robotic arm) to the inspectable object, and can cooperate with other DROIDs or equipment for interactive control, manipulation, feedback looping, and data archive interpretive statistical assessment of rapidly changing conditions.

Need for Rapid 3-D Quantitative Data Acquisition in a DROID

Humans see, understand, recognize, and react to visual data in a manner still largely unexplained by scientists. Computers are far too primitive to do anything beyond comparing digital (number) data. Until the visual field is reduced to sets of numbers, the computer cannot even begin. Two-dimensional number data from digitized videocamera input has been explored for over 25 years by scientists in a largely unsuccessful and frustrating effort to transform flat images into 'objects' that a computer can recognize and "understand". A pencil on the table, for example, is immediately isolated, recognized, and understood by a human as a discrete entity, and its implication as a writing implement available on the table, and distinct from the table, is stored in memory. Thus humans understand the environment within the image and have a perception of place, orientation, and context.

The human can now remember the pencil and locate and identify it from any perspective angle, even if partially covered, sort through any clutter (other objects later introduced onto the table top), plot a non-interfering path around obstructing objects in the foreground, and reach out and pick up the pencil. To even begin to emulate this ability in a computer will require a set of machine visual capabilities not yet available. The DROID will begin to provide the tool set by creating 3-D quantitative data in a continuous real-time fashion, and provide it without the sensor being in a fixed jig mode. The DROID visual probes can emulate a person moving his head to get a look around an object to better observe it. While instinctive to humans, and easily accomplished, this ability to look around and compile a 3-D memory, or data base, from multiple aspects, is revolutionary. The 3-D data will also provide an entire extra dimensional data set for isolation, storage, recognition, and orientation of objects. Since the DROID will acquire data so quickly, continuous real-time measurement of human organ or industrial mold deformation characteristics could provide parametric input for virtual reality simulation programmers to add realism to computer simulations. This could dramatically improve the effectiveness of training and educational instruction in almost all fields, and greatly enhance present manufacturing technologies involving semi-molten materials, plastics, composites, and ceramics.

Thus the DROID is a breakthrough research tool in close to medium range inspection and will become the basis for future intelligent visual robotics. It will eventually enable eye/hand (sensor/end-effector) coordination in an autonomous fashion by the "brilliant" robot of the future. The DROID will enable a decade or more of fundamental research in intelligent

robotics and robotic (machine) visual perception, while providing near-term practical utility for industry and medicine.

Other 3-D scanning systems presently available require fixturing of the sensors or object to obtain precision measurements. In the DROID, the capability exists for multi-axis mobile close-range 3-D probing with articulated sensors, and even cooperative multi-DROID sensing. Without the DROID capability, only the current quagmire of 2-D image analysis, circa 1970's concepts, is available for true production line industrial applications. 2-D vision, however, has failed to match the promised dream of visual intelligence in robotics, whether for surgery, industry, or other applications.

Introducing "Smart" DROID concepts in futuristic health care for cost-savings in surgery

The technology to reliably transmit high-resolution visual imagery over short to medium distances in real-time has led to the serious consideration of the use of telemedicine, telepresence, and telerobotics in the delivery of health care. Telemedicine implies the electronic transfer of data and imagery to a remote site where a health care provider can effectively interpret data. Telepresence implies there is real-time transfer of imagery and other sensory data and interactive information flow between patient, on-site provider, and off-site health care providers. Telerobotics implies that the off-site, the on-site, or both providers are physically manipulating or interacting with the patient indirectly through some form of robotic control mechanisms.

These concepts may involve, and evolve toward: consultation from remote expert teaching centers; diagnosis; triage; new virtual reality teaching methods for interns; real-time remote advice to the surgeon, real-time remote surgical instrumentation manipulation (tele-robotics with virtual reality). Further extrapolation leads to tele-design and tele-replication of spare surgical parts through quantitative tele-imaging of 3-D surfaces tied to CAD/CAM devices and an artificially intelligent archival data base of "normal" shapes. The ability to generate "topograms" or 3-D surface numerical tables of coordinate values capable of creating computer generated virtual holographic-like displays, machine part replication, and statistical diagnostic shape assessment, is critical to the progression of telemedicine. Any virtual reality simulation will remain in "video-game" realm until realistic dimensional and spatial relational inputs from real measurements in-vivo during surgeries are added to an ever-growing statistical data archive. The challenges of managing and interpreting this 3-D data base, which would include radiographic and surface quantitative data, are considerable. As technology drives toward dynamic and continuous 3-D surface measurements, presenting millions of X,Y,Z data points per second of flexing, stretching, moving human organs, the knowledge base and interpretive capabilities of "brilliant robots" to work as surgeon's tireless assistants becomes imaginable. The brilliant robot would "see" what the surgeon sees--and more, for the robot could quantify (measure) its 3-D sensing, would 'see' in a wider spectral range than humans, and could zoom its 'eyes' from the macro world to long-distance microscopy. Unerring robot hands could rapidly perform machine aided suturing with precision micro-sewing machines, splice neural connections with laser welds, micro-bore through constricted vessels, and computer combine ultrasound, micro-radiography and 3-D mini-borescopes to quickly assess and trace vascular

problems 'in-situ'. Special pneumatic bone 'nailers' would reattach splintered bones and spare-part replicas, gaining positional knowledge in real-time. The spatial relationships between organs, robotic arms, and end-effector diagnostic, manipulative, and surgical instruments, would be constantly monitored by the robot 'brain' using inputs from its multiple 3-D quantitative 'eyes' remote sensing, as well as contact and proximity force measuring devices.

Where 'virtual reality' is the generation, display, and computer representation of a spatial world (real or simulated) to the eyes of the human for human visualization, 'digital reality' is the creation, measurement, assessment, tabulation, and manipulation of digital spatial data actually sensed by artificial means of the real world and input into the host computer for computer 'visualization', or, more accurately--'3-D machine perception' of the environment and spatial relationships and solids in the robot's sensory field of view.

To achieve real-time 3-D digital reality in a dynamic sense, measurement tools must be available to rapidly acquire and calculate quantitative 3-D data from the spatial environment. Rapid, remote, non-destructive imagery using at least one structural laser light pattern generator and one or more passive videocameras off-axis to it can create accurate 3-D numerical maps of any scene. Previous research has demonstrated success with non-scanning, non-mechanical, electro-optic means to generate structured light patterns. With improved and spatially programmable light modulators capable of 1000 Hz, and parallel output video cameras of similar (1000 frame/sec) capabilities, the concept of 30 frame/second continuous capture of 3-D scenes with tens of thousands of numerical coordinate points becomes possible (buildable with current technology within a year).

The DROID has achieved accurate quantitative mapping of up to 1000K data point sets per perspective per half second, of objects varying in size from 1 cubic centimeter to objects several meters large (using conventional 30 Hz videocameras). Ranges can vary from 2 centimeters to hundreds of meters, and the field of view can be varied by suitable zoom lenses.

The DROID in Rapid Prototyping Applications

A victim of an automobile accident may have damage to critical facial bones. Restoration of form and function by surgical intervention may require a series of steps that could be automated in the future, but involve objects (bones) in which no *a priori* knowledge exists of the damage, displacement, repairablility, and precise shapes involved. Everyone has similar yet unique size and morphology of their bony skeleton. Comprehensive numerical inspection, compared to a 3-D data base, and with interactive consultation and CAD/CAM manipulation by the surgeon, could lead to the replication of damaged bones through rapid prototyping which could then be inserted as spare-part replacements, using bone grafts or bone substitute matrices. In a similar manner, prostheses involving limb replacement could be more effective for individuals by being manufactured to more closely match individual patient morphology and needs. End effectors with advanced tooling would allow the 3-D dynamic input measurement data to be directly applied to precision micro-suturing, where a cut nerve requiring tens of minutes of delicate hand repair under stereo microscopy by the surgeon using present methods could be accomplished in the future by the surgeon computer interactively designating the DROID to effect the repair with

a miniature sewing machine -- accomplishing the same task in seconds. In industrial tool and die making, the DROID could continuously inspect a prototype part during initial test use, assess wear optimization, and redesign the die to improve the final part. Where moldable materials are used that may shrink or deform upon mold release, the DROID would enable an interactive inspection process to continuously improve a prototype mold, or refine the manufacturing process interactively, so that the mold shape would be purposely and iteratively deformed to compensate for the final shape. Upon release from the deformed mold, the plastic part would cool and stress relieve to the exact final shape desired. Costs would be saved by elimination of most net shape machining, especially useful where complex shapes or exotic materials are used. The DROID could watch machining processes as they occur and perform precision 3-D analysis of the process. Most expensive hand forming labor involving the shaping of composites would be eliminated by sets of DROIDs working in unison.

FUNCTIONAL DESCRIPTION

The DROID, in its eventual full implementation would consist of the following 8 main systems:

System 1: An optical laser structured light generation system housed in a cylindrical tube on an elevation mount and connected to a vertical pedestal with an azimuthal turntable. (Fig.1)

System 2: Initially two videocameras at approximately eye separation distance mounted on a horizontal axis rotation platform mounted in turn on a vertical linear translation table mounted rigidly to System 1 tube assembly. The two videocameras elevate or depress simultaneously to follow the beam path of the laser (but off-axis to it). These videocameras have computer adjustable controls for independent focus, zoom, interpupillary distance, roll, and convergence to suit the virtual reality individual teleoperator's preference. Additional videocameras will be similarly mounted with such features as: color; infra-red; low-light intensification; and high resolution; as deemed necessary per application. High speed and gating capabilities will also be added as needed.

System 3: Four robotic arms (2 for the videocameras), each independently capable of 6 axis movement to +\- .002 inches positioning accuracy, and attached to the System 1 pedestal (such that the entire DROID can rotate on its azimuthal turntable--allowing the arms to track with the laser beam array). Each arm holds a videocamera or a specialized and interchangeable robot end-effector manipulator. Initially, the videocameras are conventional monochrome, with computer controlled focus, zoom, and gain, but will be replaced in some models with high-speed parallel videocameras. End effectors to be placed on one or more robot arms, and their selection, will be determined per application. (Fig. 2).

System 4: The computer and servo-control system to calculate data, control the robot arms and other automated systems, and integrate all functions, and provide a data base and display.

System 5: Display and interactive controls for human interfacing, virtual reality, telepresence, and teleoperations.

System 6: End-effector manipulators and auxiliary diagnostic probes, special surgical tools, etc. developed and integrated into the DROID.

System 7: Telecommunications to off-site for data display, data transfer, and control, to include modem, and satellite capability, if required for an application.

System 8: CAD/CAM replication milling software and machine tools capable of rapid prototyping light duty fabrication from DROID input.

THEORY AND DISCUSSION

Quantitative measurement by 3-D robotic vision is obtained by this DROID using patented technology developed for and by the Department of Defense by the author and his colleagues under what is now a joint US Air Force/ US Army project. The method and theory used has been previously reported in the scientific literature as a laser structured light triangulation stereo method and is well accepted [1,2,3,4,5]. The current DROID testbed is available at the DRD/WRAIR laboratory as a working research tool to test various (primarily medical/dental) applications. The barest functional DROID could consist of a single active/passive stereo pair (one structured laser projector or 'active camera', and any one videocamera or 'passive camera'), to produce the coordinate numerical values as calculated by the computer. The active camera consists of a customized doubled YAG diode laser producing green light. The polarized laser beam is expanded and passed through a patented prism to obtain a large array of simultaneously projected orthogonally patterned gaussian subbeams. This laser array as generated is subsampled by suitable optics and projected through dual ferro-electric polarized crystals oriented 90 degrees apart and each containing individually programmed rows that may be modulated independently or collectively at 1000 Hz. Through suitable computer control, the net effect is a capability to select columns or rows of electro-optical material to be activated or inactivated, rotating the polarization of the material to transmit or block corresponding columns or rows of the polarized beam array. The result is a predictable pattern of laser light projecting through the programmable spatial light modulator (SPLM). Programming a sequence of different patterns may be readily accomplished. No moving parts are required to achieve the structured light changes-- a vast improvement over conventional scanning galvanometers. Also, since the SPLM acts only as a window or blocker and does not generate the beams (i.e. it is not an aperture set) there is none of the diffraction fall-off typically found in aperture derived systems. We have successfully used this system to project patterned beam array sequences to distances greater than required for any close to medium range medical or industrial application.

The current working prototype uses a 128 column/ 128 row SPLM. Since the laser array can be greatly expanded, a 256 column/ 256 row SPLM or greater can be used as an upgrade. The present SPLM is a 6 inch diameter electro-optic component with a 2 inch square aperture, is 3 inches thick, battery operable, and can be synchronized to standard video. The initial active projector was built as an in-line optical system with no attempt at folding or minimization of the optical train length. A planned optimized redesign of the optics will cause a substantial reduction of the current 7 inch diameter by 56 inch size of the present testbed. The newest laser for this system, for example, will be cigarette case sized rather than the current foot long cylinder. Other miniaturization capabilities exist for substantial size reduction of the overall 'active camera'.

The structured light is projected onto an object. Various lenses can change the light array size to match the desired volume space to be measured. A set of binary patterns is projected onto the object in a known sequence. At least one videocamera, which must be off-axis to the laser, views the spatial pattern set and records each pattern in the set as a separate video image. Using standard 30 Hz RS-170 video, the SPLM changes patterns during the vertical blank time,

making the pattern change invisible to the image collection. A series of images is collected. Signal and image processing takes place in an image processor to refine these to a single collective space-encoded image which is transferred to the minicomputer. Numerical calculations then take place to create the 3-D data. Typically tens of thousands of X,Y,Z points are created per viewing perspective. An ASCII file of these points is generated and can then be used for CAD|/CAM, interactive 3-D graphics display, etc. This method involves triangulation, so the videocamera must be off-axis to the laser projection. By placing the videocamera on a robot arm, an object can be examined from multi-viewing directions. Depending on the lens arrangement selected for the videocamera at a particular mapping session, an object can be mapped in a microscopic or macroscopic mode, where a long-range macroscope or microscope can effectively map an object from relatively large distances away. An additional similarly equipped videocamera on another arm can allow simultaneous and independent viewing. One camera can hold a wide area lens and another can hold a macroscopic or microscopic viewing lens, for example. A unique feature is the calibration of a camera viewing an object by another pre-calibrated camera. The active-passive stereo method allows the mapping of featureless surfaces, an important requirement for mapping unknown and arbitrary anatomic parts, and a decided advantage over pure passive stereo, which cannot map featureless structures.

The DROID has two semi-fixed auxiliary cameras, also useful for 3-D mapping, located on the 'head' of the robot. They can be rotated to 'look down' along the path of the laser beams, adjusting for range to the object, and can be telescoped as a unit using the vertical translation positioner to bring the two 'eyes' closer to the laser or further from it. The two 'eyes' are at eye distance apart, and serve not only as independent videocameras for 3-D quantitative mapping when paired with the laser, but also provide non-quantitative (human visual stereo) for human virtual reality real-time telepresence. Since the DROID has pan/tilt capability and robot manipulators, the 'virtual reality eyes' can be used by the human tele-operator to assist in locating objects to grasp, etc., in telepresence mode. Additional videocameras can easily be added as needed, including color, low-light intensified, and infra-red, to enhance various application sensory needs. The robot arms chosen approximate the human arm in reach envelope. The DROID was made somewhat anthropomorphic to ease human interfacing. Two sensor arms and two manipulator arms are planned for the present testbed.

A series of binary encoded spatial images are required for each data set. At standard 33 millisecond per frame, using 128 rows, 8 images are required at a minimum, or .25 seconds. If column data is also collected, the time for data collection is doubled. For future upgrade, since the system is binary, every time the number of rows of modulated beams is doubled, only one additional image is required per data set. Thus for 256 rows only 1/30th second additional time is required, for 512 rows 2/30th second increase in time over the 128 row data set is needed. Thus as modulation optical crystal density is modernized and upgraded, very slight time for data collection increases occur. This is not the case for straight scanning systems. Using parallel videocamera 1000 Hz capability, the achievement of 33 millisecond (2-D data rate) continuous 3-D data collection is near-term. The ability for continuous 3-D data input at TV data rates will advance the capabilities of robotic visual sensing dramatically. Typically 50K to 100K data points are collected per video camera for an object as viewed from a single viewing perspective. Objects the size of teeth have been mapped to 25 micron resolution in 3-D. Head sized objects

to one-quarter millimeter. Larger objects can be mapped to less resolution if only a single view is required, or to higher resolution if multiple views are obtained. The collection of views from multiple directions can create 3-D reconstructions of a solid representation of the object. The end product of the DROID is large data files of 3-D coordinate points. These files can be used in a variety of ways: to position probes, tools, lasers, implants, etc.; to direct manipulators; to determine positioning or inspection criteria; to perform microsurgery. They can be used and transferred to CAD/CAM tools for 'rapid prototyping' shapes while-u-wait as almost a 3-D fax machine/ replicator. It will enable medical technologies in spare-part surgery, diagnosis, and training, and provide the input for the gradual building of a data-base sufficient to begin the creation of quasi-autonomous robot assistants to the surgeon.

SPECIAL FEATURES

The DROID uses no moving parts to 'scan' the object to be measured. It can be calibrated once in a fairly simple semi-automatic method, then can map anything in the field of view. It is calibrated to the scene at the scene, rather than to some restricted space or fixed jig arrangement. Thus the videocamera and active cameras can be placed arbitrarily (within some bounds) and be immediately calibrated. It obtains data fast, and can be upgraded easily to parallel architectures that will enable continuous data collection for 3-D data at current 2-D videodata rates, for dynamic tracking and measurement. More than one videocamera can obtain data simultaneously, and one videocamera can calibrate itself against another, allowing for cooperative DROID (i.e. multiple platform) capabilities. Range and field of view are dependent primarily on the lenses used. The system can map anything from the size of an ant to objects larger than a truck. Objects can be mapped in ambient light. Higher level resolution can be obtained as modulation and videocameras increase densities, with little memory or speed holdup. Currently over 100K data points can be obtained from one view and one videocamera.

REFERENCES

1.B.R.Altschuler, J.Taboada, M.D. Altschuler, "Laser Electro-Optic System for Three-Dimensional Topographic Mensuration", SPIE Proceedings Vol.182:192-196, April 1979.
2.M.D.Altschuler, B.R.Altschuler, J.Taboada,"Measuring Surfaces Space-Coded by a Laser Projected Dot Matrix", SPIE Proceedings Vol. 182:187-191, April 1979.
3. M.D.Altschuler, B.R.Altschuler, J.Taboada, "Laser Electro-Optic System for Rapid 3-D Topographic Mapping of Surfaces", Optical Engineering, Vol.20:6:953-961, Nov/Dec 1981.
4. M.D.Altschuler, K.Bae, B.R.Altschuler, G.Dijak, L.Tamburino, B.Woolford "Robot Vision by Encoded Light Beams" chapter in: Three-Dimensional Machine Vision ed. Takeo Kanade, ISBN 0-89838-188-6, Kluwer Academic Publishers, Norwell, MA, 1987.

Notes: This is a work of the US Government and may be reproduced for its needs. The opinions expressed are solely those of the author. Acknowledgment to: ADLAS, Inc. for the loan of a customized laser device; Hospital of the Univ of Pennsylvania Dept of Radiation Oncology for software support; NIST fabrication shops for calibrated geometric objects; US Army Medical and Dental Corps and US Air Force Dental Corps for tenacious joint support of this effort over many crucial embryonic years of struggling research and investigative efforts.

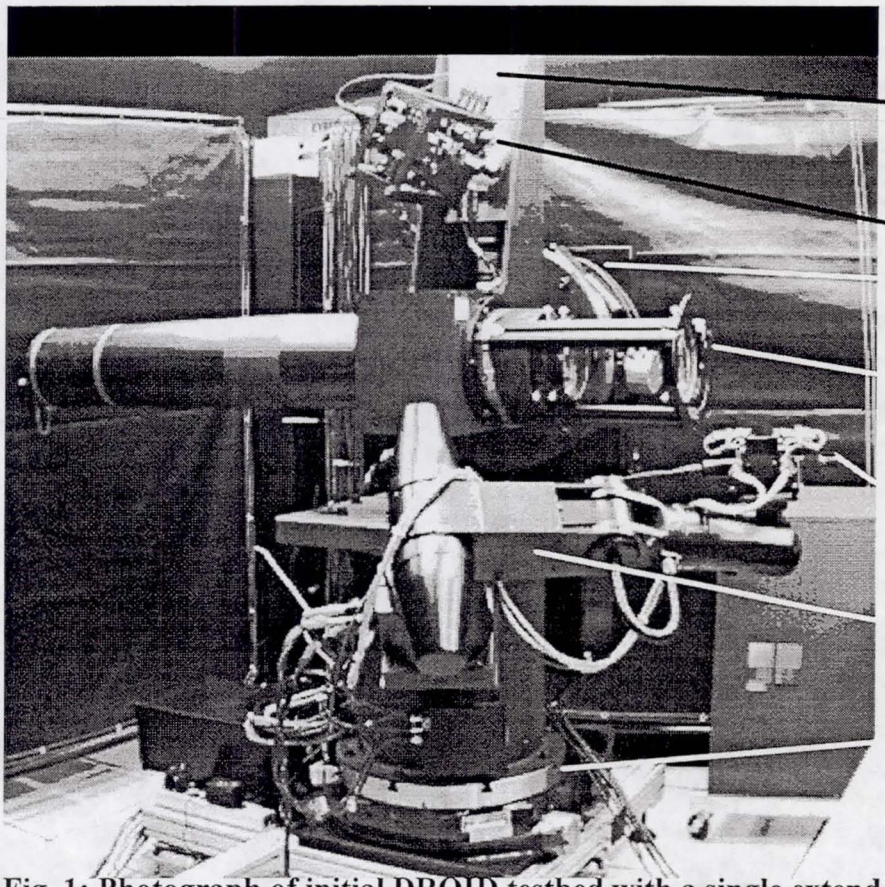


Fig. 1: Photograph of initial DROID testbed with a single extended robot arm.

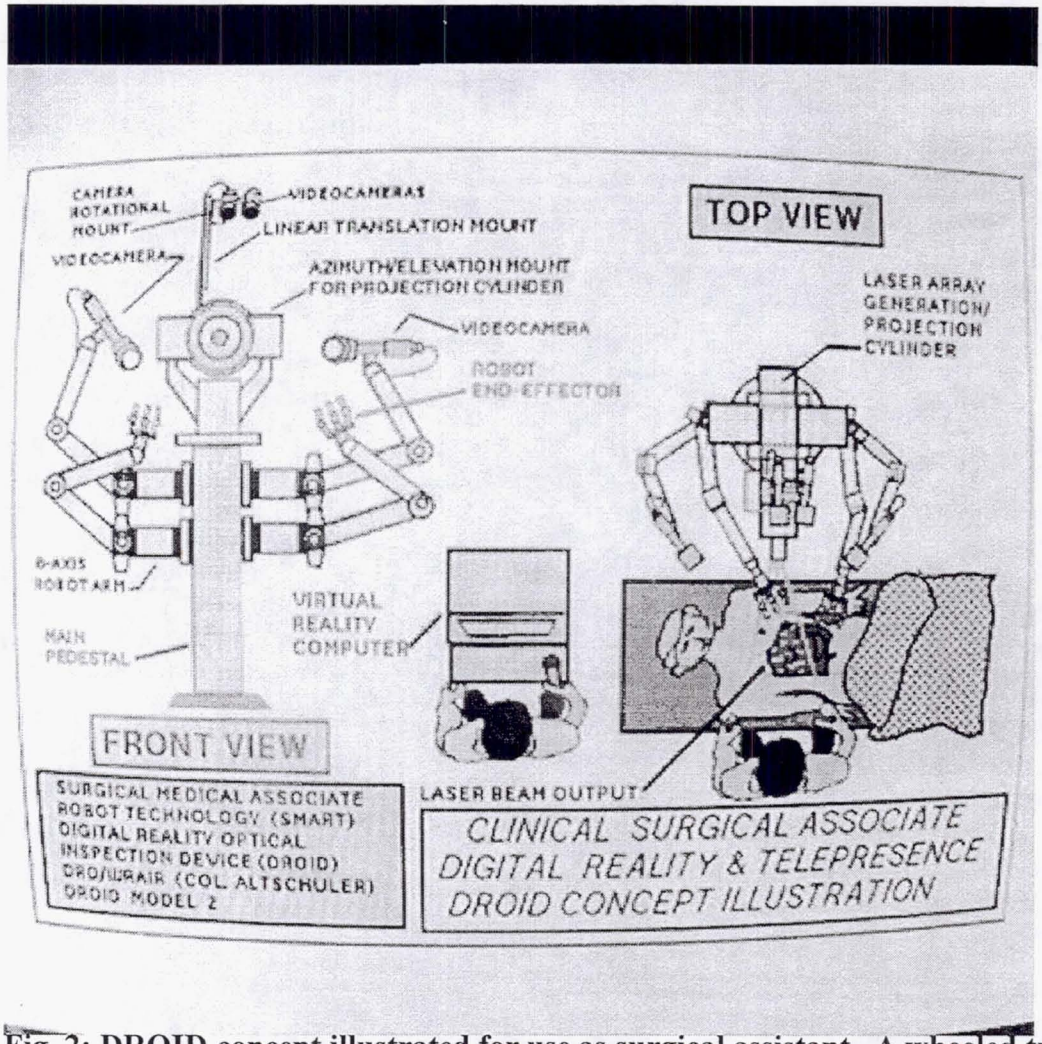
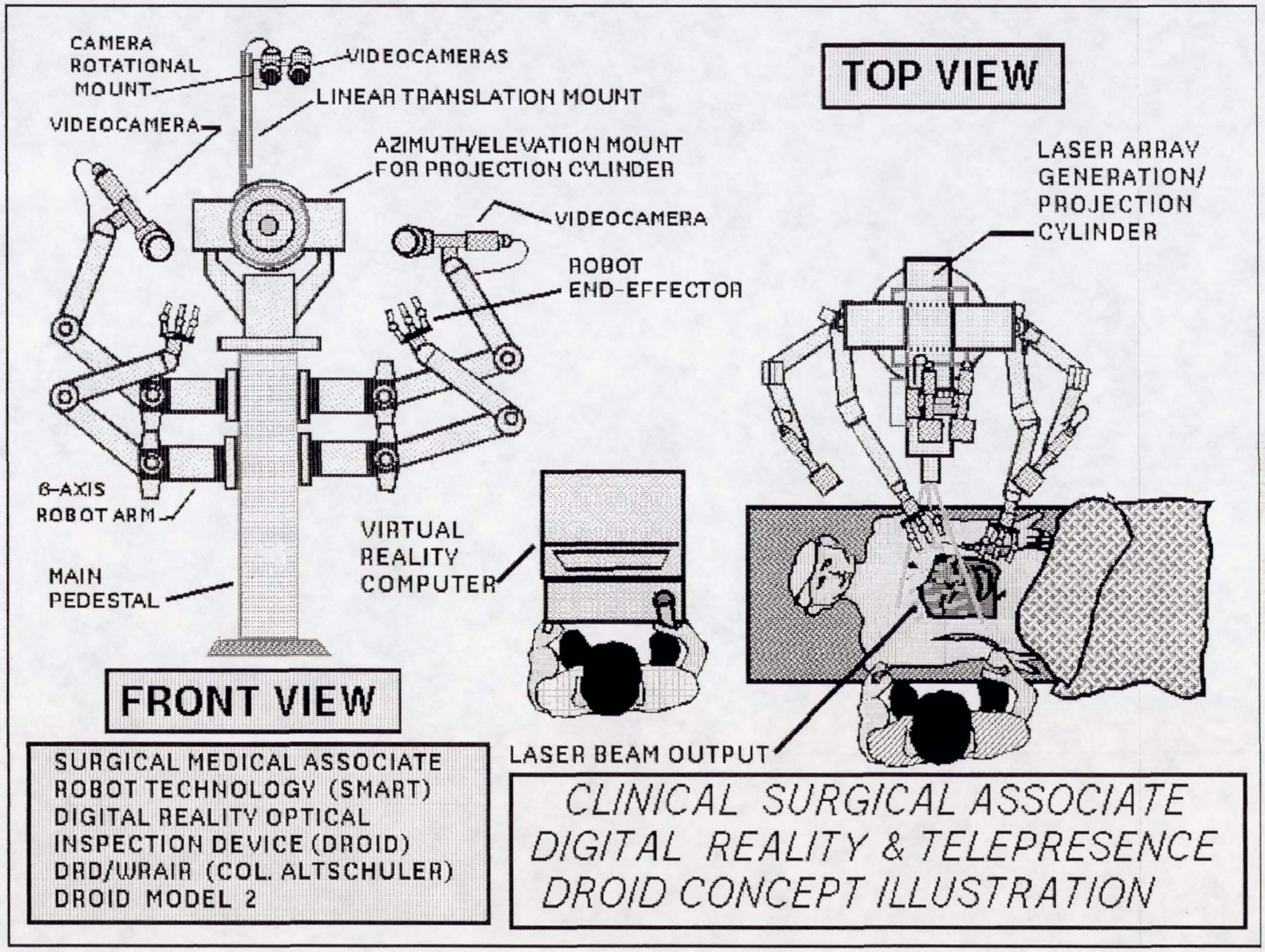


Fig. 2: DROID concept illustrated for use as surgical assistant. A wheeled tractor tread can be added as required to achieve mobility. Note consulting surgeon at virtual reality station.


CAMERA ROTATIONAL MOUNT
VIDEOCAMERAS
LINEAR TRANSLATION MOUNT
VIDEOCAMERA
AZIMUTH/ELEVATION MOUNT FOR PROJECTION CYLINDER
VIDEOCAMERA
ROBOT END-EFFECTOR
6-AXIS ROBOT ARM
MAIN PEDESTAL
VIRTUAL REALITY COMPUTER
FRONT VIEW
TOP VIEW
LASER ARRAY GENERATION/ PROJECTION CYLINDER
LASER BEAM OUTPUT
SURGICAL MEDICAL ASSOCIATE ROBOT TECHNOLOGY (SMART) DIGITAL REALITY OPTICAL INSPECTION DEVICE (DROID) DRD/WRAIR (COL. ALTSCHULER) DROID MODEL 2
CLINICAL SURGICAL ASSOCIATE DIGITAL REALITY & TELEPRESENCE DROID CONCEPT ILLUSTRATION

Medical Technology and Life Sciences

MAMMOGRAPHIC COMPUTER ASSISTED DIAGNOSIS USING COMPUTATIONAL STATISTICS PATTERN RECOGNITION

Richard A. Lorey, Ph.D.
Jeffrey L. Solka
George W. Rogers, Ph.D.
David J. Marchette
Advanced Computation Technology Group
Naval Surface Warfare Center, Dahlgren Division
Dahlgren, VA 22448

Carey E. Priebe, Ph.D.
Assistant Professor, Mathematical Science
Johns Hopkins University
Baltimore, MD 21218

ABSTRACT

Research begun for target identification utilizing pattern recognition has been applied to mammographic computer assisted diagnosis. The research has utilized the discipline of Computational Statistics. Feature extraction based on fractals and incorporating segmentation boundaries led to probability density estimation and classification based on discriminant analysis. The results of applying these techniques to mammography are very promising and are reported herein. The commercial application of this emerging technology include commercial development of medical software and the establishment and maintenance of large scale data bases. Extension of these techniques to other medical imagery is discussed.

BACKGROUND

The issue of locating and identifying potential targets has historically posed problems in war fighting scenarios. Modern warfare, with its rapid deployment, quick strike capability, and smart weapons usage has exacerbated this situation as evident in Desert Storm. Future war making capabilities with its need for faster, or even on the fly, mission planning will exceed the limits of todays technology for target identification. It is in this light that the research, sponsored by the Office of Naval Research, was undertaken at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD).

The technology necessary to identify man-made objects as distinct from natural objects is the same technology that can be used to identify any class of object. Thus, with the advent of Technology Transfer it was a natural segue for NSWCDD to apply its expertise to other areas. The Research Triangle Institute and the Federal Laboratory Consortium Demonstration Project on Critical Industry Needs opened the door for this application. Specifically, the August 1992 National Cancer Institute

problem statement called for software for Computer Assisted Diagnosis (CAD), image processing, and pattern recognition for use in digital mammography systems. It is in this light that our research has been directed toward application of this technology to mammographic CAD. Successful efforts in this endeavor, potentially a more difficult pattern recognition problem, could well result in further advances in the state-of-the-art.

COMPUTATIONAL STATISTICS PATTERN RECOGNITION

Our method for solving the pattern recognition problem involves the use of Computational Statistics [1]. This theory involves very large data sets and does not incorporate assumptions about the parametric behavior of the data. Seemingly intractable problems can sometimes yield to these techniques.

Here, we consider gray scale digital images. Each class of object in the image is characterized by a pattern or texture. We wish to analyze images and determine where changes in the pattern or texture (class) occur. Such detections enable us to distinguish targets from non-targets, man-made from natural objects, or tumors from healthy tissue.

Features

We are concerned with local texture features. We wish to categorize the features belonging to a given pattern in order to sort them into various classes. This is done by deriving features using the theory of fractal dimension [2]. The fractal dimension D (as distinguished from the normal Euclidean dimension d) can be estimated using Richardson's Power Law [2]

$$M(\varepsilon) = K\ \varepsilon^{(d-D)}, \tag{1}$$

where $M(\varepsilon)$ is the measured property of a fractal at a scale ε and K is a proportionality constant. This equation and the technique described in Solka et al. [3] allows us to extract three features that describe the texture. Thus, in a digitized image each pixel can be characterized by a three dimensional feature vector $\underline{x} = [x_1, x_2, x_3]^t$ based on a small neighborhood of the principal pixel. One feature is directly related to the fractal dimension, one is a measure of how well the fractal model fits the data, and one is related to the local degree of contrast in an image. Further, from a single image M we have available a large sample of observations $X_M = [\underline{x}_1^t, \ldots, \underline{x}_{n_m}^t]^t$. Using these features we construct probability density functions for different classes and use these for discrimination.

Probability Density Estimation

The types of problems amenable to these techniques are not those whose probability density function (pdf) can be represented by usual statistical models (e.g. normal distributions). A digitized image can easily represent a data set of up to 10^7 local observations and our work indicates this data is not well represented by a normal distribution. We estimate the pdf using a technique such as adaptive mixtures [4,5]. It is a hybrid approach which maintains the best features of the kernel estimation model [6] and the

finite mixture model [7] and does not make strict assumptions about the data distribution. The general mixture density can be given by,

$$\hat{\alpha}(x;\theta,\pi) = \int_{\Omega} \phi(x|\theta)\, dF_{\pi}(\theta) \;, \tag{2}$$

where $\hat{\alpha}(x)$ is the estimate for the true pdf $\alpha(x)$ underlying the sample X_M, ϕ is a fixed known function and F is the mixing distribution.

Segmentation Boundaries

As described by Priebe et al. [8], we can incorporate segmentation boundaries into the calculation of the fractal dimension features and hence into the pdf. Incorporation of segmentation boundaries provides for significantly more discriminatory information in the texture features and the associated pdfs. This reference further describes the two texture patches from Brodatz [9] shown in Figure 1. The three regions shown in Figure 1 (numbered 1 through 3 from the left) shows a pure texture (D17 from Brodatz) in 1 and a pure texture (D24 of Brodatz) in 3. Region 2 straddles the boundary between the two textures. Figure 2 shows the results of a pdf calculation (single feature) of the regions of Figure 1. α_1 and α_3 are the pdfs of regions 1 and 3, respectively. The two plots of α_2 (region 2) shows the effect of incorporating or not incorporating the boundary. Clearly, incorporating the boundary gives a truer picture of the pdf of the region.

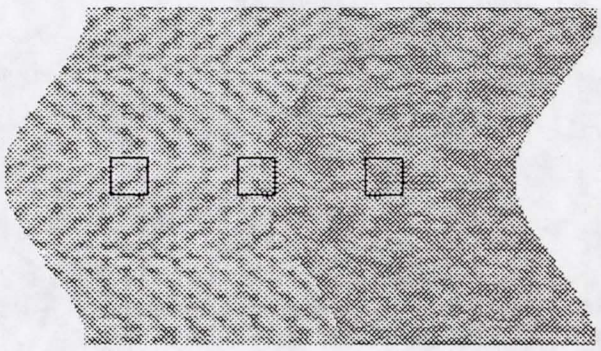


Figure 1. Two adjacent texture patches and three regions numbered 1 through 3 from the left.

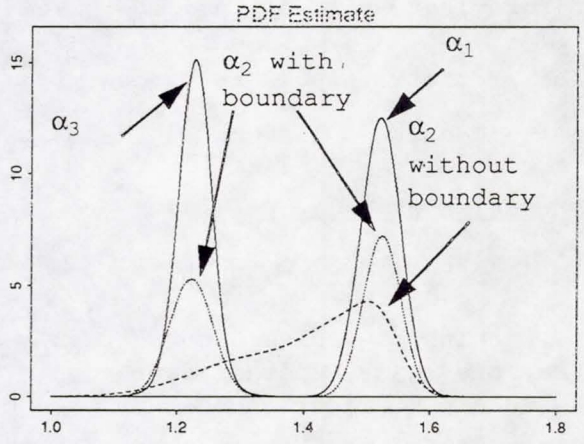


Figure 2. Single feature pdfs for the three regions from Figure 1.

Computational Complexity Reduction

For each observation the extracted fractal features are represented by $\underline{x} = [x_1, x_2, x_3]^t$. While it is true that more information is often contained in

higher dimensional feature space, the computational complexity increases dramatically with any increase in the dimension [10]. To reduce this complexity and simplify the computations, we use the Fisher Linear Discriminant (FLD) [11]. The FLD projects the three dimensions to the one dimension that is in some sense best for discrimination. The method and results have been described in Priebe et al [12]. As shown there, using all three features and the FLD yields better correlation with class than any single feature alone.

Discriminant Analysis

The probability density function characteristics are used to discriminate among the classes [13,14] by a relatively straightforward application of Bayes' rule [11]. Here we consider

$$X_M = \bigcup_{a \in A_M} X_{M_a}, \tag{3}$$

where A_M is a set of one or more classes. That is, observations from each image may be drawn from more than one class. In the simplest case, $A_M = \{1,2\}$. Hence, with estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for two classes based on observations X_{M_1} and X_{M_2} (from image M), the likelihood ratio test statistics, $LR(\zeta) = \hat{\alpha}(\zeta)/\hat{\alpha}(\zeta)$, is used to indicate the proper classification for the observation ζ drawn from another image. Generalization issues of utilizing estimates from observations from one image for discriminating classes in another image need to be addressed [15]. At a minimum, to discriminate classes in image k (classify the observations in X_M^k) a large number of training observations from images X_M^i (i = 1,..., p; i k) will need to be used to build the estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ for the two classes.

Change Point Analysis

Spatial Change Points

With the assumption that an image consists of observations from more than one class, another approach is to investigate the homogeneity of the texture. Considering whether or not the probabilistic structure of an image is uniform throughout may be construed as a spatial change point detection problem [16]. The hypothesis is that there is a region in an image whose probabilistic structure differs from the norm. The investigation of this hypothesis begins by considering small sample regions, $Y_{M_i} \subset X_M$, i = 1,...,M. These small sample regions may or may not intersect. Each small sample yields a pdf estimate $\hat{\alpha}_{M_i}$. From these, we can form a distance function

$$f\left(\hat{\alpha}_{M_i}, \hat{\alpha}_{M_j}\right) = KL\left(\hat{\alpha}_{M_i}, \hat{\alpha}_{M_j}\right) = \int \hat{\alpha}_{M_i} \bullet \log\left(\hat{\alpha}_{M_i}/\hat{\alpha}_{M_j}\right), \tag{4}$$

and the statistic

$$T = \sup_{i,j} \int\left(\hat{\alpha}_{M_i}, \hat{\alpha}_{M_j}\right) . \tag{5}$$

The integral is the Kullback-Liebler (KL) information between the two distributions and can be used to indicate nonhomogeneity [15]. This is done by estimating the probability density of the KL statistic and using T to distinguish between the homogeneous or nonhomogeneous class. T greater than some τ indicates nonhomogeneity and estimating distribution of the T statistic allows a computation of an empirical p-value. This procedure fits into the spatial change point detection framework when each Y_{M_i} is considered to be a spatially connected region. An appropriate value of τ is determined through training, that is, we wish to determine the relationship between T values and the likelihood that an observation deviation indicates a nonhomogeneity.

Spatio-temporal Change Points

This technique is also useful for detecting changes over time. We can consider images of the same scene or object produced at different times. The characteristics of the regions of the images are modeled by pdfs. A nonhomogeneity in a like region of sequential images indicates a spatio-temporal change point.

Proposed CAD System

Figure 3 shows a proposed system [17] incorporating the items discussed above. This flowchart represents a very high level schematic.

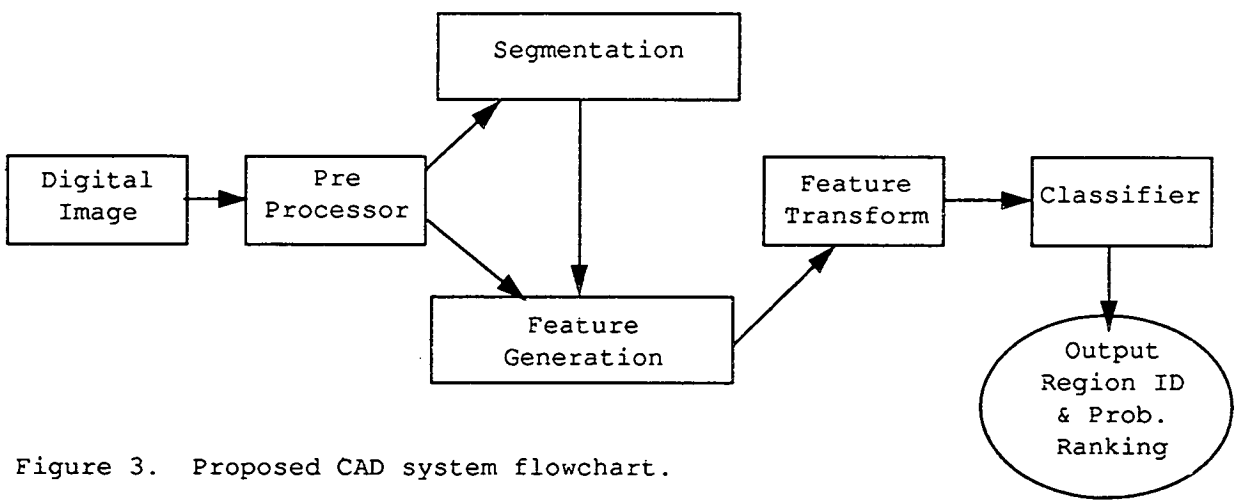


Figure 3. Proposed CAD system flowchart.

EXPERIMENTAL RESULTS

Mammographic PDFs

We conducted this study using images provided by the H. Lee Moffitt Cancer Center and Research Institute and the Department of Radiology of the University of South Florida [12]. All tumorous regions were biopsy proven. The mammograms were digitized at ~220 microns/pixel and 8 bit/pixel. Figure 4 shows regions of healthy and tumorous (~10 mm malignant stellate mass) tissue from a mammogram A. 10,000 healthy tissue observations and 500 tumorous tissue observations were used for training data. A mammogram B (not pictured) containing a ~6 mm malignant stellate mass was used for testing (10,000 healthy tissue observations and 300 tumorous tissue observations).

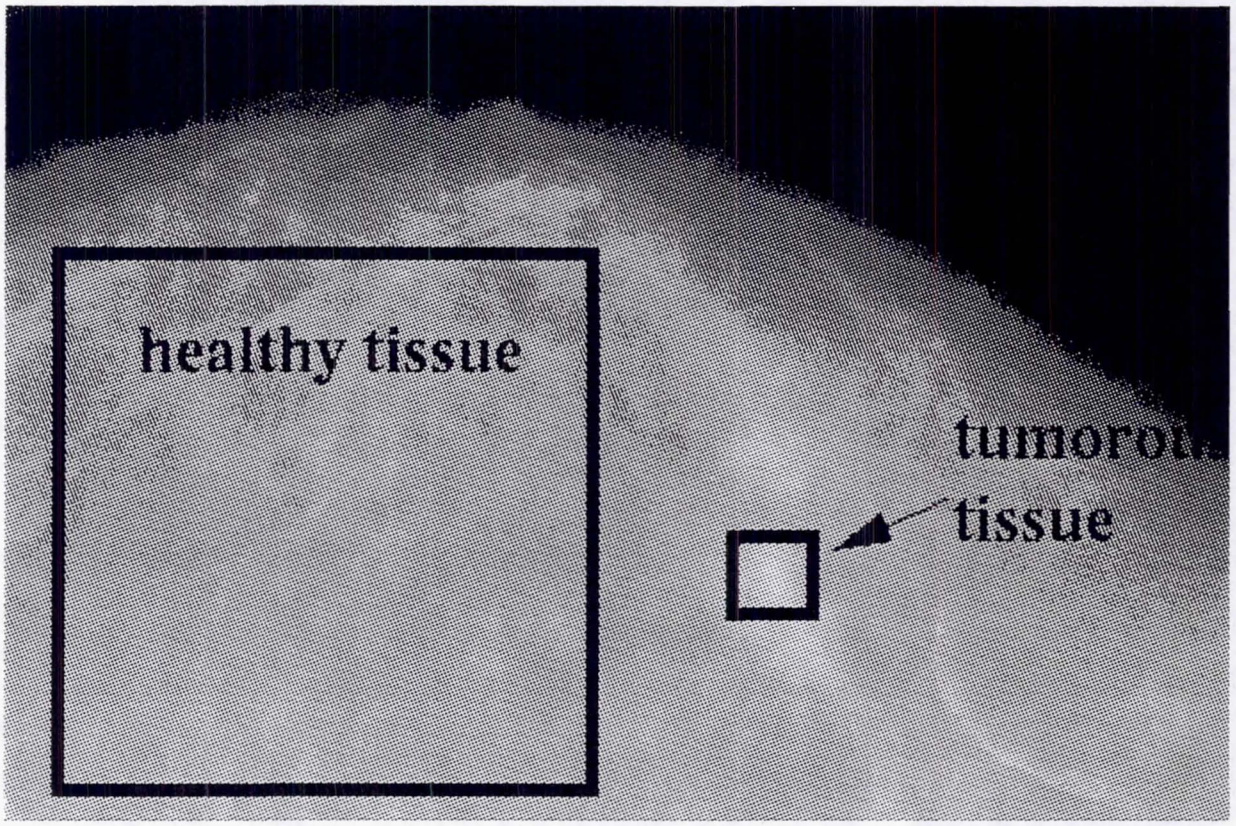


Figure 4. Regions of interest in mammogram A. This image has been enhanced for presentation

Figure 5 is a plot of the pdfs of the projected data showing the separation of the healthy and tumorous classes for mammogram A. The FLD and transformation from A is applied to B and the results are shown in Figure 6. The discriminant boundary is clearly evident and appears to be invariant. When the roles of A and B are reversed, the plots exhibit the same behavior but with a different discriminant boundary. Based on this limited study, the results indicate the possibility that once a projection is chosen the discriminant boundary is invariant from training to testing data. Thus a discriminant boundary obtained from training images can be successfully applied to new test images.

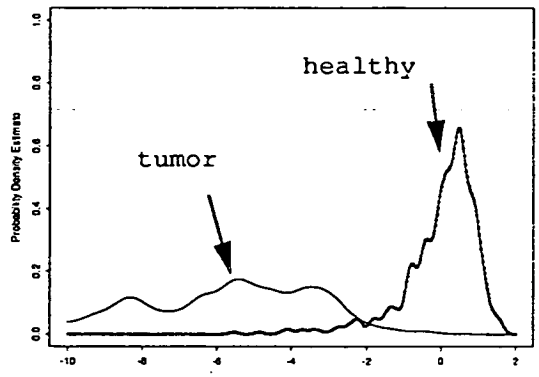


Figure 5. Fisher Linear discriminant pdfs for mammogram A.

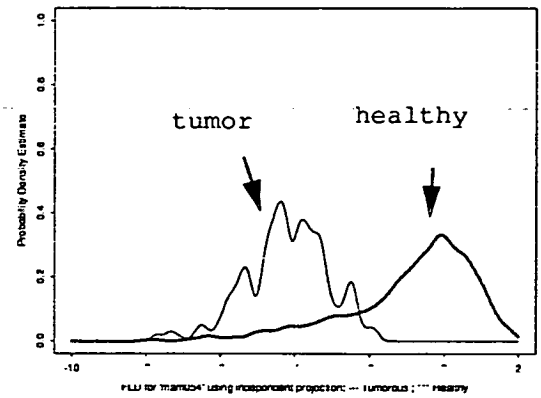


Figure 6. Fisher Linear discriminant pdfs for mammogram B using the independent projection.

Wolfe's Patterns

Wolfe distinguished four tissue patterns (labeled as N1, P1, P2, DY) corresponding to increasing breast tissue density and different morphology [18]. To determine the applicability of this technique to the discrimination of Wolfe patterns, we analyzed an additional eight mammograms from the set provided above. We used two patterns for training data and two others for testing data. Figures 7, 8, and 9 show the pdfs of the patterns indicated. The combination shown were chosen simply for illustrative purposes. In all cases, the ability to discriminate exists and the discriminant boundaries generalize from training to testing data. If these results can be extended to nonmalignant abnormal tissue the technique might be useful in distinguishing these types.

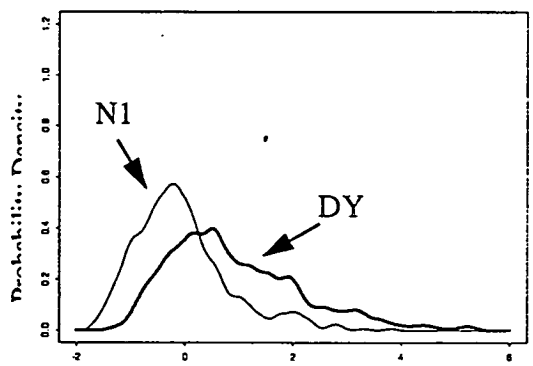


Figure 7. FLD pdfs for mammogram N1 vs. mammogram DY.

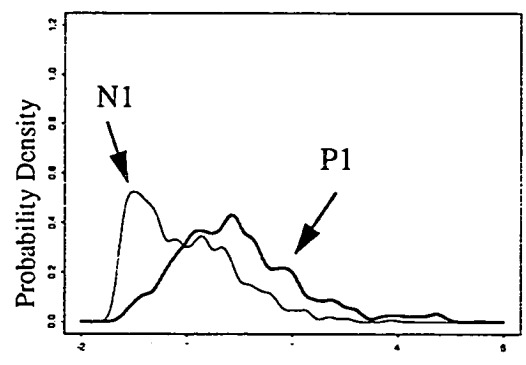


Figure 8. FLD pdfs for mammogram N1 vs. mammogram P1..

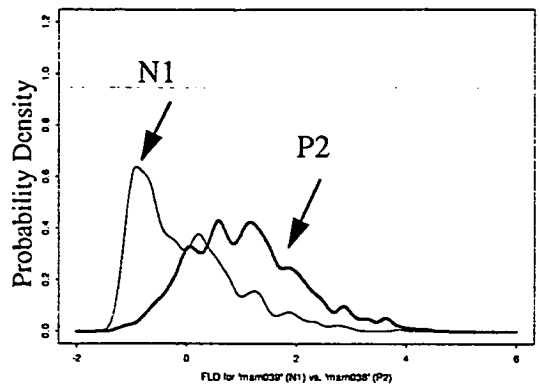


Figure 9. FLD pdfs for mammogram N1 vs. mammogram P2.

Mammograms and Change Point Analysis

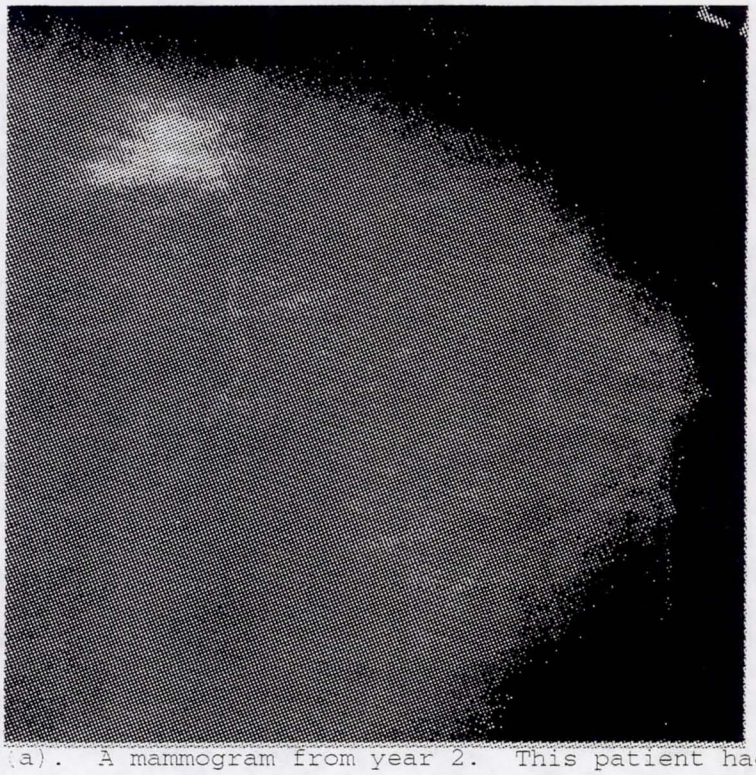
The results to be discussed next involve six patients followed for three years in which a biopsy proven anomaly was detected in the third year in three of the patients. We used at least two views of each breast for each patient for each year for a total of 81 images. The images were digitized at 600 dpi (~42.3 microns) and 8 bit grayscale. The images were provided by Kaiser-Permanente Research, Portland, Oregon.

We show pictures from only one patient. As will be discussed, we were able to detect an anomaly in the second year. We were not able to do this for the other two cases. However, it may be possible that this technique can result in earlier detection in some cases. We did not detect any false positives in the other three cases.

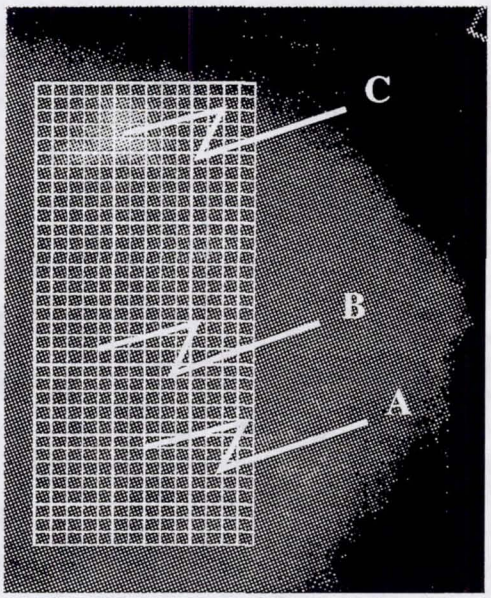
Figure 10 shows an image in (a), a grid in (b) showing the subregions, and in (c) the KL surface, KL ($\hat{\alpha}_{ref}, \hat{\alpha}_{i,j}$), for a reference healthy tissue tile against the other tiles i,j. As mentioned, this image is from the second year. The KL surface appears to be quite homogeneous except at the top where the tumor was detected in the third year. The KL values are significantly greater here, indicating a region of anomalous tissue.

The pdfs from tiles in the healthy region exhibit similar pdfs while those in the anomalous region have a shifted mean and a rather heavier tail.

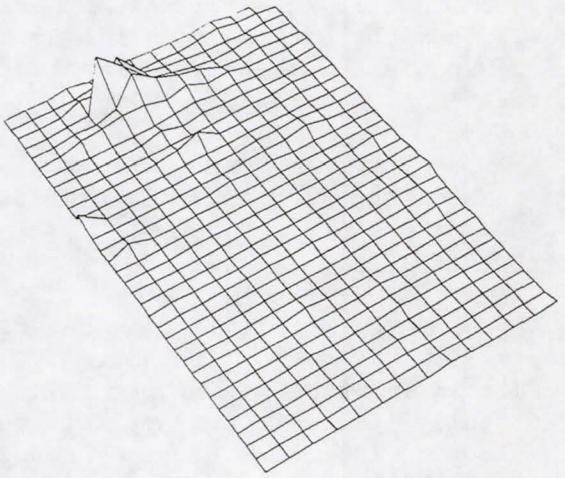
If histograms of the KL values are constructed for this patient over the three years, an estimate of a τ value can be made. Using the first year as a baseline healthy set, a τ = 2.83 (maximum KL value) is obtained. For the second year, four detections are obtained, that is, T exceeded τ for four tiles (T_{max} > 4.5) For year three the number of detections was much greater (T_{max} = 13.67) which clearly shows the nonhomogeneity detected.



(a). A mammogram from year 2. This patient had a tumor detected in the third year of the study.



(b). The mammogram from (a) with the grid overlayed. Observations are drawn from each grid tile. Tile A is healthy tissue used as the reference tile. Tile B is another healthy tile. Tile C is in the anomalous region.



(c). Kullback-Leibler surface for the grid shown in (b). The region of large KL values corresponds to the area in which the tumor was detected the following year.

Figure 10. Mammographic change point analysis.

COMMERCIAL APPLICATIONS

Software

The algorithms discussed above exist as research and engineering tools. They have been written in the C programming language and executed on a series of networked Silicon Graphics machines. The emerging technology represented by the work described here is an opportunity for development of medical quality software. Although extensive further testing needs to be done, the development of commercial grade software would enable this to be done at a more rapid pace. The goal of the researcher has been to implement ideas and produce software that is functional for research. Attempts were not made to optimize the software operation. Thus it is ripe for commercial development.

This software, used in conjunction with digitized film mammograms or digitally produced images would be a boon to the radiologist. While never intended as a replacement for the radiologist's expertise, it could be used to screen many of the normal mammograms the radiologist usually reads. It could also draw attention to questionable cases or to questionable regions of a given mammogram [19]. Obviously, the software could be made to specifications most desired by the practicing radiologist. Sales to hospitals, radiological centers, and individual physicians represents a viable market. Additionally, upgrades are inevitable.

Data Bases

In conjunction with developing commercial grade software is the necessity of producing large scale mammographic data bases. Actually, it is probably more commercially viable to retain the data bases and market the results. That is, the classifiers associated with all the various tissue types. It might even be possible to have the software collect data from each new patient screened, add that data to the data base, and market updates to the classifiers.

Some scheme to update and upgrade these data bases and classifiers would seem to keep this commercially viable for years to come.

Other Medical Applications

We have done a limited amount of work applying this technology to other medical imagery. Specifically, we have attempted to detect plaque deposits in arteries imaged by a MRI. Our limited results show great promise. Additionally, prior to our application to mammography we applied this technology to images as diverse as aerial photographs and pictures of barnacles on ship hulls. We have been successful in all these attempts. Hence, there is no reason to believe that the application to other forms of medical imagery would not be equally successful. From a commercial viewpoint, this is a opportunity to produce and market software with wide ranging applications.

REFERENCES

1. Wegman, E. J., "Computational Statistics: A New Agenda for Statistical Theory and Practice", J. Wash. Acad. Sci., 78, 310-322 (1988)
2. Mandelbrot B., "The Fractal Geometry of Nature", New York: W. H. Freeman & Co., (1977)
3. Solka, J. L., Priebe, E. E., and Rogers, G. W., "An Initial Assessment of Discriminant Surface Complexity for Power Law Feature", Simulation, 58, 311-318 (1992)
4. Priebe, C. E. and Marchette, D. J., "Adaptive Mixture: Recursive Nonparametric Pattern Recognition", Pattern Recognition, 24, 1197-1209 (1991)
5. Priebe, C. E and Marchette, D. J., "Adaptive Mixture Density Estimation", Pattern Recognition, 26, 771-785 (1993)
6. Silverman, B. W., "Density Estimation", Chapman and Hall, New York, NY, (1986)
7. Titterington, D. M., Smith A. F. M. and Makov, U. E., "Statistical Analysis of Finite Mixture Distributions", John Wiley and Sons, New York, NY, (1985)
8. Priebe, C. E., Julin, E. G., Rogers, G. W., Healy, D. M., Lu, J., Solka, J. L., Marchette, D. J., "Incorporating Segmentation Boundaries into the Calculation of Fractal Dimension Features", Proceedings of the 26th Symposium on the Interface, Research Triangle Park, NC, (1994)
9. Brodatz, P., "Texture: A Photographic Album for Artists and Designers", Dover. New York, (1966)
10. Scott, D. W., "Multivariate Density Estimation", John Wiley and Sons, New York, NY (1992)
11. Duda, R. O., and Hart, P. E., "Pattern Classification and Scene Analysis", John Wiley and Sons, New York, NY (1973)
12. Priebe, C. E., Solka, J. L., Lorey, R. A., Rogers, G. W., Poston, W. L., Kallergi, M., Qian, W., Clarke, L. P., Clark, R. A., "The Application of Fractal Analysis to Mammographic Tissue Classification", Cancer Letters 77, 183-189, (1994)
13. Priebe, C. E., Solka, J. L., Rogers, G. W., "Discriminant Analysis in Aerial Images Using Fractal Based Features", Proc. SPIE, Vol. 1962, pp. 196-208, (1993)
14. McLachlan, G. J., "Discriminant Analysis and Statistical Pattern Recognition", John Wiley and Sons, New York, NY, (1992)
15. Priebe, C. E., Lorey, R. A., Marchette, D. J., Solka, J. L., Rogers, G. W., "Nonparametric Spatio-Temporal Change Point Analysis for Early Detection in Mammography", Proceedings of the 2nd Int'l Workshop on Digital Mammography (SIWDM), York, UK, (1994)
16. Ripley, B. D., "Statistical Inference for Spatial Processes", Cambridge University Press, Cambridge, (1988)
17. Rogers, G. W., Priebe, C. E., Solka, J. L., Lorey, R. A., Julin, E. G., "A System and Method for Incorporating Segmentation Boundaries into the Calculation of Fractal Dimension for Texture Discrimination", Patent Application, Navy Case No. 75,998 (1994)
18. Wolfe, J. N., "Breast Patterns as an Index of Risk for Developing Breast Cancer", Am. J. Radiol., 126, 1130-1139, (1976).
19. Chan,H.P.,Doi,K.,Vyborny,C.J.,Schmidt,R.A.,Metz,C.E.,Lam,K.L.,Ogura,T., Wu,Y.Z.,Macmahon,H., "Improvement in radiologist's detection of clustered microcalcifications on mammograms.The potential of computer-diagnosis", Investigative Radiology, 25(10), pp. 1102-1110. (1990).

Technology 2004, November 8-10, 1994, Washington, D.C.

A NEW ROBOT FOR HIGH DEXTERITY MICROSURGERY

Paul S. Schenker, Hari Das, and Timothy R. Ohm

Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive / MS 198-219
Pasadena, CA 91109

Email: schenker@telerobotics.jpl.nasa.gov

ABSTRACT

Drawing on our prior NASA work in high-fidelity teleoperation/telepresence, we are developing a new robotic system applicable to micro- and minimally invasive surgeries. The goal product is a dexterity-enhancing master-slave telerobot and controls that will refine the scale of current microsurgeries, and minimize effects of the involuntary tremor and jerk in surgeons' hands. As a result, exciting new surgeries of the eye, ear, brain and other critical faculties should become possible, and the positive outcome rates in conventional procedures will improve. In its nominal configuration, this new *Robot Assisted MicroSurgery* (RAMS) system has a surgeon's hand controller immediately adjacent to the robot. The RAMS system is also potentially applicable to "telesurgery" -- surgeries to be carried out in local-remote settings and time-delayed operating theaters -- as considered important in field emergencies and displaced expertise scenarios. As of August, 1994, we have developed and demonstrated a new 6 degree-of-freedom robot (slave) for the RAMS system. The robot and its associated Cartesian controls enable relative positioning of surgical tools to approximately 25 microns within a non-indexed and singularity-free work volume of ~20 cubic centimeters. This implies the capability to down-scale hand motion inputs by two to three times, and the consequent performance of delicate procedures in such areas as vitreo-retinal surgery, for which clinical trials of this robot are planned in 1996. Further, by virtue of an innovative drive actuation, the robot can sustain full extent loads up to three pounds, making it applicable to both fine manipulation of microsurgical tools and also the dexterous handling of larger powered devices of minimally invasive surgery. In this paper, we overview the robot mechanical design and controls implementation, and we summarize our preliminary experimentation with same. Our accompanying oral presentation includes a five minute videotape display of some engineering laboratory results achieved to date.

INTRODUCTION

Medical applications of robotics are beginning to attract significant interest of both researchers and commerce. Several different application thrusts are being aggressively explored, as reported in recent special interest meetings and workshops [1 - 4]. These thrusts include robot-assisted stereotaxic interventions (imaging-guided biopsy), orthopedic preparations by robot (precision joint emplacements), endoscopic & laparoscopic assists (minimally invasive procedures),

teleoperative remote surgeries ("telesurgery"), and recently, robotically-enhanced microsurgery (high dexterity, scaled operations under microscopic viewing). Our primary interest is the last area, its connections to precision imaging guided surgical interventions -- and possibly later, telesurgeries. Building on our prior NASA-JPL work in dexterous teleoperations and telerobotics at conventional scale, we have begun development of a robotic micro-dexterity platform with potentially important new medical applications. This *Robot Assisted MicroSurgery* (RAMS) workstation targets new and improved procedures of the eye, ear, brain, nose, throat, face, and hand. The resulting technology developments are planned for evaluation in clinical microsurgery procedures circa 1996 -- we are working to this end in engineering collaboration with MicroDexterity Systems, Inc. (Chief Officer, Steven T. Charles, M.D.), under a NASA Technology Cooperation Agreement, with the goal that successful technology developments be commercialized.

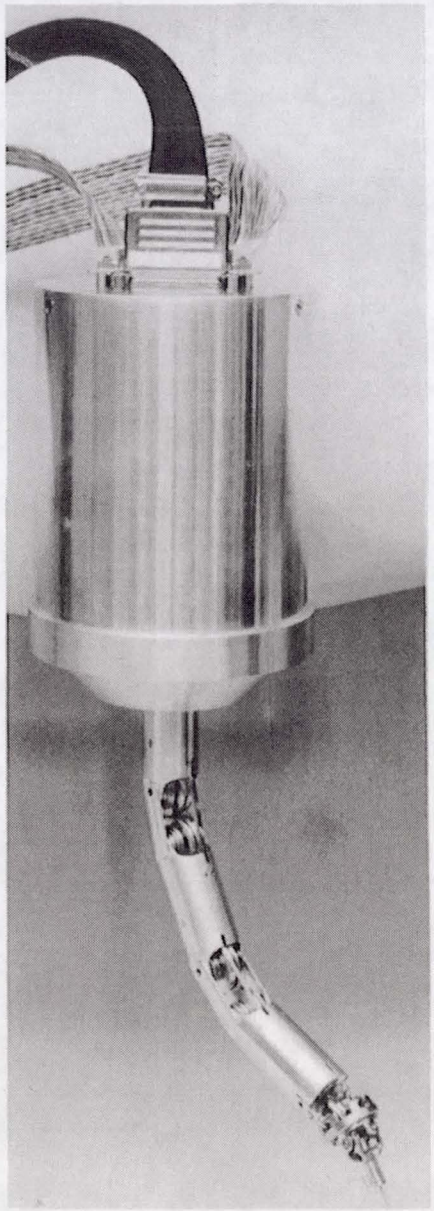
The RAMS workstation is conceived as a dual-arm 6-d.o.f. master-slave telemanipulator with programmable controls, one arm handling primary surgical tooling and the other as auxiliary (suction, cauterization, imaging, etc.). The primary control mode is to be teleoperation, including task-frame referenced manual force feedback and possibly a cross-modal textural presentation. Later sensor-related developments include *in situ* imaging modes for tissue feature visualization and discrimination. The operator will also be able to interactively designate or "share" automated control of robot trajectories, as appropriate. RAMS is intended to refine the physical scale of manual microsurgical procedures, while also enabling more positive outcomes for the average surgeon during typical procedures -- e.g., the RAMS workstation controls include features to enhance the surgeon's manual positioning and tracking in the face of the 5-10 Hz band myoclonic jerk and involuntary tremor that limit many surgeons' fine-motion skills. The first RAMS development, now completed and undergoing engineering evaluation, is a small six-d.o.f. surgical robot ("slave"), the configuration of which is a torso-shoulder-elbow (t/s/e) body with non-intersecting 3-axis wrist. This robot manipulator is approximately 25 cm. full-extent and 2.5 cm. in diameter. Robot actuation is based on a new revolute joint and cable-drive mechanism that achieves near zero backlash, constant cable length excursions, and minimized joint coupling. The robot design and controls currently allow non-indexed relative positioning of tools to within 25 microns and a work volume of ~ 20 cm^3 -- a resolution some two or three times better than that typically observed in the most highly skilled microsurgeries.

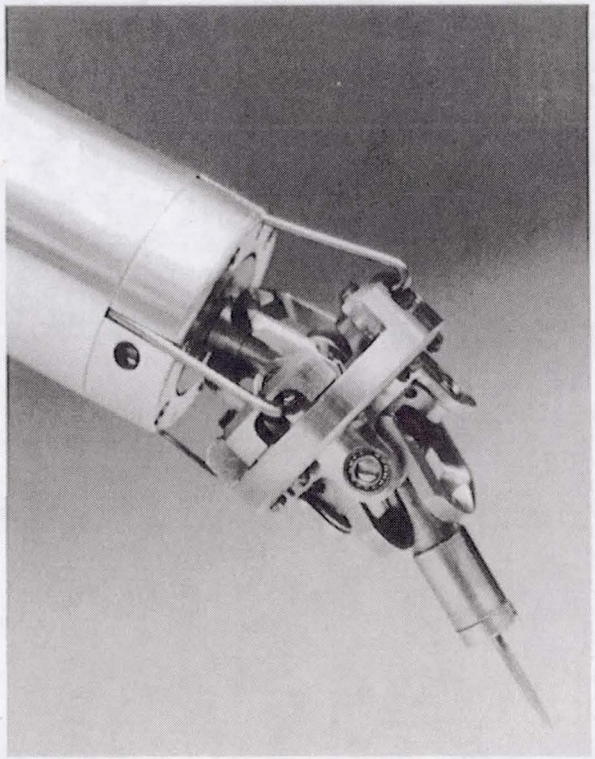
ROBOT MECHANICAL DESIGN

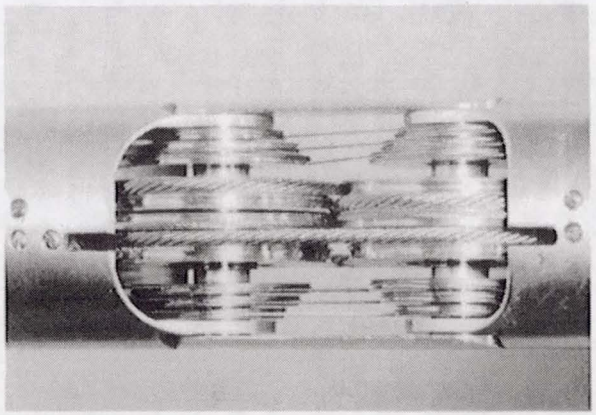
We describe the robot design in this section, briefly outlining related design requirements, as motivated both by kinematic control objectives and robot suitability to re-usable and safe application in a sterile medical environment. **Figure 1** highlights some recent robot mechanical developments, e.g., the integrated six-d.o.f. robot slave (manipulator and motor-drive base), a 3-d.o.f. wrist (close-up view), and the highly novel double-jointed tendon drive rotary joint mechanization used in shoulder-and-elbow actuation. Figure 1 also lists at page-right some key robot features, as further elaborated below. The general model for the presentation that follows is: we list a design objective (*in italics*) and its definition; we then provide a brief technical description of the technical approach we took to meet the objective. Where appropriate, and known to date, we give quantitative information.

JPL

Figure 1 : Robot Assisted MicroSurgery (RAMS) system six-d.o.f. robot slave (manipulator and motor-drive base), 3-d.o.f. wrist, and double-jointed tendon drive rotary joint







Key Features

- 6-d.o.f. serial arm (t/s/e + 3-axis wrist)
- Torso can roll 165 degrees
- Shoulder/elbow rotate a full 360°
- Singularity-free wrist design
- Wrist pitch/yaw=180°, roll=540°
- Arm-wrist: L=25 cm, OD=2.5 cm
- Base: 12 cm OD, 17.75 cm long
- Weight (incl. base/motors): 5.5 lb.
- .25 cm center pass-through
- Quick-disconnect drive (sterilization)
- Cable-driven, decoupled joints
- Zero-backlash in five joints
- Low-stiction, custom bearing drive
- Stiffness ~15 lb./in at tip
- Full-extent arm force limit ~ 3 lb.
- Designed for 10μ positioning
- Commercially-vended electronics
- PLD based-control, power/braking
- Optically-isolated control interface
- Watchdog timer on processors

PSS, 8/28/94

Mechanical Design

1. *Drive Unit Separability*: Autoclaving of the robot is possible by removing the motor/encoder units at the base prior to sterilization. The motor/encoder units can be re-attached in a quick and simple procedure.

This is done by integrating the motors/encoders into two distinct sets of three on a common mount and registering these packages via alignment pins. The resulting two motor packages can be easily removed by undoing two screws and one connector on each set. The mechanism can then be autoclaved. The two motor packages can be reinstalled quickly by reversing the removal procedure. In normal operation the motors are contained inside the robot's base, protecting anything they may contaminate. An added advantage obtained with this design is that debugging of servo- and kinematics control systems can be done while the motors are not attached to the robot, thereby sparing the robot damage during software development and validation.

2. *Zero/Low Backlash*: Low backlash (free play) is essential for doing fine manipulation, especially since the position sensors are on the motor shafts.

Five of the robot's six degrees of freedom have zero backlash and the sixth has about 20 microns. Zero backlash is achieved by using dual drive-trains that are pre-loaded relative to one another. These dual drive-trains are coupled together at only the motor shaft and the joint output. The steel cables which actuate each joint also act as springs to pre-load the gear-train. The drive-train's pre-load can be easily adjusted by disengaging the motor, counter-rotating the dual drives until the desired pre-load is reached, and re-engaging the motor. This also allows for easy cable adjustment as the cables stretch with time. The one axis that does not have zero backlash is a result of the wrist design which makes low backlash possible but zero backlash difficult, especially if stiction is a concern as with this robot.

3. *Low Stiction*: Stiction (stick/slip characteristic) must be minimized to achieve small incremental movements without overshooting or instability.

Stiction was minimized by incorporating precision ball bearings in every rotating location of the robot (pulleys, shafts, joint axes, etc.), so as to eliminate metal-to-metal sliding. Due to severe size and loading constraints, some of these bearings had to be custom designed. (Indeed, there is only one location in the wrist where such direct contact exists, because size constraints therein restricted use of bearings. In this location, backlash was allowed to reduce stiction -- see item 2.)

4. *Decoupled Joints*: Having all joints mechanically decoupled simplifies kinematics computations as well as provides for partial functionality should one joint fail.

Developing a six axis, tendon-driven robot that has all joints mechanically decoupled is very difficult. Decoupling requires driving any given joint without affecting any other joint. The shoulder and elbow joints incorporate a unique double-jointed scheme that allows passage of any number of activation cables completely decoupled from these joints. The three axis wrist is based on a concept (as originated by Mark Rosheim) that not only decouples the joints, but also has no singularities. Further, he torso simply rotates the entire robot base to eliminate coupling. *If any one of the joints were to fail mechanically, the remaining five would be unaffected.*

5. *Large Work Envelope*: A large work volume is desirable so that the arm's base will not have to be repositioned frequently during tasks.

To achieve a large work envelope, each joint needs to have a large range of motion. The torso was designed with 165 degrees of motion while both the shoulder and elbow have a full 360 degrees. This high range of motion in the shoulder and elbow is attained by the unique double-jointed scheme mentioned above. The wrist design (utilizing the Rosheim concept) has 180 degrees of pitch and yaw with 540 degrees of roll. Such large motion ranges greatly reduce the chance of a joint reaching a limit during operation, thus increasing the work volume.

6. *High Stiffness*: A stiff manipulator is necessary for accurate positioning under gravitational or environmental loads, especially when position sensing is at the motor drives.

When a robot changes its orientation relative to gravity, it will deflect due to its own weight. Likewise, if a force acts on the arm, it will also deflect. Furthermore, if position sensing is done at the motor drive, this deflection will not be known. Therefore, such deflections must be minimized by increasing stiffness. The stiffness of RAMS arm is about 15 lbs/inch at the tip. This high stiffness is achieved by using high spur gear reductions off the motors, combined with large diameter, short path length stainless steel cables to actuate each joint. The pitch and yaw axes also include an additional 2:1 cable reduction inside the forearm (near the joint) for added stiffness.

7. *Compact/Lightweight*: In some applications, a restricted work-space warrants a small serial manipulator to minimize both geometric and visual interference.

The physical size of the arm is about one inch in diameter and about 25 cm long. The robot base, containing the motor drives and electrical interfaces, has a 12 cm diameter and is 17.75 cm long. The entire unit (arm and base) weighs about 5.5 lbs. All electrical cables connect to the bottom of the base so as to not protrude into the robot's work-space.

8. *Fine Incremental Motions*: Human dexterity limitations constrain surgical procedures to feature sizes of about 50-to-100 microns. This arm is designed to achieve 10 microns relative positioning.

By combining many of the features mentioned above (low backlash, low stiction, high stiffness, etc.), this arm is designed to make very small incremental movements. This means that the manipulator can make incremental steps of 10 microns. Note conversely this does not necessarily mean that the arm is repeatable to within 10 microns absolute position accuracy.

9. *Tool Wiring Provisions*: Some tools require electrical or pneumatic power which can be routed through the arm in some cases.

The arm is designed to allow running a limited amount of wiring or hoses from the base to the arm's tip (where the tool is mounted). This passageway is about .35 cm in diameter through the wrist and exits through the center of the tooling plate. The wiring can be passed out the base of the robot so that it does not interfere with its work-space -- as would be the case if such wiring was routed externally.

Electronics and Servo-Control System

1. *Configuration management and control electronics*: It is necessary to sense, monitor, and control basic failure conditions (e.g., to implement corrective motion control/braking actions)

A Programmable Logic Device (PLD) controls power and braking relays through an optically isolated interface, and allows fault detection and error recovery. Features of this system are:

- power up and down button
- manual start-stop buttons to switch motor power from a brake mode to control mode
- panic button to stop motors
- amplifier fuse fault detection
- brake relay fault detection
- watchdog timer fault detection to insure control processors are functioning
- amplifier power supply fault detection
- amplifier fuse fault detection
- PLD logic fault detection.

2. *Commercially available components*: All components of the servo-control system are commercial items available off-the-shelf from vendors with reliable product support.

Use of commercial products and industry standard interfaces in both the servo-control system and computing architecture was critical to rapid prototyping *(The development of the robot electromechanical design, electronics, control & computing, followed by a first integration-test-debug-demonstration was done in less than nine months)*. In particular, rapid vendor product support allowed us to quickly overcome hardware failures and reduce software development cycles by comparison to our past experiences in custom robot computing design & implementation.

RAMS ROBOT COMPUTING AND CONTROLS

System Organization

Figure 2, next page, sketches the RAMS computing and control hardware organization. The major hardware implementation features are as follows: the graphics user interface (GUI) software resides on a UNIX workstation, which also serves as host to a VxWorks real-time control environment. The VxWorks-based control functions are in turn implemented on a MC 68040 board installed in a VME chassis. A Delta Tau Data Systems PMAC board, also on the VME chassis, controls the six axes of the robot by directly reading the robot sensor outputs and driving the motors through amplifiers.

Graphics User Interface

The GUI is based on the X Windows and OSF/Motif libraries. We have developed a number of demonstration modes within the GUI to show and evaluate the capabilities of the robot. These modes are:

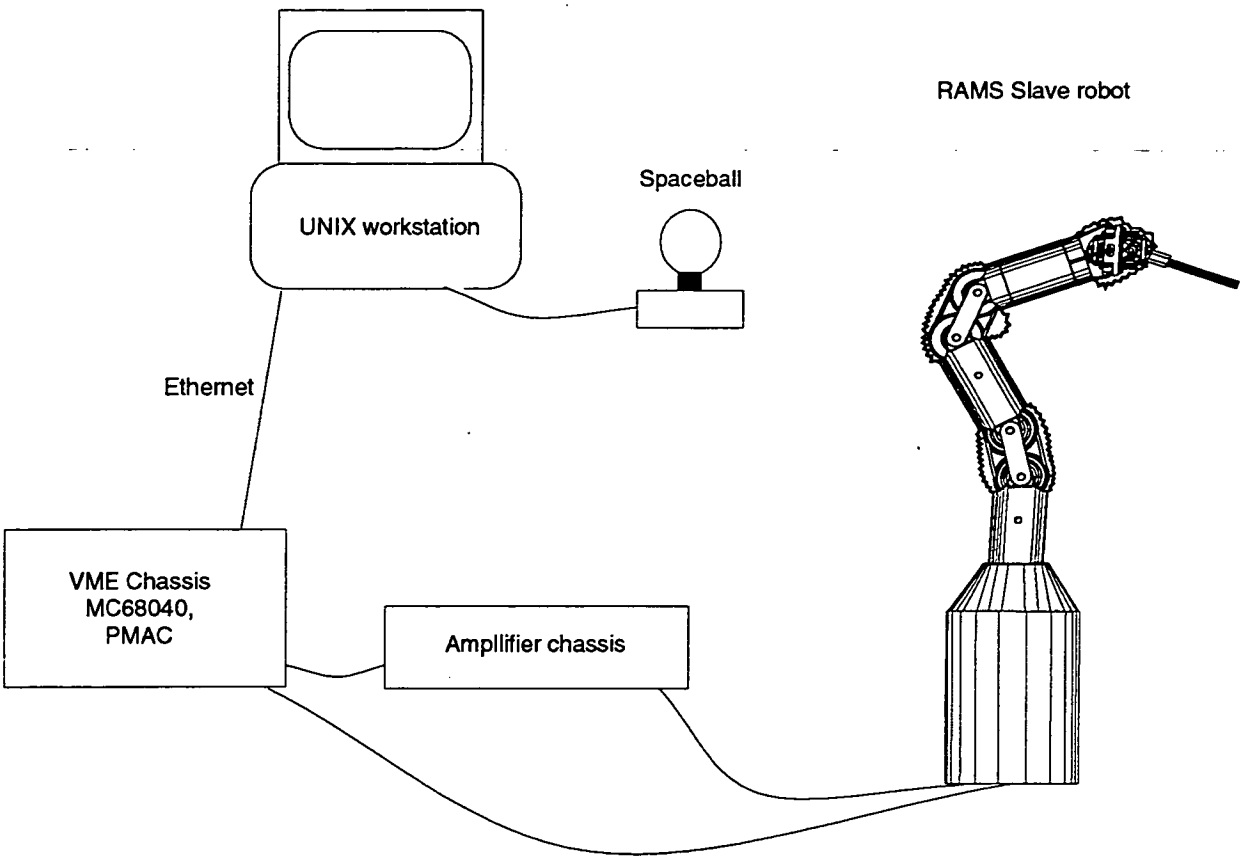


Figure 2: RAMS system hardware organization

(cont'd.)

- a *manual joint control mode* wherein the user moves individual joints manually by selecting buttons in a control window, incrementing and decrementing a desired joint position
- an *autonomous joint control mode* demonstrating the workspace of the robot. In this mode, the robot simultaneously moves each of its joints in a sinusoidal motion between set limits
- a *manual teleoperated mode* in which the robot is controlled either by using a mouse (or by selecting buttons on a display), incrementing or decrementing motion along single axes of a world-referenced coordinate frame, or by using the spaceball input device to simultaneously move all six axes of the robot
- an *autonomous world space control mode* in which the robot moves its end effector in a sinusoidal motion about one or more Cartesian-defined axes simultaneously.

Kinematic and Joint Control

The control software of the robot resides on the VME-based system. **Figure 3** sketches the control flow for the manual and autonomous world coordinate frame-referenced control modes. The general scheme by which the operator currently commands forward control to the robot is as follows: he inputs to the system from the GUI and this input is passed forward using the UNIX socket facility over the Ethernet link. Data thus passed into the control system is specified as desired changes in the robot tip position. We relate these world frame tip coordinate changes to commanded robot joint motions through a Jacobian inverse matrix, which is computed using a JPL-developed Spatial Operator Algebra [14, 15]; this inverse is then multiplied with the input tip displacement vector to determine a corresponding joint position change vector. The primary advantage afforded by the Spatial Operator Algebra for this application is its concise recursive formulation of the kinematics equations, allowing rapid software development and testing -- a simple addition of the joint position change vector to the actual position of the joints results in the desired joint positions for the robot. The desired joint positions are then downloaded to the PMAC controller board wherein joint servo control is performed using a PID loop for each joint axis. In the manual and autonomous joint control modes, the PMAC controller correspondingly receives the joint position change vector as its input. The vector is added to the actual joint positions of the robot and the resulting vector is the desired joint position vector sent to the servo controller.

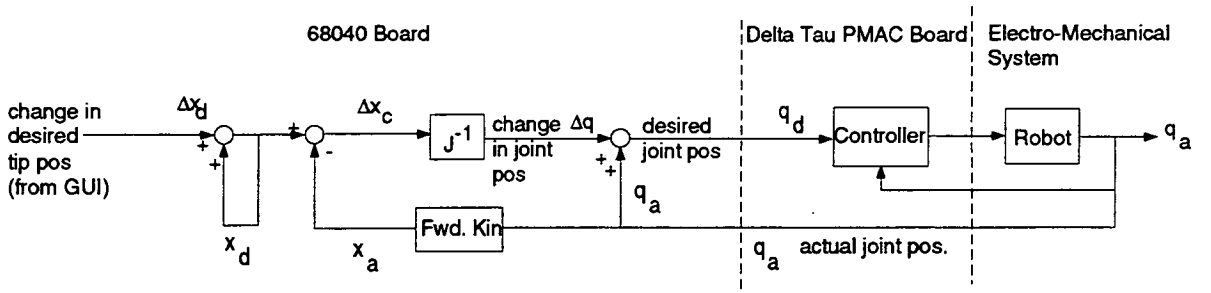


Figure 3: RAMS control flow diagram

RESULTS AND FUTURE PLANS

As of August, 1994, we had integrated the slave robot system described above and demonstrated its successful operation in all control modes. On initial integration, without benefit of significant mechanical tuning or refitting of the robot mechanisms, we achieved repeatable relative positioning of the robot tip to 25 microns or less. This measurement, verified in a number of calibrated and videotaped [16] experiments, was performed both mechanically and optically. In the former

case, we utilized calibrated mechanical dial indicators on three orthogonal axes of a wrist-tip-mounted needle; for the latter, we utilized a calibrated viewing field microscope with integrated CCD camera, and programmed and visually monitored a number of different free space, small motions within a 800 micron full-extent reticle. Cumulatively, we observed that both small (micron) and large (centimeter) free space motion trajectories are smooth. Impromptu tests in which a leading microsurgeon compared his free hand motions with that of the robot indicate that the desired scaling will be possible and highly beneficial, given an appropriate hand master interface. Development of such a non-replica master is one immediate project focus, as is also continuing, more quantitative evaluation of the robot, including its loaded (contact) motion performance. Another planned activity is development of control compensation techniques to reject feed-forward "disturbances" arising from the surgeon's involuntary tremor and jerk.

ACKNOWLEDGMENTS

The Robot Assisted Microsurgery task is being carried out by Edward C. Barlow (Chassis & Cabling), Curtis D. Boswell (Electronics), Hari Das (System Design, Teleoperator Controls & User Interfaces), Sukhan Lee (Sensors), Timothy R. Ohm (Mechanical Design & Integration), Eric D. Paljug (Computing), Guillermo Rodriguez (Kinematic Controls) and Paul S. Schenker (Principal Investigator) at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Related New Technology Notices have been filed, and a JPL-Industry Technology Cooperation Agreement with MicroDexterity Systems (MDS), Inc., exists. At MDS, Chief Officer Steven T. Charles, M.D., has been instrumental in defining dexterity enhancement system operational requirements, elements of engineering conceptual design, and strategies for subsequent medical evaluation. Charles R. Weisbin and David B. Lavery, respectively program managers for telerobotics R&D at JPL and NASA Headquarters, have aggressively supported the application and commercialization of NASA robotics technology in medicine, thereby making this project possible.

REFERENCES

1) Proc. Medicine Meets Virtual Reality, June 4-7, 1992, and Proc. Medicine Meets Virtual Reality II, Jan 27-30, 1994, both at San Diego, California, sponsored by the Univ. Calif. San Diego (Publisher: Aligned Management Consultants, San Diego, CA.).

2) Report on NSF Workshop on Computer Assisted Surgery, February 28-March 2, 1993, Washington, D.C. (Orgs., R. H. Taylor and G.A. Bekey).

3) Proc. First Intl. Symp. Medical Robotics and Computer Assisted Surgery (MRCAS'94), September 22-24, Pittsburgh, PA (Eds., A.M. DiGioia, III, T. Kanade, and R. Taylor).

4) NCI-NASA-SCVIR Workshop on Technology Transfer in Image Guided Therapy, August 5, 1994, San Francisco, CA (Chr., H. Y. Kressel, M.D.)

5) P. S. Schenker, "Intelligent robots for space applications," pp. 545-591, in Intelligent Robotic Systems: Analysis, Design, and Programming (S. Tzafestas, Ed.). Marcel Dekker: New York City, NY, 1991.

6) P. S. Schenker, A. K. Bejczy, W. S. Kim, and S. Lee, "Advanced man-machine interfaces and control architecture for dexterous teleoperations" in Proc. Oceans `91, pp. 1500-1525, Honolulu, HI, October, 1991

7) H. Das, H. Zak, W. S. Kim, A. K. Bejczy, and P. S. Schenker, "Operator performance with alternative manual modes of control," *Presence*, vol. 1, no. 2, pp. 201-218, Spring 1992; H. Das, P.S. Schenker, H. Zak, and A. K. Bejczy, "Teleoperated satellite repair experiments," in 1992 IEEE-RSJ Intl. Conf. IROS, Raleigh, NC, July.

8) E D. Paljug and P. S. Schenker, "Advanced Teleoperation Control Architecture," in Telemanipulator Technology and Space Robotics, Proc. SPIE 2057, Boston, MA, September, 1993.

9) P. S. Schenker, A. K. Bejczy and W. S. Kim, "Advanced teleoperation: technology innovations and applications," Proc. Technology 2003, Anaheim, CA; December, 1993 (NASA Conf. Publ. 3249).

10) P. S. Schenker and W. S. Kim, "Remote robotic operations and graphics-based operator interfaces," in Proc. 5th Intl. Symp. on Robotics and Manufacturing (ISRAM `94), Maui, HI, August 14-17, 1994; W. S. Kim and P. S. Schenker, "Teleoperation training simulator with visual and kinesthetic force reality," in Human Vision, Visual Processing, and Visualization, Proc. SPIE 1666, San Jose, CA, February 1992.

11) W. S. Kim, "Virtual reality calibration for telerobotic servicing," in Proc. 1994 IEEE Intl. Conf. Robotics and Automation, San Diego, CA, May.

12) P. S. Schenker, W. S. Kim, and A. K. Bejczy, "Remote robotic operations at 3000 miles -- dexterous teleoperation with time-delay via calibrated virtual reality task display," in Proc. Medicine Meets Virtual Reality II, San Diego, CA, January, 1994; W. S. Kim, P. S. Schenker, A. K. Bejczy, S. Leake, and S. Ollendorf, "An advanced operator interface design with preview/predictive displays for ground-controlled space telerobotic servicing," in Telemanipulator Technology and Space Robotics, Proc. SPIE 2057, Boston, MA, September, 1993.

13) P. S. Schenker, S. F. Peters, E. D. Paljug, and W. S. Kim, "Intelligent viewing control for robotic & automation systems" in Sensor Fusion VII, Proc. SPIE 2355, Boston, MA, October, 1994.

14) G. Rodriguez, "Kalman filtering, smoothing and recursive robot arm forward and inverse dynamics," *Journal of Robotics and Automation*, Vol. 3, No. 6, pp. 624-639, 1987.

15) G. Rodriguez, K. Kreutz, and A. Jain, "A spatial operator algebra for manipulator modeling and control, *International Journal of Robotics Research*," Vol. 10, No. 4, pp. 371-381, 1991.

16) "Robot Assisted Microsurgery project accomplishments for FY94 -- demonstration of robot joint motion, Cartesian control, and precise tip control," Production AVC-94-228 (VHS Videotape), Sep 1, 1994, Audiovisual Services Office, Jet Propulsion Laboratory.

Mr. Leonard Ault
Attn: TECHNOLOGY 2004
NASA Headquarters, Code CU
300 E Street, SW
Washington, D. C. 20546

Phone: 202-358-0721
FAX: 202-358-3938

I enclose hard copy and floppy disk copy (Word 6.0/PC-Windows) for the Technology 2004 paper entitled

A NEW ROBOT FOR HIGH DEXTERITY MICROSURGERY

Paul S. Schenker, Hari Das, and Timothy R. Ohm

Jet Propulsion Laboratory
4800 Oak Grove Drive / MS 198-219
Pasadena, CA 91109

The floppy disk copy is exactly the same format as hard copy but does not include graphics, which were separately produced. Should you have any questions, please contact me at:

Phone: 818-354-2681
FAX: 818-393-5007
Email: schenker@telerobotics.jpl.nasa.gov

Correspondence should be sent to the JPL address shown above. I look forward to our participation in Technology 2004, and appreciate the opportunity to present what I believe many will find exciting new work with an important commercialization focus.

Sincerely, Paul Schenker, Ph.D. PSS 10/4/94
Group Supervisor, Man Machine Systems
JPL

COMBINED SUPERCRITICAL AIR BREATHING AND BODY COOLING DEVELOPMENT

H. L. Gier, Ph.D.
Aerospace Design & Development, Inc.
Boulder, Colorado 80301

ABSTRACT

Aerospace Design & Development, Inc. (ADD) has developed a system to achieve body cooling from the use of supercritical (cryogenic) air carried for breathing. The breathing air is contained in a cryogenic dewar which replaces the standard self contained breathing apparatus (SCBA) using compressed air. Due to its higher fluid density and lower pressure the cryogenic system is both smaller and lighter than the compressed air SCBA. The cryogenic air is in a single phase state which allows air supply in all attitudes and prevents oxygen enrichment during loading and standby conditions. Because of the reduced temperature of the cryogen it is capable of absorbing a large fraction of the body heat produced by the user. The heat can be transferred from the body to the breathing air by a fluid transfer loop(s) between a body suit and heat exchangers near the dewar. A loading unit for the dewars has been developed which requires only compressed air and liquid nitrogen to accomplish the filling. The goal of this development is to produce an ensemble for NASA and the Air Force firefighter and rescue purposes. Further commercial development of the SuperCritical Air Mobility Pack (SCAMP) SCBA would make this ensemble available to the firefighting and hazardous materials communities for the operational advantages which it will offer.

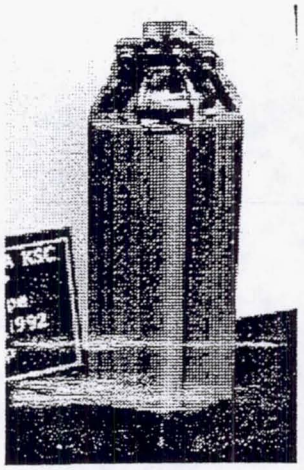
INTRODUCTION

The SCAMP SCBA was originally developed under a NASA/KSC SBIR contract to produce a compact tank which delivers air independent of tank attitude and eliminates oxygen enrichment during tank loading and storage. NASA/Kennedy Space Center (KSC) rescue personnel in a launch pad emergency must be able to crawl through a 20" square opening. Therefore a compact air system is needed which has a real lifetime in excess of 30 minutes and a backpack thickness of less than 5 1/2 inches. The system developed for NASA/KSC is shown in Figure 1 and includes the dewar, the backpack, and the loading system.

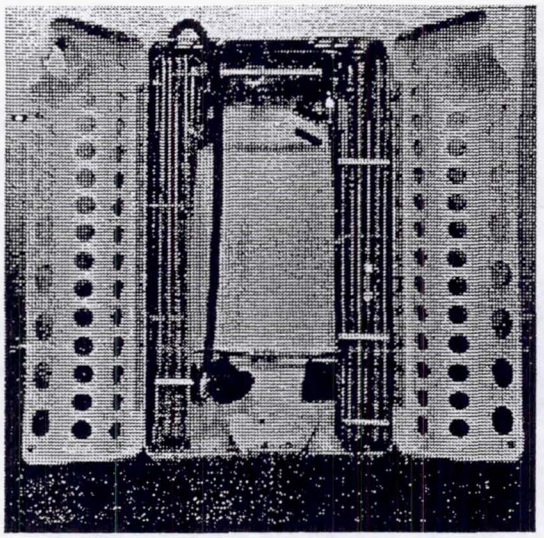
The Air Force is currently providing support with an SBIR Phase II to use the supercritical cryogenic storage of air in an SCBA to provide personnel cooling as well as breathing air in a closed firefighter's suit. The cooling provided by the cryogenic air flow will range from 100 to 500 watts depending on air flow rate. The cooling that is provided for the person will be heat absorbed by the cryogen exiting the tank. Because the person will breathe more as they exert more effort, the air flow rate which provides cooling will increase at approximately the same rate as the cooling requirements. The cooling suit is also shown in Figure 1.

The use of supercritical cryogenic air in the SCAMP came from the technology developed for the life support and fuel cell support systems on both the Apollo lunar program and the space shuttle. This particular fluid recovery system which works in the low-gravity of space will produce a flow at any attitude of the supply tank in the gravity of earth. The SCAMP breathing system application leads to a use of space technology for groundbased purposes at NASA, and the potential of use for commercial purposes in emergency services, industry, and recreation.

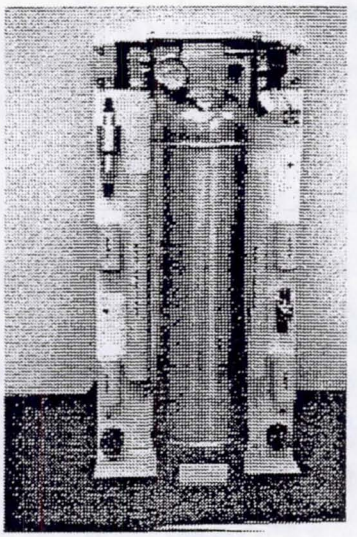
The heart of the SCAMP SCBA system (Fig. 1) is the dewar which contains the supercritical air. The SCAMP one-hour dewar is approximately the size of a standard 30 minute SCBA compressed air bottle but stores twice the quantity of air due to the high density. A two hour unit will increase the diameter and weight by about 40%. The top of the dewar has a plumbing manifold with disconnects, relief valves, quantity sensor readout, and pressure gauge. The quantity sensor is a capacitance type as the single phase storage will not have a liquid-vapor interface. The external fin heat exchangers on the dewar provide the heat for expulsion of the air from the dewar. For use the dewar is placed in a backpack which contains quick disconnects for the dewar and mask, the pressure regulator(s), and the fluid conditioning heat exchangers. The backpack provides an interface between the user and the dewar and is used to protect the extended components. A special loading system is used to hold the dewar during filling and to cool the air from the ambient temperature provided by the compressor/purifier down to cryogenic temperatures (about 90 K) using liquid nitrogen.



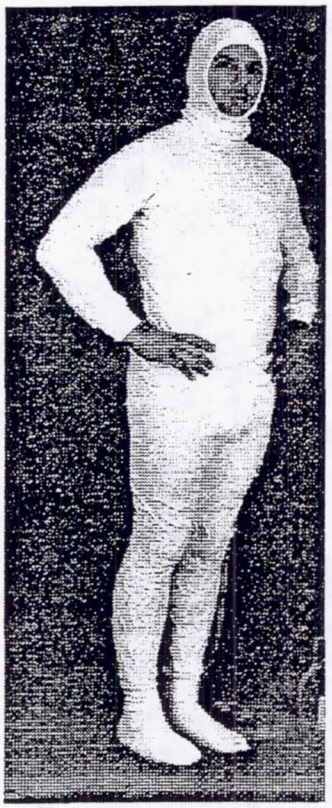
Breathing Only Dewar



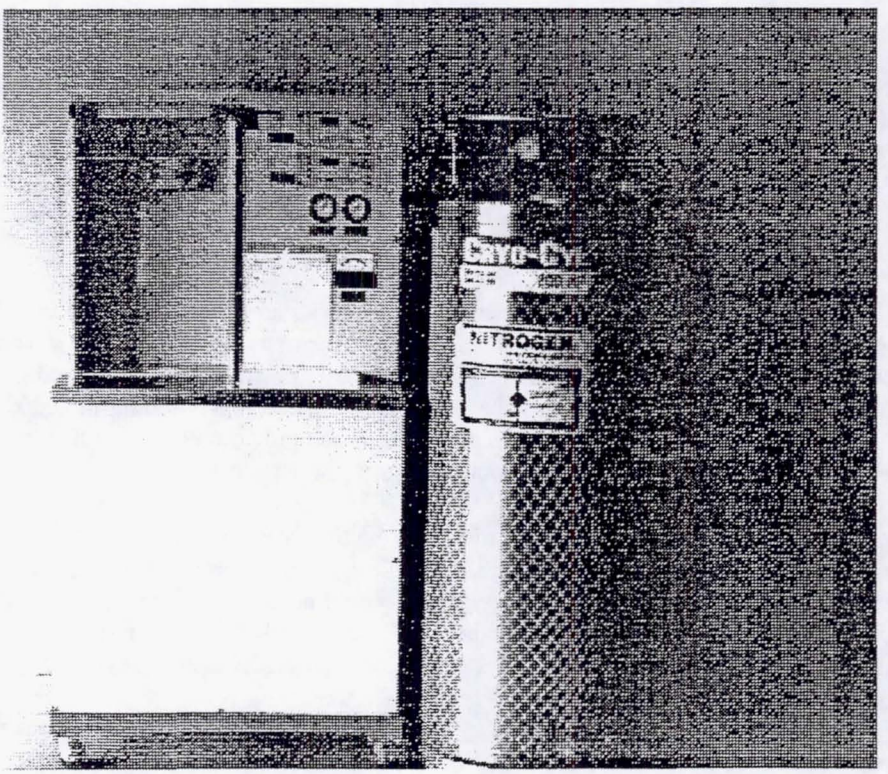
Breathing Only Backpack (Open)



Cooling Backpack
with Dewar



Cooling Suit



Loading System with LN2 Dewar

Figure 1: Supercritical Self-Contained Breathing Apparatus

SCAMP BREATHING SYSTEM DESCRIPTION

By maintaining the air in the supercritical condition the air does not become liquid at any time during the operation of the SCAMP dewar. This has several advantages on the operation of a cryogenic air system. The first of these is that when the fluid is supercritical it is always a single phase fluid. This single phase means that the air state is identical at any point in the pressure vessel (PV) and thus the dewar may be operated in any attitude and still provide the same delivery of air to the mask. The second is that the air does not change composition when it is stored for a long period of time, as it does when liquid (two-phase) air is used. When liquid (two-phase) air is stored, the natural heat load from atmospheric temperature to cryogenic causes the more volatile nitrogen to boil off first and thus leaves an oxygen enriched liquid in the tank. The single phase fluid loses oxygen and nitrogen in proportion; since there is no separation effect in a compressed gas the boil-off will not change the mixture ratio of the stored or supplied air. A Mollier diagram for air is shown in Figure 2. The lines labelled A through C are analytic traces of the fluid state of the air in the SCAMP dewar as it is being used from different initial density conditions. As shown in Figure 2, the air state will be supercritical in that the entire operating process is above the critical pressure of air.

The supercritical cryogenic fluid expands to fill the bottle that it's stored in, if sufficient heat is added, and thus delivers from a single point exit in any bottle attitude. The expulsion energy, designated dQ/dM for the energy (Q) required to expel a given mass (M), is shown as a function of temperature in Figure 3. This energy for the NASA/KSC SCAMP expulsion is provided by ambient air through natural convection to the fins on the dewar.

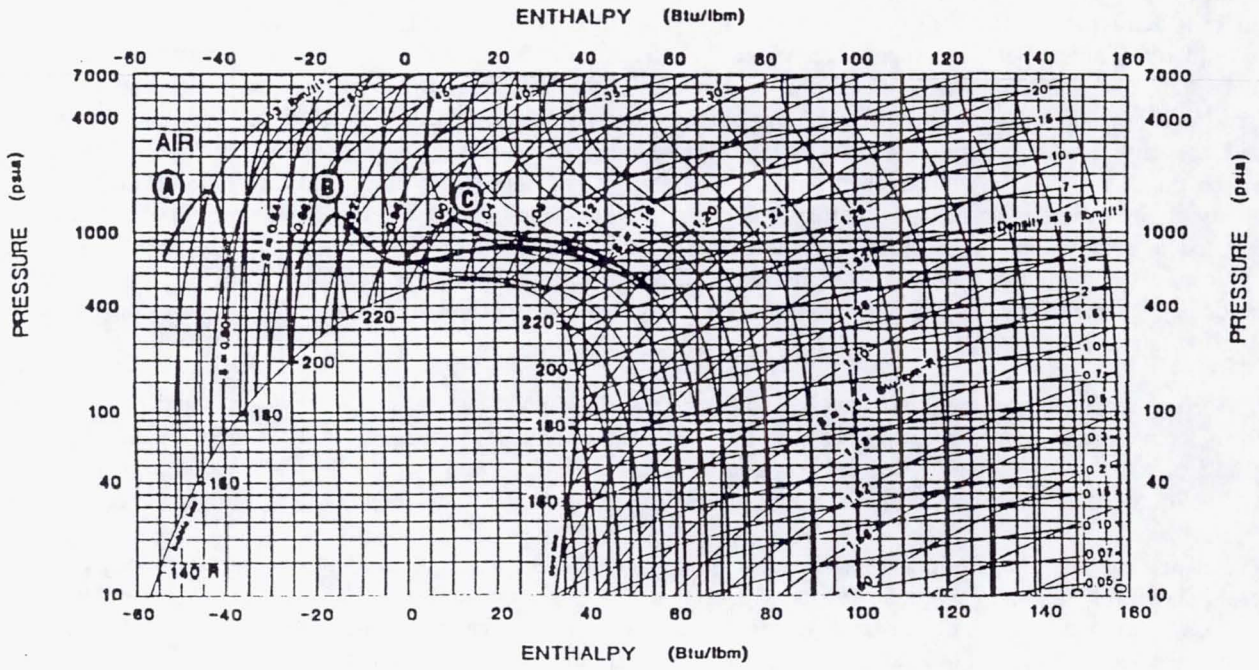
Because the air is supercritical there is no liquid-vapor interface and no discontinuity in density at any point. There is, however, a rapid change in density with temperature as shown in Figure 3 which occurs around 250° R (-135° C). This means 50% of the air is supplied with a temperature change of about 30 degrees. As the air is cooled below about 250° R (140 K), the density increases rapidly so that much more air can be stored in a given volume. However, in this supercritical cryogenic state, below 200° R (110 K), the air becomes incompressible so that increasing the pressure does not significantly increase the density. This is the reason for choosing a cold (145° R or 80 K) temperature for loading the SCAMP, but only going to a high enough pressure (750 psi or 50 atm) to reliably hold the supercritical condition. As the air in the tank is depleted by breathing, the density decreases and the temperature increases as shown in Figure 3. Most of the lifetime of the dewar is spent with temperatures at or below 250° R (140 K).

There is a penalty exacted for the increased use time or decreased bulk with the cryogenic storage in that the air bottle must be refilled on a regular basis (every 24 to 48 hrs maximum) if it is to be ready for immediate use.

The operational schematic for the SCAMP dewar when mounted in the backpack is shown in Figure 4. There are four different modes of operation for this dewar. First there is the fill mode in which the dewar is attached to the loading system. The second is a standby mode in which the dewar is not attached to any other apparatus. The other two occur when mounted in the backpack; the SCAMP may be in either the standby mode with the valve "off" or in the air supply mode with the valve "on".

The loading system selected for use with the SCAMP is as shown in Figure 1. The loading system requires a supply of clean, dry air. This may be from a cascade system of high pressure bottles or from a compressor which is used to fill any SCBA. The grade of air depends upon the purification system which is built into the compressor and should be a minimum of Grade D, or preferably Grade E. The air pressure into the loading system should be slightly above the intended setting of the back pressure regulator (Fig. 4, PR-1) or about 1000 psi (70 atm) to allow for an adequate airflow. This keeps the loading process in a purely supercritical regime and thus does not allow the compressed air to separate or change composition during the dewar loading.

After loading, the SCAMP dewar may be removed from the loading system, and either stored separately or placed in the backpack. For use in the NASA/KSC environment, which requires breathing air only, the SCAMP dewar would be mounted in a backpack as shown in Figure 1. The backpack provides a mounting interface to the user as well as providing protection for the heat exchanger (Fig. 4, HX-5) which brings the air from cryogenic temperature to close to ambient temperature for breathing.



(A) 100% Loading Density (B) 80% Loading Density (C) 60% Loading Density

Figure 2: Mollier Properties Chart for Air

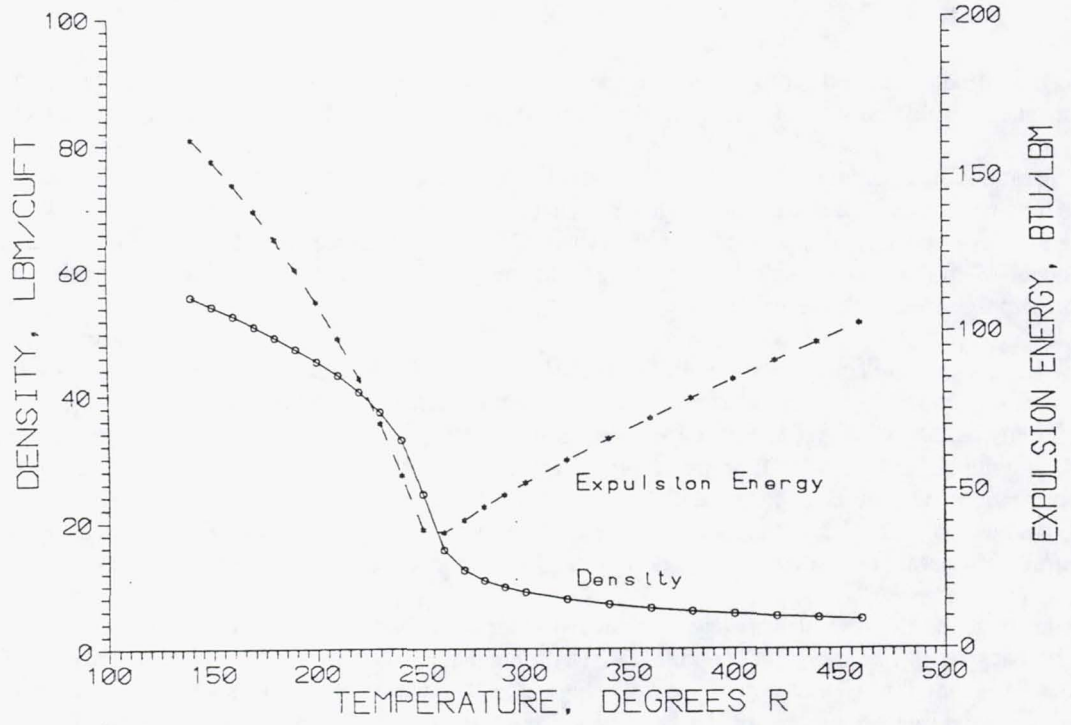


Figure 3: Air Density and Energy of Expulsion (dQ/dm) as a Function of Temperature

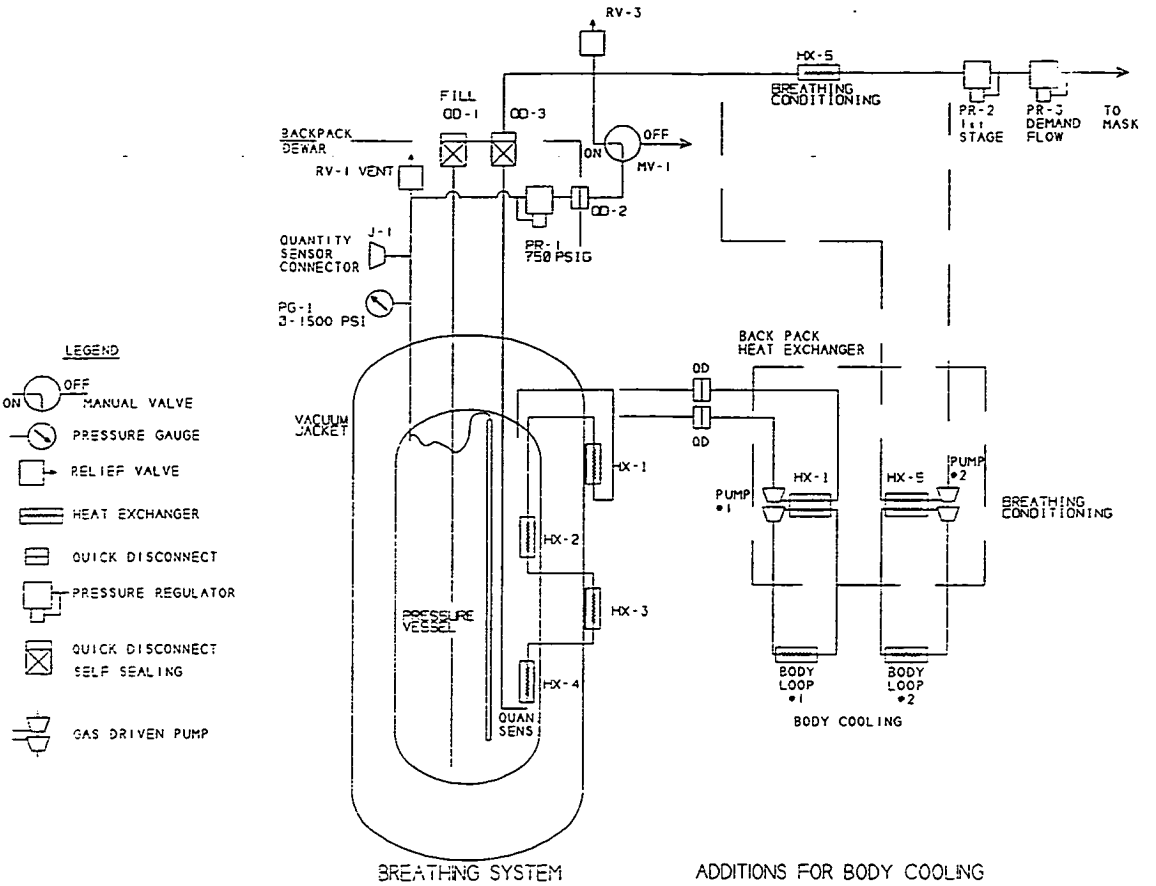


Figure 4: SCAMP Dewar Schematic with and without Cooling

DESCRIPTION OF SCAMP COOLING CONCEPT

During the development of the SCAMP SCBA for NASA it was obvious that there was an opportunity to provide cooling for the user as well as breathing air. The source of the cooling for the user will be the heat exchangers (Fig. 4) which provide the expulsion and conditioning heat. In each of these heat exchangers the fluid is brought from cryogenic temperature to just below ambient temperature by the addition of heat from some source. In the NASA/KSC (breathing air only) configuration this heat is supplied by free convection from the atmosphere; in the cooling configuration this heat will be supplied by the body of the user. Since the fluid temperature, into the heat exchangers from the dewar, starts at near liquid nitrogen temperature (90 K or -300° F), there is a large amount of cooling available for the human body. This will provide cooling to the user, while supplying the heat which is required by the SCAMP dewar. Maximum cooling will be necessary in a hot environment and fire conditions whereas in cold weather the user may become too cold when the activity level is low.

The cooling available from supercritical cryogenic air storage is shown in Figure 5. There are three sources of cooling: 1) loop 1; from the energy required to expel the air (expulsion) from the dewar including the energy required to warm up the titanium pressure vessel, 2) loop 2; from the energy required to bring the air up to ambient temperature for breathing (conditioning), and 3) from the saturation of the breathed air in the lungs. For supercritical air, the cooling available varies with the energy of expulsion (Fig. 4) and the storage temperature which are both dependent upon the remaining fluid in the dewar (Fig. 3).

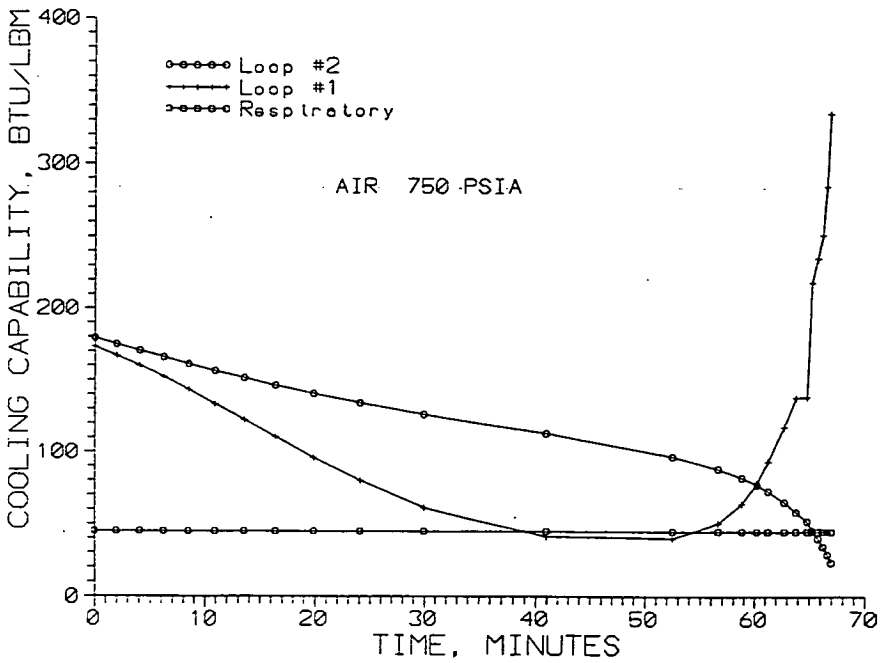


Figure 5: Cooling Capability for Each Loop

Cooling Requirements Of User

The total energy which a person can produce metabolically and which must therefore be removed by a coolant system depends upon both the person and the activity being performed. Goldman (1) gives values of metabolic heating for firefighters as shown in Figure 6. The heat energy which must be removed by the coolant, during rest and exercise, is the metabolic heat release reduced by work performed and modified by environmental conditions..

Human physiology (2,3) gives a relationship between air consumption and energy produced as follows: 1) that the human body in consuming one (1) liter of oxygen releases heat energy of 4.825 kcal and 2) that 21% oxygen is inhaled and 17% oxygen is exhaled. The relationship between oxygen consumption and energy release seems to be well enough established that the oxygen consumption is used to define energy release. The consumption of 4% oxygen seems to be much less accurate due to individual differences, breathing rate, aerobic vs. anaerobic activities, and metabolic level. Using the above values gives a metabolic energy production rate of 284.3 Btu/hr (83.3 watts) for each pound of air used per hour. For a NIOSH rated air use of 40 SLM (6.466 lbm/hr) this corresponds to a metabolic energy release of 537.3 watts.

The summary of the cooling capability of the SCAMP SCBA and the cooling deficit is shown in Figure 7. This shows the cooling deficit with cooling and without any cooling. By 20 minutes the system has over-cooled the user by about 20 Btu/lbm/hr (32.6 kcal at 40 SLM) and at the end of the hour the user has been under-cooled by approximately the same amount. For a 180 lb user with the rated consumption of 40 SLM this will amount to less than one-half (0.5) degree Celsius (one (1) degree Fahrenheit) variance in the core body temperature of the user in the consumption time. Without cooling, the user will nominally have a cooling deficit of 83.3 watts for each one lbm/hr air of flow rate. For the NIOSH rated flow of 40 SLM this totals 539 watts or summed over the hour the user has generated 483.5 kcal (1,838 btu) in energy which must be stored or dissipated. The dewar analysis shows the SCAMP cooling removes 94% of the metabolic energy produced with a final deficit of 32.6 kcal. This presumes that the dewar which is 100% full at the beginning of the analysis and thus has a use time of 72.9 minutes at rated flow. If the dewar has been on standby for the maximum time and starts with only 60 minutes at rated flow, the initial overcooling is reduced by 15 Btu/lbm and the subsequent deficit in cooling is approximately 35 Btu/lbm. This means that the cooling deficit may run to about 10% (48 kcal) if the dewar starts at rated minimum useable time.

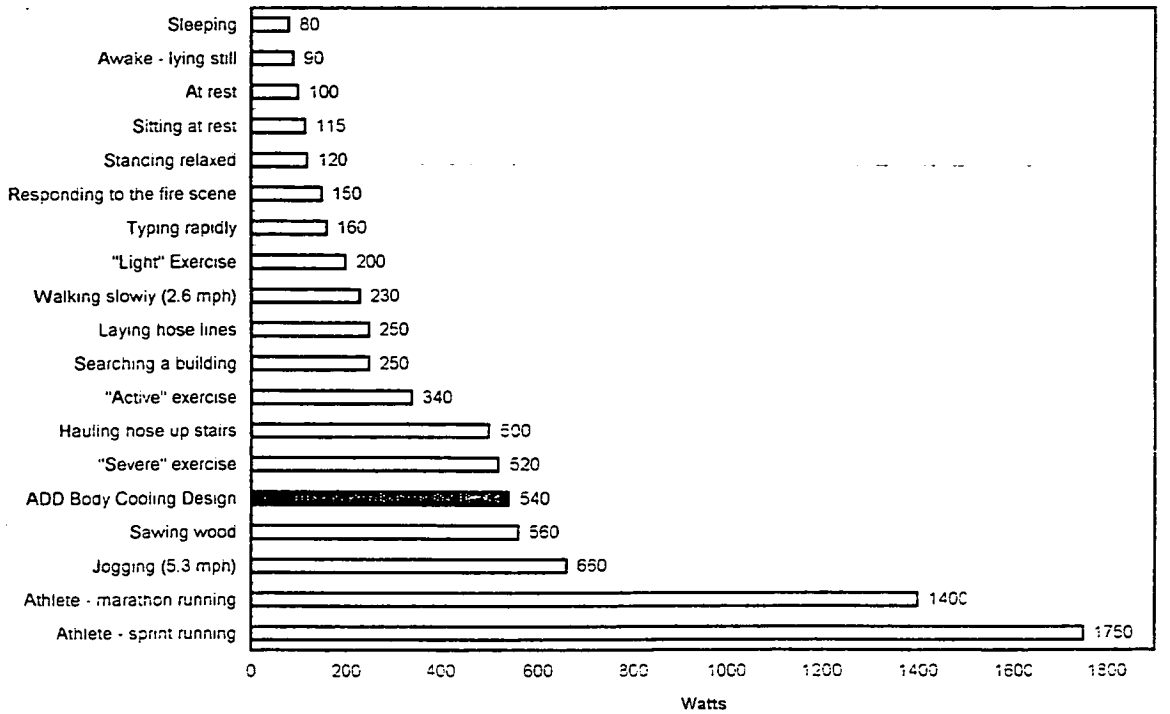


Figure 6: Heat Production in Fire Fighters (Goldman (1))

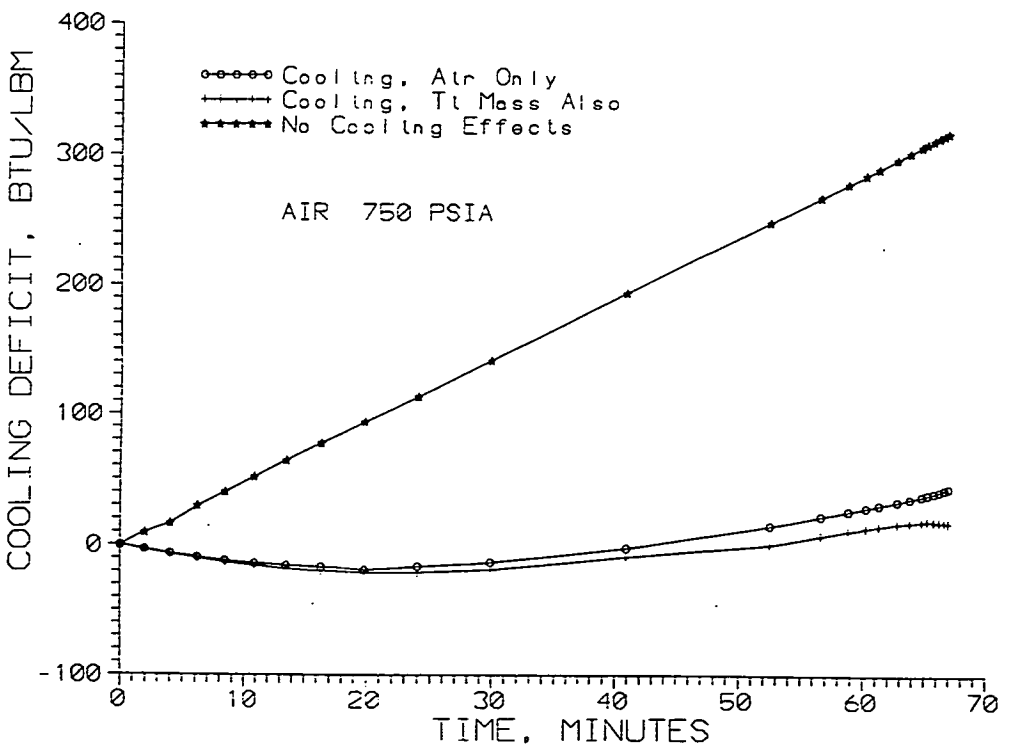


Figure 7: Cooling Deficit Compared to Uncooled

Higher or lower use rates of the breathing air will change the rate at which heat builds up in the user's body but not the final cooling deficit. If the user breaths at 80 SLM the deficit rate will double by shortening the time period of Figure 7 to 30 minutes but the final cooling deficit will remain 32.6 kcal. Only by doubling the lifetime using a NIOSH two hour version of the SCAMP would the cooling deficit be doubled. This is about the cooling deficit experienced in an impermeable suit in a 15 minute time period at a moderate activity level (40 SLM air use).

This SCAMP cooling is far better than any of the current phase change material, such as ice or blue ice devices in which the coolant material is stored in the user's vest and receives heat either by circulation or conduction. Those units are limited (4) both in cooling rate and total absorption of energy. In addition, the phase change units add to the weight of the systems which the user must carry.

The SCAMP cooling system is an outstanding improvement in the life support field. Because the SCAMP cooling source is the same air which is being used for breathing, nothing additional is required to provide the cooling, and thus the SCAMP with cooling will be lighter than without cooling due to reduction of heat exchanger and backpack requirements.

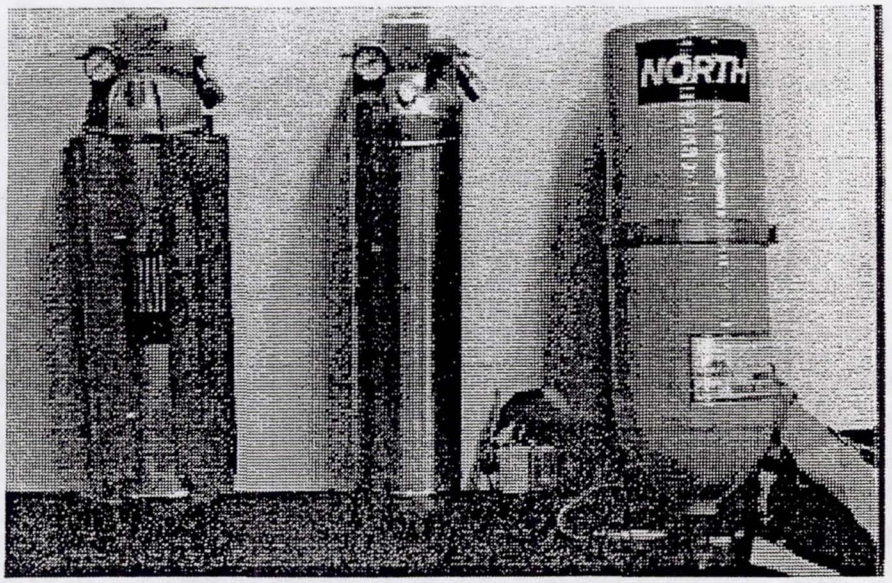
The empty weight is approximately 13.1 lbs for the prototype and full loaded weight is about 21 lbs without backpack. If the dewar is used with body cooling then the external heat exchanger fins can be eliminated, thus reducing empty weight to 9.3 lbs. The backpack may be reduced to a frame to carry the mounting bridge, heat exchangers, and other components for body cooling giving a total weight for a one hour rated (40 SLM) of approximately 22 lbs. This is about half the weight of current high pressure one-hour systems as well as being both thinner and narrower. The dewar for stand alone service and for body cooling service are both shown with a one hour compressed air bottle in Figure 8. A two hour SCAMP bottle is drawn in the same scale for comparison. The projected weights for these configurations are given in Table I. The most certain of the weights is the current "stand-alone" SCAMP SCBA because the prototype has been built and weighed; no credit is given in this table for future weight savings. The weights given are a combination of fixed numbers for the fluid, calculated numbers for the pressure vessel and outer shell portion of the dewar, and extrapolations from the current SCAMP for the backpack and components on the dewar.

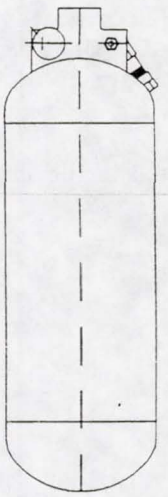
It is envisioned at this time that the ideal cooling system will keep a user's body temperature within one degree Celsius of normal for a period of four hours at rest, two hours at a moderate activity level, or one hour at a high activity level without having to resort to exterior sources of cooling. This is also the approximate limit of air which can be carried (at the current time) without exceeding the 35 pound backpack weigh limit imposed by military standards and NIOSH. To minimize dehydration, provide for maximum comfort, and to allow a person to stay in a hazardous or contaminated area for the longest possible time it would be reasonable for this ideal cooling system to limit sweating as far as is possible.

Extending the lifetime of the SCAMP to 3 NIOSH hours appears to be readily achievable within current configuration constraints. NIOSH allows a 40 lbm maximum if over 25% of the initial weight is expended during the use period so a 3 hour SCAMP with cooling is possible.

A possible configuration of a cooling suit is shown in Figure 9. The SCAMP body loop #2 is the higher capacity loop and provides the cooling for the torso and head . This body cooling loop (#2) is connected to the backpack with low loss quick disconnects in the plastic lines. This covers the region of maximum heat production at rest and during low levels of activity. The body loop #1 will be the primary coolant for the lower body.

The possible cooling suit configurations are as various as the needs for cooling. The high level of cooling possible with the SCAMP will make it necessary to further develop the cooling suit to use the SCAMP cooling effectively.





Stand Alone Breathing Body Cooling Usage One-Hour Compressed Two-Hour Cooling

Figure 8: SCAMP Dewar Outlines

TABLE I: Tentative Comparative Weights of Different Configurations

	NIOSH	WEIGHT (lbm)			
	LIFETIME	DEWAR	PACK	FLUID	TOTAL
COMPRESSED	1	18	10	7	35
SCAMP AIR	1	10	7	8	25
COOLING	1	7	5	8	20
COOLING	2	12	7	16	35

POTENTIAL POST APPLICATIONS

When ADD first described the SCAMP to marketing it was seen as a viable commercial venture. A need was seen in the emergency services for a lighter and/or a longer lifetime SCBA. The longer time provides a greater margin of safety so that a firefighter or haz-mat worker can work longer without having to return for a refill. This would aid in both the reduction of damage and increasing the possibility of life saving operations. In addition there is the greatly reduced logistics support at a fire scene because fewer air bottles would be needed.

The SCAMP can be used anywhere that a standard compressed air SCBA can be used. However its unique capabilities will be as a lower profile system (KSC use) where space or clearance distance is critical or as a longer life unit for particular situations. Fire fighting in large or high-rise buildings is one of the particular applications seen. The SCAMP may be used in Haz-Mat working and cleanup to reduce the number of times a worker will have to leave the work area, and reduce the amount of equipment which must be decontaminated. For the military use, sealed chemical warfare suits need to include both air supply and cooling. Also there might be a particular application for damage control where the area is large and contains toxic and radioactive materials.

It is only occasionally that the chance to make a direct change in peoples' lives is offered in the engineering profession. ADD views the cryogenic air SCBA as one of those opportunities. It is certain that there will be lives saved by the use of the SCAMP in the years to come due to the increased time that emergency personnel can work without refill and the increased confidence which they will have because of additional air reserve time.

Multiple domestic and foreign patents on this technology have been applied for and are pending.

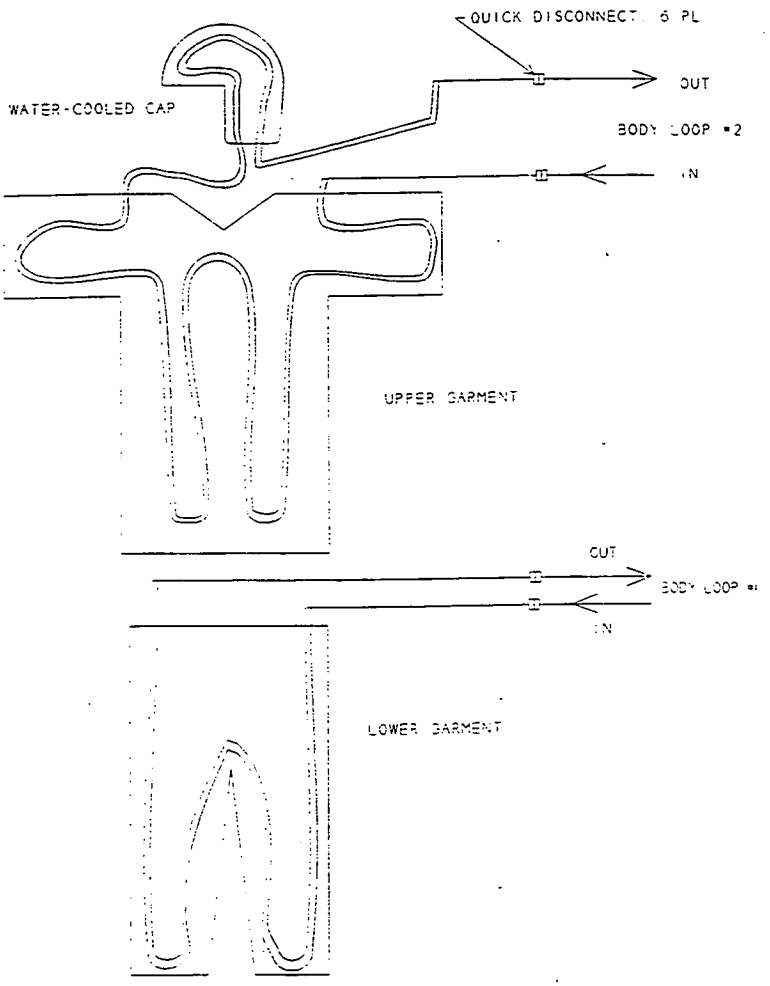


Figure 9: Basic Configuration of Cooling Suit

References

1. Goldman, Ralph F.; Heat Stress in Firefighting, Fire Engineering, May 1990

2. Guyton, Arthur C.; "Physiology of the Human Body", Saunders College Publishing, Philadelphia 1984

3. Hardy, James D.; et al; Editor; "Physiological and Behavioral Temperature Regulation", Charles C Thomas - Publisher, Springfield, Ill. 1970

4. Glenn, Stephan; Eley, W. David; and Jansen, Paul A.; "Evaluation of Three Cooling Systems Used in Conjunction with the USCG Chemical Response Suit" Hazardous Materials Control September/October 1990

Robotics and Artificial Intelligence

TRIPOD OPERATORS FOR REALTIME RECOGNITION OF SURFACE SHAPES IN RANGE IMAGES

Frank Pipitone
Navy Center for Artificial Intelligence,
Naval Research Laboratory,
Washington, DC 20375-5337

ABSTRACT

Tripod operators (TO's) are a versatile class of feature extraction operators for surfaces. They are useful for recognition and/or localization based on range or tactile data. They extract a few sparse point samples in a regimented way, so that N sampled surface points yield only N–3 independent scalar features containing all the pose-invariant surface shape information in these points and no other information. They provide a powerful index into sets of prestored surface representations. A TO consists of three points in 3-space fixed at the vertices of an equilateral triangle and a procedure for making several "depth" measurements in the coordinate frame of the triangle, which is placed on the surface like a surveyor's tripod. TO's can be imbedded in a vision system in many ways and applied to almost any surface shape. Here the focus is an experimental study in which individual TO's are used to search a cluttered range image for one of 25 known shapes, typically in milliseconds, with very few false detections. We believe that this simple way of using TO's, in conjunction with existing triangulation range sensor technology, can be effectively applied to industrial parts recognition tasks, and with additional research, to other applications.

1. INTRODUCTION

This work is motivated by the long-standing observation that a small set (e.g., six to twelve) of point samples of the surface of an object is highly informative, and that it ought to be possible to construct a procedure for mapping such data into the identity and/or pose of an object in essentially constant time, for a significant range of cases. We have largely succeeded in doing this, using a geometric procedure called the tripod operator (TO). A typical TO is applied to a range image in approximately 2 milliseconds, as currently implemented on a Sun SPARCstation 10, resulting in a hypothesis about the surface under the operator. Software optimizations are expected to reduce this to well below 1 millisecond. A range image can be searched for a shape by repeatedly applying TO's at random places on the image. Potential applications include industrial parts recognition, target recognition, mobile robot vision, and face recognition.

In order to rapidly recognize objects based on surface shape, especially if the library of known objects is large and/or the average complexity of each object's surface shape is large, one needs to make feature measurements which are sufficiently informative, despite noise, that the reduction in the candidate set per unit computation time is acceptable. For example, one might reasonably measure this by the reduction in the Shannon Entropy of the set of identities and/or poses. By such a measure, steady progress has been made in previous work. Grimson [4,5] and others [6,7,8,9] extensively developed the idea of searching for associations between image features and model elements consistent with geometric constraints among the model elements, using *interpretation trees* to represent the consistent hypothesised associations (interpretations). However, interpretation trees require quadratic time processing per model. This is mitigated by using particularly informative features. We have argued that TO's can be used efficiently as such features [1,2]. A second connection is that a TO can be regarded as precompiled prepruned interpretation trees having sparse range pixels as the image features. This is their original inspiration. Lamdan and Wolfson [10] contribute to efficiency in model-based vision by providing precompiled geometric pointers among local features. This requires the ability to detect a reasonably small number of reasonably stable *interest points* and to define informative features there, whereas TO's are to be used anywhere on a surface and their informativeness can be looked up. Stein and Medioni [11] describe local operators called "splashes" with attractive invariance properties, but they have high

computational cost and depend on unoccluded and valid range pixels on certain geodesic lines. The RANSAC method [12] uses sparse samples economically to test a fit to a specific class of functions, but indexing is not provided; one must sequentially try function classes. Many kinds of local feature detectors or matchers have been explored for range images. For example, [13] concentrates on dihedral edges and [14] on the two principle curvatures of smooth surfaces. The principle limitation of most of them appears to be their discriminating power per unit computation. For example, estimators of the two principle curvatures either provide us with two real numbers worth of indexing information (and direction information), or the decision that the surface is not a good fit to a quadric in the current neighborhood. The former case allows discrimination of roughly q^2 local surface shapes if we can resolve q curvature values in noise. The latter case requires us to continue looking for local feature information (perhaps a dihedral will fit here or a quadric patch elsewhere). TO's provide one operation at a place on a surface, yielding a feature vector of any dimensionality d, applicable to nearly any surface, and potentially discriminating as many as roughly q^d hypotheses about the object (and/or its pose) on which the TO lies.

Tripod Operators are "somewhat global" and can sometimes straddle many surface undulations with its point samples, and span a large proportion of an object. They can operate on sparse regions of a dense range map, sparse data acquired actively from a sequential random access range scanner (such as in [15]), or via a tactile version of a TO. In earlier publications, we argue that the TO should allow very fast recognition [1] and present supporting experimental evidence by discriminating 1 object from a library of 10 using in some cases only *one* TO placement, using synthetic range data [2]. In [3] we extend this to the case of noisy LIDAR range images of isolated real objects, using a Bayesian approach to obtain reliable recognition using a small number (5 to 10) of low order (order 4) and high noise (1/10 the TO's edgelength) TO placements.

We have been studying TO's using a software system called TRIPOD, which allows various experiments to be performed involving the application of various kinds of TO's to real or synthetic range images, and the use of various representation and matching methods on the resulting feature space point sets. Our overall research goal is to determine the limits of performance of a vision system based on TO's, and to realize that performance in prototype vision systems. Performance measures of interest to us include speed, classification error, tolerance of noise and occlusion, library size, storage requirements, and ease of representing new shapes. Variables in such a vision system that effect performance include

1. Edge length of operator
2. Order of operator
3. Efficiency of the algorithm that computes the operator
4. What hypothesis verifier is used, if any
5. Representation of the TO invariant signatures
6. Indexing method used to assess proximity of TO measurement to signature
7. Method for relating multiple TO's on the same object
8. Method for representing pose constraints
9. Use of probabilistic reasoning

The focus of this paper is the use of isolated TO placements to rapidly recognize instances of a set of 25 typical manufactured surface shapes in range images containing a variety of known and unknown shapes. Items 4,7,8 and 9 above are outside the scope of this paper. Our two-part research strategy is to first learn how to obtain the greatest possible discrimination in the shortest time using individual TO placements, and in other work to exploit the relative pose of multiple placements to further increase performance.

2. REVIEW OF TRIPOD OPERATORS

2.1 Definitions and Properties

A tripod operator consists of three points in 3-space fixed at the vertices of an equilateral triangle of fixed edgelength e, and a procedure for making several "depth" measurements in the coordinate frame of the triangle, which is placed on the surface like a surveyor's tripod. These measurements take the form of

arc-lengths along "probe curves" at which the surface is intersected. Figure 1 shows three examples of TO's. Figure 1a shows a very simple TO with one line probe fixed symmetrically with respect to the rigid triangle ABC. The single scalar feature is the distance from the plane of ABC at which the probe intersects the surface. This resembles a mechanical optician's tool called a *spherometer*. We call the number d of scalar features the *order* of the operator. Figures 1b and 1c show TO's that can be viewed as a set of equilateral triangles hinged together so that all all d+3 points can be made to contact a surface. The angles of the d hinges are the features. We prefer this type (called *linkable* TO's) because of their symmetry and uniform sensitivity to noise. A planar surface yields $\theta=180^\circ$ for all d feature components. We will sometimes use $\phi \equiv \theta - 180^\circ$ instead of θ for convenience. Many variations of these TO's could obviously be constructed. Feature noise is related to range noise n by the approximate expression $n_\phi \approx 51 \times n/e$, where n_ϕ is the feature error in degrees, and n is expressed in the same distance units as the edgelength e.

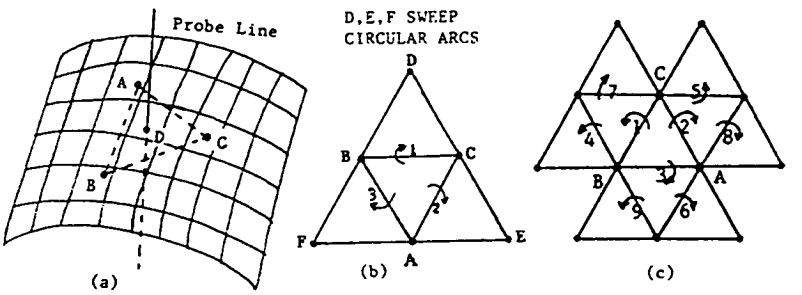


Figure 1. Examples of Tripod Operators: (a) Simple order 1 TO with linear probe, (b) Order 3 linkable TO, (c) Order 9 linkable TO.

For an N-point TO, the N sampled surface points yield only N – 3 independent scalar features (the *order* d is N – 3). These features contain all the surface shape information in the 3N components of the points, since they suffice to reconstruct the relative positions of the N points. They contain no other information; For example, they have complete six DOF invariance under rigid motions (the group $R^3 \times SO(3)$). Thus, they depend on where the tripod lies on the surface, but on nothing else. A key property is that only a 3-dimensional (at most) manifold of feature space points can be generated from a given surface, for any dimensionality d of feature vector, since the tripod can be moved only in 3 DOF on a surface. This allows objects to be densely sampled with TO's at preprocessing time with a manageable number of operator applications (typically, a few thousand) to obtain almost all the feature vector values obtainable from any range image of the object. This set is a kind of *invariant signature*. For brevity, we will call it the *signature* of the object or surface (with respect to a particular type of TO). It can be stored in bins (e.g., of dimension 3 or 4) for later efficient access of near neighbors to TO features measured at recognition time. These bins can optionally contain precomputed probability densities, analytic expressions for distances to nearby signature manifolds, and partial or complete descriptions of the relative poses of tripods and models, all to serve various purposes in a recognition system.

2.2 Computing a Tripod Operator Placement

Since in some applications of the tripod operator, the computation consists only of placement and a little indexing, the cost of placing the operator should be kept small. This can be done by efficiently implementing a procedure similar to the following. Consider placing the TO's of Figs. 1b or 1c on a dense range map. Point A can be chosen as any point on the image surface. Interpolation is to be done locally as needed (e.g., using piecewise triangular facets). Point B can be found by moving along a line at orientation α in image coordinates (pixel indices) until the 3D distance $|AB| \equiv e$. This can be done in logarithmic time (essentially constant here) using binary search. Then we search the circle of radius $.5\sqrt{3}\,e$ oriented coaxially around the center of the segment AB, using binary search, to find a point C close to the surface. A similar circular search yields each remaining point. A key step in the circular search is the mapping (specific to a range scanner's geometry) from a point (x,y,z) to the indices of the range pixel

whose ray (x,y,z) lies on. This allows the front/behind decision required by the binary search. In the case of a sequential random access range scanner, it may be efficient to monotonically search elliptical paths in image coordinates until the two distances being enforced (e.g., |AC|=e and |BC|=e) are both correct. The ellipses here are the projections of the previously described circles onto image coordinates. Finally, in the case of a tactile TO, the computation is mechanical; the feature values are to be read from position transducers (e.g. from linear potentiometers by an A/D converter).

2.3 Symmetries of Surfaces and of Tripod Operators.

Surfaces with one symmetry, such as extrusions, surfaces of revolution, and helical projections produce only a 2-dimensional manifold in feature space. Cylinders, having two symmetries, produce only a nearly circular 1-dimensional curve, and spheres a single point. Scaling a TO by changing its edgelength does not effect the signature of surfaces swept by a line with one point fixed (e.g., cones, planar n-hedral vertices, and planar dihedral edges). Regardless of the surface, an operator with a 3-fold symmetry (e.g., those in Fig. 1), produces signatures unchanged by cyclicly permuting each triple of corresponding features. In Fig. 1c, the three 3-cycles (1,2,3), (4,5,6), and (7,8,9) show this property, for features ϕ_1 through ϕ_9, respectively. This allows a 3-fold storage reduction, e.g., by permuting the features so that ϕ_1 is the largest. If the TO, in addition, has handedness symmetry (as our examples do), the signature can be modified by a procedure that allows recognition of the "other side" of any surface already recognizable. We call this *inversion* of a signature. It is done by by transposing certain pairs of corresponding features (e.g., (7,5), (1,2), (4,8), and (6,9) in Fig. 1c) and replacing each feature value ϕ with $-\phi$. Also, the signature of the opposite-handed (reflected) version of a surface can be found by performing those transpositions without negating the features.

2.4 The Structure of the TO Signatures of Some Simple Shapes

We have been studying the shapes of TO signatures [17] in order to understand how they can overlap and to find ways to approximate them with algebraic and semi-algebraic expressions. Such approximations are expected to greatly reduce storage requirements for large libraries. The signatures of order 3 operators (Fig. 1b) were rendered as a rotating cloud of points on a computer; selected 2D snapshots are shown in Fig.2. In the special case of "smooth" surface regions, the signature is nearly a circular ring coaxial with the diagonal axis. The offset and radius of the ring can be readily used to compute estimates of the principle curvatures and other differential geometric parameters [17]. Surfaces with C_1 or C_2 discontinuities tend to produce signatures with similar numbers and kinds of discontinuities (e.g., Fig. 2c,d), and have roughly commensurate complexities of description. Thus, this umbrella-shaped 2-manifold can be well approximated with a few polynomials, whereas the discrete signature might need 20,000 points (see Fig. 4) for thorough saturation.

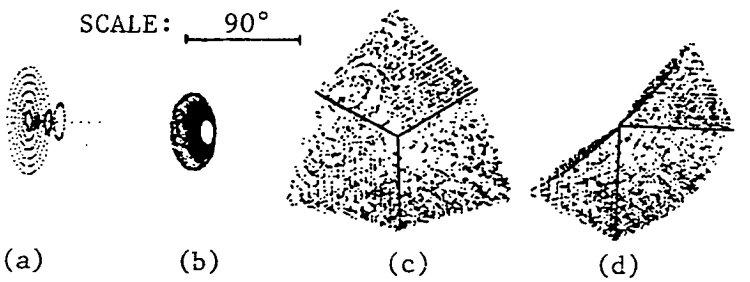


Figure 2. 2D projections of TO signatures taken with the TO of Fig. 1b. (a) Superimposed signatures of six hyperbolic paraboloid patches (large rings), four elliptic patches (rings lying on a cone), and 10 spheres (the points). (b) A torus; the signature is a piece of cone in $\phi_1\phi_2\phi_3$ space. (c,d) A 90° planar dihedral, viewed diagonally and along ϕ_1, respectively. All signatures of this TO have *at least* a 3-fold rotational symmetry about the diagonal $\phi_1=\phi_2=\phi_3$; all signatures in (a) and (b) are surfaces or curves of revolution.

3. CONDITIONS FOR RELIABLE RECOGNITION USING A SINGLE TO PLACEMENT

The low dimensionality of TO signatures (three, at most) typically allows the computation and storage of signatures containing (to a reasonable resolution) all feature vectors obtainable from a given surface shape, regardless of viewpoint. Moreover, since the feature space can have high dimensionality (d=9 in these experiments) the signatures of different objects' surfaces frequently have little or no intersection, allowing recognition of some objects with only one placement of a TO on the image of the object. Our experiments show that this circumstance occurs frequently with common shapes, and also that signature overlap can usually be dealt with. A deterministic viewpoint is taken here (When range error is a large fraction of TO edgelength, a probabilistic approach is essential).

3.1 False Positives

We will now derive sufficient conditions for precluding any false positive detections. Let us denote by **A** the set of all feature-space points obtainable by applying a certain TO to surface shape A. We call this the exact signature of shape A. Let $\mathbf{A}^{\delta}$ denote some signature of shape A such that the greatest L_2 distance from any point in A to the nearest point in $\mathbf{A}^{\delta}$ is δ. We call this a signature of A saturated to δ. This kind of signature can be obtained in practice by applying a TO a finite number of times to a surface. We similarly define **B** and $\mathbf{B}^{\delta}$ for shape B. Now let $\mathbf{A}^{-}$ denote the set obtained by deleting from $\mathbf{A}^{\delta}$ all points within an L_2 distance ε of any point in $\mathbf{B}^{\delta}$. We call this procedure *overlap removal* and speak of *subtracting* one signature from another. Now let ν be the maximum L_2 distance that sensor (and other) error can introduce, and $\mathbf{B}^{+}$ the set of points within ν of **B**. Then $\mathbf{B}^{+}$ includes all points actually obtainable by placing a TO on shape B. Summarizing key statements from above,

1. If $(\mathbf{b}^{+} \in \mathbf{B}^{+}), \exists\, (\mathbf{b} \in \mathbf{B})$ s.t. $||\mathbf{b}^{+}-\mathbf{b}|| < \nu$.
2. If $(\mathbf{b} \in \mathbf{B}), \exists\, (\mathbf{b}^{\delta} \in \mathbf{B}^{\delta})$ s.t. $||\mathbf{b}-\mathbf{b}^{\delta}|| < \delta$.
3. If $(\mathbf{a}^{-} \in \mathbf{A}^{-})$ and $(\mathbf{b}^{\delta} \in \mathbf{B}^{\delta})$, $||\mathbf{a}^{-}-\mathbf{b}^{\delta}|| > \varepsilon$.

Now consider a placement of a TO on shape B, producing the noise-corrupted feature point $\mathbf{b}^{+} \in \mathbf{B}^{+}$ instead of the corresponding exact point $\mathbf{b} \in$ B. Suppose $\mathbf{a}^{-}$ is the nearest point in $\mathbf{A}^{-}$ to $\mathbf{b}^{+}$. Then from 1., 2. and 3. above, $||\mathbf{b}^{+}-\mathbf{a}^{-}|| \geq \varepsilon-\delta-\nu$. This means that we can never mistake a TO measurement taken from shape B for one taken from shape A using a threshold τ if $\varepsilon \geq \delta+\nu+\tau$. That is, if we index into the stored signature $\mathbf{A}^{-}$ using a measured feature value **f**, and find that $\mathbf{a}^{-} \in \mathbf{A}^{-}$ is within τ of **f**, then if $\varepsilon \geq \delta+\nu+\tau$, we are sure that **f** is not in the set $\mathbf{B}^{+}$, and thus was not obtained from shape B. Figure 3 makes the inequality relation clear geometrically.

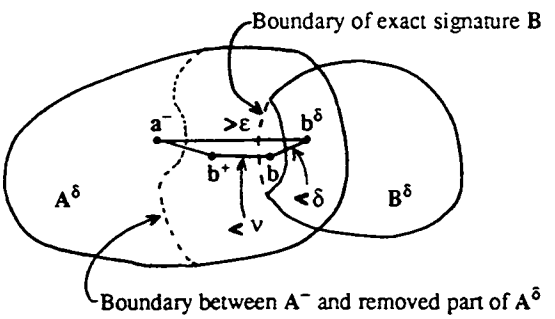


Figure 3. Schematic illustration of overlap removal for TO feature space signatures

3.2 False Negatives

If $\tau<\delta+\nu$, note that a TO placement can fail to detect a shape due to insufficient saturation. This is generally of less importance than false positives, because negative classifications are simply the deferring of a decision, resulting in extra expended time to find an instance of the shape. If we want to be sure that

every TO measurement from shape A will lead to detection (allowing false positives from other shapes), we could test for nearness of the measured point to the signature A^+ using a threshold $\tau > \delta+\nu$. In section 4 we will see that the results of section 3 are overly stringent from a statistical point of view, e.g., we can violate $\varepsilon \geq \delta+\nu+\tau$ by a significant margin and still have very few false positives.

4. EXPERIMENTS

The purpose of these experiments is to study the discriminating power of an individual TO placement. Therefore, we use no preprocessing (except range rectification) and no hypothesis verification here. Nevertheless, this "pure" approach is quite powerful in many circumstances. In the experiments TO signatures were generated for 25 surface shapes. Next, overlap removal and analysis was done, followed by recognition experiments in which a specified shape is searched for until found.

4.1 Obtaining The Signatures

TO signatures were generated for 25 surface shapes by randomly placing an order 9 TO (Fig. 1c) on synthetic range images of each shape 50,000 times. The resulting signatures were stored as discrete feature-space points, with a numerical precision of 1°. Duplicate feature vectors were removed, reducing the 50,000 points to as few as 61 points for the large cylinder and as many as 36,000 for the outside trihedral corner. Then the 3-fold symmetry of this TO (see section 2.3) was used to slightly increase the density of the signatures. These signatures correspond to A^δ of section 3, although δ was not directly controlled. The 25 shapes were chosen to include various discrimination challenges, e.g., cylinder vs torus with the same minor radius, and the hemisphere/cylinder (with C_2 discontinuity) dihedral region vs the cylinder or sphere. The following are the names and descriptions of the shapes (e ≡ TO edgelength, r ≡ radius):

0 plane	plane
1 cyl2e	cylinder; r = 2e
2 cyl2p5e	cylinder; r = 2.5e
3 sph2e	sphere; r = 2e
4 sph2p5e	sphere; radius = 2.5e
5 outcorner	outside 90° trihedral corner
6 ballcyl2e	hemisphere-capped cylinder; r=2e
7 ballcyl2p5e	hemisphere-capped cylinder; r=2.5e
8 incorner	inside 90° trihedral corner
9 pcyl2e	plane-capped cylinder; r=2e
10 pcyl2p5e	plane-capped cylinder; r=2.5e
11 dh270	270° planar dihedral (convex)
12 dh90	90° planar dihedral (concave)
13 tor2e4e	torus; r=2e, R=4e
14 tor2p5e4e	torus; r=2.5e, R=4e
15 phole2e	plane-bottomed hole; r=2e
16 phole2p5e	plane-bottomed hole; r=2.5e
17 dh225	225° planar dihedral (ramp down)
18 dh135	135° planar dihedral (ramp up)
19 peg2e	cylinder perpendicular to plane; r=2e
20 edgehole2e	cylindrical hole in plane; r=2e
21 peg2p5e	cylinder perpendicular to plane; r=2.5e
22 edgehole2p5e	cylindrical hole in plane; r=2.5e
23 thshelf	planar trihedral; 90°, 90°, 270°
24 thnotch	planar trihedral; 90°, 270°, 270°

Some are inversions of each other; (5,8), (23,24), (9,15), (10,16), (11,12), and (17,18). In these cases we generated the latter by inverting the data from the former (see section 2.3). We see in Table 1 that most pairs of the 25 shapes' initial signatures were already entirely disjoint (separation > 5°) including a cylinder (2) and the torus (14) with the same minor radius. Most ambiguous points were from shared

parts; an inside trihedral corner (8) contains an inside dihedral edge (12). Later, we will use overlap removal to make the final signatures (nearly) disjoint by design.

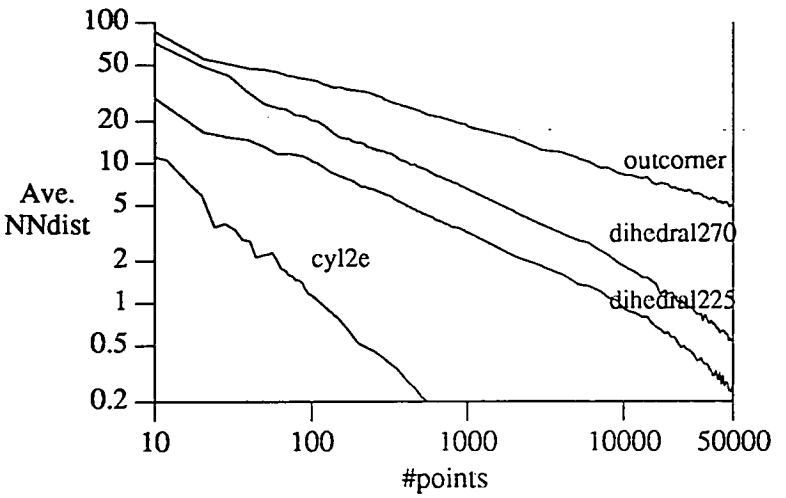


Figure 4. Saturation graphs for four representative shapes.

The results of section 3 show the importance of highly *saturated* signatures (small δ). Therefore we have studied the dependence of the degree of saturation on the number n of randomly placed order 9 TO placements, for various shapes. Figure 4 shows log/log plot of the average L_2 distance σ (in degrees) of a feature space point to its nearest neighbor versus n. We found that the dependence is approximately $\sigma = c/n^{1/k}$, where k is approximately the dimension of the signature manifold. k = .952 for the cylinder, whose manifold has dimension 1. k = 1.89 and 1.92 for the 225° and 270° dihedrals, respectively, whose manifolds have dimension 2, and k = 2.94 for the fully three dimensional outside corner. The k values are slightly lower than the corresponding dimensionality primarily because of low-dimensional subshapes (e.g., the plane (k=0) is in the n-hedral shapes). These empirical results are consistent with the observation that the density of n random points on a k-manifold is approximately proportional to n^k. Note that σ is not the same as δ; e.g., for pcyl2p5e, σ=3.4, while about .1% of new points from this shape were farther than 10° from their nearest neighbor in our signature. Thus $\delta > 10°$.

4.2 Signature Overlap

Next, pairs of signatures were processed to remove overlap ($\varepsilon = 5°$) with other shapes' signatures. Certain of these "set subtractions" were forbidden; e.g., we did not allow shapes that are parts of other shapes to be deleted. For example we did not "subtract" dh90 from plane. The full set of forbidden pairs is (5,8,9,10,11,12,15,16,17,18,19,20,21,22,23,24) from 0, (6,9,19) from 1, (7,9,10,21) from 2, 6 from 3, 7 from 4, (5,23,24) from 11, and (8,23,24) from 12, referring to the list above. Table 1 was computed before overlap removal, showing the percentage of shape A left after subtracting shape B, for all 25^2 pairs. Note that most of the signature pairs have little or no overlap, allowing easy discrimination.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0 :	--	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1 :	0	--	0	0	0	0	25	0	0	9	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
2 :	0	0	--	0	0	0	0	28	0	6	18	0	0	0	0	0	0	0	0	0	0	4	0	0	0
3 :	0	0	0	--	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4 :	0	0	0	0	--	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5 :	--	0	0	0	0	--	0	0	0	8	9	--	0	0	0	0	0	36	0	0	17	0	18	12	43
6 :	0	--	0	--	0	0	--	0	0	10	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
7 :	0	0	--	0	--	0	0	--	0	6	18	0	0	0	1	0	0	0	0	0	0	4	0	0	0
8 :	--	0	0	0	0	0	0	0	--	0	0	0	--	0	0	7	9	0	36	17	0	18	0	43	12
9 :	--	--	--	0	0	7	26	26	0	--	30	11	0	0	0	0	0	18	0	3	8	4	9	1	6
10:	--	0	--	0	0	8	0	31	0	33	--	12	0	0	2	0	0	17	0	0	9	4	10	1	6
11:	--	0	0	0	0	49	0	0	0	7	9	--	0	0	0	0	0	32	0	0	17	0	19	12	43
12:	--	0	0	0	0	0	0	0	49	0	0	0	--	0	0	7	9	0	32	17	0	19	0	43	12
13:	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	0	0	0	0	0	0	0	0	0
14:	0	0	16	0	0	0	0	3	0	0	3	0	0	0	--	0	0	0	0	0	0	1	0	0	0
15:	--	0	0	0	0	0	0	0	7	0	0	0	11	0	0	--	30	0	18	8	3	9	4	6	1
16:	--	0	0	0	0	0	0	0	8	0	0	0	12	0	0	33	--	0	17	9	0	10	4	6	1
17:	--	4	10	0	0	3	2	2	0	5	4	5	0	0	0	0	0	--	0	0	2	1	3	1	2
18:	--	0	0	0	0	0	0	0	3	0	0	0	5	0	0	5	4	0	--	2	0	3	1	2	1
19:	--	--	0	0	0	0	26	0	13	9	0	0	21	0	0	7	8	0	23	--	0	34	0	14	2
20:	--	0	0	0	0	13	0	0	0	7	8	21	0	0	0	9	0	23	1	0	--	0	34	2	14
21:	--	0	--	0	0	0	0	28	13	6	18	0	22	0	1	7	8	0	21	33	0	--	0	13	2
22:	--	0	0	0	0	13	0	0	0	7	8	22	0	0	0	6	18	21	0	0	33	0	--	2	13
23:	--	0	0	0	0	43	0	0	49	5	7	88	99	0	0	7	9	19	30	29	13	29	14	--	51
24:	--	0	0	0	0	49	0	0	43	7	9	99	88	0	0	5	7	30	18	13	29	14	29	51	--

Table 1 Overlap percentages (-- denotes 100%) for raw signatures, before overlap removal; shape indexed at left *subracted* from shape indexed at top, with separation threshold $\varepsilon=5^{\circ}$.

4.3 Recognition

Each signature was stored in bins in a three-dimensional array (using the first 3 feature components) to facilitate near-neighbor lookup. At recognition time our system randomly placed TO's on the synthetic range image of Fig. 5a, which contains instances of all 25 shapes, and labeled the locations of the TO's as ambiguous or unknown (white) or as the shape currently being sought (black). The decision rule was to note whether the distance from the TO feature vector to the nearest point in the signature at hand was less than τ, which was set to 5°. For each of the 25 shapes, with range noise initially zero to help isolate error sources, we applied the TO enough times to obtain 50 correct detections of the shape, and recorded various results such as the mean time (in TO operations) between detections (MTBD) and data on any false positive detections. The MTBD can be regarded as the ratio of the image area to the "effective area" of the shape sought, for a particular TO size e.

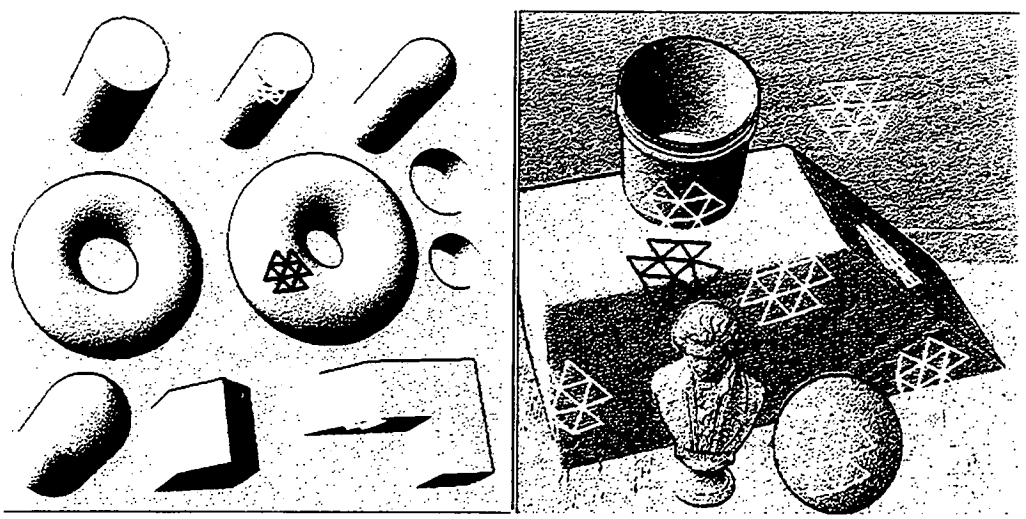


Figure 5. Noisy range images showing reliable detection of shapes by TO's; (a) tor2p5e4e is detected in 8 placements in a synthetic image. (b) dh270 is detected in 8 placements on a LIDAR image with TO edgelength = 7 cm. Both took approximately 50 milliseconds.

About 63% of the TO placements on the image of Fig. 5a *aborted* due to contact of a probe point with a jump boundary, which is locally detected by pixel disparity. This is typical for cluttered scenes and is a highly efficient substitute for image segmentation. The following results are described for non-aborted placements. The smallest MTBD values were for plane (2.74) and the tori (both about 5). The largest was for phole2e (483), the small plane/cylinder dihedral at the bottom of the small hole.

The estimated mean time between false positive detections (MTBF) was ∞ (none observed in several thousand placements) for 12 shapes and varied from 17,088 placements for incorner to 127 for pcyl2e (pcyl2p5e was falsely detected). Overall, the results showed very few false positives, which were primarily due to the lack of sufficiently exhaustive signatures, leading to failure to delete some point common to two shapes. False positives due to unknown objects are more difficult to prevent, but are fairly rare. The false positives are all due to violation of $\varepsilon \geq \delta+\nu+\tau$. If δ were 0 (exact signature), $5^\circ \geq 0+0+5^\circ$ would hold, precluding false positives. However, δ exceeds 10° in some of the signatures used, due to small portions of the exact signature B being farther than 10° from the stored signature. This causes no trouble for most pairs, because they are already separated by much more than the imposed ε before overlap removal. However, our pcyl2p5e signature has both high δ and high overlap with pcyl2e, causing the above problem. We found that we could drive sharply down the false positive incidence by either increasing ε, which had the side effect of increasing the MTBD by introducing false negatives, or by sampling more to decrease δ, particularly at the low density places in the signature. The latter is more attractive, because it does not compromise the MTBD rates significantly. We plan to pursue the construction of uniform density signatures with tightly controlled δ to address this issue.

Having discussed how to avoid shape confusions in the absence of noise, we ran some recognition examples with added range noise of peak value e/40 (edgelength e=.2, noise = .005). This yielded a peak displacement $\nu=6.5^\circ$ in feature space. For example, for the two tori, with $\tau = 5$, we found that at this noise level, well within the capabilities of various existing range sensors [15,17], there were no false positives in thousands of trials, and only 10% of the TO's falling on the tori failed to detect them. Their signatures are only 7° apart at their nearpoints. This violates $\varepsilon \geq \delta+\nu+\tau$ ($\varepsilon=7$, $\delta=1$, $\nu=6.5$, $\tau=5$), but the probability of the vectors in Fig. 3 aligning just right to cause a false positive appears small, both considering the geometry of Fig. 3 and the experiments. Repeating this with range noise e/20, we found that the large torus was mistaken for the large cylinder 10% of the time (*all* of cyl2p5e lies within 6.8° of tor2p5e4e), but all tor2e4e detections were correct. In figure 5b, we search a LIDAR image for the dihedral dh270 with range noise $\approx$ e/23 (~3mm). The MTBD is about 15 placements, and we saw no false positives. Our next step will be to seek a systematic way to set ε, τ and δ, given the average noise, for optimal performance.

5. CONCLUSIONS AND FUTURE DIRECTIONS

We have studied the ability of individual order 9 TO's to discriminate 25 surface shapes in a cluttered range image and concluded they can in many circumstances do so rapidly and with very few false positive detections. Conditions for guaranteeing this were derived. A TO can be applied and interpreted in less than 2 milliseconds on a Sparc Workstation. We plan to reduce this time by software optimization, and to study analytic approximations of TO signatures, combining pose constraints using multiple TO placements, probabilistic approaches, and other topics aimed at finding the limits of their performance.

We believe that one of the most promising potential applications of TO's is the recognition and localization of industrial parts, because (a) TO's provide the speed frequently required for economic viability, (b) range scanner technology has been rapidly achieving the requisite resolution, speed and cost, and (c) the market is broad, including automatic assembly, inspection and materials handling. Other potential applications include automatic target recognition with LADAR, a portable "seeing eye stick" for the blind that names obstacles such as "step down", pole 20° left, etc., and AGV navigation by recognizing existing indoor features. The latter might be worthwhile if the same vision system is used to recognize objects in "gopher" tasks. Also, we plan to study the effectiveness of TO's to face recognition, using a statistical approach similar to that in [3]. In general, TO's are applicable to the recognition of a wide variety of

objects for which reasonably accurate range images can be economically obtained. Exceptions include thin wire-like objects, on which the sample points of the TO are unlikely to fit, and objects for which the reflectance markings contain essential information. Some of the latter may yield to a registered range and intensity approach that unifies tripod operators with existing OCR methods.

REFERENCES

[1] Pipitone, F., "Tripod Operators for the Interpretation of Range Images", Naval Research Laboratory Memorandum Report #6780, February, 1991.

[2] Pipitone, F., and Adams, W., "Tripod Operators for Recognizing Objects in Range Images; Rapid Rejection of Library Objects", Proc. IEEE International Conf. on Robotics and Automation, pp1596-1601, Nice, France, May 1992.

[3] Pipitone, F., and Adams, W., "Rapid Recognition of Freeform Objects in Noisy Range Images Using Tripod Operators", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, New York, NY, June, 1993.

[4] Grimson, W.E.L., and Lozano-Perez, T., "Model-Based Recognition and Localization from Sparse Range or Tactile Data", The International Journal of Robotics Research, Vol. 3, No. 3, pp 3-35, Fall 1984.

[5] Grimson, W. E. L., "The Combinatorics of Object Recognition in Cluttered Environments Using Constrained Search", MIT AI Memo No. 1019, February, 1988.

[6] Gaston, P. C., and Lozano-Perez, T., "Tactile Recognition and Localization Using Object Models: The Case of Polyhedra On A Plane", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6 (3):257-265, May 1984.

[7] Oshima, M. and Shirai, Y., "Object Recognition Using Three-Dimensional Information", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(4):353-361, July, 1983.

[8] Faugeras, O.D., and Hebert, M., "A 3-D Recognition and Positioning Algorithm Using Geometrical Matching Between Primitive Surfaces", Proc. Eighth Int. Joint Conf. Artificial Intelligence, pp 996-1002, August, 1983.

[9] Bolles, R.C., and Cain, R.A., "Recognizing and Locating Partially Visible Objects: The Local-Feature-Focus Method", International Journal of Robotics Research 1(3):57-82, 1982.

[10] Lamdan, Y., and Wolfson, H.J., "Geometric Hashing: A General and Efficient Model-Based Recognition Scheme", IEEE 2nd International Conf. on Computer Vision, 1988.

[11] Stein, F., and Medioni, Gerard, "Structural Indexing: Efficient 3-D Object Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-14 (2):125-145, February, 1992.

[12] Fischler, M.A., and Bolles, R.C., "Random Sample Consensus, a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Comm. of the ACM, v24,n6, pp381-395.

[13] Besl, P.J., and Jain, R.C., "Invariant Surface Characteristics for 3D Object Recognition in Range Images", Computer Vision, Graphics, and Image Processing 33, 33-80, 1986.

[14] Bolles, R.C., and Horaud, P., "3DPO: A Three-Dimensional Part Orientation System", International Journal of Robotics Research, 5(3): 3-26, 1986.

[15] Rioux, M., Blais, F., Beraldin, J., and Boulanger, P., "Range Imaging Sensors Development at NRC Laboratories", Proc. of the Workshop on Interpretation of 3-D Scenes, pp 154-160, Nov. 27, 1989, Austin, TX, IEEE Press.

[16] Besl, P.,"Active, Optical Range Imaging Sensors", in *Machine Vision and Applications*, Springer-Verlag, Vol. 1, pp 127-152, 1988.

[17] Pipitone, F., "Extracting Elementary Surface Features Using tripod Operators", Internal Report AIC-93-036, Navy Center for Artificial intelligence, US Naval Research Lab, Washington, DC, 1993.

DEVELOPMENT OF A COMMERCIALLY VIABLE, MODULAR AUTONOMOUS ROBOTIC SYSTEMS FOR CONVERTING ANY VEHICLE TO AUTONOMOUS CONTROL

David W. Parish
Robert D. Grabbe
Omnitech Robotics, Inc.
Englewood, CO. 80110

Dr. Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena , CA. 91109

ABSTRACT

A Modular Autonomous Robotic System (MARS), consisting of a modular autonomous vehicle control system that can be retrofit on to any vehicle to convert it to autonomous control, and support a modular payload for multiple applications is being developed. The MARS design is scalable, reconfigurable, and cost effective due to the use of modern open system architecture design methodologies, including serial control bus technology to simplify system wiring and enhance scaleability. The design is augmented with modular, object oriented (C++) software implementing a hierarchy of five levels of control including teleoperated, continuous guidepath following, periodic guidepath following, absolute position autonomous navigation and relative position autonomous navigation. The autonomous vehicle control system design uses a stochastic map, and cascaded Kalman filter to fuse numerous position sensor groups, including an inertial sensor suite, a differential GPS sensor, several landmark detection sensors, and a cost effective, random access 360o scanning laser rangefinder, or LADAR. The LADAR also doubles as a high precision obstacle detection sensor. Operational capability of a rapid prototype ATV has been demonstrated including the LADAR, machine vision, and inertial sensor suite based dead-reckoning. The present effort is focused on producing a system that is commercially viable for routine autonomous patrolling of known, semi-structured environments, like environmental monitoring of chemical and petroleum refineries, exterior physical security and surveillance, perimeter patrolling, and intra-facility transport applications.

INTRODUCTION

Numerous autonomous robotic vehicles and control systems have been developed in recent years, by universities, government labs, and commercial companies. Some of these designs have found commercial applications, although broad based commercial application and market acceptance has been illusive for all but the simplest approaches like Automated Guided Vehicles (AGVs). Based on market analysis, applications for autonomous robotic vehicles are apparently plentiful, if a satisfactory system and life-cycle cost effectiveness can be met. We hypothesize that with a properly modularized, scaleable, and reconfigurable autonomous

vehicle control system architecture, that a commercially viable system can be obtained, providing sufficient cost effectiveness and return on investment to justify substantially increased market acceptance for a variety of applications. This paper will introduce our technical approach for producing such a system, followed by an overview of some of the candidate applications targeted.

TECHNICAL APPROACH

The fundamental basis of our approach for the development of the "Modular Autonomous Robotic System" or MARS.is based on recognizing that it is possible, and highly desirable, to separate the vehicle being controlled from the autonomous control system itself. This allows the use of any vehicle, whether general purpose or specially built, and by subsequently adding on a set of sensors, actuators, and "black box" electronics and control computers, an intelligent autonomous robotic vehicle can be produced for virtually any application. This approach forces a generalized architecture, and limits the use of simplifications for some applications. However, it is anticipated that the economies of scale and extra flexibility provided by having a general approach will ultimately be more advantageous. Modularizing an autonomous robotic vehicle control system requires consideration of various aspects, including:

- computing hardware, both centralized and distributed
- actuation hardware
- sensing hardware
- communication hardware and software
- control algorithms and software
- software infrastructure

The scope of these technologies is extensive, and creating an architecture that handles all aspects can seem overwhelming. However recognizing that the architecture will ultimately be embodied as one or more physical instantiations, we approached the problem by starting with a formal product specification sheet for both a low end indoor AGV-like autonomous robotic vehicle, and a high end outdoor security and surveillance autonomous robotic vehicle. These two designs were then generalized, and the common functions and interfaces defined, to provide a scaleable architecture that could meet both design targets cost effectively. We will describe some of the preliminary approaches taken and results of this effort, by describing the hardware architecture first, followed by the software architecture.

MARS Hardware Architecture

Figure 1 shows an overview of our hardware architecture for the MARS development. One of the main features of this hardware architecture is the use of a serial control bus for acquiring sensor information and controlling actuators. This "sensor / actuator bus" will now be introduced.

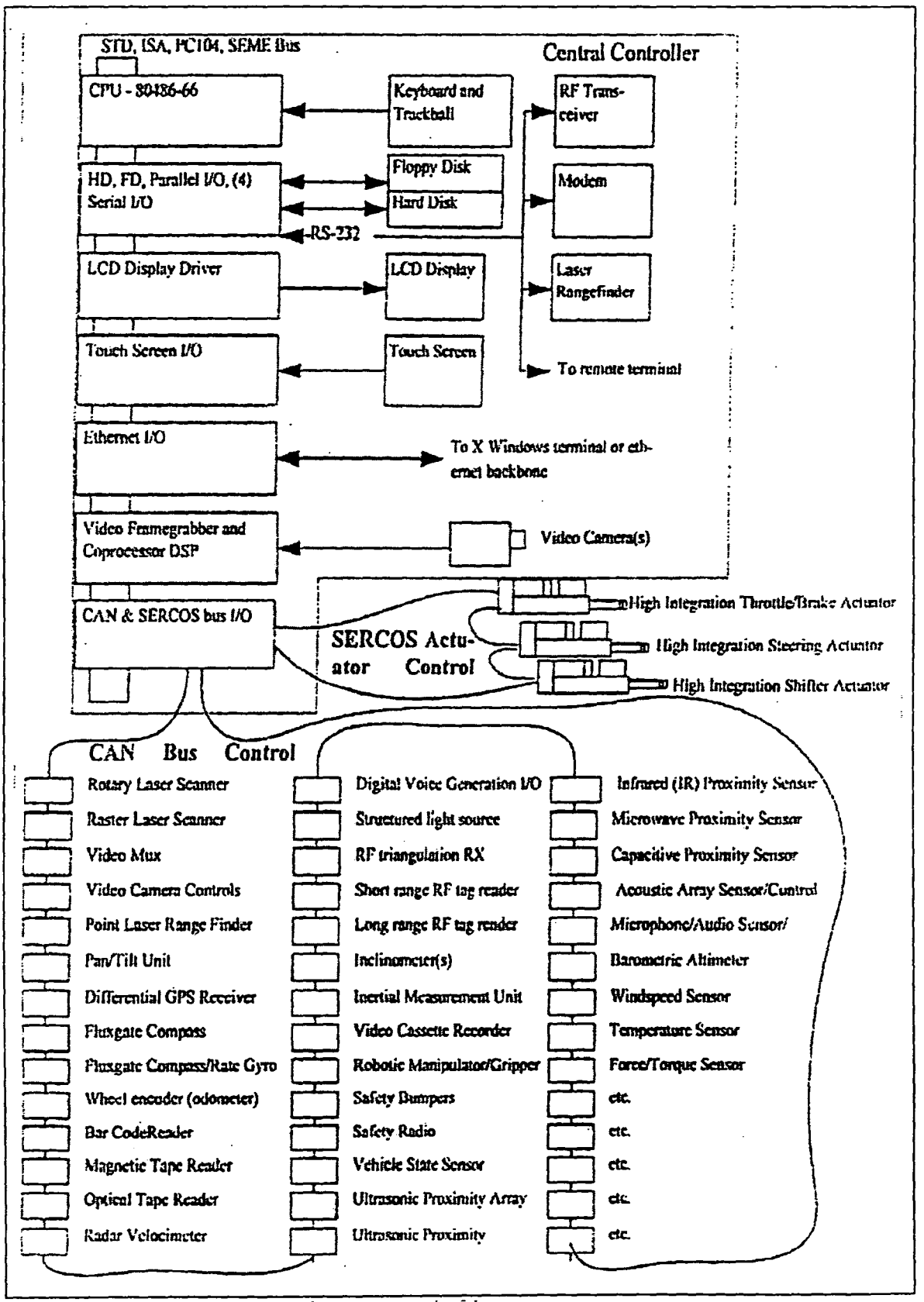


Figure 1: Overview of the MARS Hardware Architecture

Serial Control Bus Technology

The concept of a scaleable controller bus has been developed to increase the level of integration involved in embedded controller applications. Specifically, by converting from a parallel hardwired electrical interconnection approach to a distributed serial bus interconnection approach, significant savings in wiring and other raw materials, and installation costs can be gained.

The serial bus interface approach to embedded controls has been spear-headed by major automotive manufacturers to reduce in-vehicle wiring cost and size for new automobiles, for the control of electric windows, lights, accessories, and similar items. The primary automotive bus standard has been Controller Area Network, or CAN, which has been promoted by Bosch and other major manufacturers.

CAN interface chips are now available from Intel, Motorola, Phillips, Signetics and others, and the application of this approach is accelerating. These hardware components provide the physical and data layer functionality, but lack a standardized application level protocol for interoperable communications. Some work at standardization of the OSI ISO level 7 application protocol for the CAN bus has been conducted, and Omnitech Robotics is currently active in this area. Omnitech Robotics has recently completed the development of it's first CAN bus interface product, called CANAMP. CANAMP provides the following features:

- Networkable motion control
- 600 watt brush motor servo amplifier
- 32 bit DSP motion controller
- Microcontroller with BASIC interpreter
- 1 Mbaud CAN interface (ISO/DIS 11898)
- Digital amplifier parameter setting
- 8 analog inputs, 10 bit resolution
- 6 digital inputs, 2 digital outputs
- Optional analog tachometer stabilization

A photograph of a CANAMP is shown in Figure 2.

High Integration Actuators

High integration actuators refers to a design approach where the servo actuator is packaged with the necessary control components into a complete stand-alone unit. Specifically these units typically incorporate a motor, transmission device, feedback element(s), power amplifier, and logic controller with interface in a single hardware package.

The advantage of using high integration actuators include the ability to completely specify the resulting actuator's performance parameters, reduction in the total system weight, size and volume due to the elimination or minimalization of ancillary connectors, cables, etc., and the unit is convenient to use, mount, test, and replace.

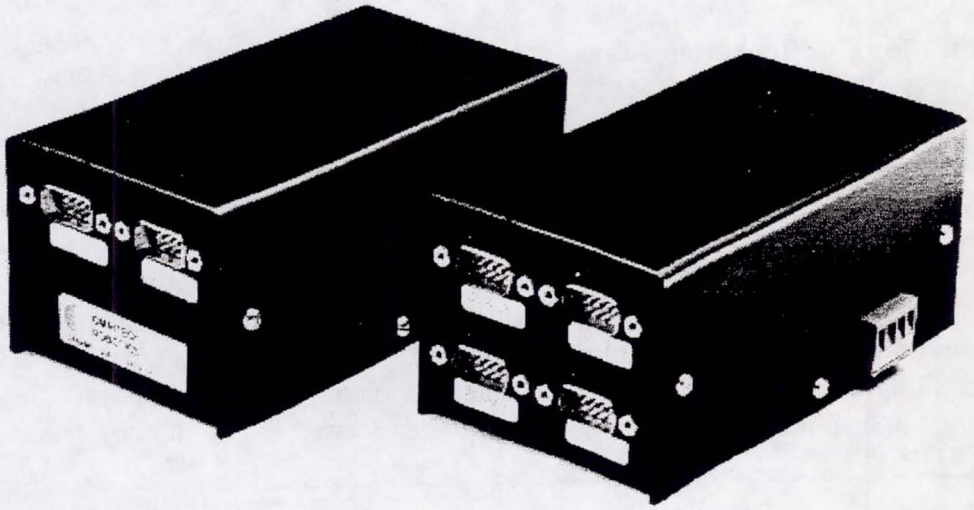


Figure 2: Photograph of CANAMP

The concept of high integration actuators is not new, in fact it has been applied to aerospace type applications for many years, due to these applications' premium on size, weight, volume, and performance. It is new to main-stream automation applications however. Figure 3 illustrates an overview of high integration actuators developed by Omnitech using the CANAMP.

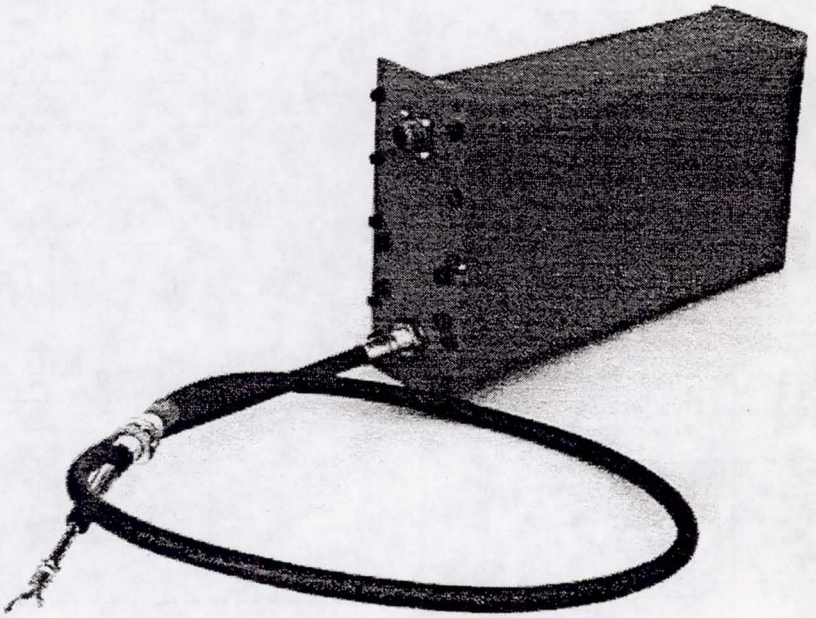


Figure 3: Photograph of the High Integration Actuator Developed by Omnitech Robotics, Inc.

MARS SOFTWARE ARCHITECTURE

Architecture for Configuration Management

The vision of this work is a development architecture that establishes and supports a managed information repository of modular hardware and software elements that can be installed into Unmanned Ground Vehicle (UGV) systems.

Figure 4 shows an overview of the combination of tools and methods to accomplish a configuration management architecture. Configuration Management will maintain documented software source code and documented hardware configuration modules available at the design, analysis, and implementation levels of system development. This would reduce the amount of re-engineering and enhance the interoperability of teleoperated / semi-autonomous systems.

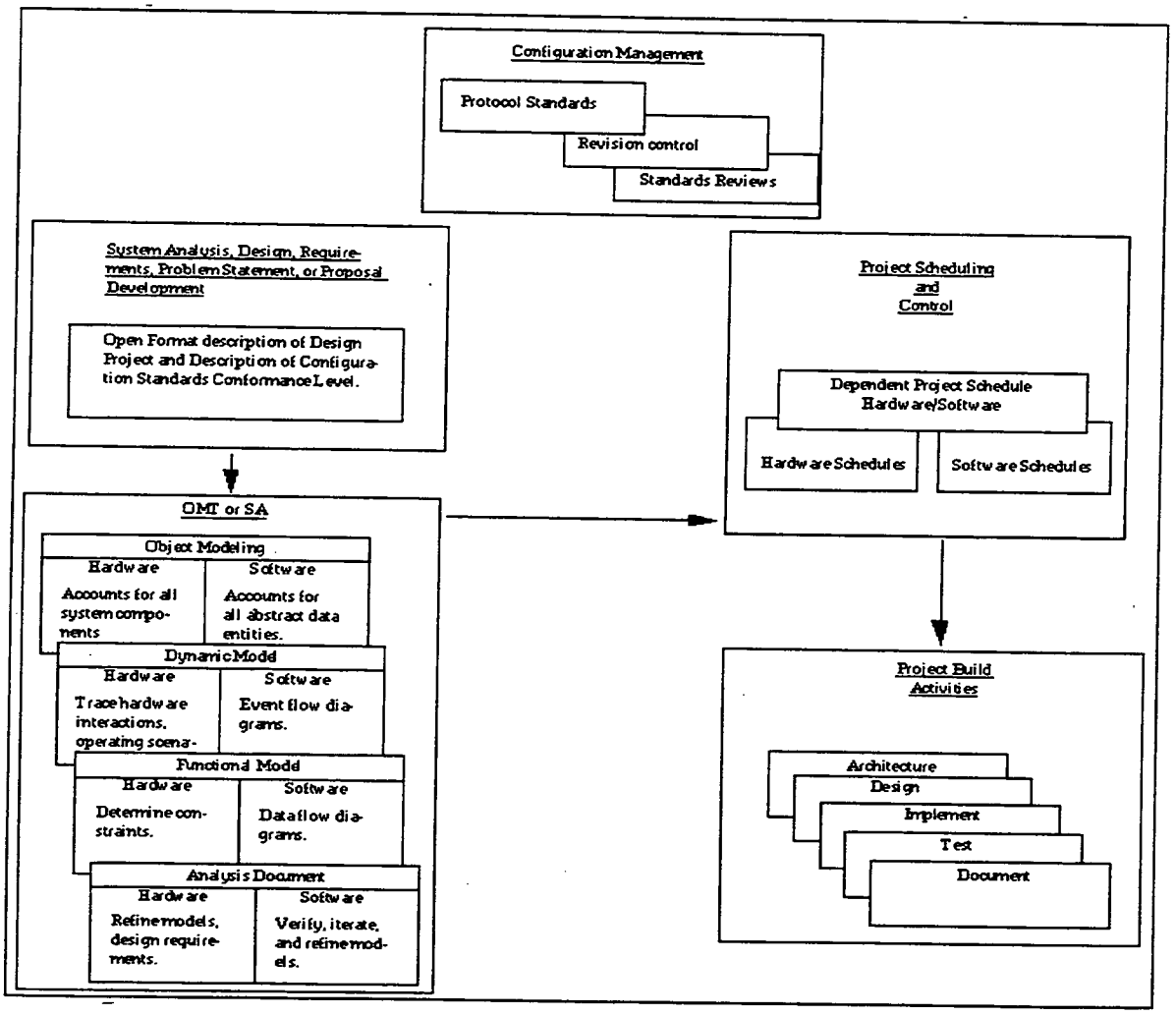


FIGURE 4: An Object Modeling Technique (OMT) emphasizes the naming of all physical objects (hardware) as well as abstract (software) objects allowing for early scheduling of the development task.

Configuration management will encompass the tasks of handling changes to software / hardware components that comprise the interface protocols that maintain intra and interoperability of UGVs developed under a joint architecture. This includes methods for evaluating changes, tracking changes, and keeping copies of the architecture that existed at various points in time. The complexity of this task requires a systematic approach that is embodied in our outline of a configuration management architecture.

Initially, our Configuration Management architecture will specify and maintain a basis set of protocol standards, software and hardware, available for the use of achieving downwardly compatible intra and interoperable UGV systems.

The support for the managed information repository will consist of automated project management tools, revision control tools, and real-time structured analysis tools such as the Object Modeling Technique (OMT) combined to manage a joint UGV configuration architecture.

Managed hardware elements will consist of controller networking such as the Controller Area Network (CAN) nodes and high integration actuators to be used on the vehicles, radio equipment, and operator control unit (OCU) equipment and standards.

Software components will consist of modularized units for communication protocols, software-hardware interface, data elements and structures, and message formats.

Documentation of the hardware / software elements will address the intended use, applied uses, and reproducible test cases and results. This type of documentation would make the evaluation of hardware / software reuse feasible.

METHODS AND TOOLS OVERVIEW

Operating Systems and Programming Software

The platform for system development would be a real-time operating system, using standardized OS services (like the IEEE defined POSIX 1003.1 and 1003.4 standards) such as the LynxOS to support threading of processes for rate monotonic process organization, and the Ada or C++ programming language.

Structured Analysis

Structured analysis tools decompose a design task into smaller, more manageable subsystems. The Object Modeling Technique (OMT) developed by Rumbaugh allows for a system decomposition based upon the objects in the system regardless of whether they are hardware components or software components.

System components from a Configuration Management repository could be evaluated and reused at this early point in the design process. Their applicability to a current system design

would be verified by the documentation giving a module's intended use, reproducible test cases, and systems currently using the module.

Project Management Tools

Automated time scheduling tools follow the object modeling process by incorporating the hardware / software objects into a scheduling process.

Revision Control Software

To maintain configuration control over systems interfacing, system modules used from a repository are compared to the originals for changes. If the changes warrant permanent inclusion into the standardized protocols while maintaining downward compatibility with other systems, then they are added to the standards and given a revision number. This aspect should not be underestimated, in fact Microsoft Corporation has stated "Version control is indispensable on team projects. It's so effective that the applications division of Microsoft has found source code version control a major competitive advantage." (Moore, 1992) [as cited from Code Complete - A practical handbook of software construction]

Object Based Design Philosophy

This section provides an overview of the documentation provided by the CADRE Paradigm Plus software's automated OMT tool that Omnitech is currently using to define our MARS and Standardized Teleoperation System software architecture.

Although the Paradigm Plus (PP) tool's most obvious application is to facilitate software engineering, it is flexible and generic enough to allow defined objects to represent hardware, software, or a combination of both. For instance, an inertial measurement unit consists of the physical hardware as well as the data structures and routines to read the data as shown in Figures 5 and 6.

An object might contain code and data structures that perform high level system tasks such as task planning. This module might be software only, and only interface with other software objects. Objects may also contain code that is designed to be downloaded onto an EEPROM and run as lower level control or monitoring software. Then the object would include both hardware and software interfaces.

At even lower levels, other objects may represent hardware functionality such as video equipment. The designer would be able to have the hardware object contain only documentation describing the hardware and how it interfaces with the system, or the object could contain code that simulates hardware behavior.

Each of these modules encapsulates data and methods (functions) used upon the data allowing more reliable code reuse.

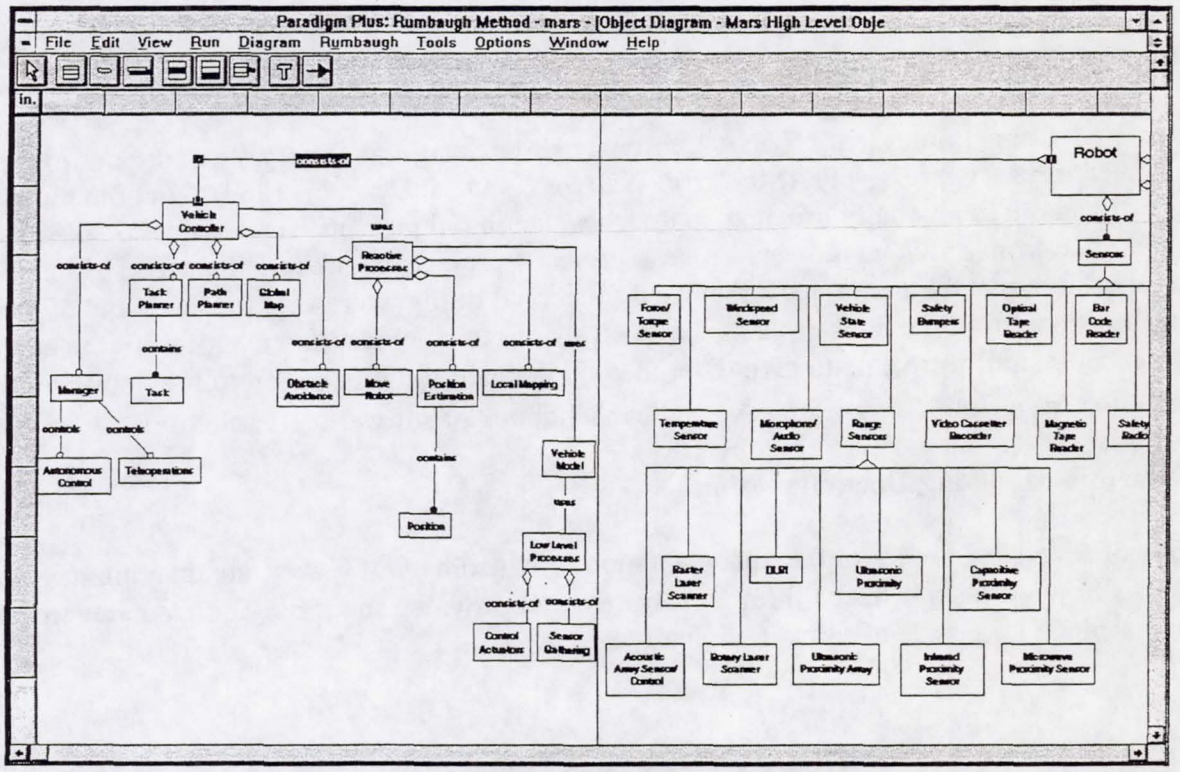


Figure 5: Sample system decomposition using OMT demonstrates the coupling of software and hardware during initial analysis.

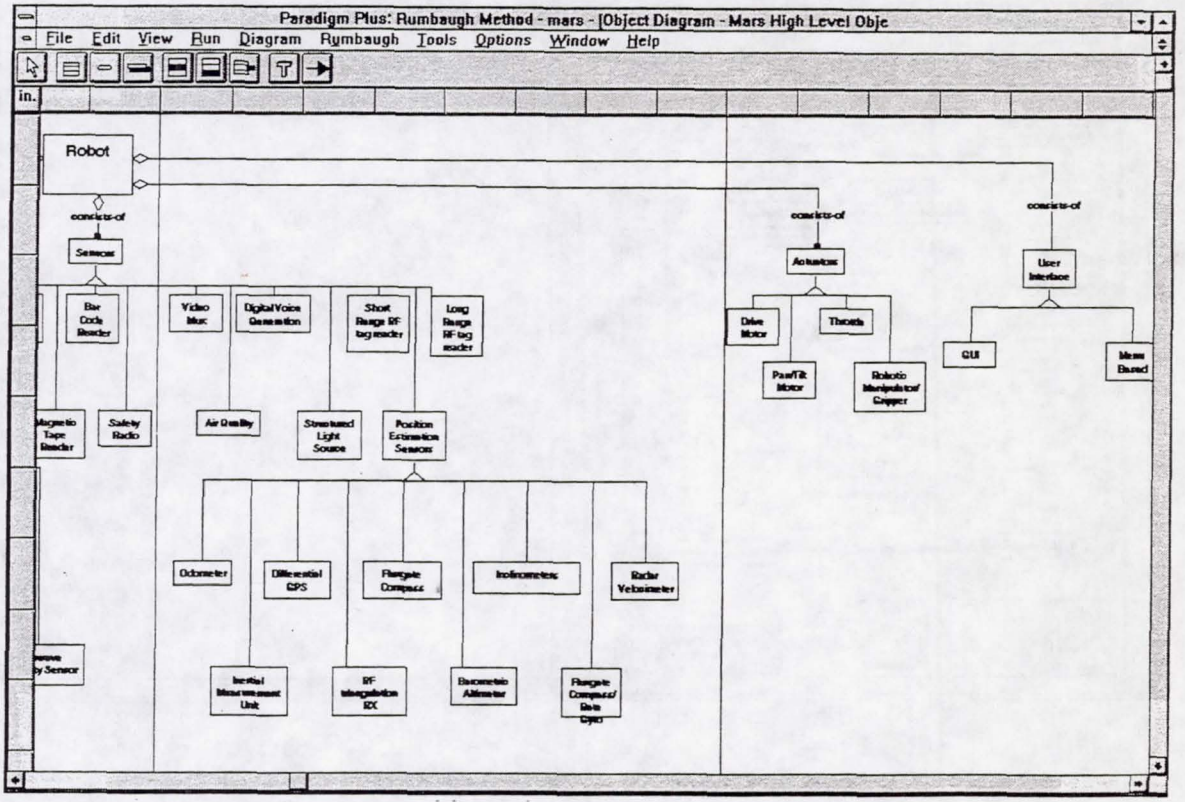


Figure 6: Sample system decomposition using OMT demonstrates the coupling of software and hardware during initial analysis.

Reusability of Code

Steve McConnel notes that "NASA's Software Engineeing Laboratory studied ten projects that pursued reuse aggressively (McGarry, Waligora, and McDermott, 1989). In both the object-based and the functionally oriented approaches, the initial projects weren't able to take much of their code from previous projects because previous projects hadn't established a sufficient data base. Subsequently, the projects that used functional design were able to take about 35 percent of their code from previous projects. Projects that used an object-based approach [the approach we recommend] were able to take more than 70 percent of their code from previous projects..." [as cited from Code Complete[1] - A practical handbook of software construction]

Dynamic Modeling Documentation

One of several examples of the documentation provided by OMT, the state diagram in Figure 7 shows how events are traced through a system and provides supporting documentation for the intended use of a piece of software or hardware.

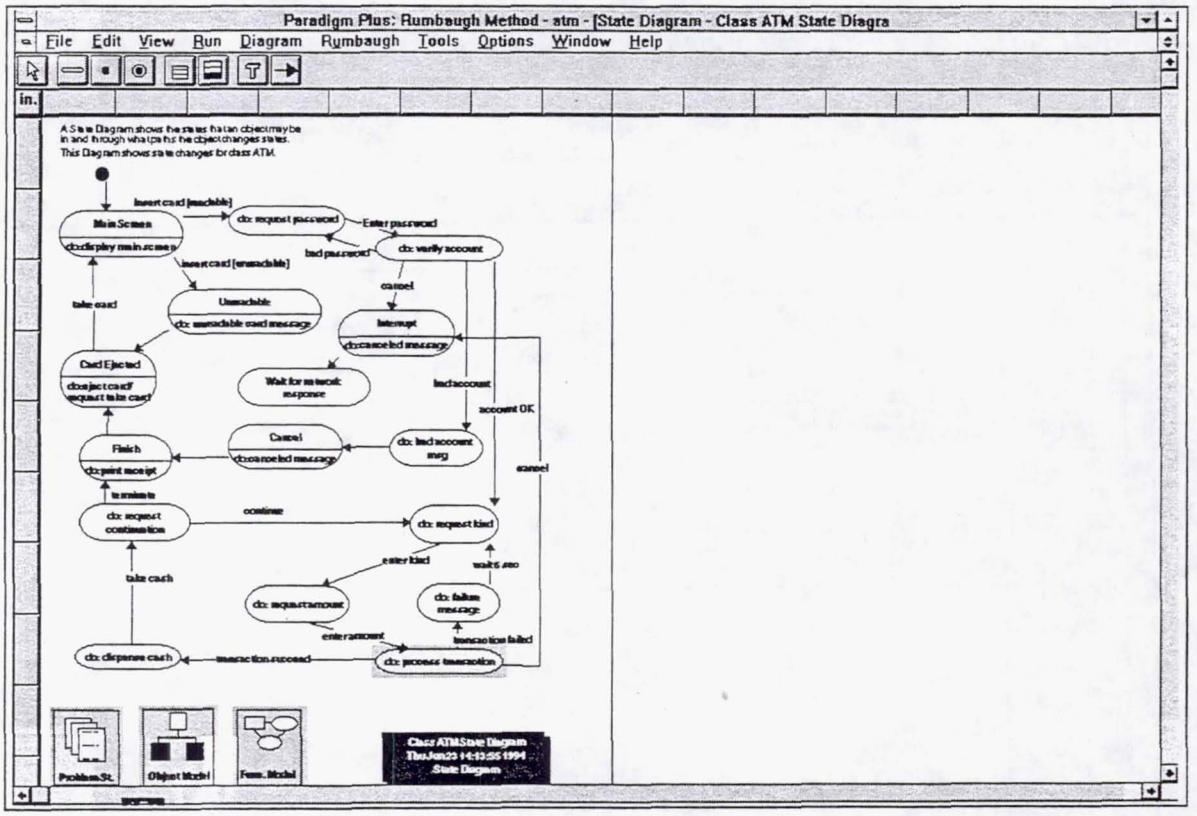


Figure 7: The dynamic modeling step in OMT provides event tracing between objects further linking the interaction between software and hardware with the use of state diagrams.

Functional System Analysis

The functional model consists of data flow diagrams and constraints as shown in Figure 8. It identifies input and output values and defines what each function does. Many structural analysis methods include this type of analysis, making a variety of methodologies interchangeable with the use of OMT. While a repository of software / hardware objects, such as communication protocols, would be documented and organized under OMT, this would not preclude other analytical methods from taking advantage of the modules documented with OMT.

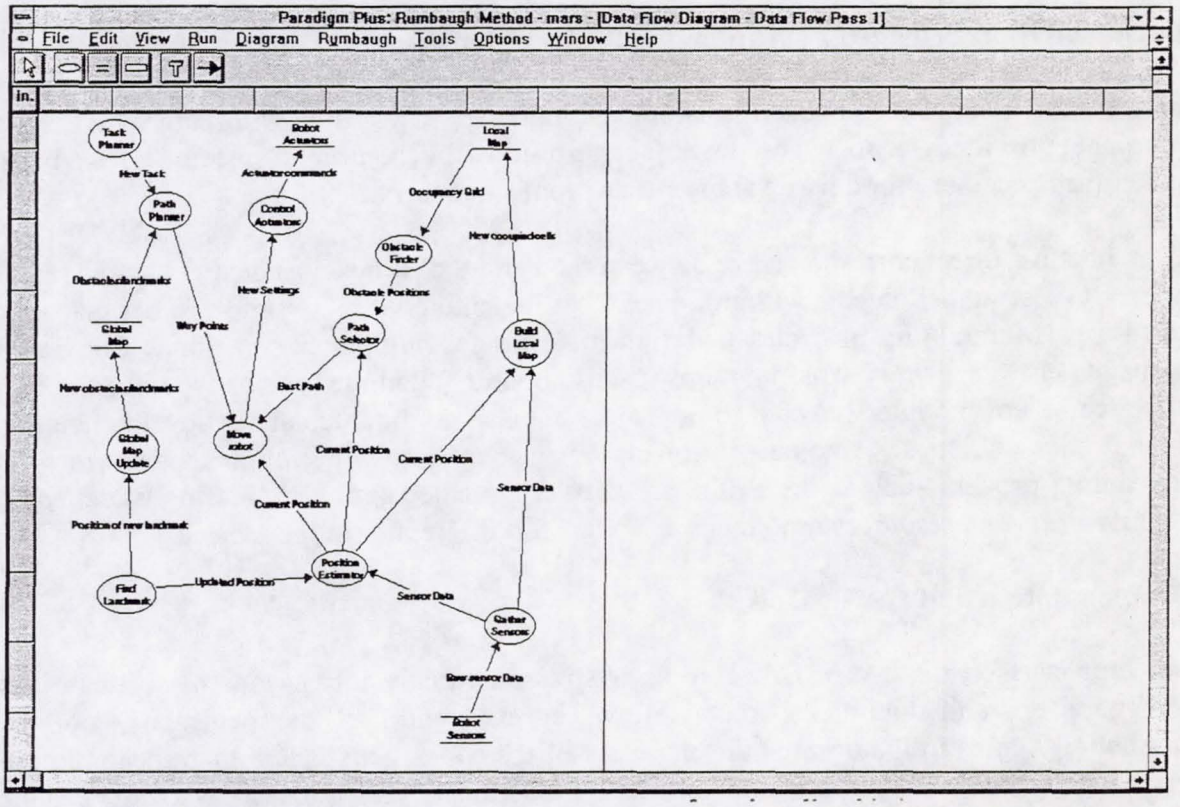


Figure 8: The OMT method then decomposes a system functionally, this is a first step in many structured analysis techniques.

Software Architecture in the OMT

Subsequent to the object modeling, dynamic modeling, and functional modeling, the overall organization of the system into subsystems comprise the system's architecture. Often the overall architecture can be based upon previous architectures. Architectures typically organize objects into a sequence of horizontal layers or and / or vertical partitions where prudent organization limits the interactions between objects, increasing modularity. An example of horizontal layering is the OSI-RM 7-layer computer communications protocol which creates definitive interfaces. An example of vertical partitioning is virtual memory management and process control in an operating system. Modularity is the key and the primary emphasis behind the plethora of high level architectures in the literature, variety of analysis tools, as well as classical programming practices.

The true test of modularity will be the reusability of components in future systems while maintaining some degree (preferably a high degree) of downward compatibility with present systems. The true test of a development architecture for configuration management will be its ability to control change in a fashion that does not stifle innovation.

Just as the evaluation of different objects within a system for reuse in future systems will be accomplished at the beginning of a design project by examining documentation, different high level architectures should be evaluated by reviewing intended use statements, reproducible test cases, and a history of implementations.

Documentation Provided by Automated Project Management Tools

In many instances, the embedded software and hardware in a system will require that a test bed setup be provided to software developers prior to the complete system. Test beds are sometimes necessary for a hardware system's proof of concept.

To maintain a direct correspondence between the variety of time dependent constraints imposed upon UGV systems, objects within a design can be entered directly into project management software . Maintaining project scheduling histories for components within a Configuration Management's repository would not only focus attention upon the time constraints and the cost effectiveness of modules selected for a project but also on modules under development having similar characteristics. Automated project management software provides many views of the scheduling process such as the PERT chart, overlaid calendars, and filtering of schedules to identify the use of resources or to make calendars for different resource sharing.

Documentation and CASE tools

Our emphasis on the use of CASE tools is stressed equally for reasons of software system configuration, and documentation to allow reuse of code. It is specifically not due to productivity gains in the original development of the code. This is less than ideal admittedly, but a reality for software development using existing CASE tools, as is evidenced in the following quote: "At the Achieving Software Quality Debates, Don Reifer reported the results of a survey on the effectiveness of CASE tools. He collected data from 45 companies in 10 industries representing over 100 million lines of code. Reifer found an average productivity gain with CASE of 9 to 12 percent, but said that not a single firm could justify the cost of CASE using the gains alone as a reason (Myers, 1992). A second report has also concluded that CASE tools have yet to meet the claims made for them in the popular press (Vessey, Jarvenpaa, and Tractinsky, 1992)." [as cited form Code Complete - A practical handbook of software construction]

Project Life Cycle and Cost Reduction

The life cycle of a project consists of several phases, each phase overlapping other phases to some extent. Good documentation during the design phase will contribute to the implementation phase and later to the maintenance phase. The cost of this documentation to support configuration management should not exceed its value however .

implementation phase and later to the maintenance phase. The cost of this documentation to support configuration management should not exceed its value however .

Justification for this approach lies in the cost advantages made through the reuse of object modules incorporated into new designs. This would allow for new design to proceed from any desired reference point in an architecture.

Figure 9: Photograph of four vehicles Omnitech Robotics is currently integrating with the MARS control system

APPLICATIONS

Omnitech intends to pursue commercialization of autonomous robotic rovers for outdore applications as a primary market area. A variety of applications exist that are suitable for the MARS approach. Figure 9 illustrates some of the candidate vehicles that Omnitech Robotics is currently working with to evaluate the use of MARS.

The initial application of MARS effort will be the development of an All Terrain Vehicle (ATV) for performing routine patrolling operations at various prisons, correctional institutions, shopping centers, schools, airports, and police support in certain neighborhoods. Other applications include surveillance and diagnostics around chemical and petroleum refineries, to support the measurement and mapping of contaminant concentrations.

A next phase of the development effort could include, loading and unloading of luggage from airplanes, loading and unloading of trucks, support to the construction industry in transporting materials and in escavation, restocking shelfs in grocery stores, identification and dismantling of explosives, forest fire fighting, road-repair vehicles, and rescue missions after tornadoes and earthquakes.

In parallel to the above commercial markets, the military application market will be pursued, as well as the components supply markets in terms of teleoperation kits for autonomous driving vehicles and tractors.

REFERENCES

1. Steve McConnell, Code Complete - A practical handbook of software construction, Microsoft Press, 1993 (ISBN 1-55615-484-4)

ACKNOWLEDGEMENTS

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, Office of Advanced Concepts and Technology. Omnitech effort was funded from NASA/JPL Small Business Innovation Research (SBIR) under contract number NAS7-1236

NAVY OMNI-DIRECTIONAL VEHICLE (ODV) DEVELOPMENT AND TECHNOLOGY TRANSFER OPPORTUNITIES

Hillery McGowen
Coastal Systems Station, Dahlgren Division
Naval Surface Warfare Center
Panama City, FL 32407

ABSTRACT

The omni-directional vehicle (ODV) development program sponsored by the Office of Naval Research at the Coastal Systems Station has investigated the application of ODV technology to the Navy shipboard environment. ODV technology was demonstrated to be applicable to the shipboard environment and shown to have potential to overcome conditions of reduced traction, ship motion, decks heeled at high angles, obstacles, and confined spaces. Under the Navy program, ODV technology was investigated and a series of experimental vehicles built and successfully tested. This paper focuses on the Navy's demonstration of the capability of the ODV to operate under demanding environmental conditions, ODV mechanical simplicity, and ODV adaptability for high level teleoperated or autonomous operations. Potential commercial applications are suggested, including applications in the industrial manufacturing and warehousing environment, remotely controlled or autonomous platforms employed in nuclear facilities or for hazardous waste cleanup, and other operations that require the movement and precise positioning of large, heavy objects. The Navy intends to implement a Cooperative Research and Development Agreement (CRADA) for further development and transfer of technology to the private sector.

NAVY ODV DEVELOPMENT

To provide material handling improvements in the demanding shipboard environment, the Navy ODV exploratory development program has investigated the application of ODV technology to cargo and ordnance handling. ODV technology was first received by the Navy in the form of the Cadillac-Gage vehicle and had the potential to overcome the limitations of existing equipment under conditions of reduced traction, ship motion, decks heeled at high angles, obstacles, and restricted spaces. Development concerns focused on omni wheel complexity, footprint pressure, and traction, vehicle control, reliability, cost, maintainability, and autonomous and teleoperated operation. Under the Navy program, a series of experimental vehicles was built and tested.

The omni wheel was originally patented by a Swedish inventor, Bengt Ilon, in 1973. The omni wheel and its operating principle is shown in Figures 1 and 2. The ODV is a four wheel drive system in which each non-steerable wheels has its own drive motor. The omni-directional wheel allows the vehicle to travel in any direction, rotate about its axis, or to do both simultaneously (see Figure 3). Omni wheels are not steered because the plane of rotation is fixed in reference to the chassis. Mounted at a 45 degree angle to the wheel plane of rotation are a series of passive elliptical rollers. When a wheel is rotated, the resulting motion tends to move the wheel on the ground at a 45 degree angle to its plane of rotation. By adding the individual motion created by each wheel, the vehicle can move in any desired direction. Vehicle speed and direction are controlled by an operator using a three axis joystick. Responding to the joystick, microprocessor-based algorithms control the rotation of each wheel to achieve the desired vehicle motion.

ODV AND ALL-WHEEL-STEERED VEHICLE COMPARISONS

Other high maneuverability drive systems, including existing Navy multi-directional all-wheel-steered (AWS) forklift trucks, were compared with the ODV. This investigation found that the AWS vehicle was the only alternative drive system capable of producing maneuverability approaching that of the ODV, and capable of being used in the shipboard environment. Two configurations of AWS sideloading forklift trucks have been used on Navy ships. Both Navy AWS multi-directional vehicles had limitations not shared by the ODV including:

a. Limited maneuvering and traction capability
b. Mechanically complex
c. Difficult to operate and maintain
d. Scuffing damage to tires and deck non-skid surfaces

Complexity. The omni wheel is more complex than a conventional wheel. However, when a conventional wheel is coupled to a suspension, drive, and steering system to produce an AWS vehicle, the vehicle is substantially more complex than the ODV. Thus when addressing complexity and reliability, complete system complexity must be considered. With the exception of the wheel, the design and fabrication of an ODV is straightforward with most components available off-the-shelf. Mechanically, the ODV is an uncomplicated system: four identical drive units; the omni wheels; a simple suspension; a battery or diesel engine power source. The relative simplicity of the ODV mechanical and electrical systems is illustrated in Figure 4.

Maintenance. Maintenance comparison of ODV and AWS vehicles must be considered on the basis of a complete vehicle rather than on the wheel alone. The 20-year old Cadillac-Gage vehicle has never suffered a failure associated with the wheel, even after having been operated under adverse conditions in sand, water, and mud. Three factors concerning omni wheel wear and failure should be noted. First, the omni wheel is essentially non-scuffing. Second, the time that the individual roller contacts the ground is only a portion of a revolution of the wheel. Third, a vehicle can be operated with a drive unit inoperative if the wheel is free to rotate or if a wheel is removed.

The ODV drive system (seen in Figure 4) consists of four identical drive units; omni wheels, suspension system, a battery, and control system. Taken together these components produce a simple, low maintenance system. The omni wheel is more complex than a conventional wheel; however, the rest of the ODV is robust and uncomplicated. Conversely, steering, drive, and control mechanisms of the AWS vehicle are complex and present a continuing maintenance burden.

NAVY ODV PLATFORMS

A family of conceptual ordnance and cargo handling vehicles based on the omni drive system have been developed by the Navy. A 8,000-pound capacity sideloading forklift truck with 18-inch omni wheels was selected for advanced development. This vehicle (Figure 5), the Omni-Directional Ordnance Handler (ODOH), was designed to transport long, heavy missiles and other ordnance down narrow Navy ship passageways.

A small ODV model (Figure 6) was fabricated to test the electronic control system and to illustrate the omni wheel principle of operation. The model clearly demonstrated the simplicity of operating an ODV, its ability to negotiate obstacles and to maneuver in extremely confined spaces.

To provide a full-scale Omni-Directional Test Platform (ODTP) (Figures 7a and 7b), the original Cadillac-Gage ODV, (Figure 7c), was modified by: removing the operator seat and other structure, leaving only a simple transport platform; adding an electronic control system to resolve the limitations of the original mechanical-hydraulic control system; and replacing the gasoline engine with a battery powered, eight-horsepower electric motor so that the vehicle could be operated below decks in enclosed spaces.

A proof-of-concept Multi-Purpose Autonomous Vehicle (MPAV) Platform (Figure 8) was developed for the Naval Air Warfare Center (NAWC) during 1992-93 to explore use as a universal platform on aircraft carriers. Currently the MPAV-ODV is operator controlled using a pendant but, in the future, high level autonomous control will be incorporated to reduce manpower, improve productivity, and relieve personnel of exposure to hazardous environments. The MPAV-ODV is 118 inches long, 50 inches wide, 21 inches high and has 18-inch omni wheels. The vehicle weighs 4,700 pounds and will transport a 5,000 pound payload. When equipped with application hardware, the MPAV-ODV will perform various missions such as cargo handling, weapons loading, jet aircraft engine handling, and deck cleaning/deicing. The ODV is also applicable to missions that require remotely controlled platforms as might be used for explosive ordnance disposal, nuclear/chemical washdown, and firefighting (see Figure 9).

Recently, the MPAV-ODV was successfully demonstrated shoreside and on an aircraft carrier. In the most demanding test, the MPAV-ODV was equipped with a jet aircraft engine handling adaptor (Figure 10). Tests showed significant improvement in precision positioning and movement compared to the standard engine handler trailer. The ODV operator successfully engaged the engine for removal within six minutes, whereas a comparative test using an existing standard engine handler trailer was terminated after trying eighteen minutes without success to engage an engine. Further tests of the vehicle with mission adapter hardware are planned for 1995.

A smaller version of the MPAV-ODV, the Omni-Directional Vehicle, Demonstration Model (ODV-DM) shown in Figure 11 was designed and fabricated for NAWC as a test vehicle for the development of High Level Control Systems and sensors. The ODV-DM, equipped with 12-inch wheels is 32 inches wide, 50 inches long, 21 inches high, weighs 550 pounds, and can transport a 250 pound payload. This design further illustrates ODV simplicity and the use of off-the-shelf components. As seen here the drive train consists of only three components, the drive motor, gear box, and omni wheel. The only unique component is the omni wheel.

OMNI-DIRECTIONAL VEHICLE TEST PROGRAM

Tests were conducted to demonstrate the capability of the ODTP and other ODVs to operate under real world conditions as follows:

a. Vehicle control and traction tests on ice in a skating rink and under cold weather conditions.
b. Operational tests on a dynamic ship motion simulator.

c. Static tilt table tests to validate the capability to operate at extreme deck angles.
d. Missile handling in a simulated shipboard environment to demonstrate the capability of an ODV to transport long, heavy loads in restricted spaces.

These tests validated the capability of the ODV to operate in confined spaces, over obstacles, and under conditions of reduced traction on wet/icy decks with ship motion and decks heeled at high angles.

Traction Test. The full-scale ODTP, small ODV model, and a conventional forklift truck were operated on ice in a skating rink (Figure 12) to evaluate the capability of an ODV to maneuver and to
retain control under low traction conditions, and to compare performance with that of a conventional forklift truck. The tests indicated that the ODVs had adequate traction to be fully controllable and capable of performing all maneuvers. The ODTP significantly outperformed the forklift truck in terms of traction and controllability.

Army Cold Weather Traction Test. Tests conducted by the Army further confirmed the ability of the ODV to operate on ice, snow, and wet surfaces. From the test results the following conclusions were drawn:

a. The omni wheel significantly outperforms a conventional, non-pneumatic tire in driving traction and control on a smooth ice surface.
b. The omni-directional wheel shows a broad peak traction region on a drawbar-pull versus slip curve, with a (desirable) slow tapering off of force after the peak value is reached. The fact that the wheel has a broad range of slip levels where peak and near-peak traction occurs makes it very "user friendly" and forgiving to operators.

Ship Motion Test. During simulated ship motion tests the ODTP, with a 5,000 pound load, was operated forward, backward, sideways, and rotated in place. Motions to a maximum of five degrees roll and three degrees pitch were induced. Neither the vehicle nor the operator experienced any control problems.

Static Tilt Table Tests. Tilt table tests, as required for shipboard forklift trucks, were conducted to validate the capability of the ODTP to operate without skidding (Figure 13). The ODTP maintained position without slip under the following extreme conditions of static tilt:

- 26 degrees, ODTP perpendicular to slope
- 23 degrees, ODTP parallel to slope
- 19 degrees, ODTP 45 degrees to slope

Missile Canister Handling. Missile canister handling demonstrations were conducted with the ODTP to validate the capability of an ODV to transport long, heavy, loads in restricted spaces (Figure 14). The ODTP performed this mission efficiently and without difficulty.

MPAV-ODV Functional Test. In September 1992, tests of the NAWC MPAV-ODV were conducted to evaluate vehicle's capability in terms of maneuverability, maintainability, and operability (see Figure 15). Fifteen different individuals operated the vehicle after a brief

description of the controls. After several minutes of practice, operators easily completed complicated movements which would have been difficult even for an experienced forklift truck operator. These tests proved that the vehicle is capable of complex motions while retaining a simple, user friendly operator interface which is one of the primary benefits of the ODV. This factor is extremely important for manual operator control, teleoperated, or autonomous operations.

SUMMARY

Under the U.S. Navy program, five ODVs have been developed and tested, ranging in size from a small model to two full scale vehicles capable of transporting 5,000 pound payloads on Navy ships. Tests indicate that the ODV outperforms existing materials handling vehicles employed for warehouse and shipboard operations. The advantages of the ODV are:

More maneuverable and capable of greater precision in obtaining a given position than other vehicle types leading to increased productivity and higher storage densities in restricted spaces.

Excellent traction, braking, and obstacle negotiation capabilities to accommodate the shipboard environment.

Mechanically simple with fewer components; easier to design, fabricate, and maintain than AWS vehicles as the ODV eliminates complicated steering and drive mechanisms.

Simpler control systems, whether implemented in autonomous, teleoperated, or manually driven vehicle.

Unconstrained omni-directional movement, coupled with ease of implementing high level control systems, will enable the coordinated movement (slaving) of ODVs to transport oversize or heavy loads as well as the coordinated movement of two or more vehicles to perform a single task.

POTENTIAL ODV APPLICATIONS

The ODV is well suited for use in the industrial manufacturing and warehousing environment. As remotely controlled or autonomous platforms, ODVs could be employed in nuclear facilities or for hazardous waste cleanup. The ODV transporter is applicable to operations that require the movement and precise positioning of large, heavy objects in a confined space. Other potential commercial applications of the ODV technology include autonomous security and inspection platforms, mobile television and motion picture platforms, and similar applications. A significant consumer application is a highly maneuverable, easily operated, omni-directional wheelchair.

The Navy intends to implement a Cooperative Research and Development Agreement (CRADA) with non-Government organizations previously involved in ODV development to continue development of the ODV and to transfer the technology to the private sector.

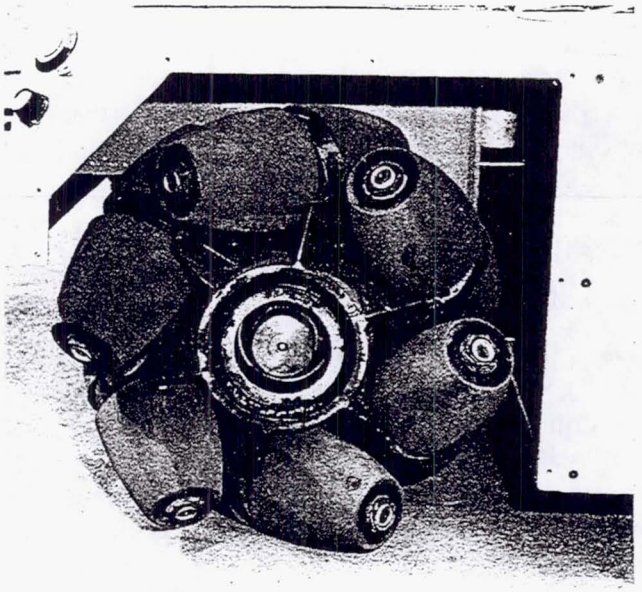


FIGURE 1. OMNI-DIRECTIONAL WHEEL

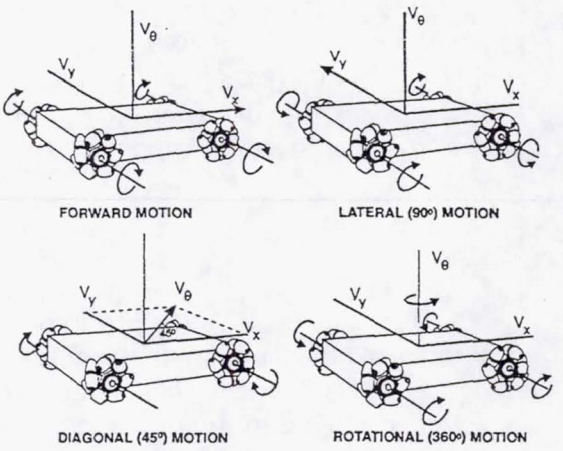


FIGURE 2. OMNI-DIRECTIONAL WHEEL OPERATING PRINCIPLE

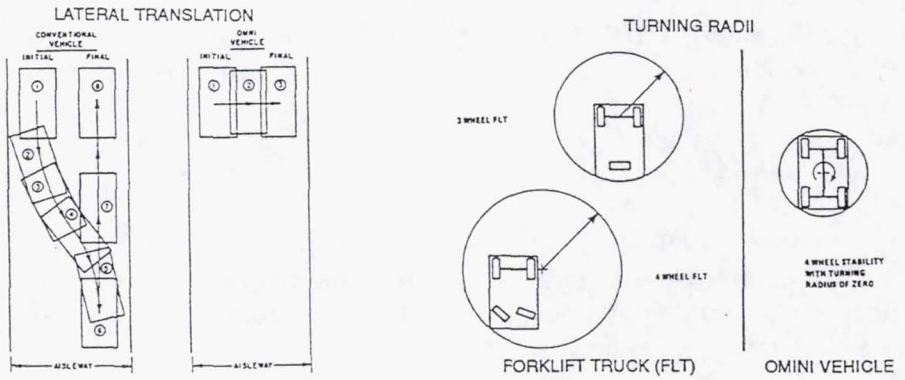


FIGURE 3. OMNI-DIRECTIONAL WHEEL MANEUVERABILITY VERSUS CONVENTIONAL VEHICLES

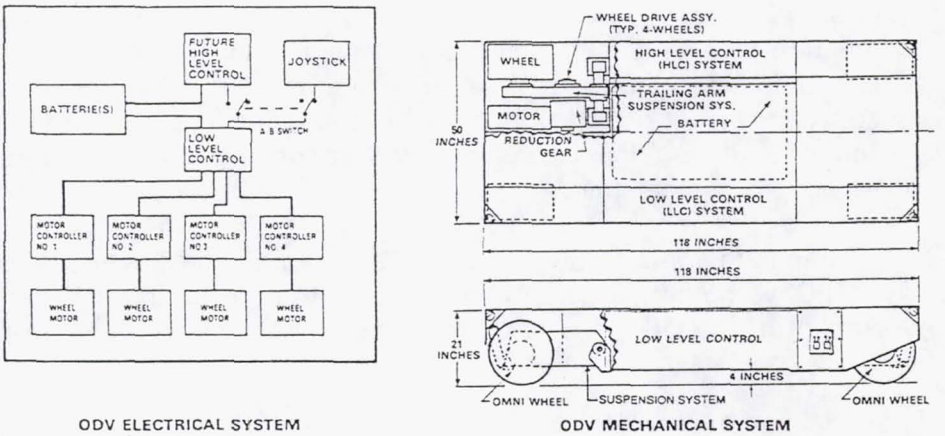


FIGURE 4. ODV MECHANICAL AND ELECTRICAL SYSTEMS

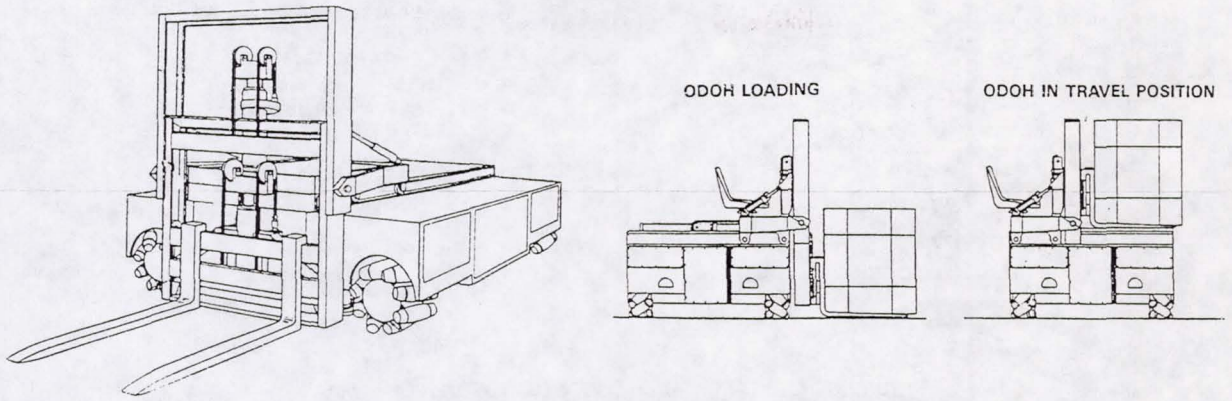


FIGURE 5. OMNI-DIRECTIONAL ORDNANCE HANDLER (ODOH)

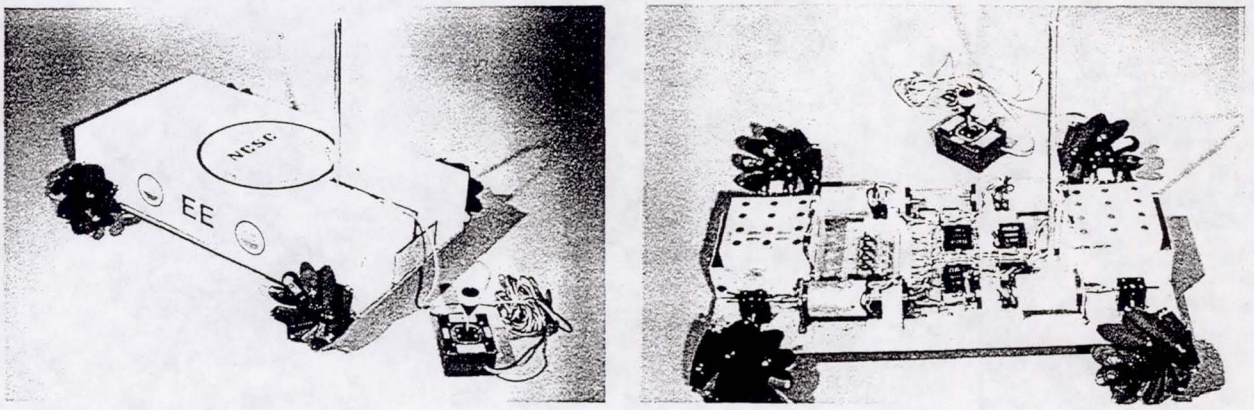


FIGURE 6. OMNI-DIRECTIONAL VEHICLE MODEL

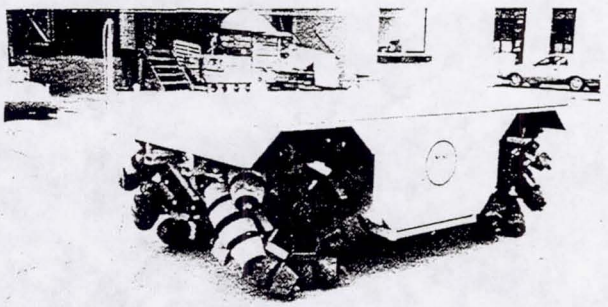


FIGURE 7A. OMNI-DIRECTIONAL TEST PLATFORM (ODTP)

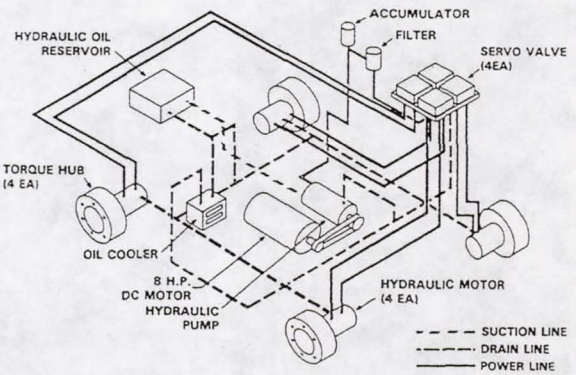


FIGURE 7B. OMNI-DIRECTIONAL TEST PLATFORM (ODTP) DIAGRAM

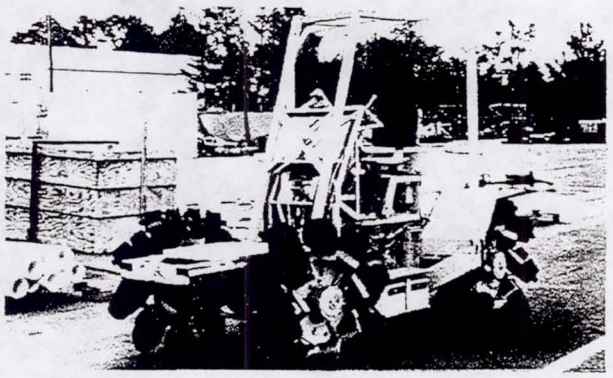
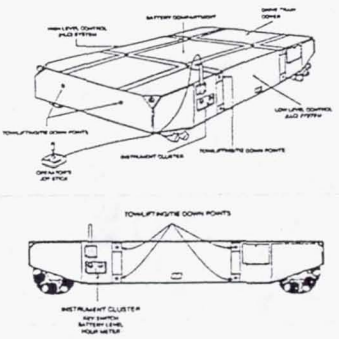
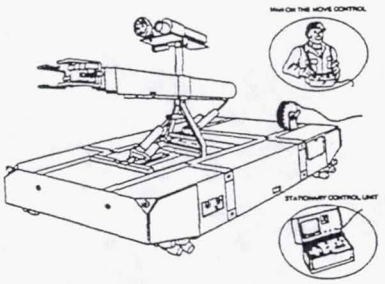


FIGURE 7C. CADILLAC-GAGE VEHICLE

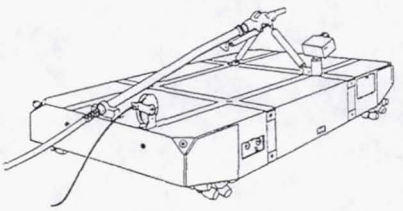


- DIMENSION - 50 INCHES (W), 118 INCHES (L), 21 INCHES (H)
- CAPACITY - 4000 LB PAYLOAD
- VEHICLE WEIGHT - 4700 LBS
- SPEED - 264 FEET / MIN (3.0 MPH)
- RAMP CAPABILITY - 15 DEGS
- OBSTACLE NEGOTIATION - 3 INCH FORWARD, 1.5 INCH LATERAL
- BATTERY POWERED WITH BRUSHLESS DC DRIVE (TRACTION) MOTORS
- ENDURANCE - 8 HRS
- VEHICLE CONTROL - PENDENT OR AUTONOMOUS (HLC) SYSTEM

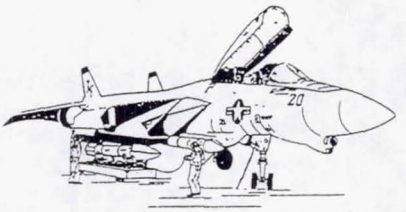
FIGURE 8. MULTI-PURPOSE AUTONOMOUS VEHICLE (MPAV)



HAZMAT PLATFORM

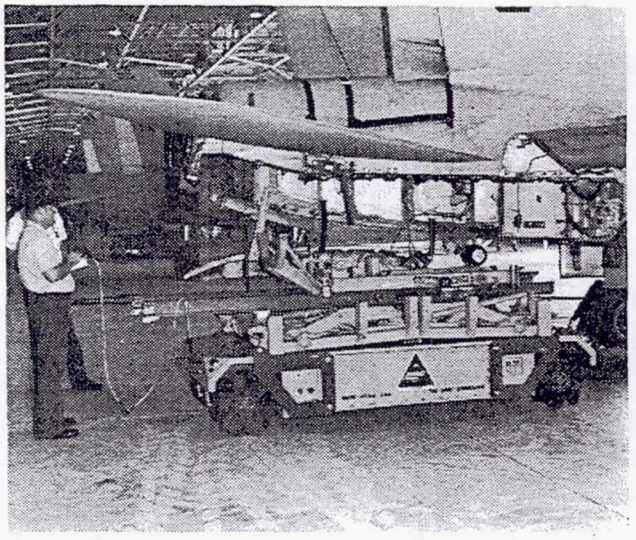


REMOTE FIRE FIGHTING



WEAPON LOADING

FIGURE 9. MPAV-ODV MISSIONS



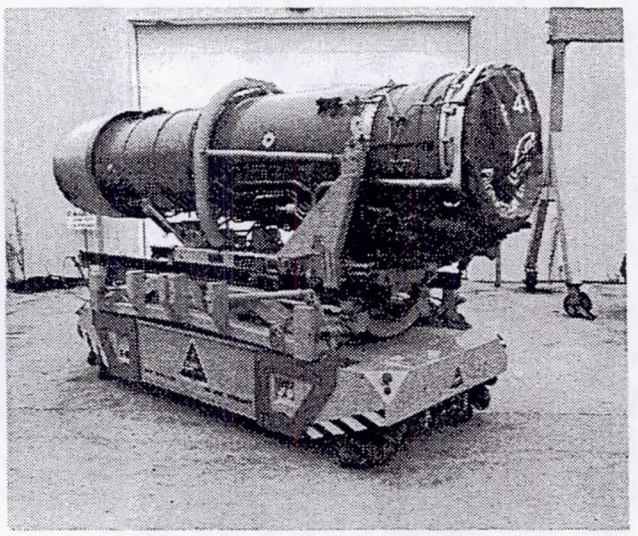


FIGURE 10. MPAV-ODV JET AIRCRAFT ENGINE HANDLING

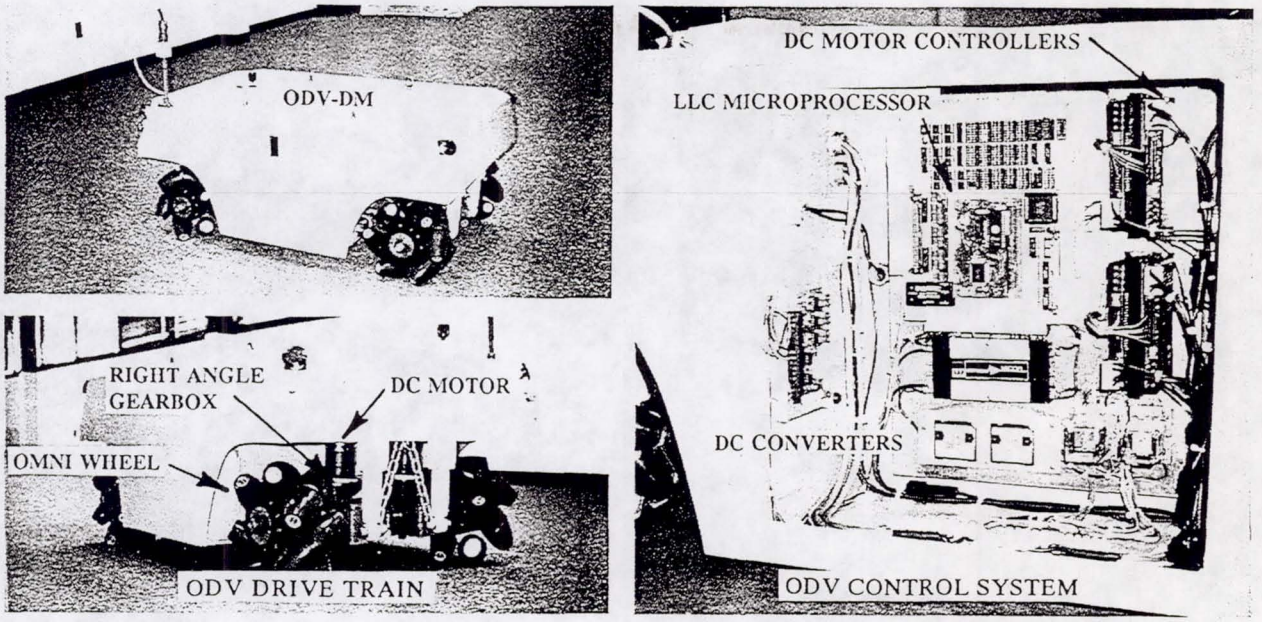


FIGURE 11. OMNI-DIRECTIONAL VEHICLE DEMONSTRATION MODEL (ODV-DM)

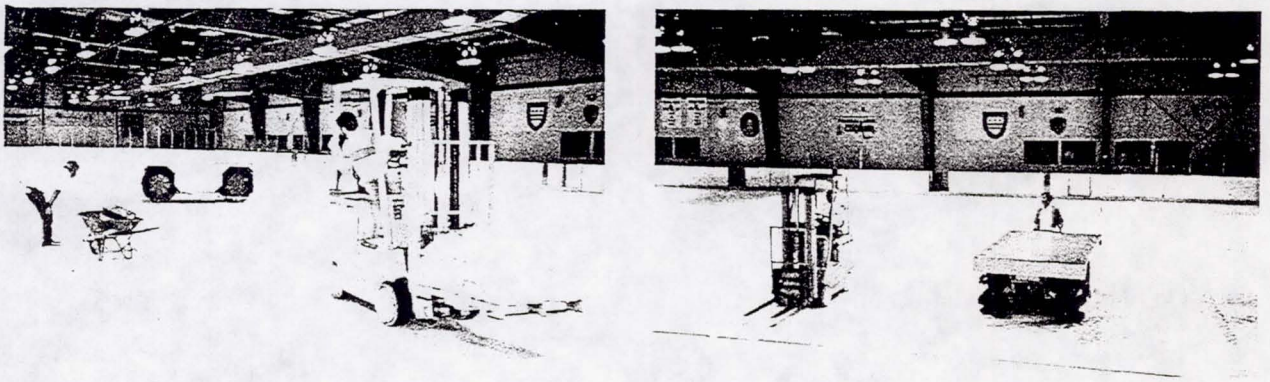


FIGURE 12. ODV ICE RINK TRACTION TESTS

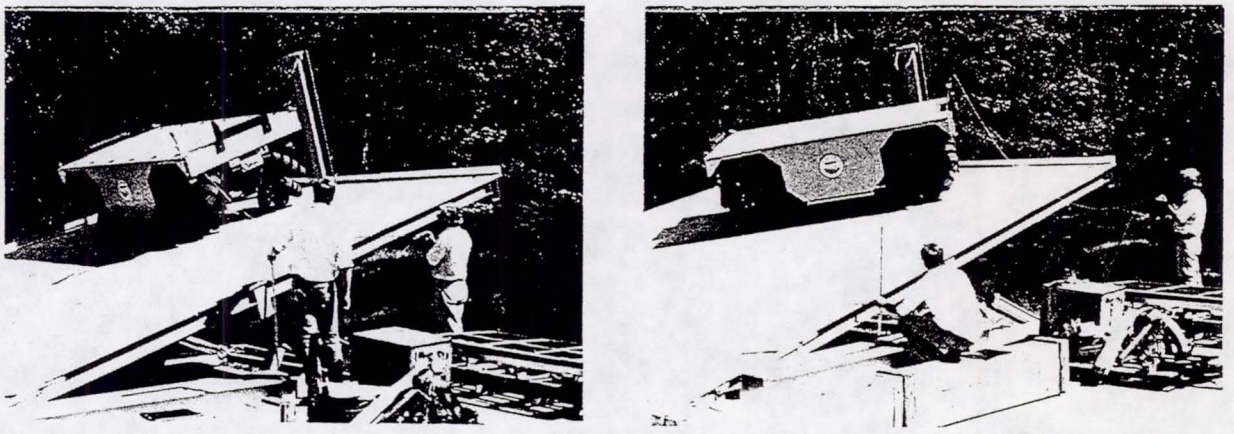


FIGURE 13. ODV TILT TABLE TESTS

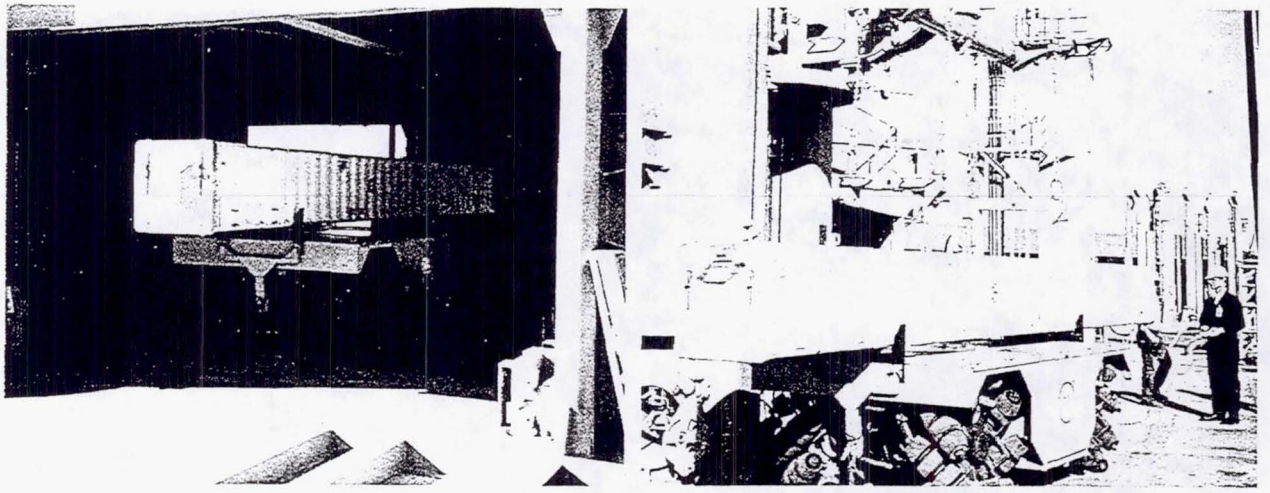
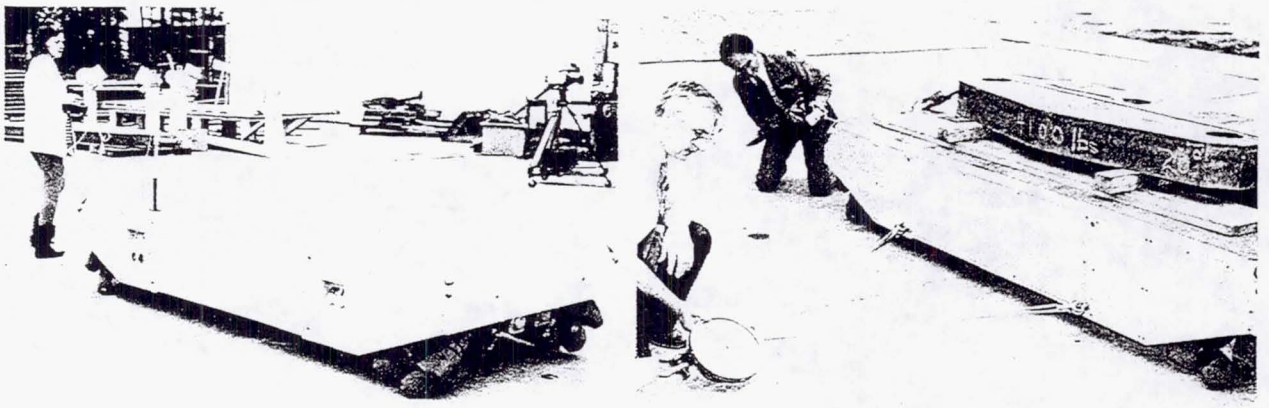
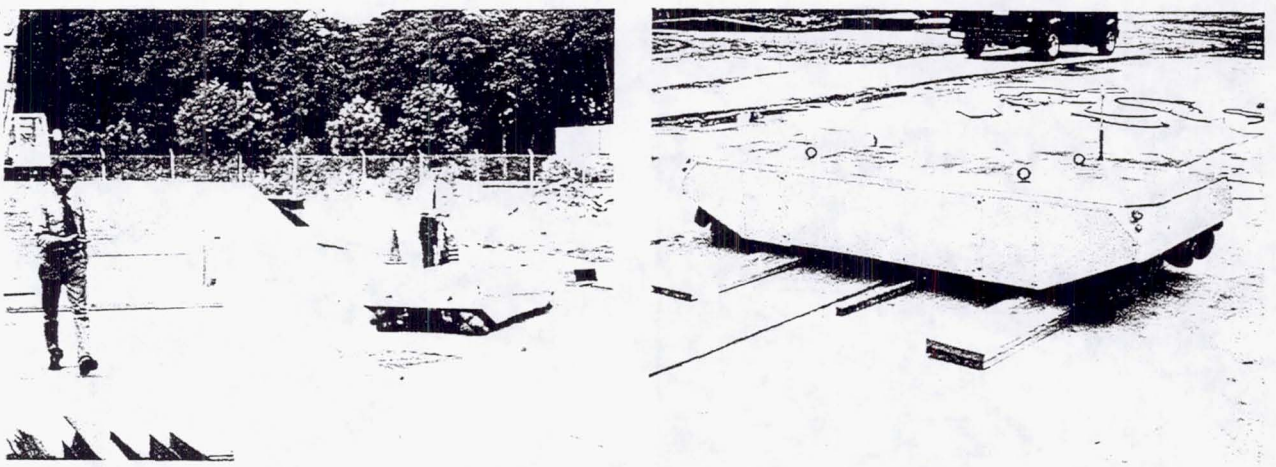


FIGURE 14. ODV MISSILE CANISTER HANDLING TEST



PRECISION MANEUVERING DEMONSTRATION

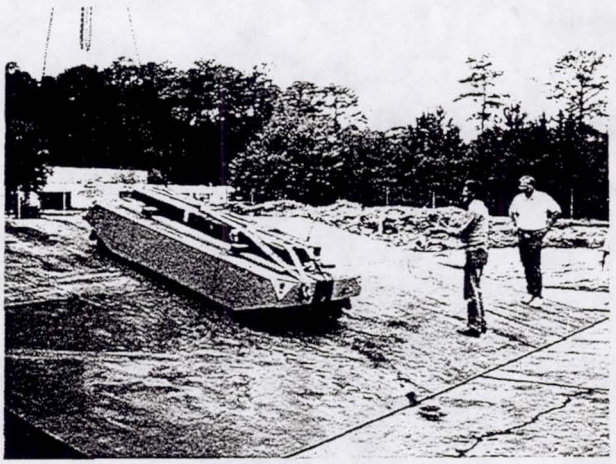
DRAWBAR PULL MEASUREMENT AT 90°



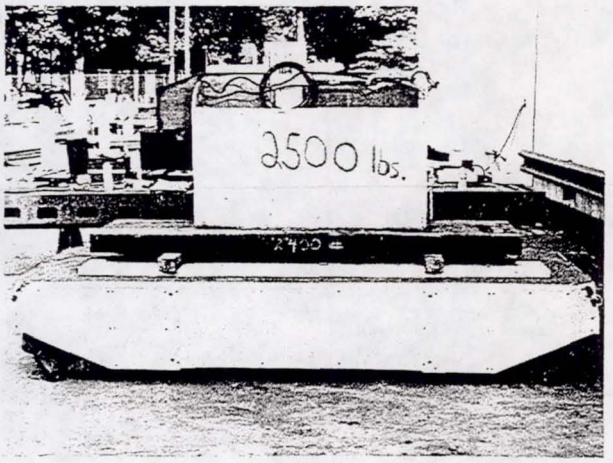
SPEED MEASUREMENT AT 45°

NEGOTIATING OBSTACLES
(3" BOARD, WIRE, ROPE, CHAIN)

FIGURE 15. MPAV-ODV FUNCTIONAL TEST



MANEUVERING LOADED ON 15° RAMP



OVERLOAD TESTS (4900 lbs)

FIGURE 15. MPAV-ODV FUNCTIONAL TEST (Continued)

An Integrated Fault Tolerant Robotic Controller System for High Reliability and Safety

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

Kam S. Tso and Myron Hecht
SoHaR Incorporated
Beverly Hills, CA 90211

ABSTRACT

This paper describes the concepts and features of a fault-tolerant intelligent robotic control system being developed for applications that require high dependability (reliability, availability, and safety). The system consists of two major elements: a *fault tolerant controller* and an *operator workstation*. The fault tolerant controller uses a strategy which allows for detection and recovery of hardware, operating system, and application software failures. It has a recovery time of less than 40 milliseconds, a period short enough for nearly all real time applications. Protection against higher level unsafe events (e.g., collisions) is provided by software resident in a separate operator workstation which includes features to predict collisions and reduce the human workload thereby reducing errors and enhancing safety. The fault tolerant controller can be used by itself in a wide variety of applications in industry, process control, and communications. The controller in combination with the operator workstation can be applied to robotic applications such as spaceborne extravehiclar actvities, hazardous materials handling, inspection and maintenance of high value items (e.g., space vehicles, reactor internals, or aircraft), medicine, and other tasks where a robot system failure poses a significant risk to life or property.

1. INTRODUCTION

Uncertain failure behavior and the potential for unsafe events have been significant concerns for the application of robots in some critical applications. Earlier work [1, 2, 3, 4] has defined two classes of potential hazards in robot contollers: those caused by *system level* failures, i.e., failure of the robot controller itself and those caused by *task level* failures, i.e., semantically valid commands input to the robot controller which in fact could result in collisions, unsafe movements, or other hazards.

Figure 1 shows a top level view of the fault tolerant robotics controller system based on this view. There are two main subsystems: an operator workstation and the controller subsystem. The operator workstation is a Silicon Graphics IRIS Crimson/VGXT running the IRIX 5.2 operating system. The controller subsystem consists of a pair of redundant VME chassis running Motorola 68040 Single Board Computers under the VxWorks Release 5.1 real time multitasking kernel. The operator workstation interfaces with the controller subsystem over Ethernet. The two redundant VME chassis are interfaced to each other using a high speed BIT 3 bus adapter, and each have a separate connection to the robot arm. For the purposes of test and evaluation, these systems are being integrated with the Robotics Research 7 DOF arm located at the Jet Propulsion Laboratory (JPL).

The approach integrates a sophisticated user interface with hardware and software fault tolerance resident in the controller. Three key technologies have emerged from this work:

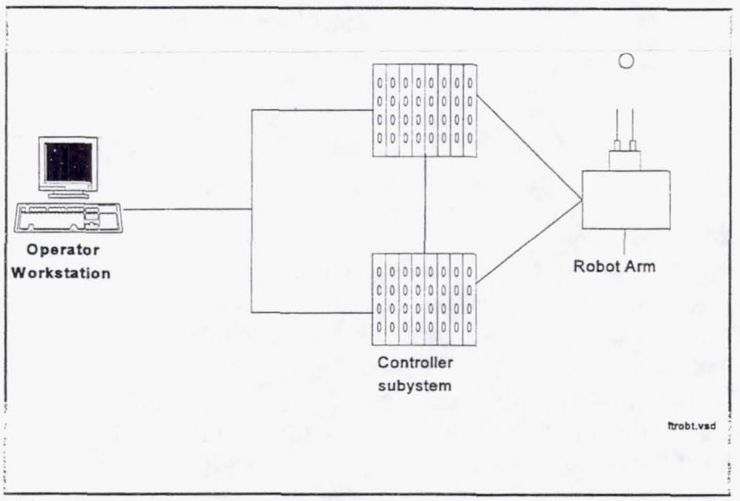


Figure 1. Top Level System View

- Comprehensive and high performance (40 msec recovery time) fault tolerance implemented as an executive layer on the VxWorks kernel.

- A robot controller implemented using Distributed Recovery Blocks, and

- A user interface allowing for the creation of complex event sequences and providing operator feedback through a simulation of the robot arm.

The first two technologies are associated with the controller subsystem and are described in the following section. Section 3 describes some aspects of the user interface. Section 4 describes potential follow-on uses of these technologies either as separate entities or as an integrated system.

2. FAULT TOLERANT CONTROLLER

The fault tolerant controller accepts Cartesian coordinates and translates them into joint angles which are then output to servo controllers within the robot. Fault tolerance for hardware, software, and communications failures is necessary in the controller because it must respond rapidly to failures. A real-time fault-tolerant distributed architecture called the Extended Distributed Recovery Block (EDRB) [5] handles controller failures. The underlying fault tolerance algorithms and mechanisms are based on the distributed recovery block [6] which is in turn based on the classical recovery block [7] with real time extensions. Figure 2 is a top level diagram of a robotic control system which incorporates the EDRB.

In the terminology of the EDRB, the replicated controller computers are collectively referred to as an operational node pair. One member of the node pair, called the active node, provides control and processing for the robot and sensors. The other node, referred to as the shadow, operates as a standby. The active and shadow nodes exchange frequent periodic status messages, called heartbeats, over redundant communication lines as both an indication of their states of health and

for state data updates. If the shadow node senses the absence of its companion active node's heartbeat, it will promote itself to the active status after verifying concurrence with a supervisor. The supervisor in this system is a task resident on the controller workstation. This concurrence is required in order to prevent a spurious takeover due to faulty communications in the shadow node or a false alarm due to a transient anomaly. After taking over, the newly promoted active node will induce a hardware reset and software reload of the failed node in the hope of restoring it to backup status. The supervisor itself need not be replicated because it is needed only to assist in recovery; the EDRB can function in steady state without the supervisor.

Figure 2 shows how distributed recovery blocks are implemented in the EDRB. Within both the active and shadow nodes are two versions of the task execution software, referred to as the primary and alternate routines. Under normal circumstances, the primary routine is run on the active node while the alternate routine is concurrently run on the shadow. The primary routine is coded to provide the greatest functionality, accuracy, and performance. The alternate routine provides less functionality and performance, but is coded to optimize reliability.

The EDRB tolerates a broad range of hardware, system software, and application failures including:

- Robotic task execution software not outputting a correct setpoint by the required deadline (detected by means of acceptance tests, timers, and recovered from using the alternate routine).

- Hardware or system software failures (detected by means of information encoding, timers, and recovered from by switching to the redundant processor).

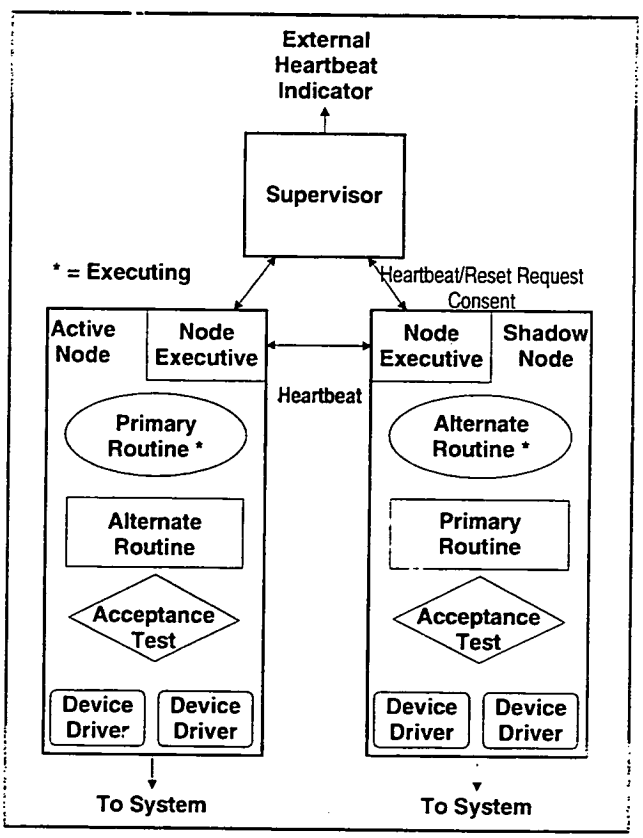


Figure 2. Software Architecture of EDRB

• Communications link failures (detected by means of encoding, and recovered from using retransmission, and redundant communication links).

• Spurious recovery actions (avoided by means of the supervisor and consideration of failure histories in the node executive).

One of the most important characteristics of the EDRB for robotic control applications is its fast response and recovery time. The algorithms used in the EDRB fault detection and recovery modules are fast because they do not require any kind of rollback. This characteristic is achieved by executing the primary and alternate routines in parallel. The EDRB provides the general framework of the primary routine, alternate routine, and acceptance test which work together to tolerate software faults. However, it is necessary to define application specific algorithms for the primary and alternate routines, as well as to define acceptance tests which dependably distinguish between correct and incorrect output.

For the controller application, diversity between the the primary and alternate routines is achieved using (1) the Jacobian pseudoinverse [8] which has good tracking but cannot handle singularity, and (2) the damped least square [9] which is singularity robust but has bad tracking near singularity. The primary routine can use the Jacobian pseudoinverse to ensure good tracking, while the alternate routine, based on the damped least square, would be used when the primary fails to handle singularity. Because many of the software failures in these routines are likely to be in the mathematical operations, the alternate routine will rely on lookup tables instead of math library functions provided by the compiler.

The acceptance test is the single most critical element of the EDRB. The two potential failure modes are rejection of a correct result or acceptance of an incorrect result. In order to avoid these failure modes, the acceptance test must be both simple so that it can be thoroughly verified and general so that it provides an adequate level of coverage and safety. While these are rigorous requirements, they are feasible in robotic applications. In the free motion example, the acceptance test will determine (1) that the next setpoint is closer to the destination than the previous, (2) the difference between the observed joint angles and the command joint angles are with in an acceptable range, (3) the command joint angles are not close to joint limits, and (4) the observed force/torque values are within an acceptable range of the gravitational force of the grasped object.

3 . OPERATOR WORKSTATION

The operator workstation provides both a user interface and additional functionality to prevent collisions and other unsafe actions. This functionality is necessary because although the fault tolerant controller can prevent unsafe actions caused by loss of controller hardware or software (i.e., *system* level hazards), it can not prevent collisions or other undesirable events which are caused by deliberate and planned motion (i.e., *task* level hazards).

The following capabilities are being developed for the operator workstation to enhance safety:

•*Graphical User Interface:* A sophisticated graphical user interface allows operators to easily command and monitor robot motions. The interface design is intended to minimize confusion and fatigue thereby prevent operator errors. Figure 3 shows an example of this interface tailored for remote surface inspection.

• Collision Prediction: The operator will have the ability to identify volumes of prohibited motion. Prior to performing a command, a robotic simulator running in the workstation will determine whether the motion will cause movement through prohibited areas.

• Position monitoring: As the robot is performing its tasks, a robot simulator processes position data generated by the robot controller to ensure that constraints are not being violated. The position is also output graphically to allow the operator to view the robot position. This is particularly important in applications where the operator is not in the immediate vicinity of the robot being controlled. Figure 3 shows an example of this display.

• Automatic Sequencing: Automated sequencing allows the development of complex trajectories based on a set of robot primitive commands. Sequence files reduce the number of required commands and hence of operator errors. They also allow an evaluation of the entire sequence of commands, not simply each command primitive.

4 . TECHNOLOGY APPLICATIONS AND UTILIZATION

We are pursuing commercialization of the fault tolerance robotics technology developed in this research in two areas:

Safe and Reliable controller for critical robotics applications: In order for robotics and its related software to be utilized in critical applications, it must be based on a safe, reliable, and dependable platform. The high performance real-time fault tolerant contoller developed under this research is being marketed to robotics systems developers, medical equipment manufacturers, and the commercial nuclear industry. The emphasis in these applications is fast response and recovery such that failures will be masked and normal operations can not be disrupted. Besides that, the systems will have high fault coverage, which means tolerance to wide varieties of failures caused by hardware, software, and communications.

Low cost high availability for widespread automation:. In non-critical but continuously operating manufacturing workcells, the fault tolerant controller system will benefit the operation through reduced downtime. A lower cost version of this platform is being developed for applications where there is a need for low cost but highly available automated systems.

5. CONCLUSION

The technological innovations emerging from this research are:

Fault tolerant platform: The high performance fault tolerant platform based on the EDRB provides broad scope fault tolerance (both hardware and software) together with fast recovery times (less than 40 msec). While intended for robotics applications, this fault tolerant platform can be used in other high performance real time control applications. The system is currently hosted to a VME-based multiprocessor system using 68040 single board computers and the VxWorks real-time kernel.

Fault Tolerant Robot Controller: The fault tolerant robotic controller uses physical redundancy and diverse application software to provide a broad scope of hardware and software fault tolerance. The fault tolerant techniques developed in this research for building dependable robotic control systems can be used in applications which require a high degree of reliability and safety, such as servicing and inspection tasks in Space Station Freedom, maintenance and waste cleanup tasks in nuclear facilities, and patient monitoring and tending tasks in medical facilities.

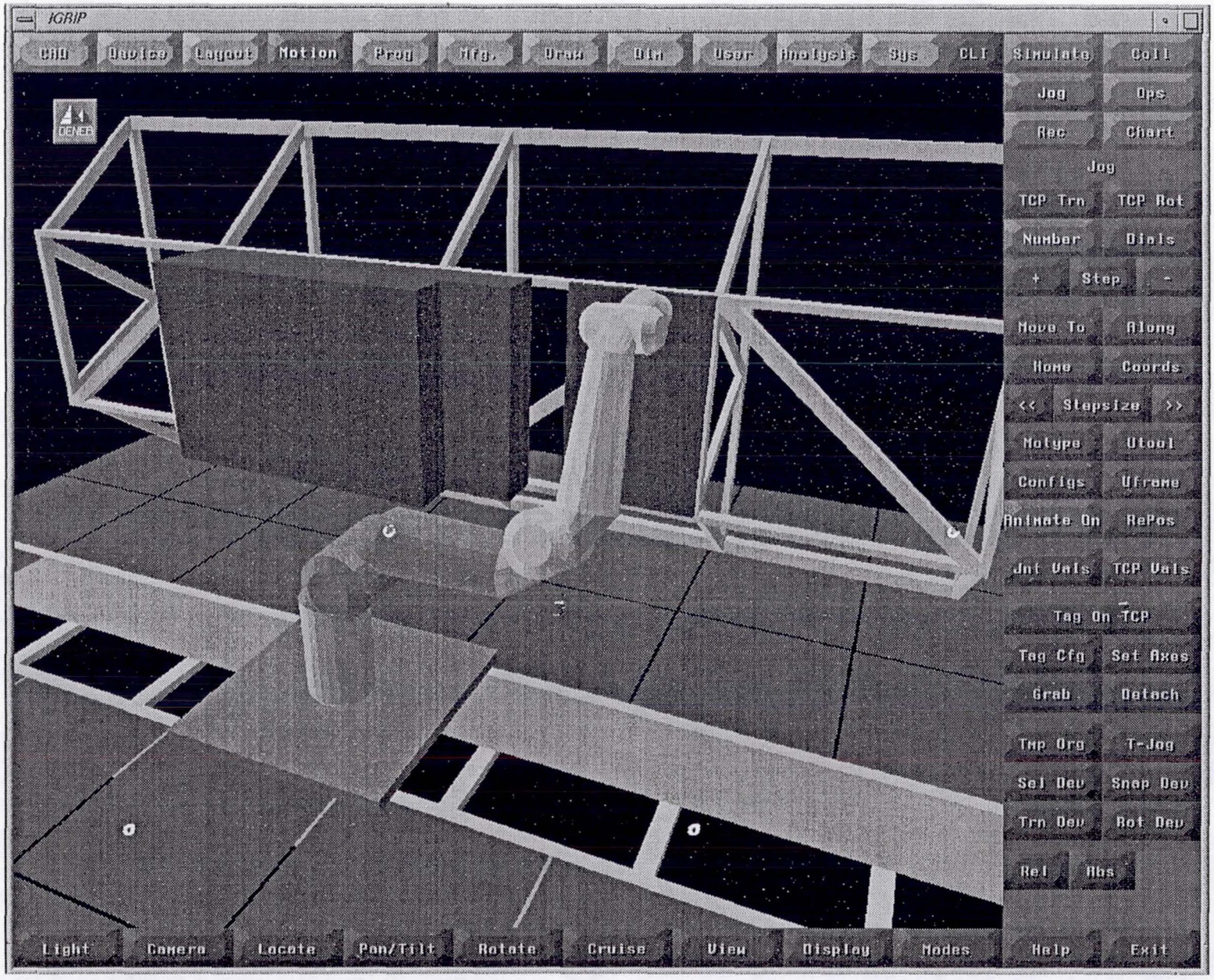


Figure 3: Typical Controller Workstation Display with Simulated Robot

Operator Workstation: The operator workstation incorporates a simulation which allows collision prediction and monitoring. In addition, an automated sequencing function allows complex tasks to be defined thereby reducing the chances of operator error.

For technology transfer and commercialization, we are modularizing these innovations so that they can be individually applied as well as integrated into a complete high dependability solution. The fault tolerant platform provides the basis for high performance broad scope fault tolerance that can be applied to process control and communications as well as to robotic control. The fault tolerant robot controller uses the services of the fault tolerant platform for operating system and hardware fault tolerance, however, it does not require the use of the operator workstation software. Similarly, the operator workstation can utilize either a fault tolerant or non-fault tolerant implementation of a robotic controller.

ACKNOWLEDGEMENTS

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, Office of Advanced Concepts and Technology. The authors acknowledge the assistance of the Jet Propulsion Laboratory (JPL) for technology transfer of portions of the Manipulator Control System (MCS) software. SOHaR Incorporated effort was funded from NASA/JPL Small Business Innovation Research (SBIR) under contract number NAS7-1247.

REFERENCES

[1] K. Tso, M. Hecht, and N. Marzwell, "A Fault Tolerant Robotic System for Critical Applications", *Proc. 1993 International Conference on Robotics and Automation,* Atlanta, GA May, 1993

[2] M. Gini and R. Smith, "Monitoring robot actions for error detection and recovery," in *Proceedings of the NASA Conference on Space Telerobotics,* vol. III, (Pasadena, CA), pp. 67-78, Jan. 1987.

[3] D. E. Wilkins, "Recovering from execution errors in SPIE," in Proceedings of the NASA Conference on Space Telerobotics, vol. III, (Pasadena, CA), pp. 79-90, Jan. 1987.

[4] E.Lopez-Mellado and R. Alami, "A failure recovery scheme for assembly workcells," in *Proceedings of 1988 IEEE International Conference on Robotics and Automation,* (Cincinnati, OH), pp. 702-707, May 1990.

[5] M. Hecht, J. Agron, H. Hecht, and K. H. Kim, "A distributed fault tolerant architecture for nuclear reactor and other critical process control applications," in *Digest of 21st International Symposium on Fault-Tolerant Computing,* (Montreal, Canada), pp. 3-9, June 1991.

[6] K. H. Kim and H. O. Welch, "Distributed execution of recovery blocks: An approach for uniform treatment of hardware and software faults in real-time applications," *IEEE Trans. Computers,* vol. 38, pp. 626-636, May 1989.

[7] B. Randell, "System structure for software fault tolerance," *IEEE Trans. Software Engineering,* vol. SE-1, pp. 220-232, June 1975.

[8] C. A. Klein and C. H. Huang, "Review of pseudoinverse control for use with kinematically redundant manipulators," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-13, no. 3, pp. 245-250, 1983.

[9] H. Seraji and R. Colbaugh, "Improved configuration control for redundant robots," *Journal of Robotic Systems,* vol. 7, no. 6, pp. 897-928, 1990.

ADVANCED GRAPHICAL PROGRAMMING ENVIRONMENT FOR ROBOTIC APPLICATIONS

John Agapakis
Vice President of Engineering, Acuity Imaging Inc.
9 Townsend West, Nashua, NH 03063

Joel Katz
Acuity Imaging Inc.
9 Townsend West, Nashua, NH 03063

Brian Avery
Acuity Imaging Inc.
9 Townsend West, Nashua, NH 03063

ABSTRACT

The welding fabrication requirements of aerospace structures and components such as the Space Shuttle External Fuel Tank typically involve significant amounts of welding and require very precise specification and control of welding torch paths and welding parameters to assure weld quality and structural integrity. Conventional programming of robotic welding process sequences can be tedious, time consuming, and error prone. Graphical off-line programming systems have been developed by several vendors to try to simplify this task, but they generally are expensive and require a high end graphical computer workstation.

The focus of the work described in this paper has been to develop a low cost advanced welding robot programming environment on an open architecture welding robot controller. This environment improves the productivity and consistency of robot motion programming and welding process parameter specification through: extensive graphical user interfacing, 3D modeling of robots, positioners, workpieces, and work cells, capabilities for off-line robot path generation and editing, simulation of robot paths and associated welding processes, and macro-level programming of robot paths and process parameters

The programming environment resulting from this effort has been a very useful prototype and has pointed the way for continued development.

INTRODUCTION

The welding fabrication requirements of aerospace structures and components such as the Space Shuttle External Fuel Tank typically involve significant amounts of welding and require very precise specification and control of welding torch paths and welding parameters to assure weld quality and structural integrity. Computerized welding process control equipment and torch/workpiece manipulation systems are widely used in aerospace welding applications to improve the consistency and productivity of such welding operations.

Conventional programming of automated welding process sequences can be tedious, time consuming, and error prone. Robot paths are commonly specified by using a hand held teaching pendant to jog the welding torch tip to a series of positions and orientations which describe the desired torch trajectory in Cartesian space. In addition, the torch speed must be explicitly specified at all points along the path together with the welding process parameters which are simultaneously controlled by the system. For welding processes such as GTA (Gas Tungsten Arc) and VPPA (Variable Polarity Plasma Arc) welding, which are widely used in the fabrication of aerospace components, a relatively large number of such parameters must be precisely sloped up, down, or maintained at certain levels along the path and in exact coordination with robot motion.

The focus of the work described in this paper has been to develop an advanced welding robot programming environment on a next-generation welding robot controller. This environment improves the productivity and consistency of robot motion programming and welding process parameter specification through:

- extensive graphical user interfacing,

- 3D modeling of robots, positioners, workpieces, and work cells,
- capabilities for off-line robot path generation and editing,
- simulation of robot paths and associated welding processes, and
- macro-level programming of robot paths and process parameters

The objectives of improved productivity and consistency are further served by providing capabilities for three dimensional (3D) modeling of workpieces, work cells, robots, and positioners, off-line path generation/editing, and robot motion and welding process simulation on the welding robot controller. Although all aerospace structures are designed using Computer Aided Design (CAD) systems, the use of such CAD models for off-line generation/editing of robot programs is only possible by using expensive off-line robot programming systems.

Macro-level programming refers to the generation of complex part programs through hierarchical calls to primitives — macros — which encapsulate all the required motions and welding parameter sequences needed to weld generic classes of parts, components, or joints. The proposed system would use macro level programming of weld procedures and robot trajectories to further enhance the productivity of the robot programmer.

The goal of this NASA sponsored SBIR effort was to improve the ability of the robot programmer to create and simulate welding robot programs on the robot controller. The developed robot programming environment represents a significant advance in the state of the art of welding robotics and, together with the next-generation welding robot controller on which they are implemented, provide a flexible base for the development and implementation of advanced robotic welding automation schemes. Although this program has focused on critical aerospace applications, the developed advanced robot programming approaches will also improve productivity and consistency in other critical low-volume or batch-manufacturing welding applications in other industries (defense, nuclear, pressure vessel, shipbuilding, heavy equipment, etc.) or non-welding robot applications such as robotic thermal spraying or surfacing.

USER INTERFACES FOR AUTOMATED WELDING

Programming robot paths and weld procedure parameter set-points is probably the most time consuming aspect of employing robots for factory automation. Industrial robots are repeatable mechanisms. However, they are not particularly accurate. The poor accuracy and high repeatability of industrial robots requires robot users to program them via a process know as lead-through teaching. The robot is moved to the desired path locations with a teaching pendant and the desired points are recorded for future playback. The process of training welding paths and validating that the path was correctly taught and performs as desired normally requires multiple iterations in which individual points are re-taught and speeds are adjusted until the desired motion profile is achieved.

Once the robot's motion is appropriately specified, it is necessary to then associate multiple welding parameters to key points or times within the motion profile so that the weld joint is correctly made. The methods used by most robot vendors to perform these attachments are generally clumsy. This is the result of the fact that most robot controllers are based on custom computer systems with custom operating software. State of the art user interface environments such as Microsoft Windows or OSF Motif are not available for many of these machines. The user interface software presented to the welding operator is often difficult to understand and use as a result of the limited user interface tools available on the robot controller. As in the case of motion programming, weld procedure programming is tedious and requires a large amount of iteration.

The extensive involvement of a human operator in teaching robot paths and welding conditions may result in a lack of consistency between programs taught at different stations by different operators or even between welds on the same part taught by the same operator. In addition, in certain applications, where strict economic justification of welding automation is required on the basis of labor savings, the extensive operator involvement during teaching significantly raises the minimum number of identical parts that must be produced on each batch before robotic automation is justifiable. These factors contribute to making the process of programming automated welding systems time consuming and pose one of the major problems in cost justifying the use of robots for limited production run parts like aerospace components.

One of the major thrusts of this program, the development of macro-level and task-level programming schemes, can allow minimizing or eliminating manual teaching and thus improving both productivity and consistency of robot motion programming and welding process parameter specification. When a robot is taught

using conventional approaches, the operator is responsible for both the form and content of the robot program. The use of a macro for a particular part geometry imposes form. The operator is then only responsible for the content of the program and this significantly reduces the time the operator needs to spend programming a part.

The objectives of improved productivity and consistency are further served by the second major thrust of this program which is the provision of capabilities for three dimensional (3D) modeling of workpieces, work cells, robots, and positioners, off-line path generation, and robot motion and welding process simulation on the welding robot controller. Virtually all aerospace structures are now designed using Computer Aided Design (CAD) systems. This makes the use of such CAD models and the off-line generation of robot programs both feasible and desirable. Currently such capabilities are typically only found on special purpose off-line robot programming systems.

Although this research effort focused on critical aerospace applications, the proposed advanced welding robot programming approaches may also improve productivity and consistency in other critical low-volume or batch-manufacturing welding applications in other industries (defense, nuclear, pressure vessel, shipbuilding, heavy equipment, etc.). As such they address a far greater segment of the total multi-billion dollar welding market than what aerospace applications alone represent and can thus further impact our national competitiveness and security because of the widespread use of welding and its importance in the fabrication of a wide variety of critical systems and structures.

THE HAWCS II OPEN ARCHITECTURE WELDING ROBOT/PROCESS CONTROLLER

In 1990, Hobart Brothers teamed with Automatix Inc. (now Acuity Imaging) to develop HAWCS II, an open architecture, standard platform based welding robot controller. This system exclusively marketed by Hobart Brothers and their affiliates for use in advanced robotic welding applications was based on robot/process programming and control technology originally developed by Automatix in several previous generations of advanced robot controllers.

The computer hardware for this open architecture controller was selected from existing third party VME Bus hardware. The VME Bus was chosen because it is an international standard, has high performance and reliability, is rugged, and hosts a large number of products from many vendors. The hardware configuration of the original implementation included a Motorola 68030 based single board computer which provided 4 serial ports, a SCSI hard disk controller, and an ethernet controller; 1 or more 6 axis servo boards; and 1 or more general purpose I/O boards which provide 48 digital inputs and/or outputs, 32 analog inputs, and 8 analog outputs. Additional boards are available including extra memory, serial ports, video display, machine vision and other kinds of I/O. Third party products are employed to minimize hardware development costs and take advantage of industry wide trends in computing. This has proved useful over the life of the controller. When single board computers based on the Motorola MC68040 microprocessor became available, they also became available for HAWCS II. As higher performance based processor boards (such as, for example, boards using the PowerPC RISC CPU jointly developed by IBM and Motorola) become available, they may be readily incorporated in the welding robot controller as well.

The software for this open architecture welding robot controller is constructed using a layered architecture which includes a third-party real-time Unix like operating system; the MIT X Window system and the Motif Window Manager for operator interface; and an interpretive programming language, RAIL, which includes extensions for robot and weld process programming. As in the case of the hardware, third party software building blocks are employed. As new board level products become available, these third party software products were modified to support them.

The hardware and software components of the system provide for a high performance robot controller with state of the art hardware, system software, and graphical user interface tools. By using these standard building blocks, a new user interface for robotic weld procedure editing can be constructed which makes liberal use of graphics to enhance the operator's understanding of the program they are creating.

THE MARSHALL ADVANCED PROGRAMMING SYSTEM (MAPS)

Even prior to the development of the open architecture controller mentioned above, engineers at Automatix Inc. began reviewing the operator interface to their robot control systems in an effort to find ways to enhance the productivity of process engineers who needed to program robots. In reviewing the operations required to program industrial robots, they speculated that the addition of a graphical user interface and robot simulation capabilities on a

robot controller would enhance the effectiveness of robot programming. They were encouraged by the success of Image Analyst®, a graphical user interface environment for the programming of the company's Autovision family of machine vision systems. Image Analyst enabled machine vision system users to program inspection tasks that used to require days of effort in a matter of minutes.

The advanced graphical programming environment discussed in this paper was conceived as a programming environment which would do for robot programming what Image Analyst® had done for machine vision. The development of such an environment was proposed to NASA MSFC (Marshall Space Flight Center) through a phased SBIR program. As the HAWCS II open architecture robot/process controller was developed it became clear that it contained the computer hardware necessary to support the graphical user interface and simulation requirements for such an environment. MAPS, the **M**arshall Space Flight Center **A**dvanced Robot **P**rogramming **S**ystem, was therefore designed and implemented as a collection of programs that could run on an open architecture robot controller such as the Hobart HAWCS II. MAPS works with Acuity Imaging's RAIL robot programming language to provide the ability to program and simulate robots. It uses the kinematic solution and robot path planning capabilities available in RAIL to perform all simulations. In this way, the simulations are guaranteed to agree with actual robot performance. In addition, this enables MAPS to simulate programs created by a hand held teaching pendant as easily as it simulates programs created within MAPS.

MAPS is intended to be a low cost software add-on package that will make some of the simulation and programming capabilities of an off-line robot programming system available on a robot controller. It is not intended to replace or supplant a full featured work-station based off-line programming system. The tools available within MAPS have been implemented in a modular fashion to enable new capabilities to be added to MAPS without requiring a complete re-write of the program.

MAPS contains two key sub-systems: tools for robot program development and simulation, and tools for weld procedure development and simulation. The robot programming tools in MAPS are used to construct graphical simulations of robot cells, move components of a cell around to perform kinematic reach and cell layout analysis, perform graphical simulations of robot paths, and graphically create robot paths. Figure 1 is a screen image of some of the tools provided in MAPS.

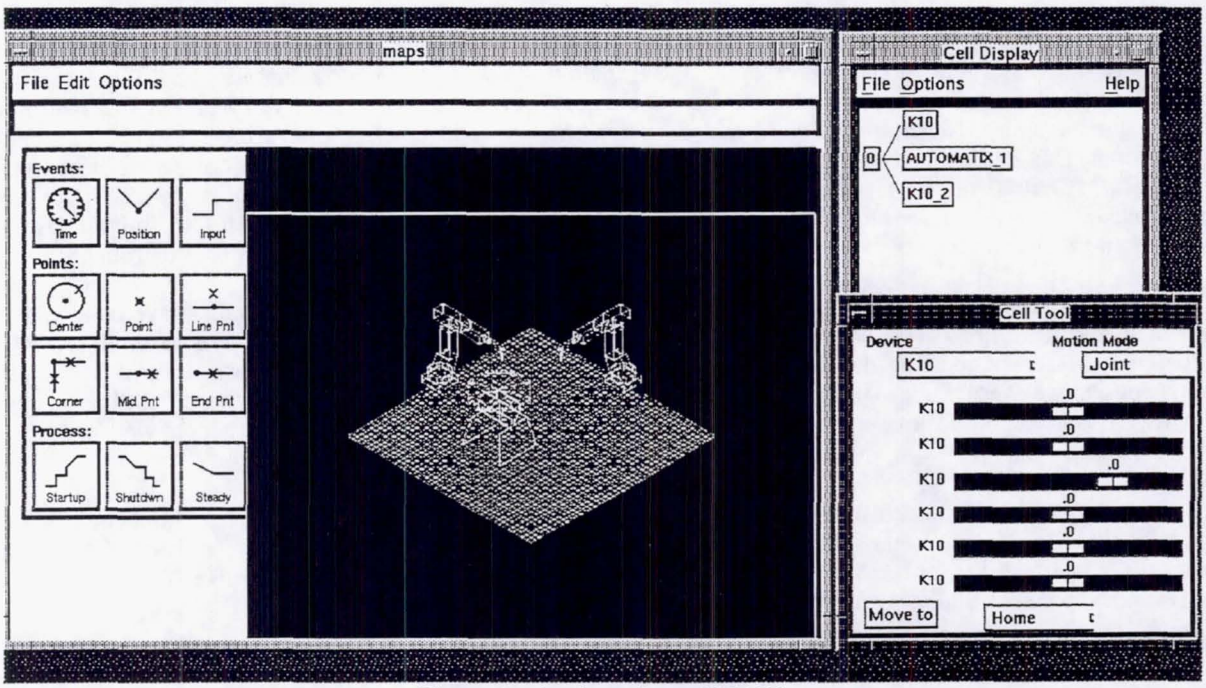


Figure 1 - An Overview of MAPS

ROBOT MOTION PROGRAMMING & SIMULATION TOOLS

The first thing a user would see after launching MAPS is a 3-dimensional drawing area called the Canvas. The Canvas displays the 3-d model of the robot workcell and any devices in the work cell. A viewing tool is available to change the viewing area. The operator can use the View Tool to Pan, Zoom, and Rotate the current view and to create and maintain multiple views of the workcell.

MAPS is built on top of a 3-D CAD package called AX developed by Cognition of Burlington, MA. Parts drawings, cell fixtures, and device linkages are all in the form of AX figure files. An IGES converter is available to convert standard IGES files into AX files so that they can be loaded into MAPS. Devices also require a text device description file which details the kinematics of the device. New device description files can be created in a straight-forward manner so that almost any device that can be built, can be simulated using MAPS.

Robots, workpiece positioners, cell objects, and parts are loaded into MAPS using the Cell Tool. The cell tool allows the operator to load CAD and kinematic descriptions of all of the different types of devices available in the workcell. Devices are loaded using pull down menus and dialog boxes. As a device is loaded, the Cell Tool creates a graph which displays the hierarchical connections between the different devices in a cell. The dependencies between devices can be re-arranged in the Cell Tool and cell configurations can be saved so that the cell does not need to be recreated every time MAPS is used.

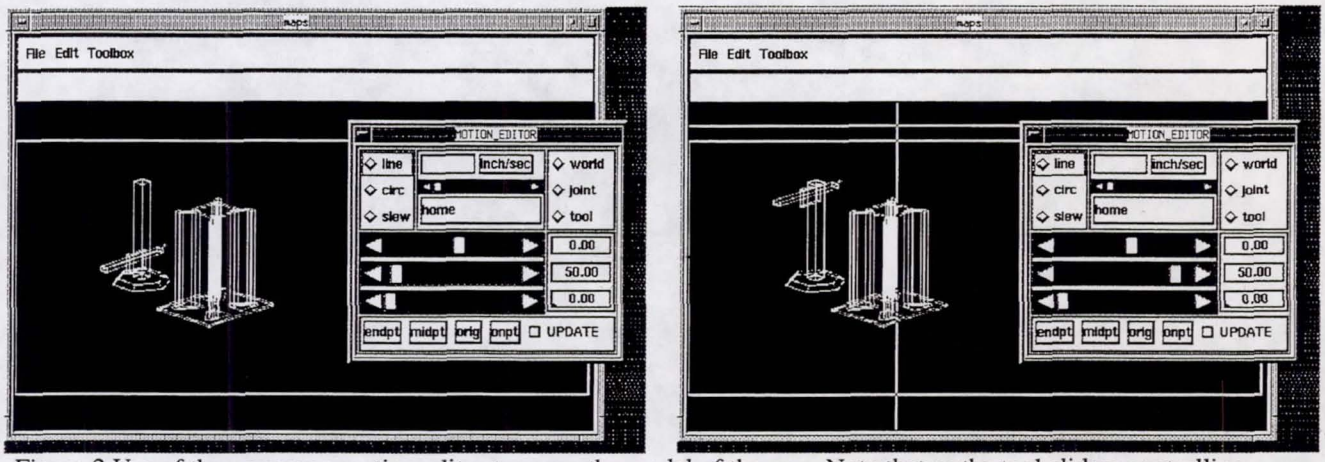


Figure 2 Use of the on screen motion editor to move the model of the arm. Note that as the tool sliders controlling joint location are moved, the graphical representation of the corresponding arm links move accordingly.

Simulated devices can be moved around in the workcell using the Device Editor, a kind of software teaching pendant. The Device Editor can be used to jog a graphical device in the joint, world, and tool coordinate systems along with some other basic capabilities (see Figure 2). It also has the ability to make a device go to named points within its workspace.

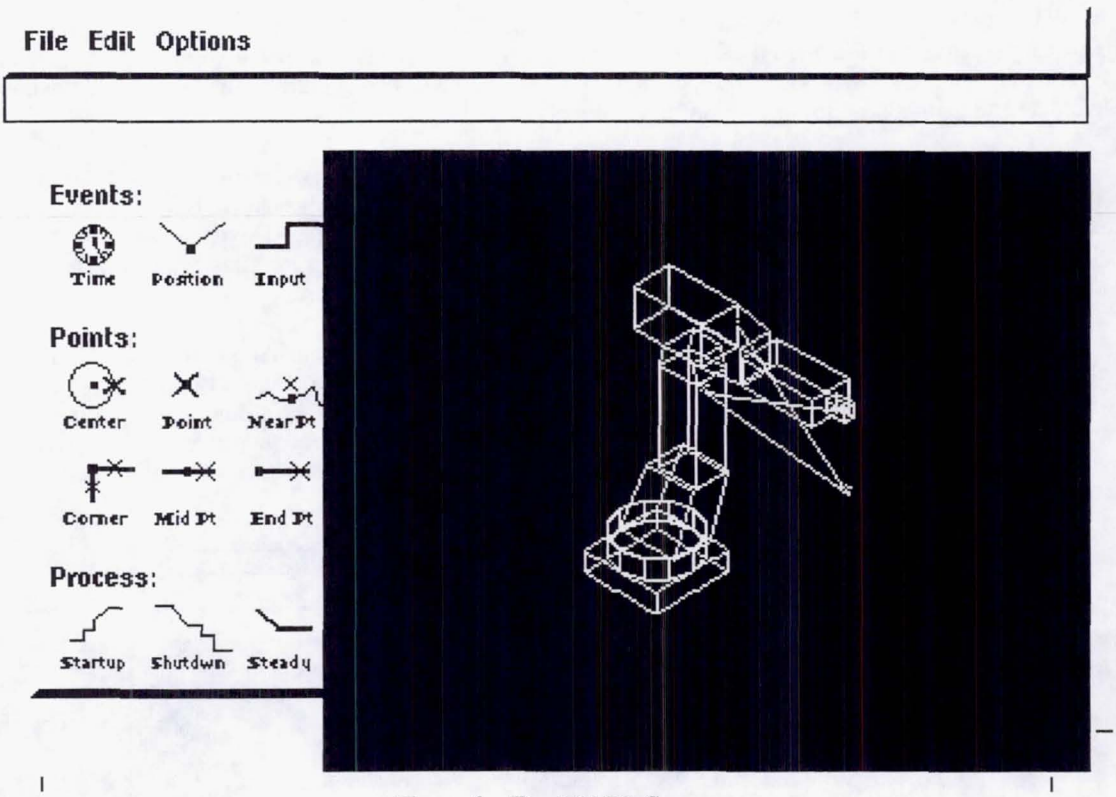
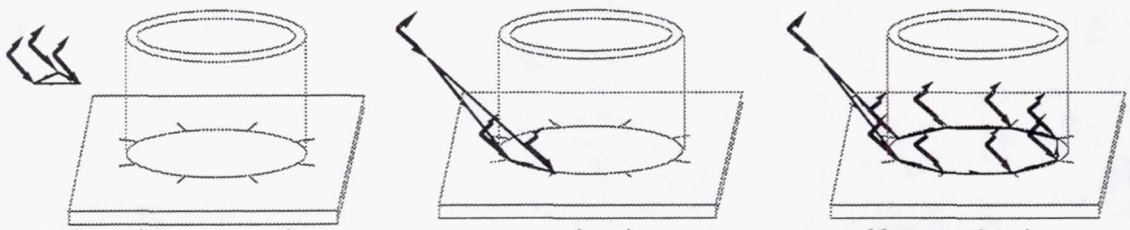


Figure 3 - The MAPS Canvas

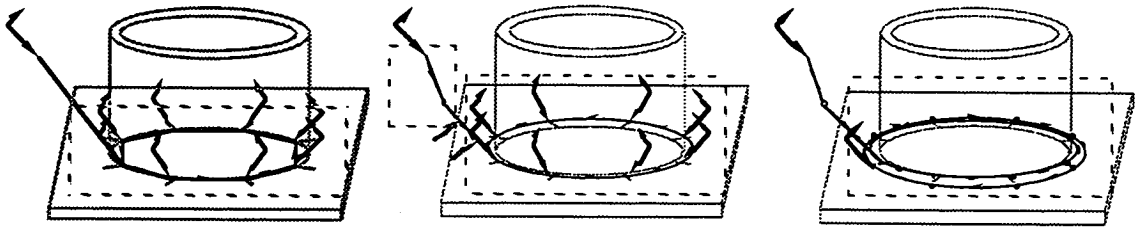
Programming robot paths from MAPS is done from the Palette which is attached to the left hand side of the Canvas in Figure 3. This allows the operator to select path training modes in a manner similar to the way many computer drawing programs work. The icon representing the desired training mode is selected from the palette and the mouse is then used to point to the location in the Canvas at which a robot point needs to be created. The designed Canvas tools allow options for:

- creating new paths by placing and manipulating in the screen an empty path stub icon
- moving the path points by dragging
- adding path segments in the new path stub
- defining part-bound coordinate systems (seam space coordinate system in the case of welding applications)
- offsetting points in a part or seam bound coordinate system
- adding approach and depart points
- setting speeds for groups of points
- Hiding point orientation indicators to reduce clutter in screen
- duplication, offsetting, cutting, & pasting a current path
- connecting indicated segments into a new path
- attaching process macros to a path

Some examples of such operations for the generation of a pipe-on-plate welding path are shown in the Figure 4.



Create/place new path move path points add new path points



Set seam space and specify orientation in seam space | Group points and set common speeds | Duplicate, inverse direction, & offset in seam space

Figure 4 Example of using the path generation/editing tools to create new part programs

This series of steps demonstrates a method of easily creating a complex weld program. As much as possible, the points and paths are defined in the most abstract way to allow for high level interaction. For example, if the diameter of the pipe is changed the program could be re-generated with no modifications. After using this procedure to create the structure of a robot program, the operator might then use a teaching pendant to show the robot the actual points to use to weld a part. However, the teaching process will be focused by the effort of creating the graphical program and using it as a template.

Programs that have been created by MAPS can be viewed graphically on the Canvas or numerically with a Path Editor dialog. The Path Editor allows the operator to select any path loaded into the system and interrogate the points in the path. This is useful for making minor changes in position and/or orientation of path points. Paths may also be graphically simulated using a control panel that looks like the controls of a VCR. The VCR Tool performs graphical simulations of the currently active path. It moves the graphical robot through the work space. If the robot path attempts to go out of the envelope, points outside of the envelope are hi-lited in red to notify the operator.

WELD PROCEDURE PROGRAMMING & SIMULATION TOOLS

A major goal of the MAPS program was to design a new approach for understanding the state of welding conditions at any point within a welding robot path and to design a graphical representation of the welding state which could be linked to the robot simulation. The most obvious representation of welding state seemed to be a multi-parameter graph showing the values of welding parameters as a function of position or time. Thus, central to the MAPS Weld Procedure editing tools is a weld parameter display window shown in Figure 5. This window displays the values of all welding parameters as a function of time in the form of a collection of graphs. The graphs can be individually scaled vertically and the overall window can be scrolled if there are more graphs than can fit on a single screen. A cursor on the display indicates the current position within the weld procedure and a set of text widgets display the parameter values of each parameter at the cursor position. Buttons are also provided which allow the operator to Zoom In or Out so that the level of resolution can be adjusted as necessary. This display is intended to represent the weld parameter values for a specific welding procedure. During simulation, the cursor position would be dynamically updated to indicate where in the weld procedure the robot is at any given time in the simulation.

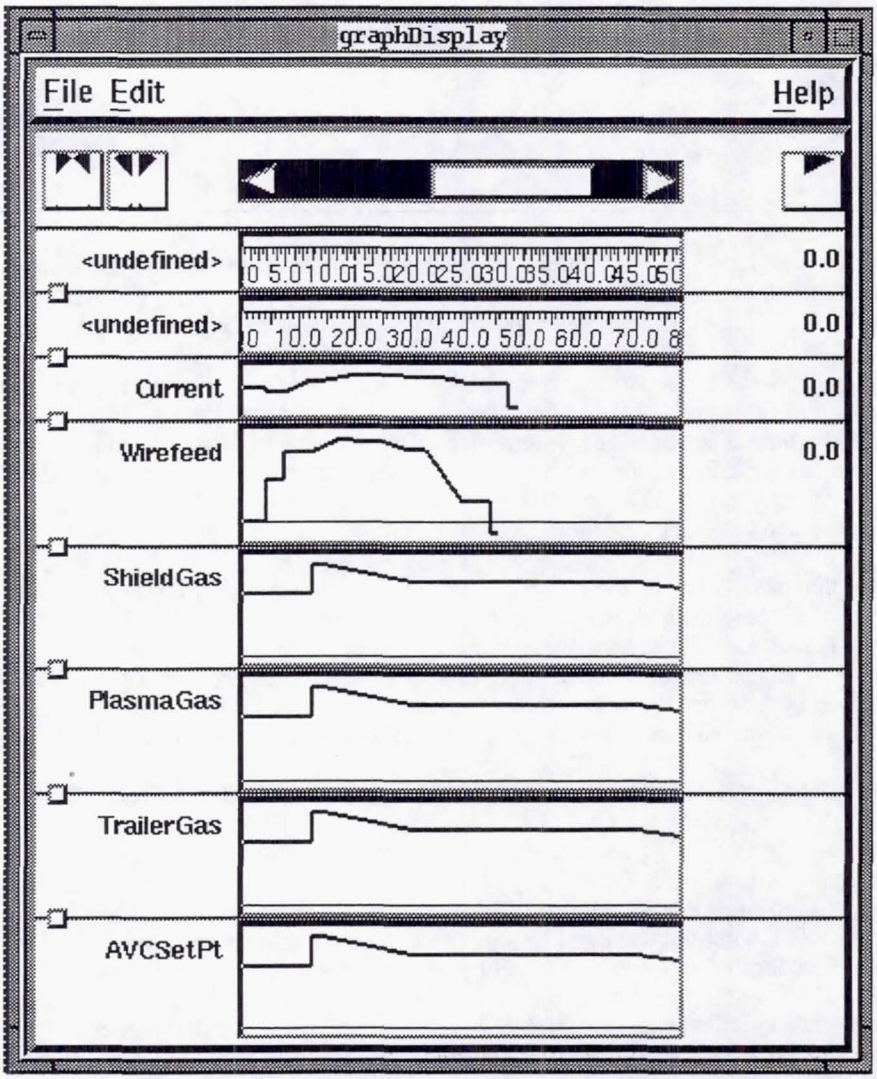


Figure 5 - The Graph Display Tool showing a Welding Procedure

MAPS constructs a welding process by combining primitive wave forms like steps, ramps, and exponentials with more complex macros which contain frequently used combinations of primitives. Every welding parameter can have its own sequence of macros. Every macro is active from a specified starting event until a specified ending event. The wave forms of different parameters are synchronized by sharing a common starting event or ending event. Events can be a wide variety of things including: an input changing state, an absolute time, a time relative to another event, an absolute position, a position relative to another event, the value of any parameter reaching a threshold, or some other well defined state.

MAPS contains two tools to edit Weld Procedure Data. The first tool is called the Process Parameter Editor. The process parameter editor is displayed in Figure 6. It provides a limited graph which shows all of the parameters of a single macro. The process parameter editor has pull down menus to let the operator choose which parameter and which macro of that parameter to edit at any given time. Once these selections have been made, the shape of the selected parameter is displayed and text widgets are presented which allow the operator to set the values of the parameters of the selected macro.

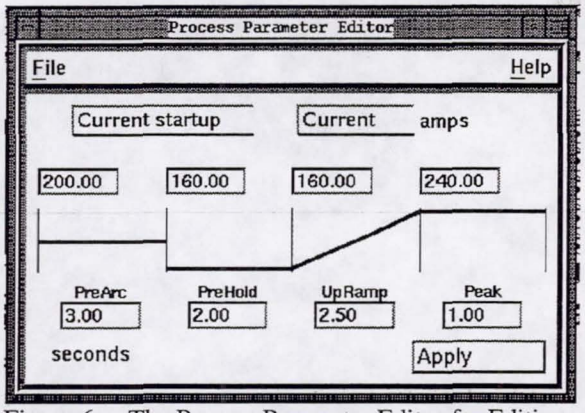


Figure 6a - The Process Parameter Editor for Editing the current startup macro of a weld procedure.

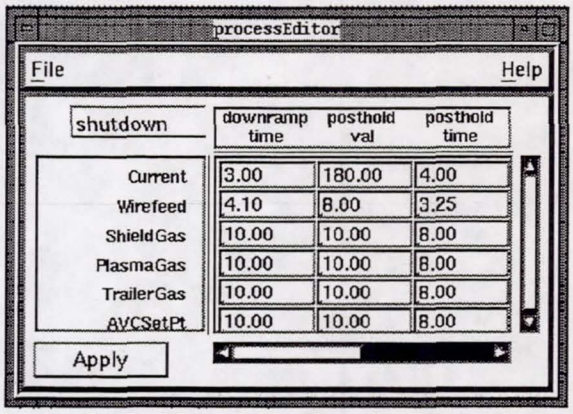


Figure 6b-The Value Editor For a Shutdown Macro

The second tool is called the Value Editor. The value editor is a window which presents the values of all parameters that use a given macro in a matrix which looks similar to a spreadsheet. The value editor is intended to provide a different way of viewing the same data that the Process Parameter Editor displays. Multiple copies of either of these tools are supported and all instances of these tools dynamically update the others if any one of them changes the value of a parameter. To support this capability, a weld parameter database which maintains all weld procedure data and which provides access to this data to multiple tasks was implemented as a central component of the Graph Tool.

The operator always has three different views of all weld procedure data available to him: the Graph Display which shows all parameters of the weld procedure from beginning to end, the Process Parameter Editor which shows the values for a single macro and parameter, and the Value Editor which shows the values of all parameters for a given macro. The Graph Display Tool has ended up being central to both parameter editing and welding simulation. The graph display tool has a cursor which can be moved to show the parameter values for all parameters at any position or time in the procedure. The cursor can be moved by using a button on the graph display or under the control of another task that sends cursor update commands to the graph display tool. During welding simulations, the welding task sends updates to the graph display window to advance the cursor to the correct time and position in the graph.

CONCLUSIONS

The MAPS prototype developed under this SBIR effort has proved useful in the process of designing a new user interface to enhance the productivity of welding robot programmers. The MAPS prototype has proved that it is both feasible and useful to incorporate 3-D graphical simulations of a robot work cell on an open architecture robot controller such as the Hobart HAWCS II. The welding procedure tools developed in this program have greatly simplified the task of setting up complex multi-parameter weld procedures. The effort begun under this SBIR program will be continued with a goal to making many of the facilities prototyped available in production robotic welding applications.

ACKNOWLEDGMENT

The graphical user interface environment for robot and process programming and simulation presented in this paper was developed under NASA MSFC sponsored Phase I and Phase II SBIR programs. Kirby Lawless and Chip Jones of MSFC's Materials & Processes Lab were the Contracting Officer's Technical Representatives.

SENSOR SKIN

Daniel G. Wegerif, Ph.D.
Merritt Systems, Inc.
Merritt Island, Florida 32953

William D. Parton
Merritt Systems, Inc.
Merritt Island, Florida 32953

ABSTRACT

An animal's skin is an incredible engineering feat. It is endowed with numerous sensory capabilities that facilitate survival in the real world. Machines built by man have, until now, possessed little ability to sense the environment in which they work. Insuring the survival of machinery is limited to keeping that machinery out of harm's way. If a machine possessed the ability to sense its working environment, that environment could be expanded, increasing the productivity and the number of applications for that machine. Merritt Systems Inc. (MSI) is involved in the research and development of sensory capabilities for robots and developed a sensing system that works much like a living skin. Prior art distributed only the sensing elements over the surface to be covered. What distinguishes our sensorCell technology is the localization of all the circuitry necessary to process sensory data at the sensor site. Relay of the sensory information is done via digital communication techniques. This allows for interspersing differing sensory media over the same skin via a generic communication command set architecture. A sensor skin was developed using this technology for articulated robots to detect potential obstacles in the workspace. This sensor system could be used for numerous robotic applications.

INTRODUCTION

In 1990 MSI became involved in whole-arm proximity sensing as a means to provide collision avoidance for robot manipulators in partially modeled, constrained or dynamic operating environments. Until recently, no commercial systems were available to provide sensor-based obstacle avoidance for robots. Under NASA's Small Business Innovative Research (SBIR) program, MSI has developed a practical full-coverage sensing methodology using multiple sensing media, a digital communications architecture, and a control algorithm that utilizes the real-time sensor data to provide collision-free motion of the robot. We have developed three generations of sensor hardware and two distinct control algorithms (1-3). MSI also developed a PC-based robot simulation software package to observe and verify the control algorithms before they are implemented on an actual robot and to provide a graphical interface for the user. When MSI completes its current work with NASA, over $1,000,000 will have been invested in the development of the whole-arm sensing system. The Department Of Energy is also very interested in this work and has provided two separate Commercial Research And Development Agreements with MSI, totaling $500,000, to assist in the commercialization of the system.

The development of Whole-Arm Proximity Sensing technology was initiated at the Kennedy Space Center (KSC) to facilitate the use of dexterous robotic manipulators near critical flight hardware. Flight hardware including the Space Shuttle Orbiter and its payloads are highly sensitive critical systems, and Ground Support Equipment (GSE) working near these systems must meet stringent operating requirements. Most GSE has redundant safety systems and fail safe or fail operational features. To meet these requirements, GSE robotic manipulators must possess safety systems that prevent the possibility of damaging or colliding with flight hardware. The Whole-Arm Proximity Sensing Systems will be incorporated into future robotic manipulators working near flight hardware at KSC to provide a redundant safety system.

The concept for the non-tactile sensor system arose from the need to detect objects in the robot's work envelope and prevent unforeseen collisions. The sensor system functions both as an emergency stop (E-stop) and as a path

planning tool. In E-stop mode, the robot will be stopped and the operator notified if an imminent collision is detected. In path planning mode, the robot uses the sensor information to traverse an obstacle-free path while seeking a pre-defined goal point.

Ongoing and future KSC robotic projects will use this sensing system. The Payload Inspection and Processing Robot (PIPR), which is a long reach serpentine inspection robot, is presently the primary system targeted for sensor application. The PIPR will be used for pre-flight inspection and verification of Shuttle payloads. Future robotic systems for Orbiter radiator inspection, main engine leak detection, and other flight hardware related automation will also use the sensor system for safe operation.

WHOLE-ARM SENSING SYSTEM

Over the last four years, MSI has developed three generations of sensing systems for robot collision avoidance. MSI based its first generation sensing system on a design based by E. Cheung (4) that uses reflected infrared (IR) sensors to determine object proximity. The sensor signals are amplified and filtered and then sent to a PC-based analog to digital converter. This method is functional, but is very sensitive to external noise and has a limited detection range. We resolved these limitations through the development of a second generation system, the sensorCell. The sensorCell converts the analog signals from multiple IR sensing elements to a digital format at their source. By doing this, the resulting data could be effectively transmitted over long distances without degradation. The sensorCell incorporates a microcontroller that conditions and previews the resulting digital outputs before transmission to the cell controller. See Figure 1. By previewing the data, only interesting data is transmitted, reducing the communications requirements. We developed a dedicated communications architecture for the sensorCell that handles up to 1024 sensorCells in a single system. The sensorCells provide a detection range of 0.4 meter and utilize standard RS-485 serial interfaces to the cell controller.

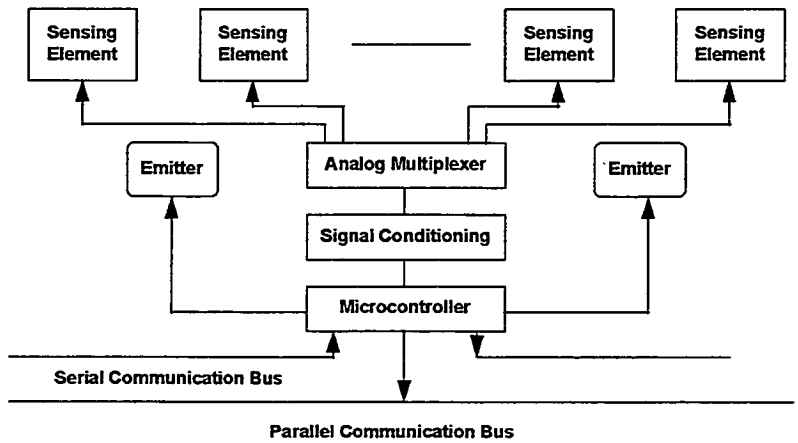


Figure 1 Block Diagram of SensorCell

We introduced the concept of the sensorCell "set". We can assign each sensorCell to one of 256 possible sets. Sets are enabled or disabled to allow, for example, only sensorCells facing in the direction of travel to perform scans. This innovation greatly improves the effective scan rate. The revised communication instruction set permits complete sensing operation using either one of the two busses alone in the event of a failure. This adds another degree of redundancy and makes the system more fault tolerant.

MSI originally developed rectangular sensorCells for a PUMA 560 robot. Later, we instrumented a Robotics Research 1207i. The physical configuration of the RR 1207 robot necessitated a new configuration for the actual sensorCells. The RR 1207 is composed of three tapered cylinders along the major axis with orthogonal cylinders between each link. We designed trapezoidal shaped sensorCells with two rows of four receiving elements for the tapered cylinders, and developed octagonal boards with outward looking components. Six trapezoidal boards are

placed around the first link, five on the second and four on the third link. Four octagonal boards are attached to the ends of the perpendicular joints. Figure 2 shows the system configuration.

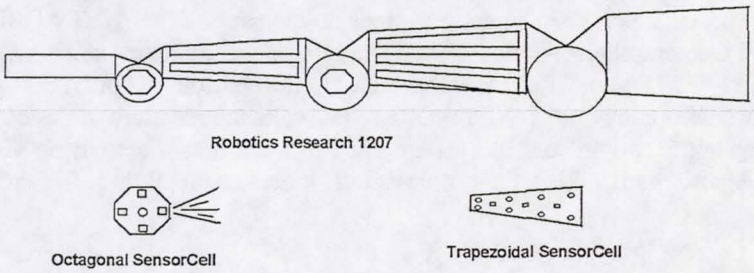


Figure 2 Robotics Research Manipulator

In this figure, the octagonal board has a single sensor represented by a circle in the middle of the board, which looks out, perpendicular to the board; and four additional sensors located on opposite sides, represented by rectangles, which detect objects in the plane of the board. This layout provides a hemispherical coverage for the end of the cylindrical components of the robot. The trapezoidal boards incorporate eight receiving components, represented by circles, and three emitters, represented by the rectangles. The geometry of the trapezoidal boards allows efficient installation on the three tapered cylinders making up the main portion of the robot. Figure 3 shows the actual installation on the Robotics Research 1207i robot manipulator.

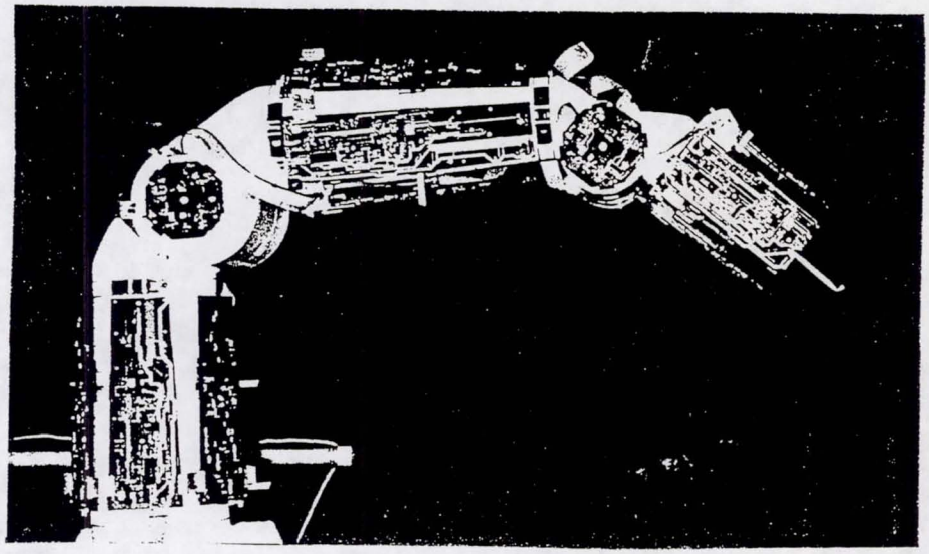


Figure 3 SensorCells on the Robotics Research 1207i Manipulator

Other robot systems may require different sensorCell configurations and sizes. We discovered that even though we developed a general method for collision avoidance, a variety of layout modifications could be required for each new application. Another limitation to this design was the ability to provide only a single sensing medium, IR. Operating environments at the Kennedy Space Center include a number of different materials with a variety of physical characteristics. In some cases, the IR-based media is not the best selection. To overcome these limitations, MSI developed a third generation system that incorporates smartSensors and sensorSkin. The smartSensor is an intelligent sensor module with an integrated digital communications architecture. The smartSensor retains most of the functionality of the sensorCell, except that it typically incorporates only a single sensor emitter/receiver pair. Advantages to this method include the ability to incorporate multiple sensing media, a smaller footprint device and a generic mounting method. The sensorSkin provides the communications and power bus to support the smartSensors. It was designed so that it can be cut and configured without losing functionality.

The communication architecture of the smartSensor uses three serial busses in this design. Each sensor module uses all three of the communications busses in a democratic fashion, it will receive equally on any of the three and transmit on all three when it is required to communicate. The "A" and "B" busses seen in Figure 4, are routed around on the panel, and provide redundant communications on the panel. The "NETWORK" bus is an extra port used to interfaced back to the controller. All data to and from the sensor modules on a panel communicate back to the controller via the "NETWORK" bus. Redundant interface to the controller is provided by connecting two lines from the controller to two sensor modules' "NETWORK" bus. Communication to the cell is picked up on the "NETWORK" bus, or one of the two "A" or "B" busses. The cell monitors for activity on all three busses and will receive on the first active port sensed. When the cell transmits, it does so on all three busses simultaneously.

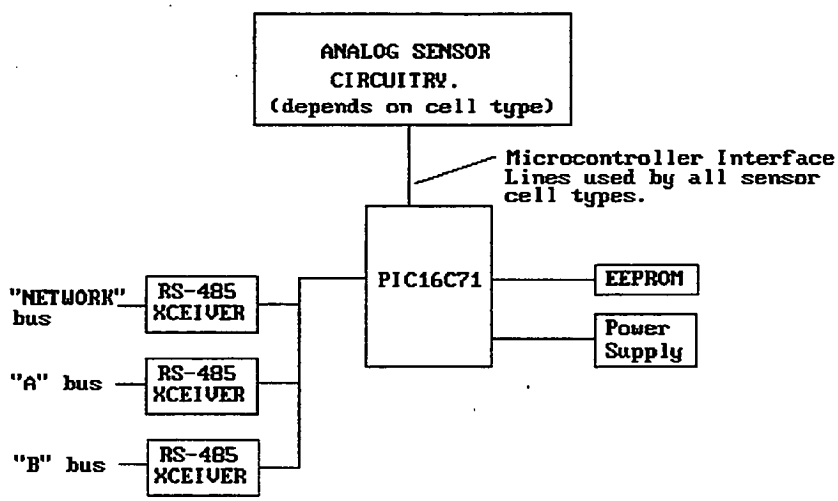


Figure 4 SmartSensor Architecture

The ANALOG SENSOR CIRCUITRY block shown represents the hardware unique to each sensor cell type. There are 5 control/data lines dedicated to interfacing to the analog circuitry for any sensor cell type. Four of these can be used as ADC inputs since the micro controller has a built in four-channel analog to digital converter. There is a linear voltage regulator for supplying power and also an EEPROM on the sensor cell for saving of calibration data.

Figure 5 illustrates the bus architecture on the sensorSkin. The sensor Skin is composed of up to 32 sensor panels, each which contains up to 32 smartSensors. This figure illustrates four sensors on the panel. The redundant on-panel busses are the ones denoted bus "A" and "B" from above and the network bus is denoted with an "N". On each panel, two sensors will be used as panel masters, which communicate to the external cell controller. We designate only one master sensor at time: the other is a backup. Note that for sensors not being used as interface boards to the cell controller that the network output remains unused.

In addition to the redundant power and communication bus, the sensorSkin also provides a fixed geometry for the sensor elements. The sensorSkin was designed to provide a fixed spacing of approximately 3 inches by 1.5 inches. Other spacing is available if required. Our mounting material provides a flexible mounting surface that we can fold and cut to meet specific contour requirements. Because the sensorSkin bus architecture provides a generic communications method, a user can place both IR and acoustical sensors on the same panel. The digital communications architecture is completely independent of the sensing media.

Under normal operating conditions, the controller periodically issues scan commands to the active panel master on each panel and handles any returned information. The panel master controls the sequential scanning of all sensors on that panel, including itself. We can perform scanning operations simultaneously on multiple panels, provided that radiated emissions from one panel do not interfere with the operation of others. When a sensor detects an

object in proximity, the device returns the ID number to the controller, where a lookup table provides the location of the detecting sensor. Communication paths on a panel also consist of redundant multi drop busses. If one of the two busses fails, either on a panel or between panels, or if the panel master fails, the backup panel master takes control and uses the backup bus to continue operation.

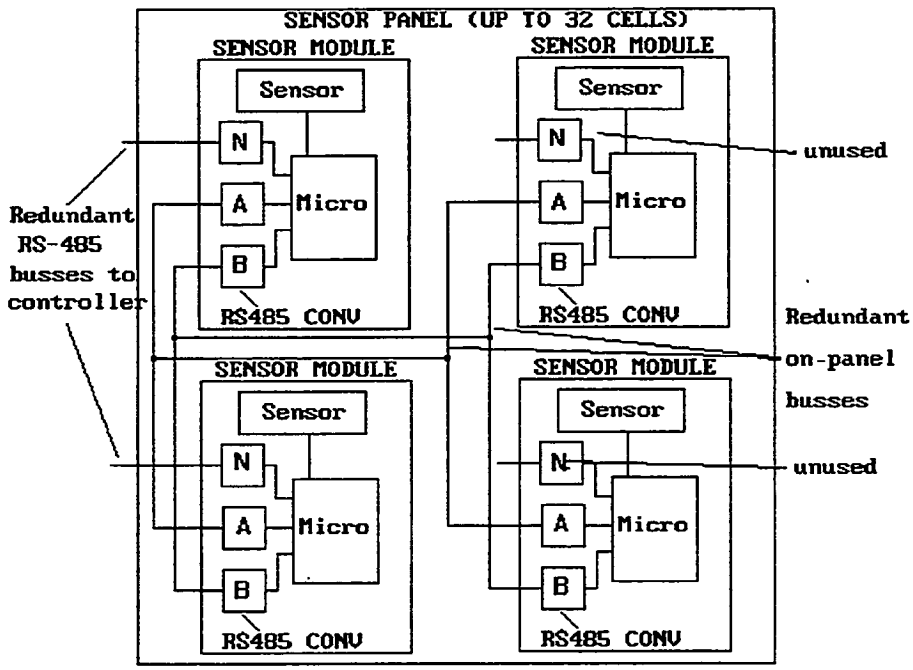


Figure 5 Bus Architecture on SensorSkin

The command architecture for both the sensorCell and smartSensor is a combination of firmware embedded in the microcontroller and software in the cell controller. This architecture enables the user to set various parameters such as gains and threshold values, to select offset values from calibration tables and to intelligently group the sensors into sets. A PC-based interface allows the user to determine the system status, set operating parameters and monitor the sensor data.

The smartSensors are manufactured using a hybrid methodology, which incorporates surface mount, die, and masked components. This provides a one inch by two inch footprint. This is as small as practically possible due to the physical constraints of the receiver and emitter components. The hybrid component utilizes a ceramic cover that protects the internal elements from exposure to airborne particles and improves the cleaning process. The coating also prevents the sensor from becoming an ignition source should it come into contact with a conductor.

In addition, MSI is currently investigating other sensing media that can be readily incorporated into our smartSensor architecture. These additional sensing elements include: capacitive-based proximity, tactile, temperature, and hazardous gas sensors. This capability would allow a user to select the sensing elements to meet various sensing application requirements. The Department of Energy developed a capacitive-based sensor, the WHole-Arm Proximity (WHAP) sensor, which we are currently integrating to our smartSensor. This additional proximity media is useful in detecting conductive materials. A unique property of the capacitive sensor is that it too can be contoured over complex shapes and maintain reasonable sensitivity. It has been successfully demonstrated in unfriendly environments and is impervious to many liquids.

SYSTEM ARCHITECTURE

In addition to the sensing components, we developed a PC-based control architecture for integrating the whole-arm sensing system with existing robots and controllers (6). It is made up of a connected set of sensors, a system control computer, and a robot device with its control processors. The system controller operates the robot (typically through a supervisory mode available in the robot controller) and the sensors. The user provides high-level motion commands through a user interface component of the system controller. Collision-avoidance motion planning is performed in the system and robot controllers.

In one demonstration system, the system controller is a 486-based PC which communicates through serial links to the sensorCells and to the controller of a Robotics Research seven-DOF robot. The system controller has a Windows user interface, and performs the motion planning for the robot. It sends joint motion commands to the robot controller at the 40 Hz. update rate. The PC-based controller is implemented in the C++ object-oriented programming language using the Borland development tools.

Figure 6 depicts the main functional components of the controller software and their interfaces. The User Interface is a graphical interface that provides control and displays of the robot and sensorCell operation or of the system simulation. The Simulation module models the operation of a system, including the sensorCells, and displays animations. The Controller coordinates control of the devices and generates motion trajectories. The SenCell Handler operates the sensorCells and processes their data. Device-specific computer controllers drive the individual robots and devices, such as camera mounts. These may include Device User Interfaces, for example, to start the robot and put it in a supervisory mode.

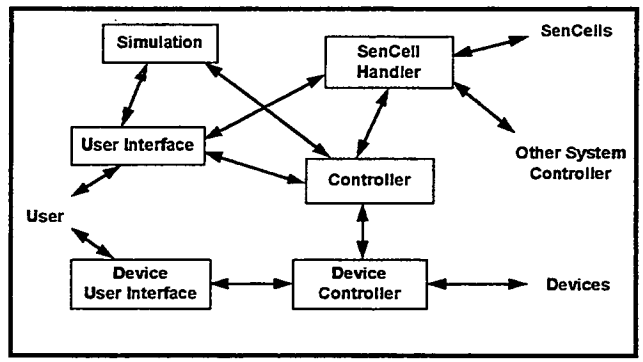


Figure 6 Robot System Controller

Some of these modules have a real-time component that must execute at regular intervals or when data is available. These functions must be handled outside the main Windows message-handling loop. Also, parts of the system, such as the SenCell Handler, can be extracted and used in other system control architectures. The software components developed specifically for this project are described further below.

The user interface is built using Microsoft Windows and supporting the standard Windows look, feel and menu structure. We developed three distinct applications: 1) a model builder, for defining robot geometries and graphic models; 2) the robot control and simulation application; and 3) a sensor handler application for operating the sensorCells by themselves or as part of another controller architecture.

Figure 7 shows an example configuration of the simulation application with one of the pop-up menus active. The command menus include standard file and edit menus, selections for creating and handling animation views, and commands that execute the simulation, control the robot system and operate the sensorCells.

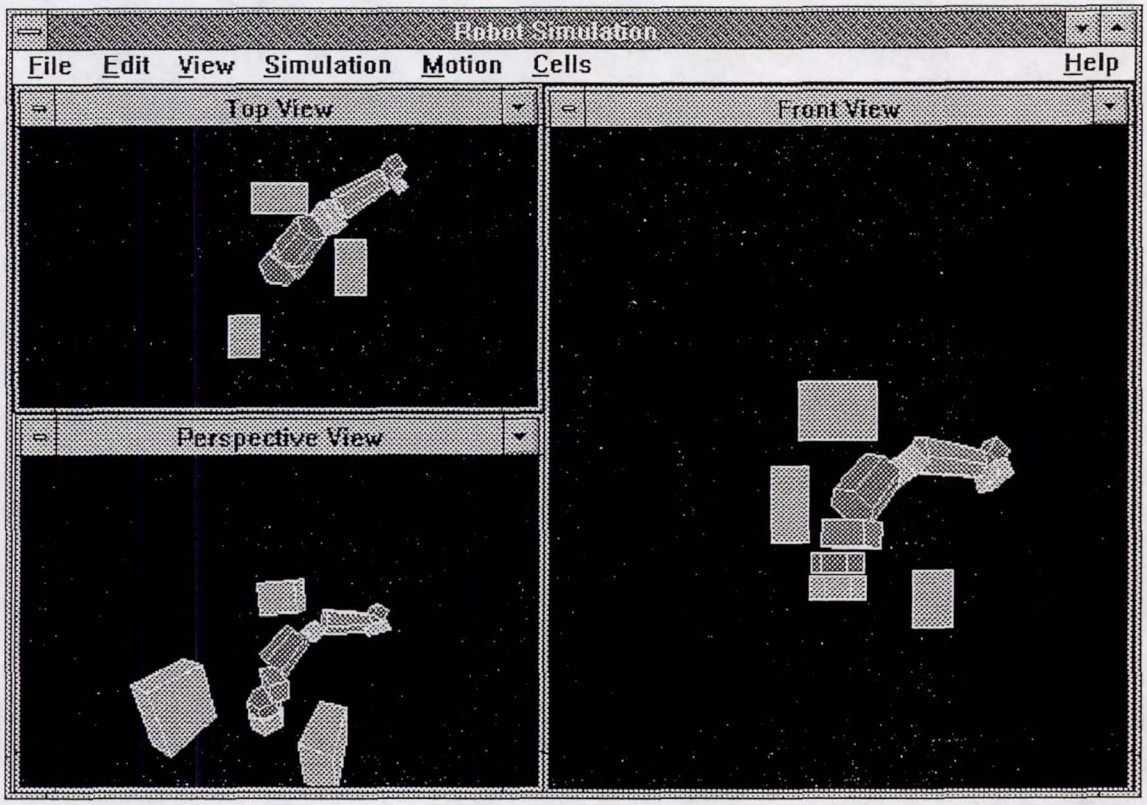


Figure 7 Robot Simulation Interface

APPLICATIONS

As mentioned previously, this technology was developed under NASA funding particularly to address several robotic applications at the Kennedy Space Center, Florida. However a number of additional applications have been identified which can benefit from this technology. The Department of Energy has identified whole-arm proximity sensing as an enabling technology to support hazardous and radioactive cleanup efforts at several government facilities. Specifically, a long reach (greater than 30 feet) robot manipulator is being considered to support waste cleanup tasks at the Hanford Site, where high-level radioactive material must be removed from old single-shell storage tanks. Whole-arm sensing would assure that the robot does not collide with obstacles in the constrained operational environment.

Several commercial robot manufacturers have expressed interest in incorporating whole-arm sensing into their product line. With the new generation smartSensors, a wider range or robots can be instrumented using a variety of sensing media. For example, temperature, radiation and hazardous gas sensors could be interspersed with the proximity sensors to give a user more accurate information about the robot environment. With this type of implementation, a user could reconfigure the type of sensors on the skin for various applications.

Other commercial applications exist in manufacturing (5). The widest application may provide enhancements to a large number of existing robotic work cells. Instead of the current intrusion protection techniques (fences, pressure sensitive pads) used to restrict humans from being struck by robots, the robots could be instrumented with sensorCells to allow humans to work safely in the same area. Integrated man and robot work cells would utilize the dexterity, finesse, and intelligence of a human with the strength, accuracy, and endurance of a robot. This capability has the potential to greatly improve the productivity of numerous existing manufacturing processes.

CONCLUSION

Several robotic applications at NASA-KSC and elsewhere require advanced sensor systems for improved safety and reliability. MSI is developing proximity sensing technologies and simulation and control techniques for modeling and using these new sensors. One product is a sensorCell system, which provides full-coverage proximity sensing and integrated collision-avoidance for existing robots. Because of certain limitations of the sensorCell, MSI developed and tested IR and ultrasonic sensing sensor modules or smartSensors, and demonstrated sensor ranges up to 0.5 meter and more. The sensorSkin allows the smartSensors to be located on all portions of the robot. Due to the design of the sensorSkin, it can be easily cut to size and placed over a wide range of contours, including both the robot and payload.

MSI is currently investigating, developing and integrating additional types of sensing media compatible to the smartSensor architecture. This includes the capacitive based sensor developed by Sandia National Labs and a tactile sensor. In the future, we plan to investigate various hazardous gas and radiation sensor elements. Additionally, MSI recently integrated the sensorCells and PC controller with a Robotics Research 1207 seven-DOF manipulator and successfully demonstrated its operation.

To complete the collision avoidance system MSI also developed a PC-based simulation and control tool for verifying algorithms and integrating the sensors as add-on components to existing robotic devices. This tool operates the sensorCells, performs collision-avoidance trajectory generation, interfaces to the robot controller, and provides a Windows user interface for robot system operation. It is implemented in C++ and communicates with the sensorCells and robot controller via serial interfaces. The software provides kinematic simulation for general serial manipulators. Although the implementation is typically slower and has lower graphics resolution than high-end packages for engineering workstations, it provides a very economical alternative for robot system analysis, design and control tasks. It incorporates many of the features of high-end tools, and the wide familiarity of PCs and Windows make it an ideal tool for introducing a large user community to robot kinematics and control issues.

REFERENCES

1. D. Wegerif, D. Rosinski, W. Parton, "Whole-Arm Proximity Control System for Articulated Robots Working Near Space Vehicles and Flight Hardware," *Proceedings of The American Institute of Aeronautics and Astronautics Space Programs and Technologies Conference*, Huntsville, AL, 1993.

2. D. Wegerif, D. Rosinski, W. Parton, "Results of Proximity Sensing Research for Real-Time Collision Avoidance of Articulated Robots Working Near the Space Shuttle," *Proceedings of The American Nuclear Society Fifth Topical Meeting on Robotics and Remote Systems*, Knoxville, TN, 1993.

3. D. Wegerif, D. Rosinski, "Sensor-Based Whole Arm Obstacle Avoidance for Kinematically Redundant Robots, *Proceedings of SPIE/OPTCON Sensor Fusion V*, Boston, MA, 1992.

4. E. Cheung, "Real-Time Motion Planning for Whole-Sensitive Robot Arm Manipulators," Yale University Ph.D. Dissertation, New Haven, CT, 1990.

5. J. Craig, *Introduction to Robotics Mechanics and Control*, Addison-Wesley Publishing Co., Reading, MA, 1988.

6. M. Thomas, "PC-Based Simulation and Control of a Manipulator With Whole-Arm Proximity Sensing," *Proceedings of 1994 Florida Conference on Recent Advances in Robotics*, Gainesville, FL, April 1994.

MOBILE ROBOTIC SYSTEM FOR SERVICING OF THE SPACE SHUTTLE ORBITER LOWER SURFACE TILES

Todd Graham
NASA Kennedy Space Center
Kennedy Space Center, FL 32899

Kevin Dowling
Carnegie Mellon University - Field Robotics Center
5000 Forbes Ave.
Pittsburgh, PA 15213

Richard Bennett
I-NET Space Services Inc.
Kennedy Space Center, FL 32899

Eric Cooper
NASA Langley Research Center
Hampton, VA 23665

Cregg Cowen
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

Davoud. Manouchehri
Rockwell International Space System Division
Downey, CA 90241

Abstract

A mobile robotic system has been developed that will inspect and rewaterproof 85% of the lower surface Space Shuttle Orbiter thermal protection tiles. This system is comprised of advanced mobile robotic, computer vision, information system, and manipulation technologies. These technologies have been developed and configured to perform critical processing tasks on space-flight hardware in an unstructured and dynamic working environment. Robot tasks include chemical injection (i.e. rewaterproofing) and inspection of the Orbiter's thermal protection system. This paper outlines tasks, rationale, facility, and operational requirements for the development of this system. A detailed look at the mobile robot, computer vision system, end-effectors, information system and computing system architecture will be provided. Salient features of the mobile robot include omnidirectionality, high reach, high stiffness and accuracy with safety and self-reliance integral to all aspects of the design. The vision system incorporates high resolution CCD video cameras, laser line protectors, and computer controlled illumination. The information system is built around a commercial relational database manager (i.e. ORACLE®) and utilizes X-VIEW toolkit for X-WINDOWS for the graphical user interface. The computing system is comprised of one off-board and four on board computer systems. The

tooling that performs the rewaterproofing process incorporates advanced force control and redundant sensing of critical parameters. The robot system is shown to meet task, facility, and NASA requirements in its design resulting in unprecedented specifications for a mobile-manipulation system. This system has potential technology spin-offs in the areas of inspection, manipulation and computing systems architectures. The system concept and architecture could be applied to inspection and servicing of commercial airlines. The inspection system could also be applied to a variety of parts inspection tasks in the manufacturing, assembly and testing of various products.

Background

Robotics and automation technologies have historically not played a role in the ground processing operations of spacecraft and space systems. In part, this has been due to skepticism regarding the viability of these technologies and a strong concern for safety of flight hardware and personnel. NASA Headquarters and KSC Robotics personnel felt strongly that the time was right to investigate both applications and technologies for ground processing. In 1990 the Orbiter Thermal Protection System (TPS) ground processing activities were investigated [NASA-TPS 90,]. This study not only concluded that there were tasks worth automating but that current robotic technologies were mature enough to make this automation possible.

Focus was placed on the TPS of the Space Shuttle since it was a task that was possible to automate within a relatively short time and it brought in flight hardware issues. It also offered high payback and could serve as a first step in providing compelling arguments for automation for both ground processing and space robotics. Beyond this there was a strong rationale for these applications including: safety, time, quality and reliability and paperwork reduction.

Thermal Protection System Tasks

Lower surface rewaterproofing and inspection are the two processes that were initially chosen for implementation because they were both technically feasible and had good payoff within a relatively short time frame. The TPS Process Study [NASA-TPS 90,] outlined additional processes that would be natural extensions to initial robot system's capabilities (e.g. non-contact tile bond verification, surface contouring).

Rewaterproofing

The Orbiter is covered with several types of TPS materials that protect the orbiter's aluminum skin during the heat of reentry. The lower surfaces are covered with silica tiles. These tiles have a glazed coating over soft and highly porous silica fibers. Water in the tiles causes a substantial weight problem which can adversely affect launch and orbit capabilities for the shuttles. Since the orbiters may be exposed to rain, the tiles must be waterproofed with a specialized chemical, Dimethylethoxysilane (DMES), which is injected into each and every tile.

In the current process, DMES is injected into a small hole in each tile by a hand-held tool that pumps a small quantity of chemical into the nozzle. The nozzle is held against the tile and the chemical is forced through the tile by a pressurized nitrogen purge for four seconds. The nozzle diameter is about 1cm and the hole in the tile surface is about 0.1cm.

Inspection

During launch, reentry and transport a number of defects can occur on the tiles. These are evident as scratches, cracks, gouges, discoloring, erosion of surfaces. These defects are examined to determine if they warrant replacement, repair or no action. The typical procedure involves visual

inspection of each tile to see if there is any damage and then to assess and categorize the defects according to detailed criteria. Later, work authorization documents are issued for repair of these defects.

Design Constraints

As described above, the tasks dictate robot system specifications. In addition, the operating environment will also impact robot design. The following issues had strong influences on design options.

Facility Issues

During a flow, the time period between landing and takeoff, the orbiters are refurbished in the Orbiter Processing Facilities (OPF) at Kennedy Space Center (KSC). These facilities provide access to all areas of the orbiters through the use of intricate platforms that are laced with plumbing, wiring, corridors, lifting devices etc. With the exception of the jackstands that support the orbiters, the floor space directly beneath the orbiter is initially clear. This is not the case for the surrounding structure. During flow the work areas can be very crowded.

It was clear that the robotic system must require minimal special conditions for deployment and operation. As a result, it was decided that the robot must be capable of entering the OPF through the personnel access doors and gaining access to the lower surface of the orbiter via several different routes. These constraints resulted in the maximum height, width and length of the stowed robot to being 1.75m (70"), 1.1m (42"), and 2.5m (100"), respectively. However, once under the orbiter the tile heights range from about 2.9 meters to 4 meters.

Additional constraints include the negotiation of jack stands, columns, workstands and overcoming cables and hoses. In addition there are hanging cords, clamps and hoses. Because the system might cause damage to the ground obstacles (i.e. cables and hoses), cable covers are used for protection but the robot system must traverse these covers.

Task Issues

People are understandably cautious about robot systems in close proximity to flight hardware such as the Orbiter. The paramount concerns are safety to the personnel and the Orbiters and process integrity. As a result, the robot was designed to be fail-safe.

Since, DMES is flammable the system must meet or exceed Class I Division II Group D requirements of the National Electrical Code (NEC).

One of the tenets of the project is to impact the current tasks and flow as little as possible. This means performing the same tasks in at least the same amount of time with minimal operational and facility impacts and providing equal or better quality.

System Design Issues

There is a direct relationship between the size of the robot workspace and the number of tiles covered. As one might expect, the larger the workspace, the greater the number of tiles covered. What is not so obvious are the effects of the workspace size on the time that it takes to process all the tiles. For example, if the robot has a small workspace, then the time to stow, move and re-deploy may dominate the overall time, not the actual processing of the tiles. In this section we show the effects of workspace and mechanism movement time on the design.

Base Tessellation

Tessellation refers to the tiling patterns of the robot workspace across the total area of the orbiter to be serviced (the task- space). An important observation about tessellating the workspaces is that

there are always some inefficiencies in coverage due to overlap. From our studies we determined that the amount of overlap did not greatly affect overall time for a workspace greater than 150-200 tiles. This gave us a bound on workspace and therefore robot manipulator reach.

Base Move Effects

If operators are required to interact every few minutes with the system for monitoring base moves then the attractiveness of the system to users is far less than one that needs only infrequent attention. A goal of approximately one base move per half-hour was set. Once per half hour translates roughly into 80 moves during the course of rewaterproofing the orbiter. This results in a workspace of 300 tiles. It is important to note that a reduction of 1 second on the tile servicing time results in an approximate 4 hour reduction in total task time.

Mobile Robot

The facility constraints provided dimension limits for the whole system and task constraints drove system specifications such as speed and accuracy as well as physical specs on reaches with the workspace.

We examined a wide variety of options for the needs of these tasks. This included classes of devices that allowed inspection from afar, large fixed but movable, manipulators and even suction-cupped walkers. These ideas became detailed examinations of a wide variety of robotic devices. Many options were rejected on the grounds of flexibility, issues of self-sufficiency, safety to personnel or flight hardware etc. As a result of these preliminary studies the system we focused on was that of a mobile base integrated with a manipulator system.

Locomotion

The size constraints of the vehicle coupled with the close quarter navigation needs for operating in the OPF required a locomotion system of high maneuverability. A wheeled system utilizing Mecanum® wheels was selected which utilizes novel roller wheels to obtain three-degree-of-freedom (DOF) motion in the plane, pure rolling contact for accurate positioning, and non-singular motions for small and precise final motions.

Wheel compliance is eliminated by the use of automated screw jacks that descend from the base and contact the floor.

The base is formed by a very rigid welded steel frame. The design was deflection driven to provide a very stiff base from which to operate the manipulator. Figure 1 shows a general outline of the sub-systems of the mobile robot.

The base also supports two enclosures for electronics and rewaterproofing equipment as well as an on-board nitrogen tank and a battery cage.

Manipulation

When the base reaches a particular work area the stifflegs are deployed. The manipulator then deploys itself from it's stowed configuration. The manipulator provides a number of motions to reach the tiles. As shown in Figure 1 the first vertical motion is termed the Major-Z.

Linear rails connect the two Major-Z actuators to give a vertically raised rigid platform that can move the rest of the mechanism along the length of the robot.

A second vertical motion (Minor-Z translate) is then used to lift the later sections of the

manipulator. The two vertical motions are used because a single telescoping device could not provide the combination of stroke length, short stowed height, payload and accuracy needed. Atop this motion is a 360 degree rotating motion (Minor-Z rotate).

From this rotate motion a boom nearly a meter in length extends to a stow-deploy link. This link only swings the wrist and toolplate into position for the work. The need for this motion stems from the height requirements and the need to package the robot within the constraints imposed by the facilities.

The wrist is a modified Rosheim wrist that provides a hemispherical non-singular workspace. It is capable of moving and accurately positioning the end-effector (25kg).

To relieve moment loads on the Minor-Z motion the boom, wrist and toolplate are counter-weighted. The counterweight, although it increases overall weight and complicates deployment slightly, simplifies issues of accounting for deflections due to tremendous off-center loading conditions.

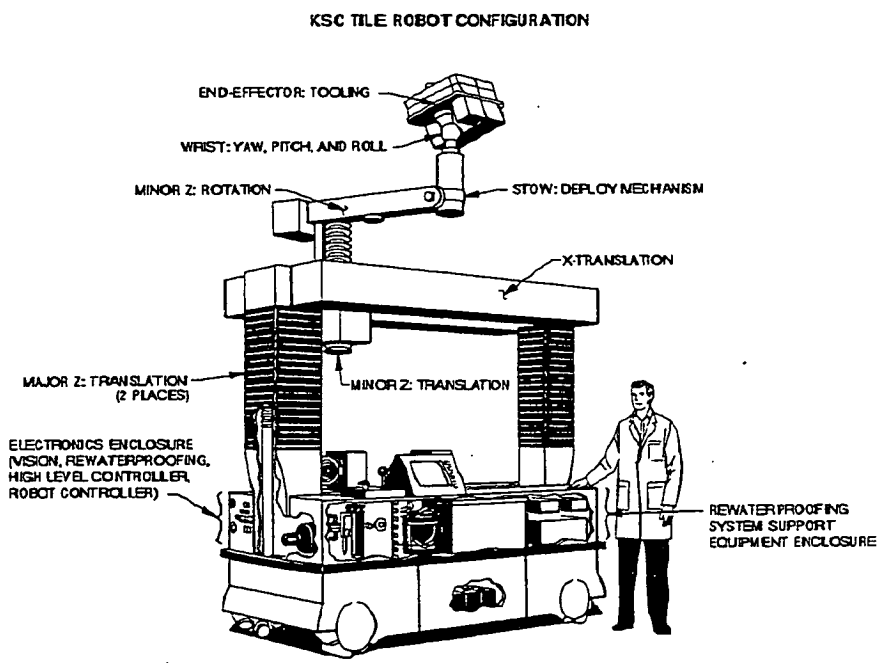


Figure 1 - Mobile Robotic System

All motions are designed to be manually operated for maintenance and component failure reasons. All motions also have absolute encoding to give position at all times, even in the event of power cycling or computer failure.

Positioning and Navigation

Precise positioning is needed to achieve accuracies of 1mm across the lower surface of the orbiter. An approach that utilizes two systems, delivers the required accuracy. A rotating eye-safe laser scanner reads bar code targets that are precisely located in the facility. Triangulation from three or more of the many targets can give us robot position within a few centimeters. This will position us

precisely enough to find a specific tile. The tile positions are known with respect to the shuttle and we can register the tile position with the vision system being used for inspection.

There are several frames of reference and corresponding transforms between them. The Orbiter is parked within some position error which is known and measured as a normal procedure. This gives us the orbiter-facility transform. Then the transform between the robot and facility is given by the laser positioning system and finally the loop is closed through the vision system which precisely identifies the position of a specific tile whose position is known on the orbiter. This finally gives the precise robot-orbiter transform.

Electronics and Power

The electronic design of mobile robot is driven by two major constraints: It must a) run un-tethered for up to 10 hours (one 8- hour shift plus setup, employment, and deployment time), and b) meet the NEC Class 1 Division II group D requirements for operating in a hazardous atmosphere.

Detailed power estimates indicated that a minimum of 15 kilowatt-hours needed to be carried on-board. Standard gelled lead acid batteries were chosen since they offer good power density, can be deeply discharged with no degradation, are simple to charge, and are rugged, cheap, and reliable. The battery pack is removable via a palette-jack, and the entire battery pack is charged off-board the robot. The robot can also be powered through a tether.

To meet the NEC requirements, all of the electronic enclosures are purged (Type Z per NFPA 496), including the battery pack. Additionally, excess heat will be removed from the main electronics enclosure with heat pipes.

Computing

The computing environment consists of four on-board computers (Figure 2) and one off-board database.

Three of the on-board computers are VMEbus based real-time systems: a robot controller which controls the base and manipulator motions and monitors the overall health and status of the robot; a vision system which performs the registration and inspection tasks; and a waterproofing system which controls the waterproofing injection system.

The two computer systems which directly control actuator motion (robot controller and waterproofing system) employ "safety circuits" between the computer servo outputs and the motor amplifiers. This piece of hardware has a large number of analog and digital inputs which monitor various safety critcal system parameters. The power to all motor amplifiers is immediately disabled should the safety circuit detect an out of tolerance parameter.

The fourth on-board system is the high level controller which resides on a Sun Sparc-2 computer. The controller is responsible for planning the course of action to complete a given task and appropriately commanding the subsystems. In the case of an error or failure in any system, primary safing is performed via the safety circuits, and the controller performs recovery actions. The controller also maintains a graphical operator interface, which allows the operator to load and update task parameters and stay informed about the progress of the current task.

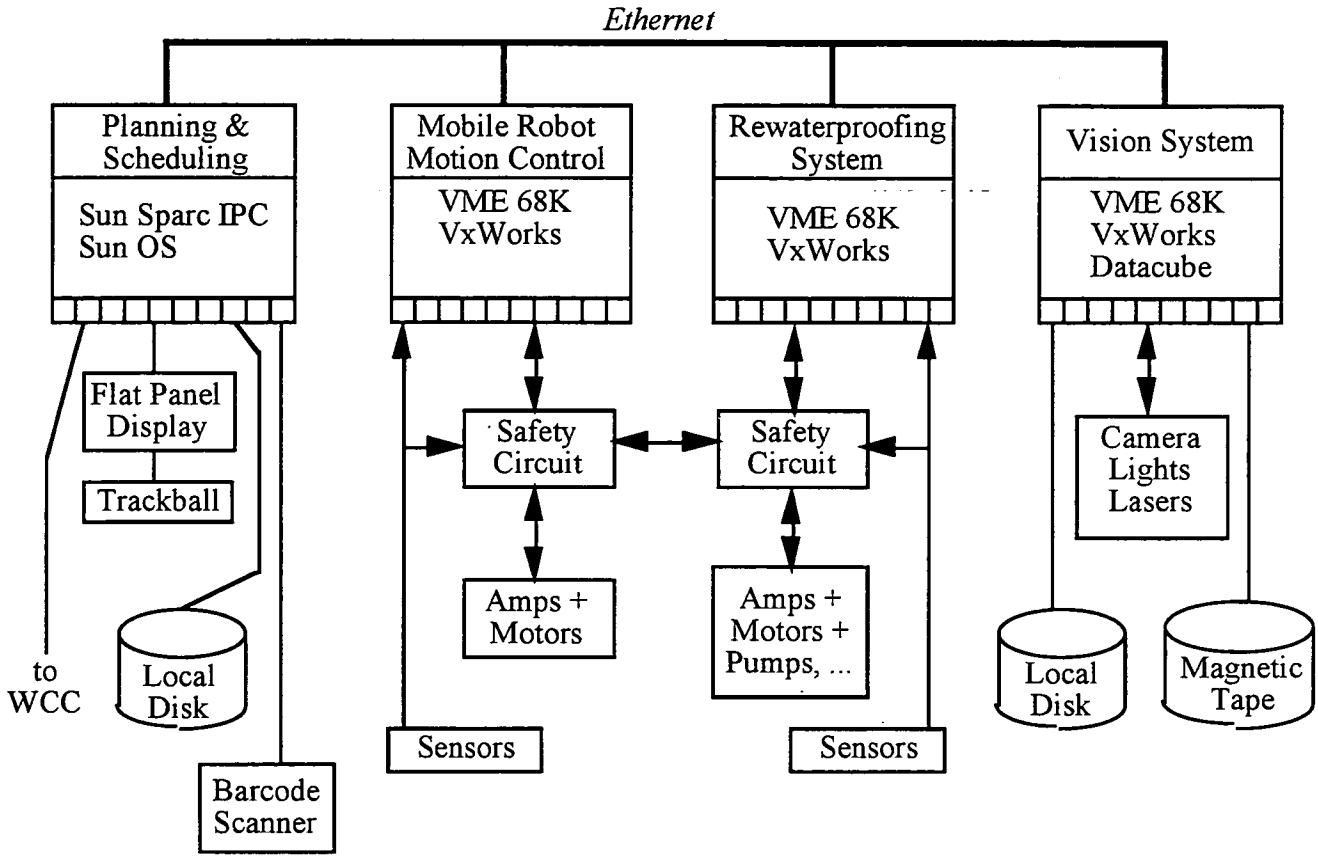


Figure 2 - On-board Computing Hardware

Software

Software comprises a large and critical portion of the entire system.. All software is written in ANSI C and utilizes ANSI compliant SQL.

Our goal from the beginning has been to allow a full upgrade path to autonomy in both hardware and software and to provide a self-reliant system. In system design this has meant provisions for all on-board power and supplies that might have been supplied through a tether. At the same time issues of safety and many 'what-if' scenarios had to be considered. The decision was finally for the operator-override option. This was no different than the full-autonomy mode but allowed the robot to be shut down at any time by the operator.

To protect both flight hardware and the robot, we are using untrasonic proximity sensing around the robot tooling. Contact bumper strips surround the base and critical sections of the manipulator. The operator is primarily accountable for robot actions during base movements and deployment and stowage of the manipulator.

Vision System

The vision system has two primary functions. One is to accurately determine the relative position and orientation of the robot tooling with respect to Orbiter tiles. The other is to perform post-flight visual inspections. To accomplish these functions, the vision system is comprised of a high resolution black and white camera, computer controlled diffuse(4) and specular strobes(4), laser light projectors, CPU, software, and image processing hardware. The accuracy requirements for

positioning are driven by the need to ensure that the 1 cm rewaterproofing nozzle is placed over the 1mm waterproofing hole. The location of this hole is calculated off-line through analysis of tile images. These images will be acquired by the robot system during an initial data acquisition run for each orbiter. The purpose of this initial data acquisition run is to effectively develop as-built geometrical data and to acquire baseline tile images for each orbiter.

Relative Position and Orientation

The vision system uses a two step process to accurately position itself with respect to a tile. First, it uses its laser light projectors to determine the perpendicular distance from the robot tool plate to the tile surface and the orientation of the optical axis with respect to the tile surface. The optical axis is perpendicular to the surface of the tool place which faces the tile. This information is used by the robot high level controller to move the vision system to the proper position and orientation so the remaining 3 degrees of freedom can be calculated. These remaining degrees of freedom are calculated through image matching techniques that utilize the current and baselined tile images.

Visual Inspections

The vision system performs visual inspections by comparing pre- and post-flight time images. It will identify areas on a tile whose visual appearance has changed. It does this by first aligning the pre and post flight images very accurately to account for any differences in vision system viewpoint location. The differences between these images are calculated. These differences are then processed to eliminate differences that are artifacts of the processing so only real differences in the tile's visual appearance are reported to operations personnel. Currently the vision system is capable of identifying missing tile coating and missing pillow type gap fillers. Plans are in work to extend the capability of the vision system to identify all lower surface tile defects.

Rewaterproofing System

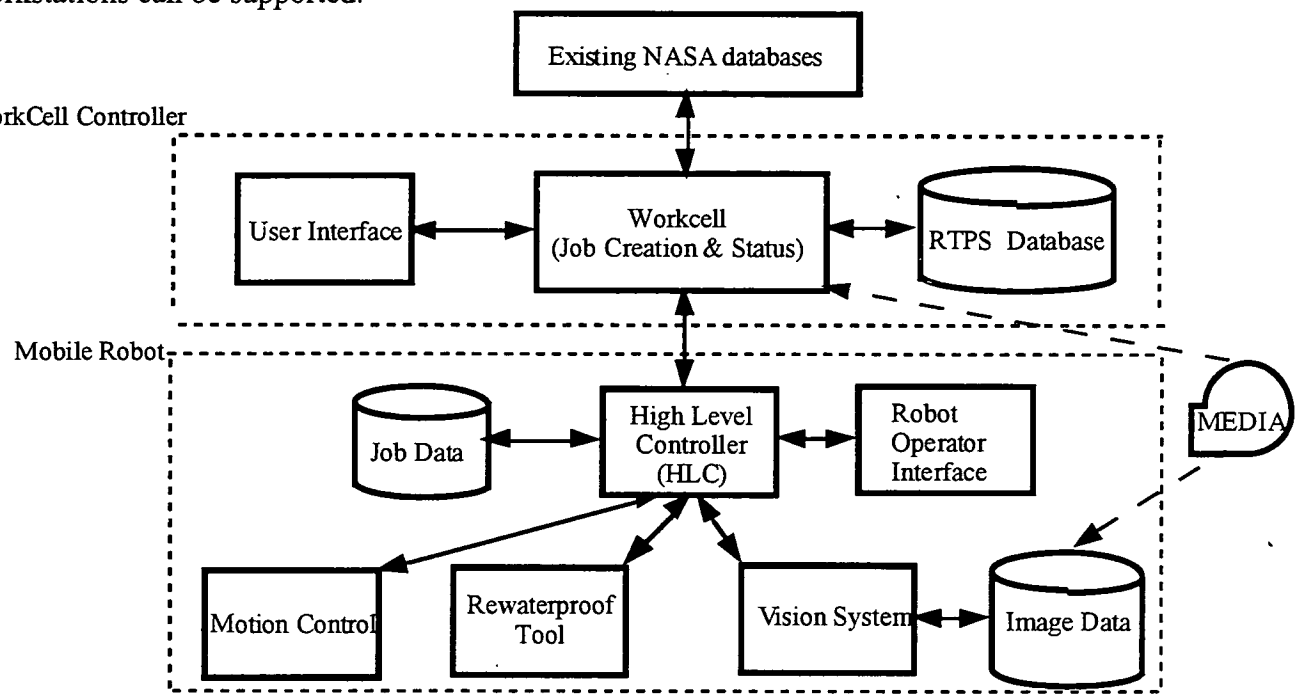
The rewaterproofing system was designed to automate the current manual rewaterproofing process. The system was designed to be fail safe to ensure that tiles were not damaged and that the proper amount of fluid was injected in each tile's rewaterproofing hole(s). It utilizes force control with redundant sensing to ensure that proper contact force is maintained between the rewaterproofing nozzle and the tile surface during the injection process. The nozzle is surrounded by a containment system seal and a slight negative pressure to capture any DMES from a failed injection. The containment system helps to minimizes unnecessary DMES from being vented to the local environment. Process completion is verified through redundant sensing of injection force and DMES injection pressure.

Workcell Controller

Considerable effort has been directed at making the robot fit the existing process. As the primary interface between operations and the mobile robot, the WorkCell controller (WCC) is central to the successful implementation and acceptance of the robotic system. Operations personnel will use the WCC to specify the robot's work and then manage the data which documents the results of that work. A user interface has been developed which allows the specification of work by graphical or textual methods. Process results will be available as textual reports or graphical tile maps which indicate status with stippling or color. Additionally, the WCC will manage and display the tile images acquired by the vision system. The requesting of a job and processing of the resulting data is done independent of the execution of work by the robot.

The WCC takes data from existing NASA databases and creates the tables which contain the data

required by the robot to complete a job. The overall information system architecture is depicted in figure 3. The WCC will interface to the Master Dimension Database (MDD) and the Tile Information and Processing System (TIPS). The MDD contains information on the geometry and location of each tile on the Orbiter. This data is used to calculate where to send the robot in order to complete a task. TIPS is a database which contains information about the Orbiter which is dynamic. The robot receives data from the WCC via a temporary ethernet link. The WCC utilizes a multitasking, distributed architecture. It is networked using TCP/IP and multiple workstations can be supported.



System Integration and Certification

All major sub-systems have been designed and fabricated in separate facilities. This has required tight control of all interfaces. Integration and acceptance test plans are being written to support system delivery to KSC and final system integration and testing at KSC.

The mobile robotic system has been designed as a "production certifiable prototype". This means that the system has been designed and built such that all relevant NASA requirements pertaining to Ground Support Equipment have been satisfied. However funding constraints have not allowed completion of a test program to verify this fact. Additionally the SR&QA, Maintenance, Spare parts, and Training plans will not be developed as part of the current program.

After delivery of the system to KSC, the robot will be thoroughly tested in KSC's Robotics Laboratory. This will be followed by demonstrations to operations personnel (KSC & JSC). Successful completion of these demonstrations will be used to justify certification of the RTPS system.

Conclusion

A prototype mobile robotic system for space shuttle servicing has been configured, designed and is currently undergoing system integration and testing. This robot system, when implemented, will

mark the beginning of a new era in the ground processing of critical space flight hardware at NASA's Kennedy Space Center.

Acknowledgments

This work has been funded through NASA's Code C Telerobotics program under the leadership of Dave Lavery and Mel Montemerlo.

This project is being done in partnership with Rockwell International, SRI International , NASA Langley Research Center , I-NET Space Services Inc., and the NASA Kennedy Space Center.

Reference

[NASA-TPS 90] NASA Kennedy Space Center, "Thermal Protection System Process Automation Study Final Report," KSC-DM-3491, October 1990

Robot Control in Dynamic and Uncertain Environments with Known Objects

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena

Thomas Peurach, Margaret Ganzberger, Douglas Haanpaa
Cybernet Systems
1919 Green Rd. Suite B-101
Ann Arbor, MI 48105

ABSTRACT

A telerobotic control system which integrates image processing, force reflection and computer graphics in an effective semi-autonomous robotic control/planning system has been developed. The system is composed of Cybernet's PER-Force 6-DOF force feedback handcontroller, JPL's operator control station (OCS) and machine vision algorithms developed at Cybernet. The system is designed to operate in uncertain environments having well defined objects. It connects physical world attributes with graphical world attributes by applying computer vision to well defined entities which can then be manipulated in 6 degrees-of-freedom . Making the appropriate connections, or establishing the correspondence between realities, facilitates efficient operations in graphical worlds while maintaining accurate manipulations in the physical world. Successful operations in the physical world depend on both visual feedback provided by video and tactile feedback supplied by the handcontroller. The system has been designed to support both semi-autonomous operations using a man-in-the-loop and full autonomy using functional scripts. The system's ability to operate in uncertain environments facilitates more dynamic utilization of robots in many applications, including material handling, hazardous materials disposal, and manufacturing applications. Many of the principles from this telerobotic control system have been applied successfully to specific applications such as teleoperation in manufacturing at the Ford Motor Company, an interactive 3-D stereoscopic viewing for molecular biologists at the University of North Carolina and bomb fuse removal. The developed system also provides unique capabilities for NASA remote space supervised telerobotic repair and maintenance operations, space platforms comprehensive inspection, and routine functions in space that could be accomplished remotely from the ground.

INTRODUCTION

Robotic systems are often used in environments with well known objects. The robot can be programmed to effectively move within the workspace and manipulate the objects. Unfortunately, when the environment changes or is uncertain, controlling the robot becomes quite difficult. However, since some aspects of the environment remain known, full human control is unnecessary. Thus, to effectively and efficiently utilize robots in these changing environments a system must be designed which optimizes the balance between the human controller and the autonomous actions of the machine controller.

In an attempt to meet the balance, we have designed a system which integrates graphical object representation, video input from the work site, machine vision, 6 degrees-of-freedom manipulations, and force feedback functionality. The graphical objects are used to represent well

defined entities in the environment, as well as provide estimates for uncertain components. In static environments, the graphical world representation can be used for robotic control. A number of commercially available packages provide this functionality, but they can not be used in dynamic, uncertain environments. To provide graphical object manipulation in dynamic and uncertain environments, the developed system constructs a bridge between the graphical world and the physical world. The physical world is represented in the system as video. Making the connection between the graphical objects and the video involves matching objects from the graphical world to objects in the video data. Once the bridge is established, graphical and physical actions correspond.

To establish the bridge, a system has been developed which uses both the human and the machine. The human is used for operations which are difficult for the machine and the machine is used for operations which are difficult for the human. Using a 6-DOF handcontroller, the human can easily provides a coarse pose estimation for objects in the dynamic environment. This operation is difficult for the machine to perform. The fine pose estimation, which may be difficult for the human, is determined by the machine using machine vision tools. Thus, by integrating the human's ability to perform coarse estimation, and the machine's ability to do fine estimation, a robot control system has been developed which can operate in dynamic and uncertain environments.

There are a number of applications which can utilize the developed robotic control system. The applications range in domain, but are consistent in that they have well known objects in dynamic and uncertain environments. The following sections will describe the vital components of the control system and will describe applications which can use the developed technology.

6-DOF HANDCONTROLLER

The developed system requires manipulation of graphical objects in 6 degrees-of-freedom for coarse pose estimation. In a previous SBIR for NASA, Cybernet developed a 6-DOF force reflection handcontroller (see Figure 1). This commercial product has received considerable accolades, and is currently being integrated into robot control and virtual reality environments .

The PER-Force handcontroller manipulates robots or objects by "feel." It can simulate a "sense of touch" by "force-reflection" with a wide motion range. This functionality greatly enhances the efficiency of operations which require manipulation and dynamic control of objects in multidimensional spaces. The PER-Force handcontroller is a small backdrivable robot which moves in 6 degrees of freedom, 3 linear positions (x-, y-, z-) and 3 attitudes (roll, pitch, yaw).

The robot levitates an aircraft-type sidearm-grip control stick which incorporates three cueing buttons, an analog trigger, and a palm-actuated deadman safety switch. An operator can use this motorized handle to precisely position other robots or graphically displayed objects to a given location (x-, y-, z-) and tool angle (roll, pitch, yaw). This is done by a host computer or a robot control system that reads the handcontroller joint or transformed position, velocity, or force.

"Force-feedback" can be generated on each axis by the handcontroller through 6 small, brushless, DC servo motors. The six axis force-reflection output and six axis orientation and position control make the manipulation and "feeling" of multidimensional objects or data sets extremely easy.

The PER-Force handcontroller can be driven by either a single card VME, Macintosh, or IBM-AT compatible computer. The device can be flexibly programmed to control either slave robots or graphical display systems. It provides position, rate, or force feedback to the operator.

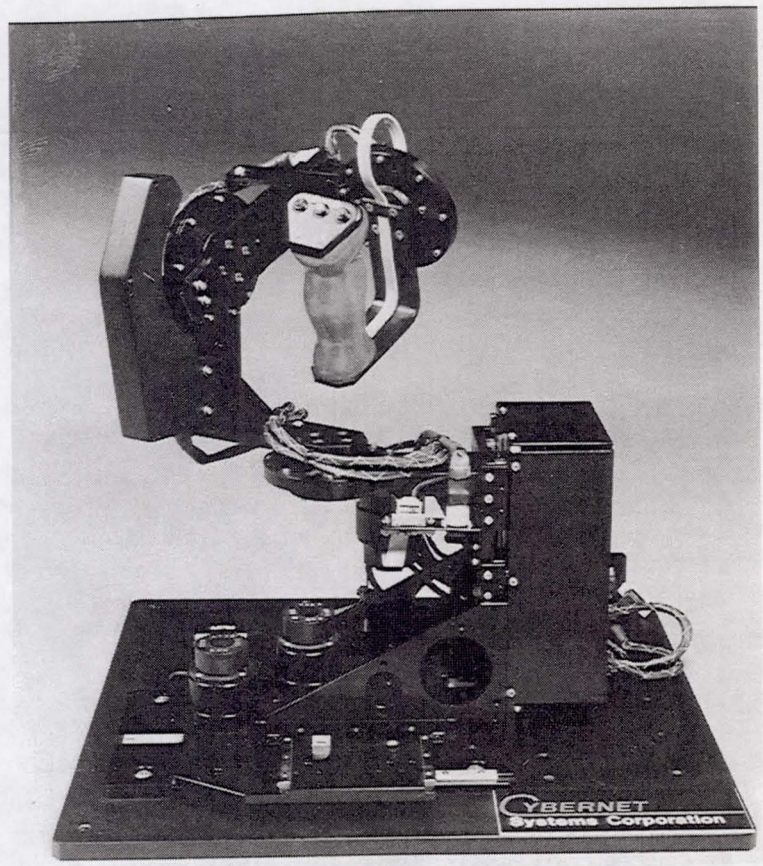


Figure 1. The Cybernet PER-Force 6 DOF Force Reflection Handcontroller

The kinematic arrangement of the PER-Force stick is designed for maximum simplicity and the best possible performance for both the electronic digital servo process and mechanical gravity compensation (Figure 2). The first two stages are a simple X-Y table (driven by a rack and pinion, and held in place by two parallel rails per stage). By convention X is side to side and Y is back and forth. Because these axes work perpendicular to gravity, no compensation is required.

The next stage is the Z axis, which is translated up and down. This axis levitates the yaw, pitch, and roll mechanisms, and the structures to which they attach. However, the Z motor and gear train themselves do not levitate (thus saving additional weight). The Z stage is gravity compensated by two constant force springs which are matched to the upper stage weight. The first revolute stage is yaw, which operates parallel to the base and therefore needs no gravity compensation. The next axis is pitch. The last axis is roll. All six axes of motion intersect at a point through the middle of the handle. We have found this to be the most comfortable pivot point for teleoperation.

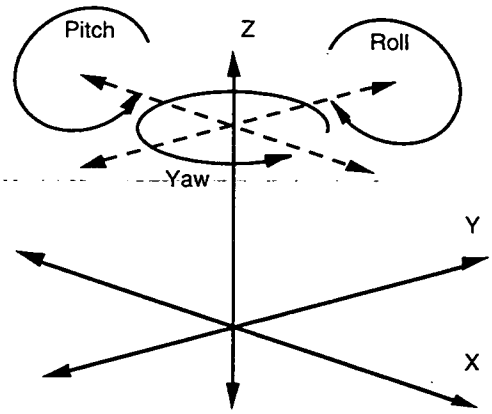


Figure 2. Handcontroller Kinematic Arrangement

Servo closure occurs in software loaded in the computer system. Thus, motor rotary position (sensed by encoders) is made available to the computer through six encoder decoding channels. The computer synthesizes velocity and acceleration from periodic position readings. Each servo loop cycle computes new motor torque values (actually drive voltages for the PWM amplifiers). These programmed torque values are interfaced to the PWM amplifiers through six digital to analog interfaces.

The PER-Force Handcontroller is completely programmable. A C library is provided with the controller software facilitating the development of advanced force reflection. In addition to the handcontroller resident software and libraries, an interface to standard UNIX machines is provided. Using the interface, programs can utilize nearly all the functionality of the resident handcontroller software. This flexibility enables the development of advanced user interfaces which use force feedback to implement new forms of machine-operator cooperative problem solving.

MACHINE VISION LOCALIZATION

Localization of 3D objects is the process of matching a graphical representation of the object (a wire frame in the described system) to a physical representation (an image in the described system). It is also known as "wire frame matching," "pose determination" and "attitude estimation." We developed a machine vision localization algorithm for use in detailed pose determination in space applications.

The localization module is conceptually simple. The following four steps best describe the algorithm:

1) Capture an image(s).
2) Overlay a wire frame on top of the image(s).
3) Rotate and translate the wire frame until it matches the edges of the object in the image(s).
4) Output the best match as the object pose.

To accomplish the four steps, the system incorporates coarse pose determination, wire frame projections, the Hough transform, and the Newton-Raphson method. The localization system is pictured in Figure 3.

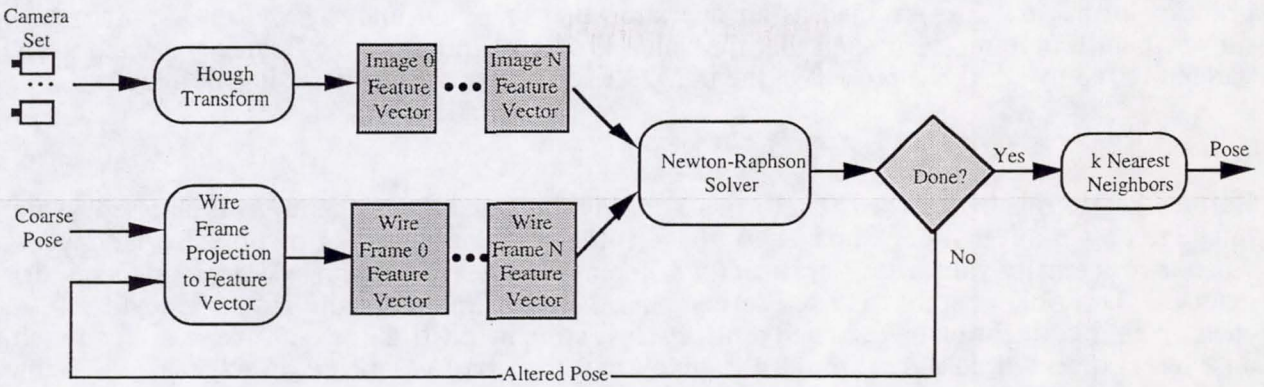


Figure 3. Cybernet's localization module.

The first step involves inputting *n* images from *n* cameras. The *n* images are of a single object in a single pose, but are captured from *n* different camera locations. These images are assumed to contain the object for which the pose is to be estimated. Each image is processed to form a feature vector. A feature vector is a set of measurements which condense the description of the image into a small, Euclidean feature space of *m* dimensions.

The localization module requires a coarse pose as input. Currently, the system uses Cybernet's PER-Force 6DOF with a human operator to obtain the information. From this initial coarse alignment, the localization module generates several wire frame projections with slightly perturbed orientations to compare to the object image. For each orientation a feature vector is used to describe the wire frame information. To avoid the combinatorial nature of generating multiple wire frames, an algorithm drawing on the Newton-Raphson Method of root-finding is used to determine which orientations to test and the best matching orientations.

Finally, a K-nearest neighbor weighted averaging of the top K best matching wire frames is used to produce a final orientation for the object. This pose is returned from the localization module to be used as the position of the object.

Each of the steps in the localization method; coarse pose determination, image feature vector construction, wire frame projection, measure of distance and the Newton-Raphson method, are described in detail in the following subsections.

Coarse Pose Estimation

The localization module depends on an initial pose estimation of the target object. This estimation allows the system to narrow its search for the correct attitude of the object. Currently, the localization system accepts user input for this coarse initial pose. Upon selecting an object in the environment to localize, the operator uses the handcontroller to position a graphical wire frame structure over the object in the image.

In some systems, the standard pose of the object is known, and the deviation of the object from this desired pose must be determined. In such systems, the coarse input pose is the standard pose.

Image Feature Vector Construction

The localization method employs the Hough transform [Duda] to describe the image of the object in feature space. The Hough transform is a common vision tool for detecting straight lines in

gray level images. It is defined as an operation that maps an image $f(x,y)$ into another two-dimensional function $H(r,\phi)$ such that the value of $H(r,\phi)$ indicates the degree to which a line parameterized by r and ø is present in the image. Each pair (r,ø) specifies a line defined by:

$$r = x\cos\phi + y\sin\phi.$$

Before the Hough transform is performed, an edge enhancement operation is applied to the image to obtain a rough edge image. An edge is the boundary between two pixels that appears when their intensity values are significantly different. In the localization system, a gradient edge detector, the Sobel operator, is used to compute the edge image for the Hough transform. At each pixel, a gradient magnitude and a direction is computed. If the magnitude is high enough, an edge is detected passing through the pixel, and the pixel is labeled an edge pixel. If the gradient magnitude is smaller than a threshold, the pixel is labeled as having no edge.

Since n images of the object are captured corresponding to n camera locations, an image feature vector is constructed for each input image. The images are first normalized to a set size, usually resulting in a reduction of size. Using an overlaid window grid (see Figure 4), the Hough transform is computed in small non-overlapping windows of the normalized image. The strongest line in each window is identified and parameterized into r and ϕ values. A feature vector is constructed such that each window contributes three values to the feature vector: confidence, r and ϕ. An example of the Hough features for an image is shown in Figure 4. The image on the left is a normalized cube image with overlaid grid windows. On the right, is an image of the strongest line found in each of the windows.

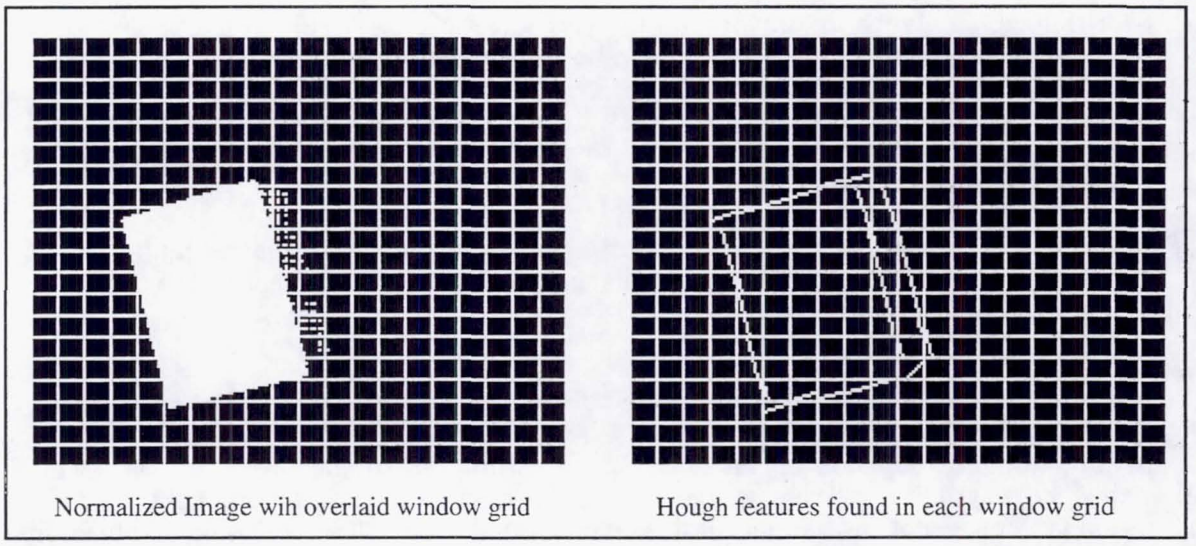


Figure 4. Normalized image and Hough features found in each window grid.

Wire Frame Construction

Once the images' feature vectors are created, the localization method creates and tests wire frames in various orientations to find the best matching pose (X, Y, Z, Roll, Pitch, Yaw) for the image. To construct a wire frame a perspective view must be created from the camera model. A camera model, as described in [Thompson], is a matrix of four vectors, C, A, H and V, describing

the location and orientation of the camera. This view determines which sides of the object are visible to the camera and which are hidden. The visible sides of the wire frame image are then projected onto the image plane.

To compare the wire frame to the object we mathematically create a feature vector for the wire frame. The wire frame feature vector and the image feature vector contain the same type of information. Construction of a wire frame feature vector begins with a desired position and orientation (X, Y, Z, Roll, Pitch, Yaw) for an object and a specific camera model. Matrix multiplication is used to transpose the vertices of the object using the camera model. Next, back face removal is used to determine which faces of the object are visible to the camera in the given orientation. The visible faces are projected into image coordinates. The image coordinates for the wire frame are then converted to polar coordinates (r,ϕ). As seen in Figure 5, the ray r is perpendicular to the edge and intersects the origin. Theta (ϕ) is the angle from the x-axis to the ray.

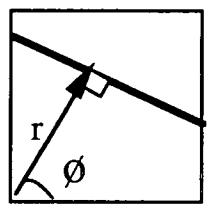


Figure 5. Polar coordinates for an edge.

With the edge converted into polar coordinates, the next step in processing the edge is to determine how it contributes to the wire frame feature vector. As in the construction of the image feature vector, a window grid is overlaid on the entire wire frame. The single wire frame edge is then tested to determine which window grids it intersects. For each window grid that the edge intersects, a triple of confidence value, r and ϕ is inserted into the wire frame feature vector at the appropriate location. If the wire frame has more than one edge intersecting a window, the additional information (edge number, r and ϕ) is kept in a separate structure so it can be used when matching the image feature vector.

Distance Measure

A measure of distance between two feature vectors indicates how well the two vectors represent the same occurrence in feature space. In the localization system, the distance measure computed between the wire frame feature vector and the image feature vector indicates how well the wire frame orientation matches to the image pose. The distance measure is based on the absolute difference between the feature vectors.

Newton-Raphson Method

The algorithm developed to locate the best matching wire frame draws from the Newton-Raphson method of root-finding. The Newton-Raphson approach requires the evaluation of both the function $f(x)$ and the derivative $f'(x)$, at arbitrary points x. The formula consists geometrically of extending the tangent line at a current point x_i until it crosses zero, then setting the next guess x_{i+1} to the abscissa of that zero-crossing. This method is depicted in Figure 6. The initial starting point for the Newton-Raphson method is very important. An initial guess far from the root can cause the Newton-Raphson formula to output grossly inaccurate, meaningless corrections.

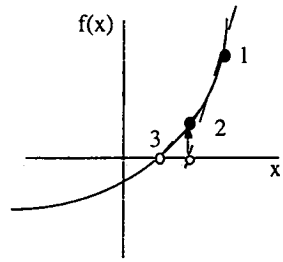


Figure 6. Newton-Raphson Method for one-dimensional root-finding.

In the existing localization system, values must be found for the six axes (X, Y, Z, Roll, Pitch and Yaw) such that the distance between the image and a wire frame generated with those axes values is minimized. Thus, the Newton-Raphson method must be generalized to handle the multiple dimensions. The problem in *N* dimensions translates to finding points mutually common to *N* unrelated zero-contour hyperplanes, each of dimension *N-1*. Identifying the neighborhood of a root, or of a place where there might be a root, as an initial starting location becomes much more important. In the system, the operator supplied initial pose information serves as a point in the neighborhood of the ultimate solution.

Unfortunately, the ultimate solution in the localization system will not have a value of zero. The features found in the image are not the same as the perfect edges of the wire frames. This is due to edge detection in images, which can be difficult due to image quality, image background and lighting.

As described, the localization system begins with a coarse input pose. Several wire frames are constructed and tested to find a best matching wire frame. This best match is then slightly perturbed to produce a new input pose and the process repeats. If the new input pose is the same as the last input pose, a local (or perhaps global) minimum has been reached. Each axis of the new pose is changed producing a different pose. This has the effect of popping out of the local minimum and allowing the system the opportunity to proceed to a new minimum.

From an input pose, wire frame vectors are generated for each of the camera models. Two values for each axis are explored, one value on each side of the input value. Currently, this step parameter on each side of the input value is 1° for the rotational axes (Roll, Pitch, Yaw) and 0.001 meters for the translational axes (X, Y, Z). Thus, as depicted in Table 1, if the input pose indicates a roll of -30°, the roll value for the generated wire frames will be either -29° or -31°. Similarly, an input value of 0.006 on the x-axis will result in x-values of 0.005 or 0.007 for the generated wire frames. All combinations of the two values for each axis are computed to produce a total of 64 (2^6) wire frame vectors for each camera model.

Each wire frame vector is compared to the image vector from the same camera model and a distance measure is computed. The distance measures from the *n* wire frames (from the *n* different camera models) with the same pose are compared and the largest distance measure is used as the measure for the pose as a whole. Thus, a pose is only as good as its worst match.

Hence, each pose (X, Y, Z, Roll, Pitch, Yaw) has one score that reflects its match for all the camera views. From the 64 poses tested, the pose with the lowest overall measure is taken as the best matching pose.

AXIS	-1° or -0.001m	Input Pose	+1° or +0.001m
X	0.005	0.006	0.007
Y	0.008	0.009	0.010
Z	0.019	0.020	0.021
Roll	-31°	-30°	-29°
Pitch	26°	27°	28°
Yaw	48°	49°	50°

Table 1. Perturbations generated from input pose.

AXIS	Best Pose	X -.001	X +.001	Y -.001	Y +.001	Z -.001	Z +.001	Roll -1°	Roll +1°	Pitch -1°	Pitch +1°	Yaw -1°	Yaw +1°	New Pose
X	0.005	0.004	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.0030
Y	0.010	0.010	0.010	0.009	0.011	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.0101
Z	0.021	0.021	0.021	0.021	0.021	0.020	0.022	0.021	0.021	0.021	0.021	0.021	0.021	0.0208
Roll	-31°	-31°	-31°	-31°	-31°	-31°	-31°	-32°	-30°	-31°	-31°	-31°	-31°	-30.6°
Pitch	26°	26°	26°	26°	26°	26°	26°	26°	26°	25°	27°	26°	26°	25.7°
Yaw	50°	50°	50°	50°	50°	50°	50°	50°	50°	50°	50°	49°	51°	48.8°
Score		43.62	44.35	69.32	40.89	39.56	48.78	52.21	47.28	42.97	44.61	39.43	56.14	

Table 2. Construction of new input pose.

From this best matching pose a new input pose is constructed (Table 2). Leaving all other axes with the value determined by the best matching pose, each axis is tested individually to determine which direction its value should be heading. A wire frame and match score is computed for a slightly incremented and a slightly decremented axis value. Again, this value is 1° for the rotational axes and 0.001 meter for the translational axes. Thus, if the best pitch angle is 13°, pitch angles of 12° and 14°, with all other axes remaining as in the best match, would be tested. If the distance measure is less for the incremented axis value, the value for that axis would be larger in the new input pose. Otherwise, the new value would be smaller.

The amount of the increment or decrement in the new value is based on the difference in the distance measures for the slightly incremented and decremented axis values. A small difference indicates that the axis is not very sensitive to the change, and thus a large change is needed for the new input value. On the other hand, a large difference in the measures indicates that the axis is sensitive to change and a small change in the value is needed. The translational axes are limited to a maximum change of .002 meters and the rotational axes are limited to 2° change. Bracketing bounds are set for each axis based on the original input pose defined by the operator and any increment or decrement of an axis must keep its value within the boundaries established.

If the new input pose is the same as the last input pose, a local minimum (or perhaps the global minimum) of the space of all permutations of wire frame values has been reached. To pop out of this minimum and continue the search for the global minimum, each of the input pose axes are incremented or decremented to form a new input pose. Again, any increment or decrement of an axis must keep its value within the original boundaries established.

The matching and refining of the input pose continues through a set number of iterations. Upon completion, the K wire frames with the lowest scores over all the iterations are weighted and averaged to determine a final pose for the image.

This algorithm for localizing three dimensional objects integrates wire frame projections, the Hough transform, and the Newton-Raphson method into a fast, robust and dependable localization module.

THE GRAPHICAL ROBOTIC ENVIRONMENT: OPERATOR CONTROL STATION

The developed system uses graphical representations of the robotic workcell for the coarse pose estimation and robotic control. The Jet Propulsion Laboratory (JPL) has developed and demonstrated a unique local-remote robot control architecture which enables management of intermittent communication bus latencies and delays as those expected from ground-remote operation. The JPL effort has focused on enhancing the technologies and transferring the control architecture to hardware and software environments which are more compatible with projected ground and space operational environments. At the local site, the operator updates the remote worksite model using stereo video and a model overlay/fitting algorithm which outputs the location and orientation of the object in free space. That information is relayed to the robot User Macro Interface (UMI) to enable programming of the robot control macros. This capability runs on a single Silicon Graphics Inc., machine. The operator can employ either manual teleoperation, shared control, or supervised autonomous control to manipulate the intended object. The remote site controller, called the Modular Telerobot Task Execution System (MOTES), is written in Ada and runs in a multi-processor VME environment on 68020 processors and performs the task sequencing, task execution, trajectory generation, and reflex motion.

The JPL robot control system utilizes a command interpreter. The command interpreter is a limited robot language which provides commanding of concurrent control from different control modules which execute based upon command parameterization. The permutations of control module behaviors are then available to the local site. This method allows a fixed software system to provide a wide range of robot control behavior. Also, the command interpreter approach has been proven successful on unmanned robotic spacecraft as Galileo [Galileo]. Utilizing the command interpreter approach, MOTES has been designed such that each module is data driven. A command to a module is a parameter set describing the desired behavior for that module. Capability is maximized by providing simultaneous control based upon various real and virtual sensors. The permutations of the behaviors of the various control modules provides the wide range of capabilities of the system. The desired behavior of each module is specified by commands from the local site which are issues by the remote site command interpreter.

MOTES Architecture

The motes system architecture is shown in Figure 7. The functionality of MOTES is similar to the Prim of the NASREM architecture. This level of telerobot system generates dynamic motion commands from a static description of the desired behavior. MOTES provides all task level control and task to actuator space mapping. The MOTES module types represent different functionality within the control system. There may be multiple modules of the same type, for example force, teleoperation, and collision avoidance control modules.

The Shared Memory Module provides access to all command parameters and system status information. The Executive Module handles communication with the local site system. It places new commands into the Task Command Queue and returns status and data. The Interpreter Module controls the transition between execution states by checking the status of the various modules and specifying the appropriate commands and parameters to the various modules via shared memory. The Monitor Modules provide monitoring of the status of execution for both intended termination conditions and unintended error conditions. The Sensor Modules provide sensor data processing. The Sensor Modules can represent both real and virtual sensors. The Control Modules provide the control associated with the various real and virtual motion sources. Each Control Module generates a task space motion command. The Fusion Module merges the

motion commands of the various Control Modules into task space motion commands for the manipulators and other physical devices. The task to Joint Map Module maps the task space command of the Fusion Module to the actuator space of the physical devices. The task space to joint space mapping for the extended task space is done using a composite Jacobian approach [Oh] using the method described in [Long]. The Device Drivers Modules communicate with the physical devices to send the actuator space commands and receive status data, as well as perform computations which are hardware specific. It is assumed that the physical devices have their own low level control to implement the actuator space commands. The M in the various boxes of Figure 7 indicates monitoring within the associated modules.

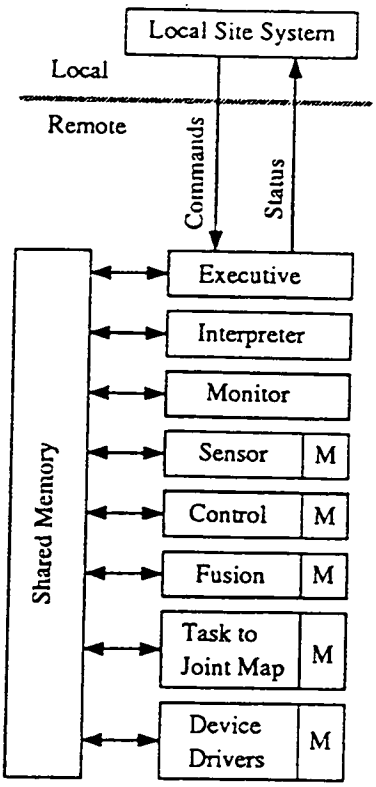


Figure 7. MOTES functional diagram

MOTES Modules

Each module interfaces to the rest of the system through shared memory with specified input and output parameters and functionality. This allows each module to be developed, tested and evolved independently. The modules operates asynchronously with respect to each other with the Interpreter responsible for synchronizing the various modules via modification of command and state parameters in shared memory. Modules on a given board may run as fast as possible or be interrupt driven, e.g. clock driven to allow fixed rate computations.

There are various types of commands that the Executive can receive from the local system. Command types may include Module, Interrupt, Reflex Table, Execution Mode, Initialize, and Emergency Stop. Additional command types, e.g., Cancel may be added in the future.

The MOTES system has been implemented. The present configuration utilizes six processor boards. The hardware diagram of the local-remote system is shown in Figure 8. The software design of MOTES was approached using the concepts of layered virtual machines and object oriented design within a consistent real-time methodology [Nielsen]. Modularity was achieved by developing various modules which could be easily configured onto different Ada tasks running on different boards, e.g., math, shared memory, trajectory generators, force control, teleoperation, impedance equation, forward kinematics, and inverse kinematics. The tasks run asynchronously from each other with some tasks clock driven and others running continuously. Communication between Ada tasks utilizes global shared memory exclusively except for an Ada rendezvous from the Monitor task to the Interpreter task to signal the arrival of the new command. Board memory between tasks and task memory between modules are language supported features but are not used since they would reduce reconfigurability of the system. Module memory is used when appropriate. The global shared memory communication is implemented via Ada generic units. The read and write utilities provided by the generic units provide protection of the data, e.g., complete record transfers. All parameters for one subtask are sent together in one command block. The command type is given as a parameter so that the Executive and Interpreter know how to process the command parameters. Thus data for all modules are placed together in one command and are then parsed out by the Interpreter. The destination queue parameter specifies which command queue to place the command in, e.g., Reflex Command Queues or Task Command Queue.

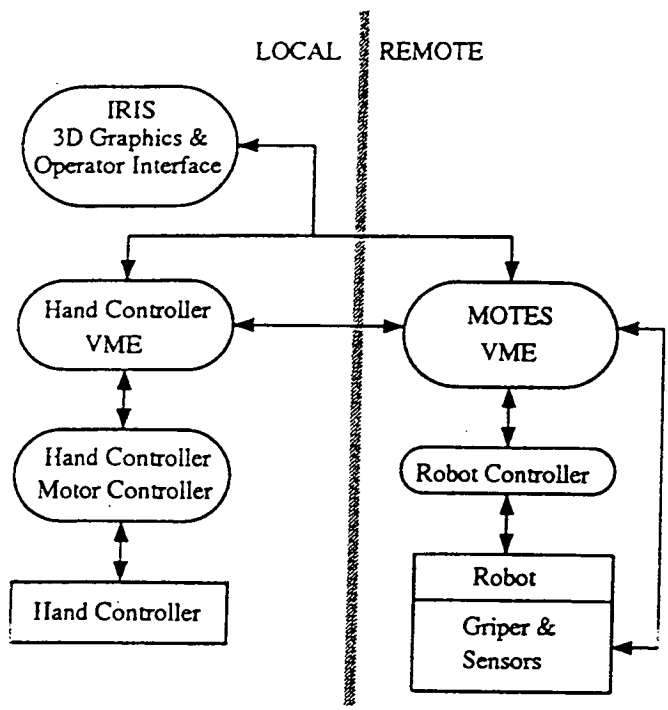


Figure 8. Local remote hardware diagram

Thus a telerobotic task execution system has been developed which provides redundant arm task execution for the remote site of a local-remote telerobot system. The system represents an approach to expanding telerobot performance beyond the baseline teleoperation capabilities into shared and supervised autonomous control. This enhanced capability is important for telerobot control with time delay for many industrial and commercial applications.

APPLICATIONS

There are a number of applications which leverage the described work. This section will describe applications in molecular modeling, manufacturing, factory automation, and fuse removal.

Manipulation and feedback for drug construction using molecular modeling. Cybernet Systems is developing an interactive 3D environment for molecular modeling. This molecular docking system is being created for the Materials Laboratory at Wright Patterson Air Force Base for the development of new materials. Integration of Cybernet's six axis force reflecting hardcontroller,a large screen projection TV, a 3-D stereoscopic viewing system, and the University of North Carolina's software - Grope III (a 3-D interactive computer graphics program), will provide molecular biologists a valuable tool to increase computational abilities. The system will simulate the interaction forces between a drug and a receptor site in a molecule and provide real-time feedback with regards to these forces in the form of visual and tactile output. The handcontroller is used to generate the forces and torques exerted on the drug molecule when it is aligned with the receptor site by the user's hand. As such, Cybernet is creating a cost-effective 3-D virtual molecular world. Figure 9 contains a picture of the handcontroller and the molecular modeling system in which the handcontroller has been integrated. The controller (Argonne ARM) in the molecular modeling system picture has been replaced by the handcontroller.

Figure 9. The PER-Force handcontroller with the molecular modeling software.

Teleoperation in manufacturing at the Ford Motor Company. The application under consideration involves loading transmission cases. These cases are very heavy and are delivered to the site in large bins in unknown orientations. Manual lifting and placing of the transmission cases is very dangerous. Our proposed approach combines an automated pick and place cycle robotics system with specific steps when an operator takes teleoperated control of the system. It includes an industrial robotic arm with an end effector capable of gripping and manipulating transmission cases. Control and manipulation is provided by a telerobotic force reflecting handcontroller electronically interfaced to the robot arm. Both teleoperated and automatic

motions are supported in the system to achieve the flexibility of teleoperation coupled with the rapid cycle speed possible through automation. The alignment of transmission cases is enhanced through operator views provided by two cameras (and a graphics enhanced video viewing system). Transmission allotments are provided through a conveyor system. All moving parts/robotic elements are surrounded by safety fences. For a complete description see [Marzwell *et al*].

Fine pose determination system for factory automation. Machine vision algorithms have been developed which localize or match, graphics and image data. Localization is essential for a number of assembly and manipulation robotic tasks. More specifically, as parts move down assembly lines, many times the coarse pose of the object is known, but due to imperfections in placement of the object, the exact pose is unknown. In these scenarios it is very difficult to perform fine assembly tasks since there exists no exact measurements. To facilitate fine assembly and manipulations, the localization modules developed were slightly augmented to produce a commercial fine pose determination product. This product integrates machine vision localization, camera calibration, error estimation and a Graphical User Interface (GUI) into a system which facilitates factory automation.

Removal of the fuse from unexploded bombs. The recovery of unexploded bombs has its obvious benefits in bomb analysis and safe disposal, but not many people are willing enter an area containing an unexploded explosive. The technologies described in this paper provide the infrastructure to remove the human from the unnecessary danger of defusing unexploded bombs. Using the infrastructure, a semi-autonomous or man-in-the-loop system which integrates Cybernet's 6-DOF force feedback handcontroller for manipulations and machine vision for autonomous manipulations can be constructed. The fully integrated fuse removal system will benefit both military and civilian organizations. Military organizations will be able to test new ordnance for failures. Civilian organizations, such as the police, will be able to defuse bombs to avoid property damage and personal injury.

SUMMARY AND CONCLUSIONS

The system described successfully integrates computer vision, 6DOF manipulation, and graphical object representations to control robots in dynamic and uncertain environments containing well known objects. The construction of such a system strongly leverages the abilities of the human controller and the machine controller by balancing the input of each. The result is an advanced robotic controller which facilitates the use of robots in applications which do not currently utilize robotic technology.

ACKNOWLEDGMENTS

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, Office of Advanced Concepts and Technology. Cybernet Systems effort was funded from NASA/JPL Small Business Innovation Research (SBIR) under contract number NAS7-1189. The authors would like to acknowledge Dr. Brian T. Mitchell and Dr. Charles Jacobus of Cybernet Systems and Dr. Paul G. Backes and Dr. Bruce Bon of the Jet Propulsion Laboratoy for their technical input and support.

BIBLIOGRAPHY

[Duda] Duda, R. O. and Hart, P.E., "Use of the Hough transform to detect lines and curves in pictures", Communications of the ACM, 15, 1972, 11-15.

[Thompson] Thompson, A.M., "Camera Geometry for Robot Vision," Robotics Age Magazine, Mar/Apr 1981, 20-27.

[Marzwell] Marzwell, N.I., Jacobus, C.J., Peurach, T.M., Mitchell, B.T., "The Use of Interactive Computer Vision and Robot Hand Controllers for Enhancing Manufacturing Safety," Proceedings of Technology 2003, Anaheim, California, December 7-9, 1993.

[Galileo] Galileo Project. Galileo program description document - command and data subsystem, phase 9.1. Technical Report 625-355-06000, D-535 Rev. G., Jet Propulsion Laboratory, May 1989.

[Oh] S.Y. Oh, D. Orin, and M. Bach. An inverse kinematic solution for kinematically redundant robot manipulators. Journal of Robotic Systems, 1(3):235-249. 1984.

[Long] Mark K. Long. Task directed inverse kinematics for redundant manipulators. Journal of Intelligent and Robotic Systems, 6:241-266, 1992.

[Nielsen] Kjell Nielsen and Ken Shumate. Designing Large Real-Time Systems with Ada. McGraw-Hill Book Company, 1988.

Electronics

Fiber Optic Communication Systems for Spaceflight and Avionics Applications

NOT AVAILABLE AT THIS TIME

The Global Positioning System (GPS) Service: A Technology Ripe for Commercial Innovation

NOT AVAILABLE AT THIS TIME

SATS: SMALL, AUTOMATED TRACKING SYSTEM -- ELEMENTS OF A BETTER SYSTEM FOR SATELLITE TRACKING AND TELEMETRY

Jeffrey M. Srinivasan
Technical Group Leader
Jet Propulsion Laboratory
Pasadena, CA 91109

Stephen M. Lichten
Technical Group Supervisor
Jet Propulsion Laboratory
Pasadena, CA 91109

Bruce J. Haines
Member of Technical Staff
Jet Propulsion Laboratory
Pasadena, CA 91109

Lawrence E. Young
Technical Group Supervisor
Jet Propulsion Laboratory
Pasadena, CA 91109

ABSTRACT

The Jet Propulsion Laboratory (JPL) has been exploring applications of precise Global Positioning System (GPS) techniques to navigation and data communication for Earth orbiting spacecraft. GPS tracking can be exploited in several different ways, depending on the orbital altitude of the spacecraft of interest, to support orbit and trajectory determination. At low-Earth orbits below 3000 km, "upwards-looking" GPS tracking—analogous to ground-based GPS tracking—can be used to provide real-time orbit determination for navigation. For these applications, GPS flight receiver architectural studies coupled with advances in GPS data analysis and estimation techniques at JPL have resulted in a wide range of GPS-based navigation capabilities that trade off orbit accuracy for instrument complexity and cost. Orbit accuracies of several hundred meters (ultra low power architecture) down to 10 meters can be routinely achieved in real-time with a properly equipped, stand-alone flight receiver. Much higher accuracies can be achieved in a post-processing mode when data from a global ground network of precise GPS receivers are differentially combined with the flight receiver data. This technique has been demonstrated at JPL, where orbits accurate to better than a few cm in height are generated automatically for the TOPEX/Poseidon spacecraft. In addition, a GPS flight receiver architecture that integrates the command telemetry receive function is being explored and shows significant cost savings potential. At altitudes between 3000 km and 8000 km, visibility of GPS rapidly decreases and it becomes advantageous to add a nadir pointing antenna in order to continuously see enough GPS signals to navigate an orbiter. For orbits above 8000 km, JPL has developed the GPS-like tracking (GLT) technique which dispenses with the on-board GPS receiver in favor of transmitting beacon (usually the existing spacecraft-to-ground link (SGL)) whose phase is tracked, simultaneously with normal GPS signals, by a ground network of "enhanced" GPS receivers. With the GPS data providing several key calibration parameters, the SGL phase data can be processed in near real-time to produce spacecraft orbits with accuracies of a few tens of meters up to geosynchronous orbits and beyond. A recent JPL experiment demonstrated that the GLT approach can be used to determine the orbits of NASA's geostationary Tracking and Data Relay Satellites (TDRS) to better than 25 meters. The systems referred to above all have the potential to provide inexpensive and autonomous navigation/orbit production and, in some cases, integrated data communications for a wide class of Earth orbiters and should be of interest to designers of NASA, military, and commercial space systems.

INTRODUCTION

GPS satellites (presently numbering 25) transmit carrier signals at 1.228 and 1.575 GHz (L-band) which are modulated by pseudo random noise (PRN) ranging codes as well as a navigation message with GPS clock and orbit information necessary for real-time positioning. In normal operation, the Department of Defense (DoD) turns on *selective availability* (SA) for most GPS satellites, introducing a clock dither and adding errors to the broadcast

ephemeris, as well as *anti-spoofing* (AS), which encrypts the precise ranging codes (P-code). Authorized users can be equipped with GPS receivers which accept keys to correct for these effects yielding about 10 meter stand-alone positioning accuracy, but other users see only 50-100 meter accuracy.

Civilian and scientific uses of GPS have led to a wide variety of applications in geodesy, surveying, navigation, and remote sensing, and have resulted in the development of sophisticated strategies which enable accuracies much greater than 50-100 meter. Examples include a cm-level non-real time positioning capability for receivers on the surface of the Earth [1], several-cm accuracy for low-Earth satellite orbit determination [2], and, in theory, several-meter accuracy for high-Earth orbiters [3]. Such high-precision applications typically require a global GPS ground network of high quality dual-band receivers and simultaneous post processing of data in estimation software which incorporates detailed physical and observation models.

JPL is developing several candidate technologies in the areas of GPS receiver hardware and data analysis software that are suitable for either direct transfer to industry or further co-development with commercial partners. These technologies are based on JPL-developed high precision, configurable GPS receiver architectures and automated, highly flexible, GPS-based multiparameter estimation algorithms that have been adapted and streamlined specifically for Earth orbiter trajectory and orbit determination and data communications. This paper will present these technologies, some of which have been demonstrated in the field while others are still in the conceptual stage, as they apply to Earth orbiting spacecraft at various altitudes.

LOW EARTH ORBITER TECHNOLOGIES

Most military and civilian ground-based GPS applications involve an upward-looking geometry where the users' receiving antennas are pointed away from the Earth towards the GPS satellites. This geometry is depicted in Figure 1, which shows the low-Earth orbiter TOPEX/Poseidon and a global network of ground stations tracking GPS satellites.

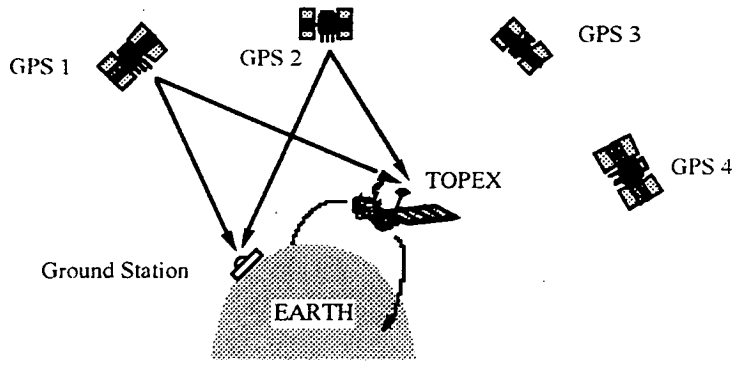


Figure 1. Upwards looking geometry for low-Earth orbiter and ground stations tracking GPS.

TOPEX/Poseidon, a joint NASA/CNES spacecraft operated by JPL, carries a radar altimeter to map the oceans' surfaces and measure global ocean circulation. The satellite also carries a flight qualified GPS receiver to demonstrate GPS-based spacecraft orbit determination. At its altitude of 1340 km, 6 to 8 GPS satellites are typically visible at a given time with a zenith-pointing, hemispherical field of view antenna. By relying at least partially on available precise force models, a dynamical fit can be performed in a sequential filter using only GPS flight data over at least several hours. The accuracy of the resulting solution (a few tens of meters) improves significantly over that achieved from stand-alone geometric solutions (50 to 100 meters). Additional accuracy improvement results from differential cancellation of the receiver and transmitter clock errors (and SA clock dither) at each measurement epoch when ground and flight receiver data are processed together. This requires common visibility of at least two GPS in the flight receiver and a ground receiver (see Figure 1). When all the GPS orbits are estimated as well, low-Earth orbit determination at the few cm-level (2-3 cm RMS, radial component, ~10 cm RMS in cross track and along track components) was demonstrated by JPL for the TOPEX/Poseidon spacecraft [2] using about 15 ground sites.

The data analysis software being used for the above demonstration was developed at JPL and is an adaptation of the GIPSY/OASIS II package of modeling and estimation algorithms [2]. The ground network of 15 stations consists of geodetic quality dual-frequency receivers [4,5] which can utilize the P-code when not encrypted, or rely on codeless techniques to recover precise ionospherically corrected observables when AS encryption is on. The GPS receiver used

to collect the TOPEX/Poseidon flight data is also a dual-frequency P-code receiver, but with moderate performance and relatively high mass and power consumption specifications [6]. In contrast to the ground stations, the TOPEX/Poseidon flight receiver reverts to L1-C/A tracking when the P-code is encrypted. The next few sections will examine more closely, and consider future improvements to, the data analysis strategies, ground network design, and flight receiver design as they apply to orbit and trajectory determination for low-Earth orbiters.

Data Analysis Strategies and Implementation

The GPS data collected by TOPEX/Poseidon are routinely processed in an automated fashion with GIPSY/OASIS II analysis software using a technique known as reduced dynamic tracking. The reduced dynamic strategy exploits the extraordinary geometric strength of GPS to minimize the dependence on dynamic force models and achieve a superior orbit solution through an optimal synthesis of dynamic models and geometric observables [2]. Other (ground-based) techniques against which the GPS-based orbits were compared, in contrast, provide measurements in just one direction at a time and may have substantial coverage gaps; they must therefore rely on models of the satellite trajectories (derived from dynamic force models) to recover three-dimensional information.

The accuracy of GPS-based reduced dynamic orbit solutions were assessed using three approaches: 1) internal consistency checks within the GIPSY/OASIS II processing system; 2) comparison with other orbit solutions, some of which were produced with independent data while others used the same data but different processing software or models; 3) external consistency checks that exploited the independent satellite height information provided by the on-board radar altimeter. These orbit quality assessment tests are described in detail in *Bertiger et al.* [2].

The GIPSY/OASIS II software consists primarily of a GPS data editor, orbit integrator, measurement model generator, and filter/smoother. An automated executive ties the modules together producing daily orbit solutions unattended. The system typically produces a reduced dynamic orbit solutions for TOPEX/Poseidon within 2 days of acquisition of flight GPS data, using less than 6 CPU hours on an HP 735 workstation. In addition, an experimental executive script was demonstrated by JPL which automatically and continuously produced near-real time TOPEX/Poseidon orbits within about 8 hours after midnight (for the day which ended at midnight). Since these quick-look orbits were determined later to be accurate to better than 1 meter even when propagated ahead in time 24 hours, the demonstration actually produced sub-meter real-time knowledge of the satellite's three-dimensional position. In altitude, the RMS accuracy of the quick-look orbits was about 5 cm, with real-time knowledge better than 10 cm.

Ground Network Design

In addition to producing post-processed high-accuracy GPS-based orbits for TOPEX/Poseidon, JPL has generated experimental quick-look orbits for TOPEX/Poseidon using data from a subset (between 0 and 12 sites) of the global ground network. This demonstration was performed over several weeks in the fall of 1994, when anti-spoofing was on continuously, and Figure 2 shows expected orbit determination accuracy with various combinations of data from the flight and ground receivers.

In Figure 2, cases 1–3 are based on actual results achieved for TOPEX/Poseidon using data from various size ground networks, and are meant to illustrate how the orbit accuracy varies as a function of the number of ground stations included in the solution. Cases 4–6 are results from simulations (these solutions were not actually determined in real time on board TOPEX/Poseidon, but were evaluated after the fact in simulations of real-time estimation procedures) and are representative of what could be achieved on low-Earth satellites such as TOPEX/Poseidon using only data from the flight receiver in an on board autonomous filtering strategy. Especially noteworthy is Case 6, as it represents a configuration in which only a small subset (5 %) of the actual GPS data from the flight receiver are used in the orbit determination solution. Theoretically, this means the receiver need only be turned on during brief data acquisition periods, resulting in drastically reduced average power consumption while incurring only a small degradation in accuracy. This receiver architecture, among others, is discussed in the next section.

Other experimental versions of GIPSY/OASIS II have been used recently to test the capability to provide real-time meter-level knowledge of the GPS orbits, clock and SA delays, and ionosphere delays. Such products could be used in real-time navigation applications, such as the various systems being considered by the FAA for aircraft positioning.

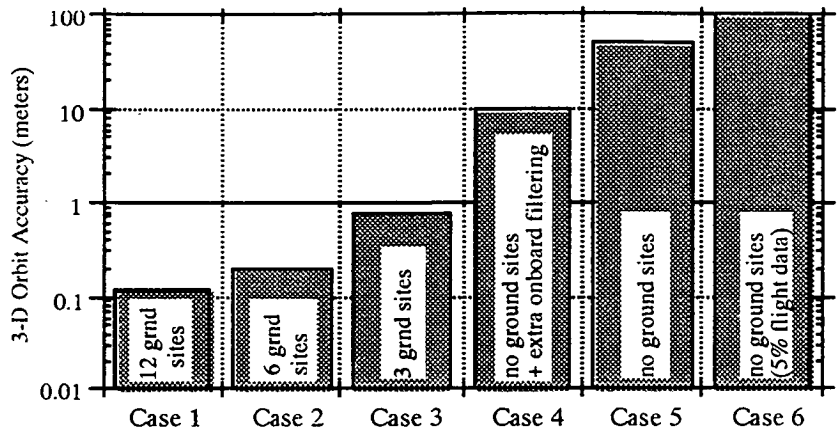


Figure 2: Comparison of low-Earth orbit accuracy for different strategies, with number of ground sites varying from 12 to zero. Cases 1-3 are from actual TOPEX/Poseidon solutions, while cases 4-6 are simulations; the fourth case incorporates extra filtering parameters to reduce errors from the fifth case (state-only filter), and the last (sixth case) includes only 5 minutes of data every 2 hours (such a strategy could save power on a future mission)

Flight Receiver Architectures

GPS receivers in low-Earth orbit can offer the following advantages:

- enables autonomous on-board navigation with 10-100 m accuracy
  - ground processing required only for Precise Orbit Determination (POD) (<10 m orbits)
- simplifies ground equipment requirements
  - require telemetry stations only; no Doppler tracking required
  - does require access to GPS ground network data for POD
- provides the following signals to the rest of the spacecraft (in certain architectures):
  - one pulse per second timing signal with 0.2 - 1.0 μsec accuracy
  - frequency standard with stability approximately equivalent to the GPS satellite clock ensemble
- integrates spacecraft uplink command receive function (in certain architectures)
- provides real-time platform attitude information (in certain architectures)
- provides "cost-free" atmospheric science opportunity with appropriately located supplemental antenna

JPL has been investigating the above advantages and the following sections report current status and results.

Atmospheric Science with GPS

Addressing the last advantage first, JPL has great interest in the atmospheric limb-sounding science potential of high precision GPS receivers in low-Earth orbit and is actively pursuing several missions of opportunity. The GPS/MET instrument, funded by the National Science Foundation (NSF) is being developed by JPL under the direction of the University Consortium for Atmospheric Research (UCAR) and is scheduled to be launched by Orbital Sciences Corporation (OSC) on a Pegasus launch vehicle. An improved instrument is being funded by NASA and designed by JPL for the Danish spacecraft, Oersted. Both instruments are high precision GPS receivers that will collect global data which will be used to assess the utility of the GPS atmospheric occultation data type [7]. Figure 3 illustrates the GPS signal occultation of the Earth's atmosphere. The time varying phase delay of the GPS signals will be recorded, downlinked and analyzed to produce global temperature and pressure profiles of the troposphere as well as ionospheric measurements. It would be advantageous to climate modeling or weather prediction to have as many occultation-capable receivers in orbit as possible; producing simultaneous, globally distributed atmospheric occultation data arcs. Since the occultation science requirements have led to a receiver architecture that also offers autonomous navigation, it may be possible to include such receivers on large constellations of commercial satellites, creating a potentially vast source of science data.

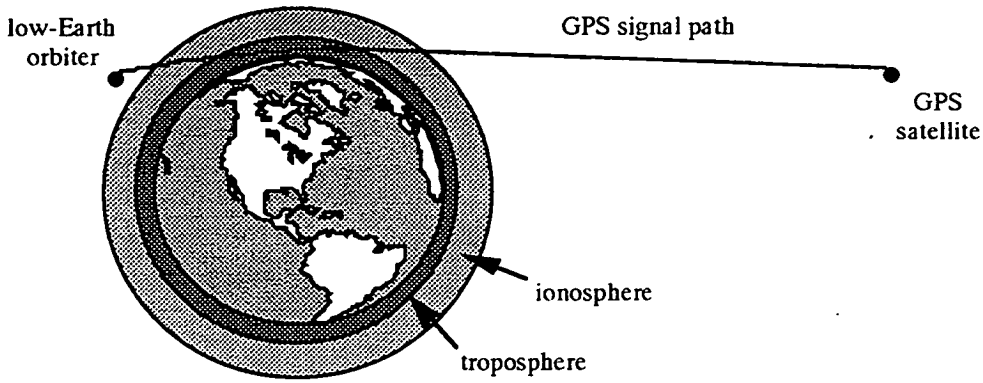


Figure 3: Atmospheric Occultation Geometry with GPS Satellite signals and LEO-based receiver

Configurable Receiver Architectures

Spacecraft and spacecraft instrument designers must select from a continuum of options that seek to balance oft-conflicting performance and "cost" constraints. Typically, the "costs" are power consumption, mass, as well as actual cost of development, fabrication and test. JPL is investigating several GPS receiver architectures that are readily scaleable and offer a convenient way of trading off power/cost/mass constraints against navigation performance requirements. Performance can refer to real-time position accuracy as well as ultimate knowledge of the orbit gained through inclusion of ground network data and post-processing. The architectures presented herein address both definitions. They are intended to offer the space system designer a wide range of choice when inclusion of a flight GPS receiver is desired.

The simplest receiver architecture is illustrated in Figure 4. It offers miniaturized receiver with extremely low power consumption in exchange for reduced accuracy. This reduced power is achieved by only collecting very short time samples of GPS data a few times per orbit (see Figure 2 for orbit accuracy degradation when sparse GPS data are used). When the optional GPS processor is included, it would operate on these stored bits and produce a point position after each collection time. Alternatively, in exchange for even more simplicity and lower power consumption, the on-board processor can be eliminated and the stored GPS signal samples telemetered to the ground for post processing. This option reduces spacecraft autonomy, increases downlink bandwidth requirements, and introduces latency in orbit updates; all of which may be acceptable for spacecraft with severe power/mass limitations. In addition, this simple architecture can offer a low cost, backup mode to any flight receiver; at any time a short interval of antenna data can be captured and downlinked to the ground for analysis and health assessment. This architecture has not yet been demonstrated, but performance has been verified in computer simulation (see text accompanying Figure 2).

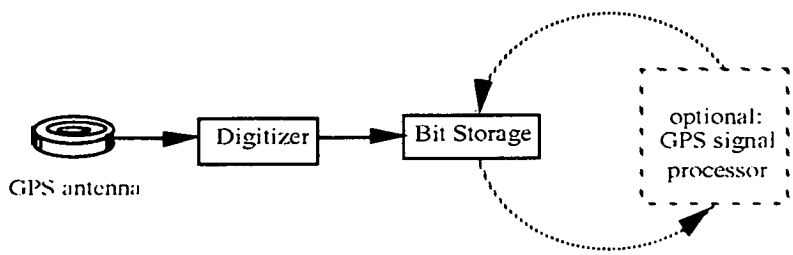


Figure 4: Functional Block Diagram of ultra-low power architecture

Figure 5 lists a sampling of receiver configurations which offer a wide range of performance/power consumption compromises. The underlying architecture for all these configurations is based on the geodetic quality, dual-frequency ground receiver developed by JPL. The GPS/MET and Oersted flight receivers development experiences have led to several of the key advances that enabled this architecture.

The first key advance was the design of satellite tracking channel that can process GPS signals with the same processing hardware, independent of whether the *anti-spoofing* (AS) function is on or off [8]. The second

development was a receiver design which allows dual-frequency tracking to be deleted for users that do not require ionospheric corrections. This results in a receiver with lower power, size and mass, but with all components already space qualified. The third development is the inclusion of power management features in the receiver architecture. Specifically, the receiver that can selectively turn off various receiver components when not in use to reduce power consumption. It can also vary the clock speed of the controlling microprocessor and the data rate of the GPS signal processor to meet a specific performance goal with the minimum power usage.

Architecture Description	Accuracy	Power Consumption
Digitize, Store, & Forward to Ground	~200 m	< 0.1 w (avg per orbit)
Digitize, Store & Process	~200 m	< 0.5 w (avg per orbit)
L1 Sparse Data & Fit to Orbit	~100 m	~1 w (avg per orbit)
L1 Continuous Data; point positions only	50-100 m	~4 w (continuous)
L1 Continuous Data; filter/fit to orbit	10-20 m	~4 w (continuous)
L1/L2 Continuous Data; ground processing w/ ground network data	subdecimeter	6-8 w (continuous)

Figure 5: Range of receiver architectures with estimated performance and power consumption

Clock Steering Capability

Properly equipped receivers, when not connected to an external frequency standard, can steer their internal oscillators using real time, in-receiver GPS solutions in order to keep the receiver clock offset from GPS time reasonably small (typically <200 ns in the presence of SA clock dither). Long term (>1000 sec) stability of the receiver clock is controlled by the ensemble of the GPS clocks being tracked. Data processing from a high-Earth orbiter demonstration employing a ground network of receivers with clock steering, showed formal errors on the clock offset estimates to be 0.2 ns over short (<1000 km) baselines [9]. These results are consistent with time-transfer capabilities demonstrated by *Dunn et al.* [10], whose results went further to show that GPS can be used to make accurate clock drift (or frequency offset) measurements. The performance of GPS-based frequency offset measurement was also assessed for the GPS Demonstration Receiver (GPSDR) on TOPEX/Poseidon [11]. These studies suggest that a properly designed GPS receiver could serve as an on-board frequency and time subsystem (FTS). Further study at JPL will include improved clock solution filtering techniques to mitigate SA effects, the dominant error source.

Integrated GPS/Telemetry Architecture

JPL is also developing a GPS receiver architecture that can offer significant power/cost/mass savings to an Earth orbiter by incorporating an uplink telemetry receiver within the GPS processor. This was accomplished in a recent lab demonstration at JPL (see Figure 6) by adding electronics to collect and down convert the telemetry bandpass and feed the digitized uplink into one of the GPS signal processing channels of a GPS receiver.

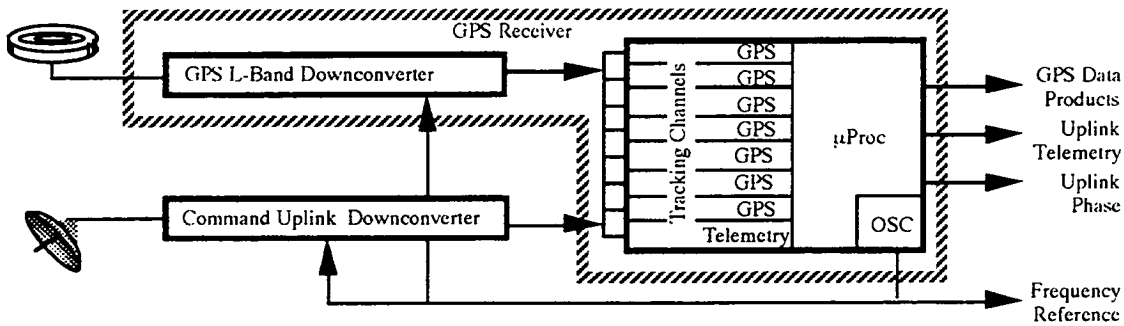


Figure 6: Integrated GPS/Telemetry receiver

With the GPS hardware channel acting as a command telemetry processor, software was then written and added to the receiver firmware to perform telemetry symbol synchronization and extraction and carrier phase acquisition and tracking. Telemetry bit streams of up to 2 kilosymbols per second (modulated using one of several standard methods) were extracted without special symbol extraction hardware. In addition, a precise phase/Doppler measurement was

made of the uplink carrier signal that can add strength to the on-board navigation solution when the frequency and location of the ground transmitter are well-known. If included in the uplink bandpass, a ranging signal could also be measured by such a receiver to add further strength (as well as validation value) to the on-board navigation solution. We estimate that an integrated GPS/Telemetry Receiver would reduce power and mass by a factor of two over a system using stand-alone telemetry and GPS receivers.

Ground Systems for Data Communications

In the area of ground systems for data communications, JPL has recently demonstrated a prototype Telemetry and Tracking System which consists of a low cost (<$200k) weather satellite ground tracking station. The purpose of the system is to provide unattended and continuous retrieval of science data telemetered from low-Earth orbiters. The demonstration showed that a workstation controlling the system can autonomously retrieve NORAD-published, km-level spacecraft ephemerides via commercial phone lines and automatically track and receive science telemetry from specified earth orbiters during overflights of the ground station location. An additional feature of this demonstration was the automatic distribution of the downlinked data to the personal computer of the cognizant Principal Science Investigator, also via commercial phone lines [12].

MEDIUM ALTITUDE EARTH ORBITER APPLICATIONS

The "upwards-looking" GPS configuration employed for low-Earth orbiters is suitable only for orbits up to about 3000 km. Above this altitude GPS signal visibility from a zenith pointing antenna begins to degrade and it becomes advantageous to employ an additional antenna in a "down-looking" configuration [13]. For the down-looking geometry, the orbiting user directs the receiving antenna towards the Earth and tracks the GPS satellites located on the far side (GPS satellites broadcast a beam which is slightly wider than the angle which the Earth subtends). This technique (for the 3000-8000 km altitude range) has not yet been demonstrated in an actual flight test. A magnified atmospheric distortion effect is also expected. Finally, GPS signals will be weaker since they must travel a farther distance and the data would be sampled from the edge of the GPS broadcast beam pattern.

HIGH EARTH ORBITER TECHNOLOGIES

At altitudes above 8,000 km, the visibility of the GPS signals degrades rapidly and the geometry becomes increasingly poor as the user satellite moves away from the Earth, regardless of the antenna pointing configuration. A need for precise positioning services at these extremely high altitudes exists among the geosynchronous spacecraft. It is for these and higher altitude applications (up to 150,000 km) that JPL has developed GPS-like tracking (GLT) [3], illustrated in Figure 7, an approach which exploits GPS in a decidedly different way than the techniques outlined above.

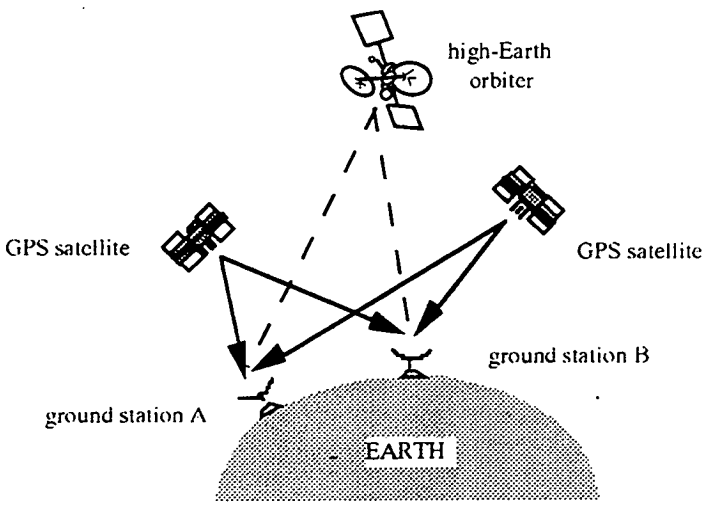


Figure 7: Differential GPS-like tracking (GLT) applied to high-Earth orbiter. After tracking GPS carrier phase and pseudorange for 12–24 hours, the GPS orbits can be determined to a few tens of centimeters. In GLT, the carrier phase of the high-Earth orbiter is also included and its orbit similarly estimated. This relationship is discussed further by *Lichten et al.* [14].

The GLT approach for high-Earth orbiters replaces the usual flight GPS receiver on the user spacecraft with an Earth-pointing transmitter beacon whose signal structure is such that it can be phase tracked, along with GPS signals, using a ground network of "enhanced" GPS receivers. The configuration illustrated in Figure 7 also shows the similarity between high-Earth orbiter and GPS satellite orbit determination. In principle, the accuracy of high-Earth satellite orbits should approach that of GPS satellite orbit determination (routinely determined by JPL and several other analysis centers to better than 50 cm). In the GLT technique, precise GPS tracking on the ground serves to calibrate media delays, synchronize clocks between stations, and account for various geophysical parameters (like station coordinates).

The concepts of interferometric phase tracking of geosynchronous orbiters and using GPS to assist in their orbit and trajectory determination have their heritage in deep space tracking techniques and were developed and refined in the mid 1980's by JPL scientists [15,16]. Recent analysis has suggested that meter level orbit accuracies can be achieved at altitudes up to 150,000 km using a globally dispersed network of combined beacon/GPS tracking stations [3]. An initial demonstration of the GLT technique using real data was performed by JPL earlier this year with NASA's Tracking and Data Relay Satellite System (TDRSS).

TDRSS, whose space segment includes 5 geosynchronous orbiters, is used by NASA to support positioning and data relay activities. Accurate near real-time positioning of the TDRSS spacecraft is fundamental to proper operation of the system. In our initial demonstration, GLT phase observations were combined with few range measurements from a single station to yield TDRS orbits with an accuracy of better than 25 meters in total position [9]. This level of accuracy is consistent with the results of studies performed prior to the demonstration. Achievement of higher accuracies for the TDRS application is limited by the quality of the TDRS ranging, and the footprint size of the existing TDRS beacon. The latter dictated that the GLT ground stations be located within 1000 km of one another, a configuration that results in weakened geometric tracking strength in comparison with a globally dispersed network. In addition, since the angular sensitivity of the tracking measurements are proportional to the baseline length, good performance with short baselines requires extremely tight control of delay errors. Remarkably, virtually all of the potential sources of delay error in a short-baseline tracking scenario can be measured with GPS: a) *Clock synchronization/time transfer:* Routine processing of GPS data for the IGS are providing clock synchronization at tracking stations dispersed around the globe to better than 1 ns [10,17]. b) *Station coordinates:* Geocentric station coordinate solutions accurate at the cm level are generated routinely [1]. c) *Atmospheric delays:* Zenith wet troposphere delays can be measured in a variety of conditions with sub-cm accuracy [20]. The dual-frequency nature of the GPS signal also allows calibration of the ionosphere delay which, when mapped to the line of sight to TDRS, can provide sub-cm accuracy at the TDRS 13.731 GHz SGL.

As in the TOPEX/Poseidon demonstration, the data analysis was carried out using an augmented version of the GIPSY/OASIS package of modeling and estimation algorithms. The ground tracking stations for the TDRS demo consisted of dual-frequency, geodetic quality GPS receivers, enhanced by JPL with special phase-tracking software and electronics necessary to receive and down convert the existing Ku-band (13.731 GHz) TDRS spacecraft-to-ground link (SGL). This signal is present whenever the TDRS is servicing user spacecraft. Determination of the TDRS orbits using GLT can thus be performed without disruption of user services (a characteristic not shared by the current operational orbit determination system, Bilateration Ranging Transponder System (BRTS), against which the GLT system was compared). The next few sections will examine more closely, and consider future improvements to, the data analysis strategies and enhanced ground receiver design as they apply to orbit and trajectory determination for high-Earth orbiters.

Data Analysis Strategies and Implementation

The orbit determination method for the TDRS demonstration draws its heritage from a technique proposed by *Nandi et al.* [18] which uses station-differenced carrier phase observations with integer cycle ambiguities and biases left unresolved. These so-called short baseline differenced carrier phase ($SB\Delta\Phi$) measurements determine the *change* in plane-of-sky position of the TDRS spacecraft. When included in a dynamical orbit determination, well-calibrated $SB\Delta\Phi$ measurements can determine 5 of 6 components of the spacecraft state vector. The longitude of the orbit—or the satellite's down-track position in inertial coordinates—is poorly determined; moreover, the orbit solution is somewhat sensitive to mismodeling of forces such as solar radiation pressure [18]. To combat these problems, a few well calibrated range measurements were needed.

The unified TDRS/GPS orbit solutions were performed using the GIPSY/OASIS II software. The epoch satellite states for the TDRS as well as all GPS spacecraft were estimated simultaneously using tracking data arcs of approximately one days length. The overall solution strategy, with a few exceptions, is the same as that employed for routine processing of GPS for the International GPS Service (IGS) global network [19]. (Particulars of the strategy that are specific to the TDRS estimation problem are summarized by Haines et al. [9].)

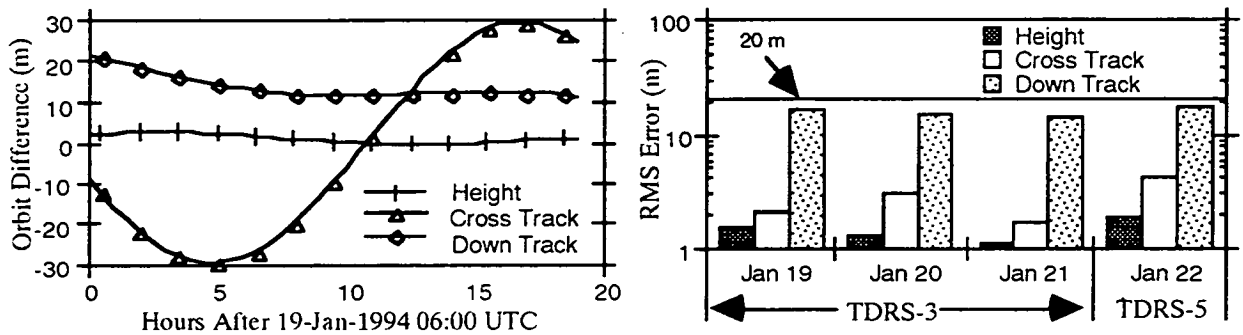


Figure 8: a) TDRS-3 inertial orbit differences between GLT and BRTS (from Goddard Space Flight Center) orbits for January 19, 1994. The RMS differences in height, cross track, and down track are 1.6 m, 22.4 m and 14.2 m respectively. b) Bar graph showing RMS formal errors of GLT-based TDRS orbit solutions. The arcs vary between 18 and 20 hours in length.

The comparisons between the GPS-based local (regional) network solutions and the global BRTS solutions (compare Figure 8a) indicate RMS agreement at the level of about 25 meters or better. It is possible that the GPS-based orbits are, in fact, more accurate than this, but at the present time the BRTS orbits are the best that are available for comparison. Note also that for the above solutions, the White Sands range bias was calibrated with the BRTS orbit, so the down track results show internal consistency (precision) rather than an independent measure of accuracy.

Enhanced Ground Receiver Design

The GPS ground receivers, in addition to providing various calibration parameters to the TDRS tracking system, also perform the precise phase tracking of the TDRS Ku-band spacecraft-to-ground link. Figure 9 illustrates this capability, which was added to an existing geodetic-quality GPS receiver by the development and inclusion of additional electronics and software. The electronics consisted of a Ku-band horn antenna (opening dimensions 17 X 14 cm) and a Ku- to L-band downconverter, both developed at JPL. The TDRS SGL, after translation to L-band, was power combined with the GPS L-band signal. The relative power levels of the two signals were set such that Signal-to-Noise Ratio losses were minimized. In addition, the receiver was augmented with software, also developed by JPL, which measures and records the phase of the TDRS SGL with the same sub-mm precision and accurate receiver time-stamp as GPS carrier phase measurements. This ability of the receiver architecture to permit integration of SGL tracking, coupled with the powerful calibration features of precision GPS tracking, contributes significantly to the simplicity of the TDRS ground stations.

Another key feature of the GLT tracking stations used for the TDRS demonstration is the ability to remotely monitor operations and download data in an automatic, unattended fashion. The long term effectiveness and reliability of such a tracking station has already been demonstrated in the autonomous and continuous operation of the IGS network of over 50 globally distributed GPS tracking stations [19]. The IGS tracking stations are identical to the TDRS demonstration tracking stations except that they lack the TDRS specific antenna and downconverter hardware. They, in fact, already contain the software necessary to track a high-Earth orbiter SGL. This software, however, is not invoked during normal GPS-only operation.

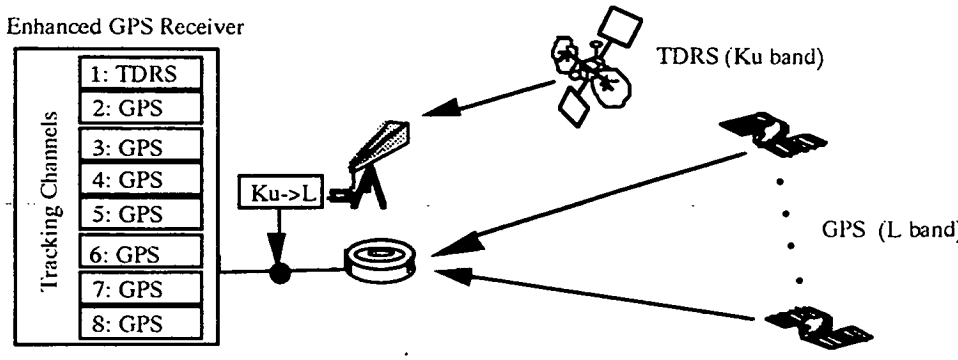


Figure 9: Component diagram for the GLT ground tracking station for TDRS, including the enhanced GPS receiver which simultaneously tracks TDRS along with GPS satellites. For the TDRS SGL, which is at 13.731 GHz, a small separate antenna with down converter was added.

SUMMARY

JPL's technologies for real-time and post processed orbit determination for Earth orbiters of various altitudes have been outlined in this paper. Depending on the altitude of the satellite, on the complexity (and cost) of the system, and on the time delay in producing results, GPS-based tracking systems can provide orbit knowledge at the cm-level or at the level of a few hundreds of meters. Experiments recently carried out at JPL have, in fact, demonstrated these capabilities for satellites in low-Earth to geostationary altitudes, with a wide breadth of processing strategies ranging from detailed post-fit analysis to complete automation. These experiments have included the development of new hardware and software technology which are now suitable for either direct transfer to industry or further co-development with commercial partners. For low-Earth orbiters, the recent TOPEX/Poseidon demo showed that orbits at the accuracy level of a few cm can be determined with a GPS flight receiver and a global ground network, while a few tens of meters can be expected without any ground stations. A demonstration with the TDRSS constellation showed that the GPS-like tracking data, combined with range data and GPS data acting to calibrate a number of critical parameters, produces geostationary orbits with better than 25 meter accuracy. In addition, JPL's development of a high precision receiver to make atmospheric radio occultation measurements has led to a configurable receiver architecture which could offer a diverse choice of flight receivers to the space system designer.

ACKNOWLEDGMENT

The work described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The authors are grateful to Willy Bertiger, George Hajj, Tom Meehan, and Tim Munson of JPL for their valuable consultations.

REFERENCES

1. Blewitt, G., Heflin, M.B., Webb, F.H., Lindqwister, U.J., and Malla, R.P., Global Coordinates with Centimeter Accuracy in the International Terrestrial Reference Frame Using GPS, *Geophys. Res. Lett.*, Vol. 19, 853-856, May 4, 1992.
2. Bertiger, W.I., Bar-Sever, Y.E., Christensen, E.J., Davis, E.S., Guinn, J.R., Haines, B.J., Ibanez-Meier, R.W., Jee, J.R., Lichten, S.M., Melbourne, W.G., Muellerschoen, R.J., Munson, T.N., Vigue, Y., Wu, S.C., Yunck, T.P., Schutz, B.E., Abusali, P.A.M., Rim, H.J., Watkins, M.M., and Willis, P., GPS precise tracking of TOPEX/POSEIDON: Results and implications, *J. Geophys. Res.*, TOPEX/POSEIDON Special Issue, 1994 (in press).
3. Lichten, S. M., C. D. Edwards, L. E. Young, S. Nandi, C. Dunn, and B. Haines, A demonstration of TDRS orbit determination using differential tracking observables from GPS ground receivers, AAS Paper 93-160, AAS/AIAA Space Flight Mechanics Conference, Pasadena, Calif., February, 1993.
4. Meehan, T. K., Munson, T.N., Thomas, J. B., Srinivasan, J. M., D. Spitzmesser, L.E. Young, and R. Neilan, Rogue: A New High Accuracy Digital GPS Receiver, *Proceedings of the IUGG*, Vancouver, Canada, 1987.
5. Meehan, T. K., J. M. Srinivasan, D. Spitzmesser, C. Dunn, J. Ten, J. B. Thomas, T. Munson, and C. Duncan, The TurboRogue GPS receiver, *Proceedings of the 6th IGS Conference on Satellite Positioning*, Columbus, OH, 1992.
6. Carson, L., Hailey, L., Geier, G.J., Davis, R., Huth, G., Munson, T.N., Design and Predicted Performance of the GPS Demonstration Receiver for the NASA TOPEX Satellite, *Proceedings of the Position, Location, and Navigation Symposium*, 422-454, November, 1988.
7. Hajj, G.A., Kursinski, E.R., Bertiger, W.I., Romans L.J., and Hardy, K.R., Assessment of GPS Occultations for Atmospheric Profiling, *Proceedings of the Seventh Conference on Satellite Meteorology and Oceanography*, Monterey, CA, June, 1994.
8. Thomas, J.B., J.M. Srinivasan, Digital GPS-Signal Processor With P-Code/No-P-Code Option, *NASA Tech Briefs*, Volume 18, Number 5, May 1994.
9. Haines, B.J., Lichten, S.M., Muellerschoen, R.J., Spitzmesser, D.J., Srinivasan, J.M., Stephens, S.A., Sweeney, D., Young, L.E., A Novel Use of GPS for Determining the Orbit of a Geosynchronous Satellite: The TDRS/GPS Tracking Demonstration, *Proceedings of the ION GPS-94*, September, 1994 (in press).
10. Dunn, C.E., Lichten, S.M., Jefferson, D., and Border, J., Sub-Nanosecond Clock Synchronization and Precision Deep Space Tracking, *Proceedings of the 23rd Annual Precise Time and Time Interval Applications and Planning Meeting*, NASA CP 3159, 89-97, December, 1991.
11. Stiver, B., Time Correlation, presentation at TOPEX/Poseidon: Satellites/Sensors: Performance Characteristics Workshop at JPL, 347-360, July, 1994.
12. Personal communication with Drs. William Rafferty and Nasser Golshan, both of the Jet Propulsion Laboratory, April, 1994.
13. Lichten, S.M., Haines, B.J., Young, L.E., Dunn, C.E., Srinivasan, J.M., Sweeney, D., Nandi, S., and Spitzmesser, D., Using the Global Positioning System for Earth orbiter and deep space tracking, paper presented at the National Telesystems Conference, San Diego, May, 1994.
14. Lichten, S. M., D. Sweeney, L. Young, B. Haines, D. Spitzmesser, J. Srinivasan, C. Dunn and S. Nandi, New ground and space-based GPS tracking techniques for high-Earth and deep space orbit determination applications, *Proceedings of the ION 1994 National Technical Meeting*, 371–380, San Diego, January, 1994.
15. Yunck, T. P., and S. C. Wu, Ultra-precise orbit determination by GPS, AAS Paper 83–315, presented at Astrodynamics Specialists Conference, Lake Placid, N.Y., August, 1983.
16. Wu, S. C., Differential GPS approaches to orbit determination of high-altitude satellites, AAS paper 85-430, presented at Astrodynamics Specialists Conference, Vail, Colo., August, 1985.
17. Dunn, C.E., Jefferson, D., Lichten, S.M., Thomas, J.B., Vigue Y., and Young, L.E., Time and Position accuracy using codeless GPS, *Proceedings of the 25th Annual Precise Time and Time Interval Applications and Planning Meeting*, NASA CP 3267, 169-179, December, 1993.
18. Nandi, S., C. Edwards, and S. C. Wu, TDRSS orbit determination using short-baseline differenced carrier phase, paper presented at Flight Mechanics Estimation Theory Symposium, NASA CP 3186, 103–115, May, 1992.
19. Neilan, R. and J. Zumberge, "The International GPS Service for Geodynamics," *Proceedings of ION GPS-94* , Alexandria, VA), Salt Lake City, September 1994 (in press).
20. Lichten, S. M., Precise estimate of troposphere path delays with GPS techniques, Jet Propulsion Laboratory Telecommunications and Data Acquisition Progress Report, Vol. 42–100, 1–12, Febraury 15, 1990.

PERSON LOCATOR SYSTEM

Frederick W. Mintz
Jet Propulsion Laboratory
Pasadena, California 91109

Brent R. Blaes
Jet Propulsion Laboratory
Pasadena, California 91109

Charles W. Chandler
Jet Propulsion Laboratory
Pasadena, California 91109

ABSTRACT

Extremely High Frequency (EHF), very narrow band, RF communications techniques have proven to be applicable to the development of a passive transponder design, capable of re-transmitting a received signal over distances greater than previously demonstrated. This proven concept may be implemented in a number of Radio Frequency Identification (RFID), micro-miniature tags useful in locating and tracking persons or materials over distances of 10 to 15 meters, within a closed facility. The Locator System comprises a transceiver array and transponders secured to the persons or things which one desires to track within a small geographic area, such as a prison or a large industrial complex. The transceivers and transponders are individually, digitally identified. The transponder is powered solely by RF energy it collects, stores, and transmits when queried. It should be noted that as of the delivery of this paper, only the laboratory bench, proof-of-concept has been demonstrated. However, design calculations and computer simulations give rise to confidence that the remaining implementations, namely: 1) design of a small, very narrow band transceiver, 2) construction of both transceiver and transponder antennae, 3) optimization and micro-miniaturization of the existing passive transponder circuits, 4) design and construction of the transceiver system control and computer interface, and finally, implementation of the graphics display software, can be rapidly and successfully carried out.

A. INTRODUCTION

The Jet Propulsion Laboratory (JPL), as an operating division of the California Institute of Technology (Caltech), funded through a prime contract with the National Aeronautics and Space Administration (NASA), has undertaken certain research and development for the National Institute of Corrections (NIC) known as The Person Locator System, hereinafter called the project.

The project is divided into three, distinct development phases, generally corresponding to the three funding periods outlined in the JPL Task Plan submitted to the National Institute of Corrections on July 28, 1993. The Scope of Work outlines six technical tasks, which comprise the three phases as follows:

PHASE I	**DEVELOP PROOF OF CONCEPT**
PHASE II	**PREPARE AND FIELD TEST A PROTOTYPE MODEL**
PHASE III	**ASSIST IN THE DEVELOPMENT OF THE COMMERCIAL PROTOTYPE SYSTEM AND ASSIST IN A N INSTALLATION IN A LOCAL JAIL FACILITY**

B. GENERAL BACKGROUND

Actual work on the project began on January 3, 1994.

On March 7, 1994, the National Institute of Correction's, National Advisory Board visited JPL to appraise work progress and was given a Laboratory Demonstration Proof of Concept (Deliverable Item 1) of a functioning passive oscillator, responding with an imbedded digital code, to a transceiver system approximately 4 meters away.

C. TECHNICAL BACKGROUND

The first two weeks in January were spent in revisiting the technical literature to appraise the project of additional technical progress for Digital Tagging Devices and associated technologies. When this project was first conceived in 1992, there was no indication that any work had been done in industry, government, or university laboratories using the concept of a passive or "ringing" oscillator to effect a transponder response at distances greater than approximately 0.3 meters. It was felt in 1992 that, given the great advances in space communications at the Jet Propulsion Laboratory, particularly in the area of narrow-band and frequency stability designs, as well as circuit micro-miniaturization, that it might be possible to design and construct a high-sensitive RF system, based upon an old technology dating back to the MIT Radar Lab days, commonly known, then, as "ringing oscillators".

Following this (unsuccessful) literature search, it was decided to apprise the project team of both domestic and foreign current research and development in this technology area, which might not yet be published. On-lab and off-lab visits were scheduled with researchers whom the literature search (as well as personal knowledge) indicated had interests in the field of RF energy designs. A variety of applicable technology was reviewed at a major industrial conference during January in San Diego. This also comprised a large exhibition of foreign and domestic manufacturers of crystal and digital controlled oscillators. The balance of January was spent visiting with and talking to over 25 individuals and organizations active in this technology field.

Further, discussions with several colleagues at JPL, and Caltech regarding the technical concept of the Person Locator System Transponder led to the conclusion that such a technology might well have "Dual-Use" applications for planetary remote sensors.

At the inception of the project it was decided to proceed down parallel paths in developing a passive oscillator which could be used as the basis of a transponder system of the type envisaged in the basic concept. For a number of technical reasons (e.g. circuit size, antenna wave length, power requirements, FCC licensing etc.), it was felt, early on, that a frequency in the 900 MHz Band was the most desirable for the design and implementation of the project concept. It was not clear at the time, however, that a design and proof of concept could be achieved at this frequency without a possibly large "research" effort.

Therefore a parallel effort was implemented, at a frequency on the order of 27+ MHz (CB Band) to take advantage of component and test equipment availability. In retrospect this decision was fruitful in that both circuit designs and breadboards were successful, unequivocally demonstrating the viability of this concept for the purposes intended.

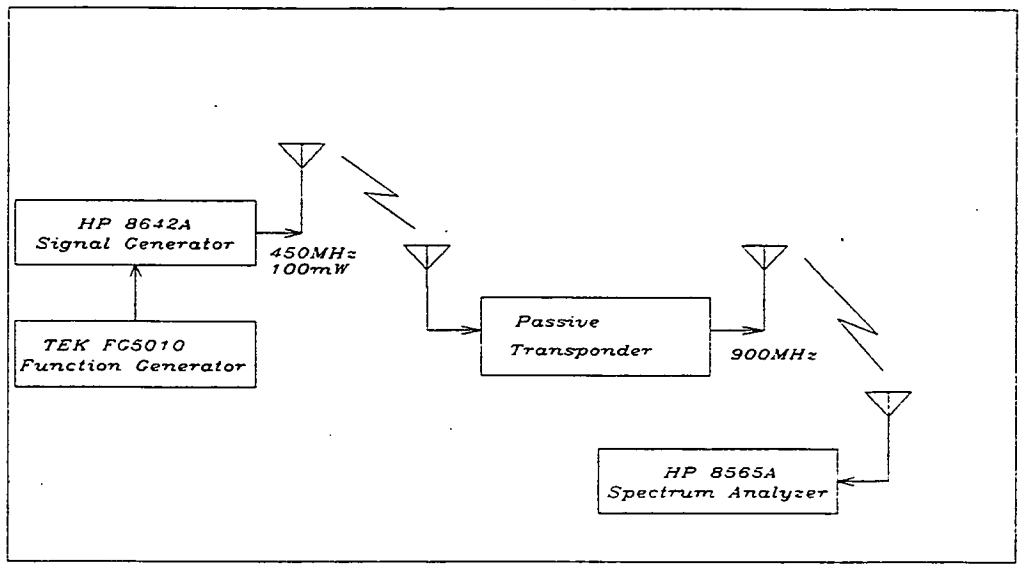
By mid-February, the calculations and designs were completed for the 27 MHz circuit. The 450/900 MHz circuit designs, which included the imbedded digital integrated-circuit (IC), were completed and implemented by the end of February. Testing and optimization of both circuits proceeded up to the Proof-of-Concept demonstration for the NIC's National Advisory Board. However, final development efforts were concentrated on the 450/900 MHz circuits because these meet the overall system requirements and are, indeed, far more precise in design and implementation.

D. PROOF OF CONCEPT

A block diagram of the experimental setup used to demonstrate the passive transponder is shown in Figure 1. A 450 MHz, 20 dBm (100 mW) RF signal supplied by the HP8642A synthesized signal generator and pulse modulated by the TEK FG5010 function generator is transmitted with a simple Yagi-Uda antenna. The passive transponder receives this signal with an end-fed 450 MHz dipole antenna and transmits a signal at 900 MHz with a similar antenna. An HP8565A

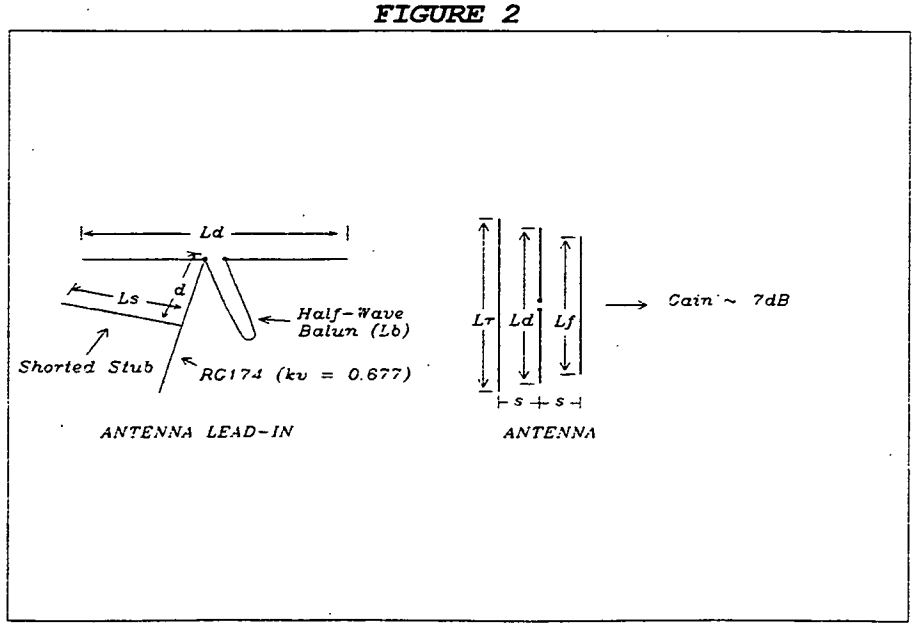
spectrum analyzer operating at a fixed frequency (900 MHz), sweep time mode, receives the 900 MHz signal with a simple Yagi-Uda antenna. The transmitter and receiver antennas were oriented at 90 degrees with each other to minimize pickup of the 900 MHz second-harmonic generated by the HP 8642A signal generator.

FIGURE 1



The transmit and receive antenna and lead-in designs are shown in Figure 2. A half-wave balun transformer is used to connect the center-fed balanced dipole antenna to coaxial cable. This also drops the impedance by 4 (70 ohm to 17.5 ohm). The shorted stub provides an impedance match to the 50 ohm RG174 coaxial cable.

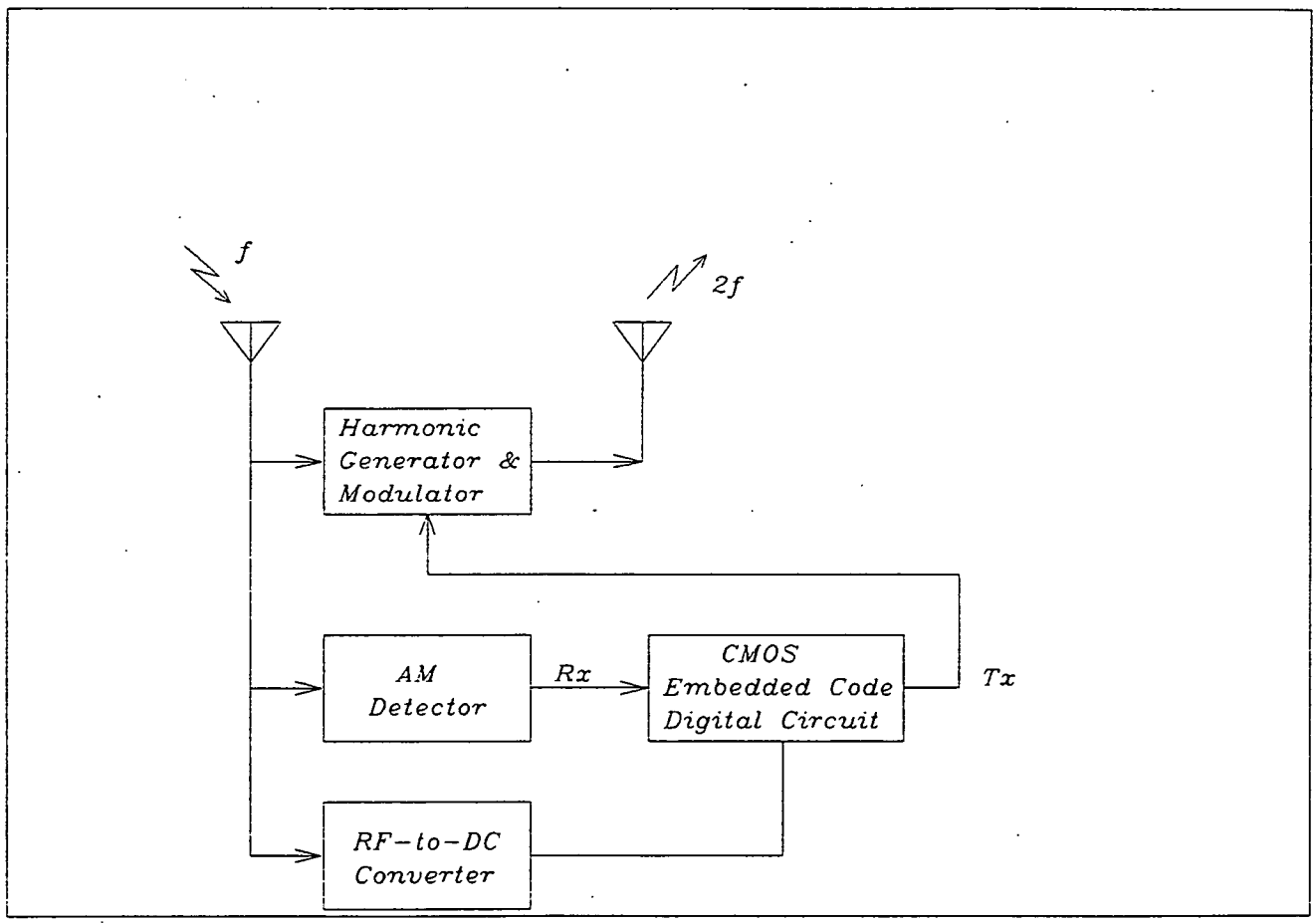
FIGURE 2



Functional Design

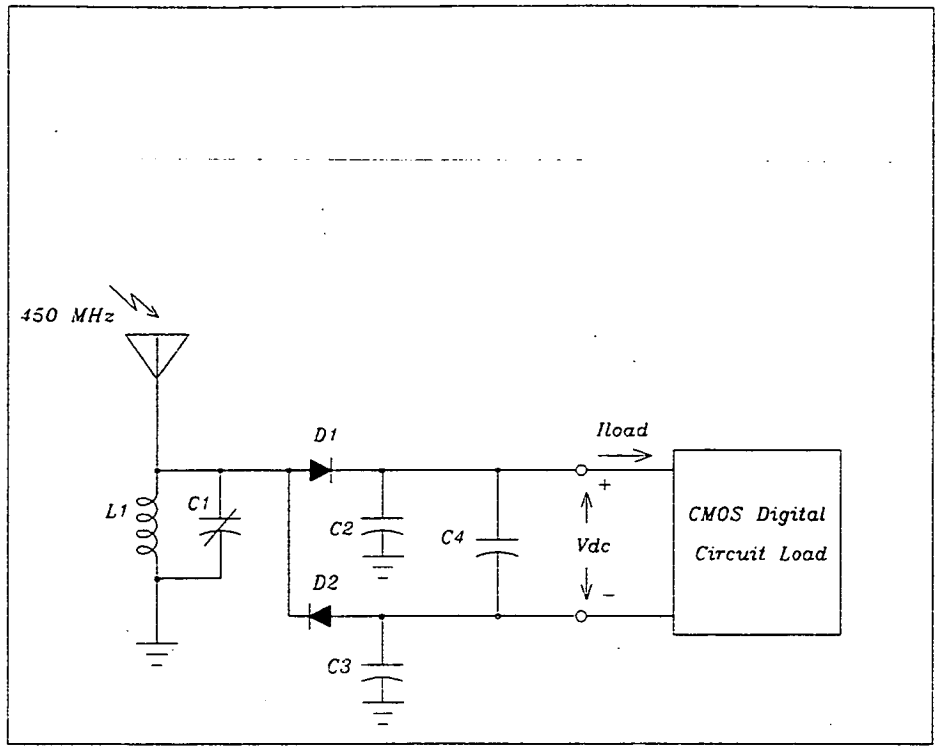
A block diagram of the passive transponder is shown in Figure 3. Three functions are performed on the 450 MHz received signal. First, DC voltage (~ 2 V) is generated to power the CMOS digital circuit; second, the signal is demodulated with an AM detector circuit generating Rx; and third, a signal twice the received carrier frequency is generated (900 MHz) and functions as the transponder transmit carrier. The received message, Rx is processed by the CMOS digital circuit that in turn generates a transmit message, Tx which is AM modulated on the 900 MHz carrier and transmitted with a half-wave dipole antenna

FIGURE 3



A schematic circuit diagram of the RF-to-DC converter is shown in Figure 4. It consists of a full-wave rectifier and storage capacitors. The positive peak of the coupled RF charges C2 through D1, while the negative peak of the RF charges C3 through D2. The full peak-to-peak voltage (minus the voltage barriers of diodes D1 and D2) is developed across C4. C4 is chosen large enough to limit the ripple on Vdc for the load current, Iload. The tank circuit consisting of L1 and C1 is connected to an end fed dipole (33.3 cm AWG 16 bus wire) and tuned for maximum Vdc.

FIGURE 4



A schematic circuit diagram of the 900 MHz harmonic generator and AM modulator is shown in Figure 5. The square-law nonlinearity of diode D1 generates a 2nd harmonic component that is resonated in the tank circuit consisting of L2 and C2 and radiated with an end fed dipole antenna (16.7 cm AWG 16 bus wire). A transmit modulation signal (Tx) is applied through an RF choke to the anode of D1. A large negative Tx voltage (2 V) is used to turn off diode D1 and hence the 900 MHz signal (100% modulation).

FIGURE 5

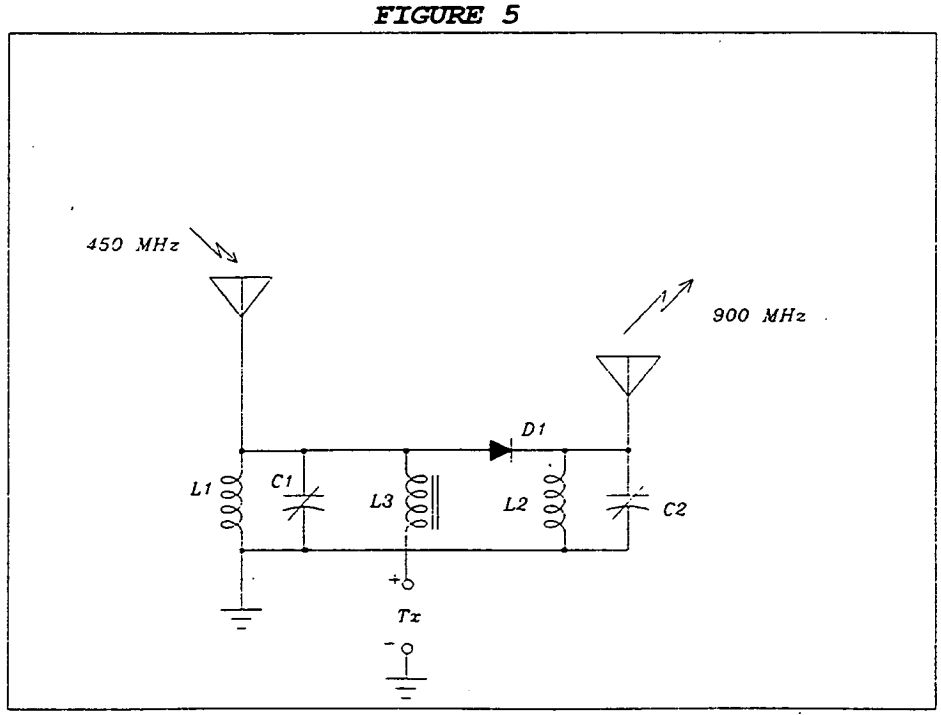


Figure 6 shows how the circuits of Figures 4 and 5 are connected together. The DC source powers a CMOS relaxation oscillator consisting of a 74C14 Schmitt trigger inverter. 74C series CMOS integrated circuits require a minimum of 2 volts to function normally. At 2 volts, this circuit oscillates at around 1 KHz. Notice that the inverter output is connected to the modulator circuit ground and the DC ground (triangle symbol) is connected to the modulator input (L3). This presents a negative effective voltage to the anode of the modulation diode. This circuit functions at a range of over 10 feet from the transmitter antenna. At 10 feet, Vdc is measured at 2.4 V and the 900 MHz returned signal is 15 dBm peak-to-peak at an average level of -100 dBm. Notice that two 450 MHz antennas were used in the circuit of Figure 6.

FIGURE 6

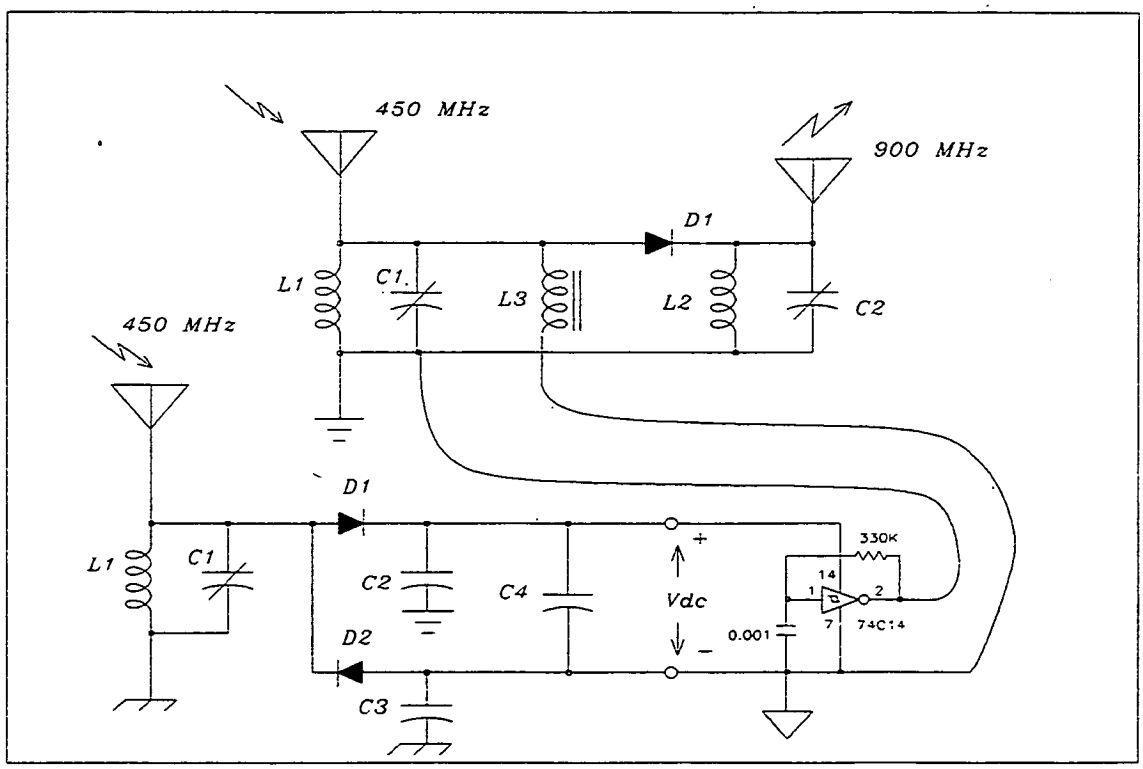
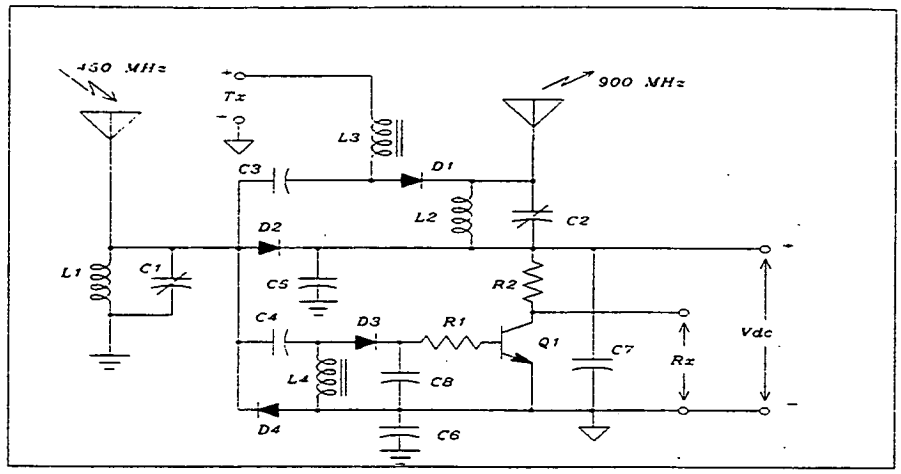


Figure 7 shows a design employing one 450 MHz antenna. An AM detector is also included to generate a receive signal, Rx. The RF-to-DC converter is direct coupled to the 450 MHz tank circuit and antenna, while the AM detector and modulator are, each, capacitively coupled. This isolates each of these circuits effective grounds. The AM detector circuit generates CMOS compatible out levels through the use of the inverting amplifier Q1. The detector time constant is set by R1 and C8. RF choke, L4, is required to shunt the DC component at the anode of D3 to DC ground.

FIGURE 7



The CMOS digital circuit shown in Figure 8 was connected to the RF circuit of Figure 7 and evaluated. This CMOS circuit accepts a receive signal (Rx) which toggles flip-flop Z4, pin 9 each time an Rx pulse is received. The state of this flip-flop determines whether the transmit signal, Tx, is at fosc = 1 KHz or at fosc/2 = 500 Hz. At 5 feet, the 900 MHz returned signal is 10 dBm peak-to-peak at an average level of -100 dBm. The reduced range (5 ft versus 10 ft for the 2 antenna design) of this integrated design is due to the shared loading on the single 450 MHz antenna and the large DC load presented by the CMOS circuit. Advanced low threshold, low power CMOS circuits should reduce this loading considerably.

FIGURE 8

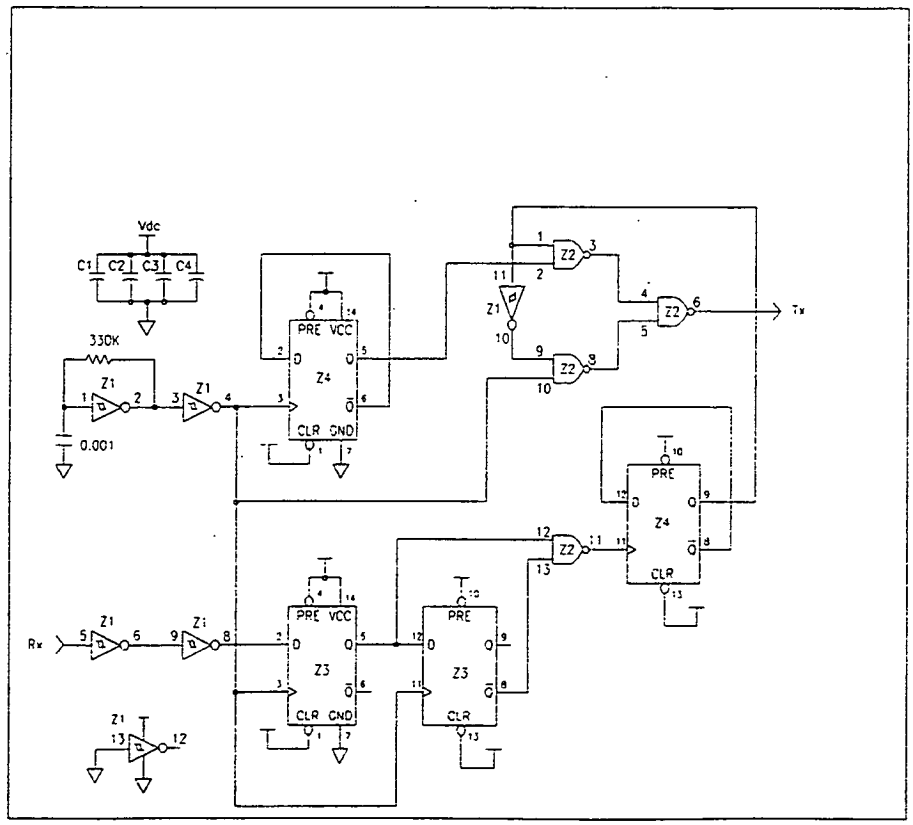
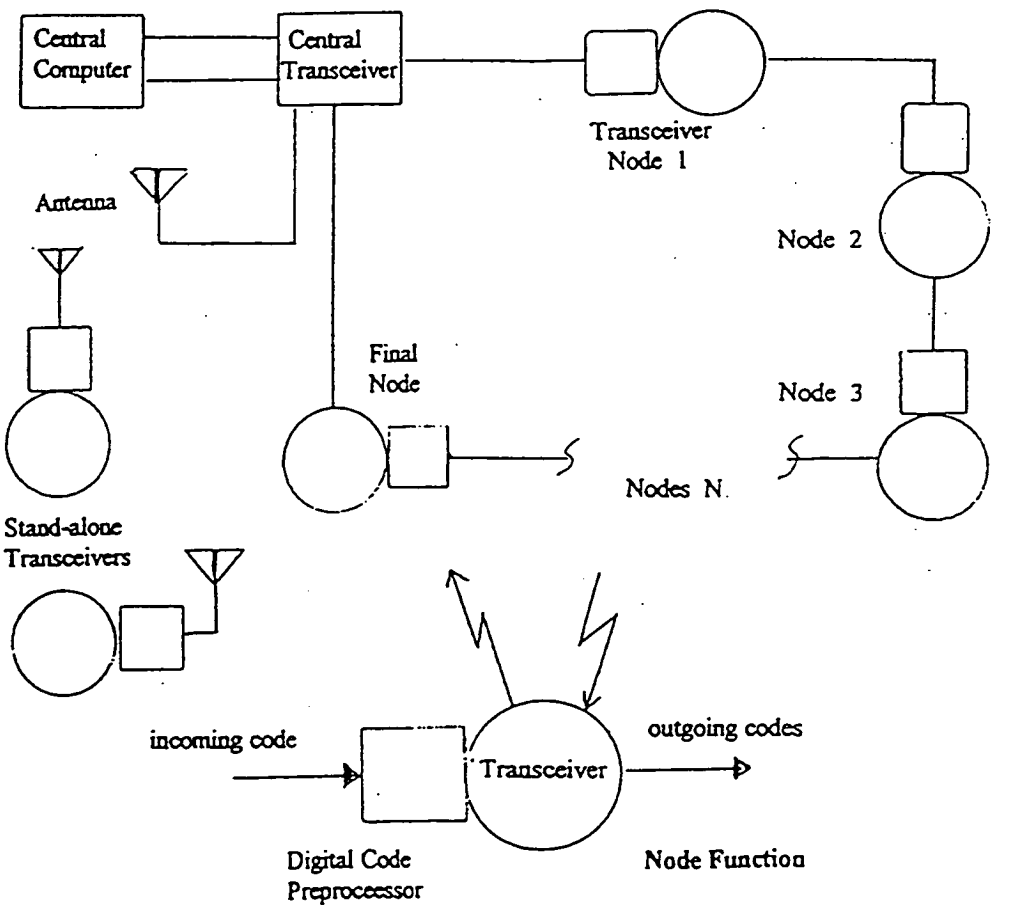


Figure 9 shows a depiction of the system concept. The transceivers are shown in both hard-wired and stand-alone configurations. Transceiver deployment is intended to provide optimal RF radiation patterns such that "dead spots" are minimized. The installation of the transceiver antennae is intended to be part of the coax cable, affixed to the outside of the insulation and adjusted for the optimal radiation pattern for that location. Transceiver power will be chosen so as to meet the boundary conditions required for site conditions as well as the constraints of EPA and OSHA regulations. While the proof-of-concept was demonstrated with 100 mW, the intent is to use only the power necessary to accomplish the transponder communication goals; probably on the order of 3 to 5 Watts.

FIGURE 9



F. CONCLUSION

Narrow band, high frequency, high stability passive RF technology has applications for locating and tracking both persons and materials within the confines of a well-bounded geographic system. In addition to the Person Locator System, this technology is

being considered by JPL for property control and for tracking hazardous materials within the 275 acre, 215 building facility.

D. TECHNOLOGY TRANSFER

JPL has filed a NASA New Technology Release (NTR) on the Person Locator System and is pursuing the filing of a Request for Patent in the name of the California Institute of Technology according to the terms and conditions of the NASA Prime Contract.

During the course of Phase I of the project, several conversations and meetings have been held with potential commercial organizations all of which have expressed considerable interest in the licensing and commercialization of the system.

Care has been taken, throughout this initial phase of the project, to ensure that no one person or organization has received information not available to all and that no organization currently has an "unfair advantage" over any other.

However, it seems clear that the pressing need for the Person Locator System and its potential usefulness in Jails, Prisons, Hospitals, Large Industrial Complexes, Mines, Warehouses (for locating "things" rather than people), etc., is being increasingly felt by JPL. Therefore, JPL will begin to pursue, more vigorously, exploration of this technology transfer and commercialization opportunity, concomitant with necessary patent and intellectual property rights protection.

JPL's Commercial Program Office will be asked to assist the Task Manager in identifying additional commercial organizations as appropriate and participate in the negotiations necessary for the orderly transfer of this technology with all deliberate speed.

(end)

HIGH-SPEED/HIGH-PRECISION ANALOG-TO-DIGITAL CONVERTER FOR ALL-DIGITAL RADIO/TELEVISION

William H. McKnight
Naval Command Control and Ocean Surveillance Center
R D T & E Division Code 573
San Diego CA 92152-5000
(619) 553-2485; (619) 553-2485 FAX; e-mail mcknight@nosc.mil

ABSTRACT

Considerable effort is currently being made in seeking to develop improved speed and precision in high speed analog-to-digital (A/D) converters, in large part to allow for direct digitization of the RF spectrum and thus provide for all-digital radio and television. One may recognize, however, that rather than directly digitize all signals within the full tuning range of an antenna, thus permitting for digital mixing and tuning, it is typically the case that one is only interested in a tiny slice of the spectrum at any given time. Thus a delta-sigma A/D converter can be designed with a tunable signal passband or tuning transformation for the quantization noise stopband to permit high precision A/D conversion and digital mixing to baseband in a single operation.

BACKGROUND

The family of delta-sigma (D-S) A/D converters (often used synonymously with sigma-delta) are in widespread use today because they are very inexpensive to produce, largely due to their circuit simplicity and forgiveness in component parts matching tolerances, and also because they offer a number of attractive technical features. One of the more compelling ones is the very limited extent of anti-alias filtering required with their use since D-S A/D converters oversample by often as much as 50-150 times Nyquist. Another is because they use a low resolution quantizer, typically a 1-bit comparator, which again makes for simplicity and low cost. D-S A/D converters can be thought of as generating high-speed low-precision digital data from an analog signal, which is easily achieved, and trading it, through a digital decimation filter, for high resolution lower speed digital data which, largely because the decimation filtering is a digital process, is also easily (and cheaply) achieved. Variants of D-S A/D converters also serve very simply and efficiently in digital-to-analog converters and in digital-to-digital converters. The innovations to D-S A/D converters presented here are equally applicable in these other two applications

The traditional D-S modulator (DSM) typically features one or more integrators with a low-resolution quantizer in a negative feedback loop. The quantizer introduces quantization (or round-off) error or quantization noise. The noise transfer function of

the DSM loop (or loops, as they are often used in multiplicity), being different from that of the signal, is such that the spectral power of the quantization noise is "shaped" by providing greatest attenuation for the noise at low frequencies, in effect, acting as a high-pass filter for the noise. Thus, if the signal resides at relatively low frequencies, it will be in a low noise environment at the DSM output and the quantization noise can be filtered out by a low-pass filter. A two-integrator (thus second-order) DSM loop is shown in Figure 1 where 1/Z is a unit of time delay. The corresponding noise spectral output is shown in Figure 2. Each integrator puts a transmission zero at d.c. in the noise transfer function of the DSM loop in the traditional configuration (where the a=0 in Figure 1). The DSM alone constitutes only a low-resolution but very high data rate A/D converter. The low-pass (digital) decimator or decimation filter which follows the DSM filters out the quantization noise, leaving the relatively low frequency signal in a very low noise environment where it is down-sampled (the data or clock rate reduced), and, in effect, its low resolution values averaged over many (high-rate) samples to yield a high-resolution lower data rate output.

DESCRIPTION

The innovation of the present concept consists of three components which, taken together, offer enabling technology for all-digital radio, TV, etc. They consist of the following: A new and novel technique is presented for generating and tuning the quantization noise stop-band throughout the (predetermined) tuning range of a receiver, thus defining the (tunable) passband of the signal. A new and novel technique for distributing the quantization noise transfer function transmission zeros to optimally define a noise stop-band (and thus a signal passband) is presented which offers heretofore unattainable stability in cascaded DSM loops.

The value of the scaling coefficient a in Figure 1 controls the position of the noise transfer function zeros such that if a=0 the conjugate pair of zeros in Figure 2 would collapse to a pair of repeated (second order) zeros at D.C. The values of the b coefficients controls the positions of the signal and noise transfer function poles and hence the overall stability of the loop. In order to tune or translate the signal passband (noise stop-band) of Figures 1 and 2, one applies the tuning or translation transformation of Figure 3 to each of the integrators in the circuit of Figure 1. The translation parameter, alpha, of Figure 3 defines the relationship between the sample frequency, f_s, and the center frequency, f_c, of the translated signal passband. The resulting circuit is shown in Figure 4 where the boxes labeled with G contain the tuning circuit with transfer function G. Figure 5 illustrates the translated noise stop-band where an artifact of applying the translation algorithm is that the effective order of the SDM loop is thereby doubled.

A technique for cascading these SDM loops with distributed noise transmission zeros which insures unconditional stability, heretofore at best difficult if not impossible to achieve, is

illustrated in Figure 6. The first two second order DSM loops follow the pattern of the single tuned second order DSM loop of Figure 4 except note that the input to the second loop is the quantization error (noise) generated in the first loop. The third loop features the same transfer function as that of the first loop but with the quantizer replaced by a linearized gain factor. Thus the third loop adds no additional quantization noise to this cascaded system. This matching of the first and third loop transfer functions allows "singley" shaped quantization noise to be cancelled at the lower summing junction so that the net resulting noise is "doubley" shaped (in effect attenuated by two DSM loops... not to be confused with traditional DSM first-, second-, third-, etc. order noise transmission zeros at d.c.).

Finally, Figure 7 illustrates an efficient decimation filter design for processing the DSM output. It basically consists of (digital) down-mixing to form in-phase and quadrature signal components. The low-pass filter that then follows consists of a series of comb filters followed by half-band filters with down-sampling between each stage. The comb filter is especially efficient because it's a finite impulse response filter (tapped delay line) where all the coefficients are equal to unity.

In summary, the combination of these three techniques solves the problem of directly and accurately digitizing the rf spectrum.

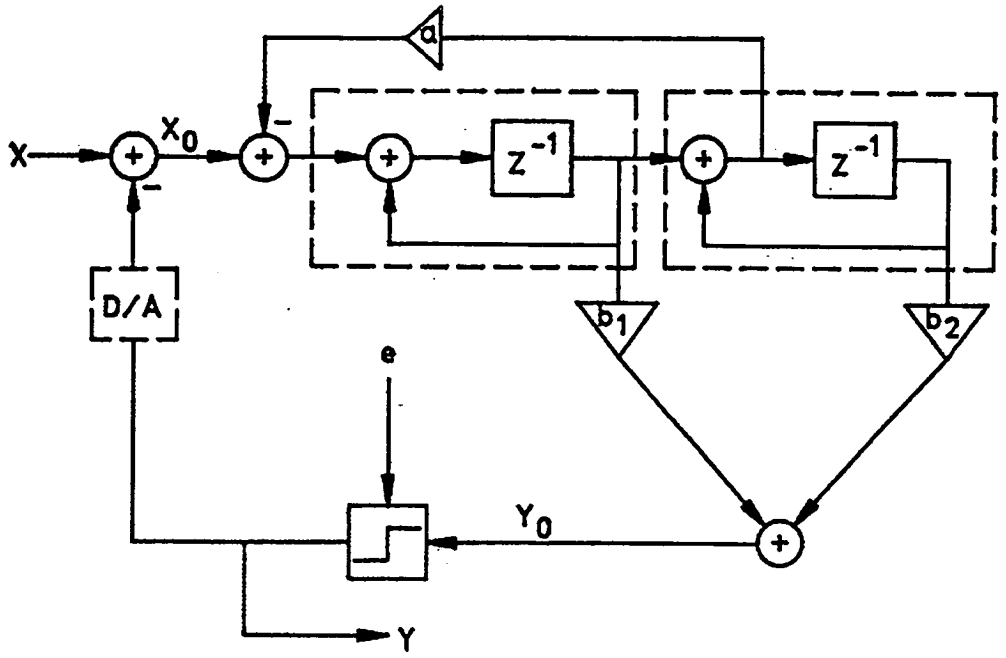


Figure 1

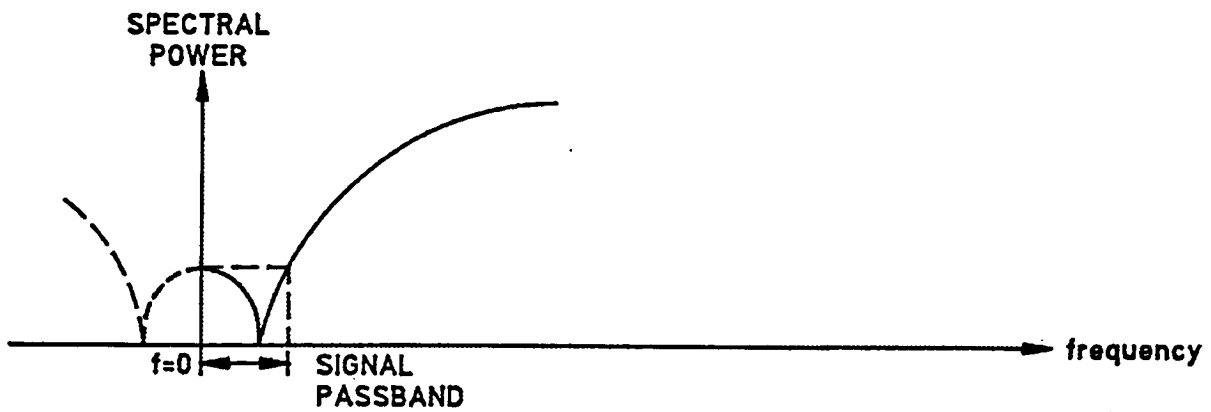


Figure 2

$$G = \frac{\alpha Z - 1}{Z - \alpha}$$

$$\alpha \equiv \cos\left(\frac{2\pi f_c}{f_s}\right)$$

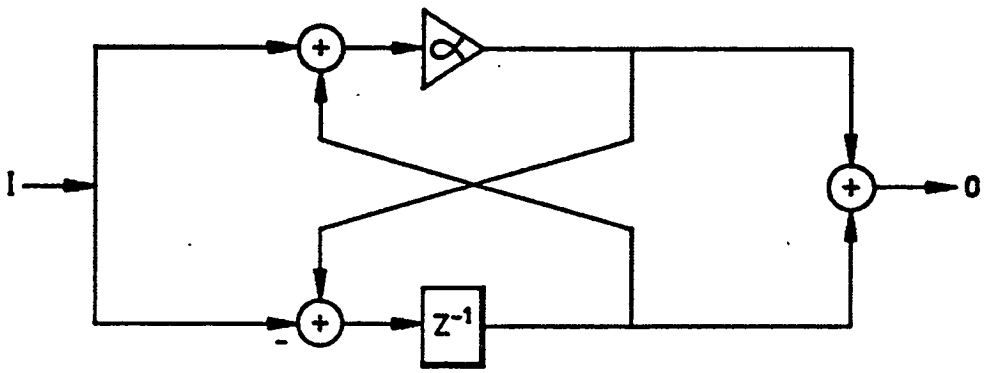


Figure 3

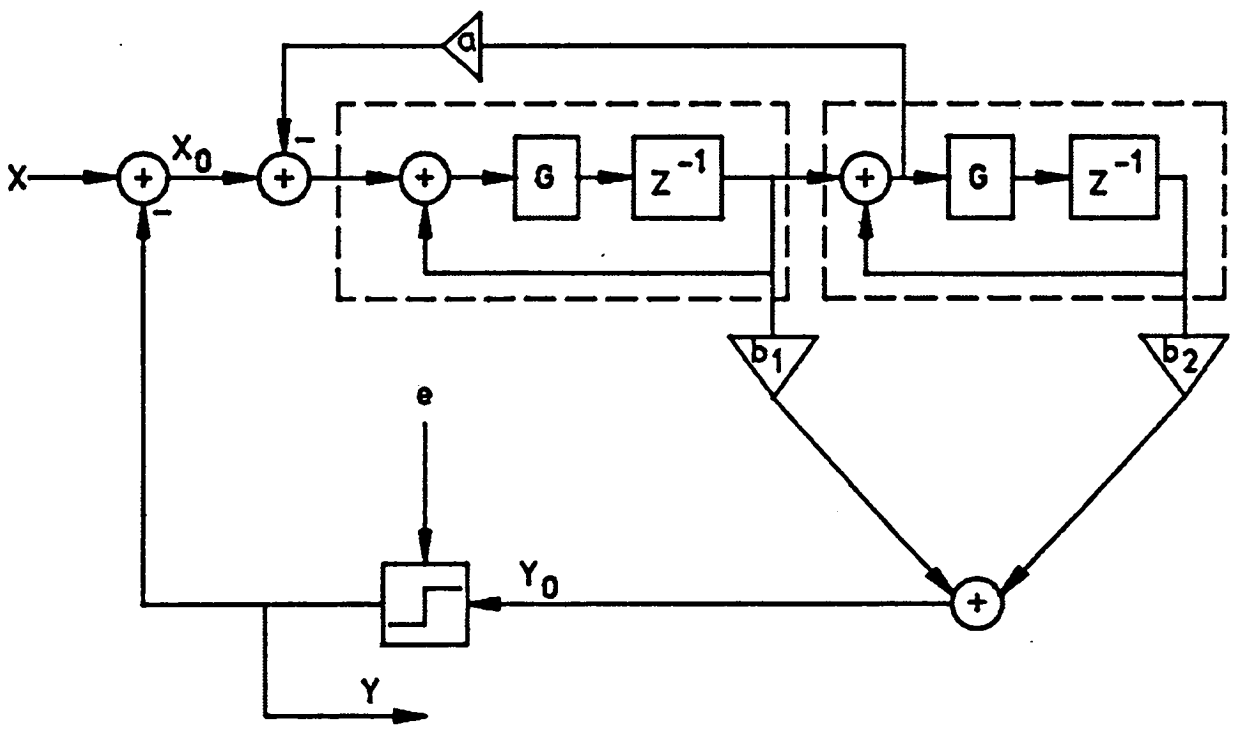


Figure 4

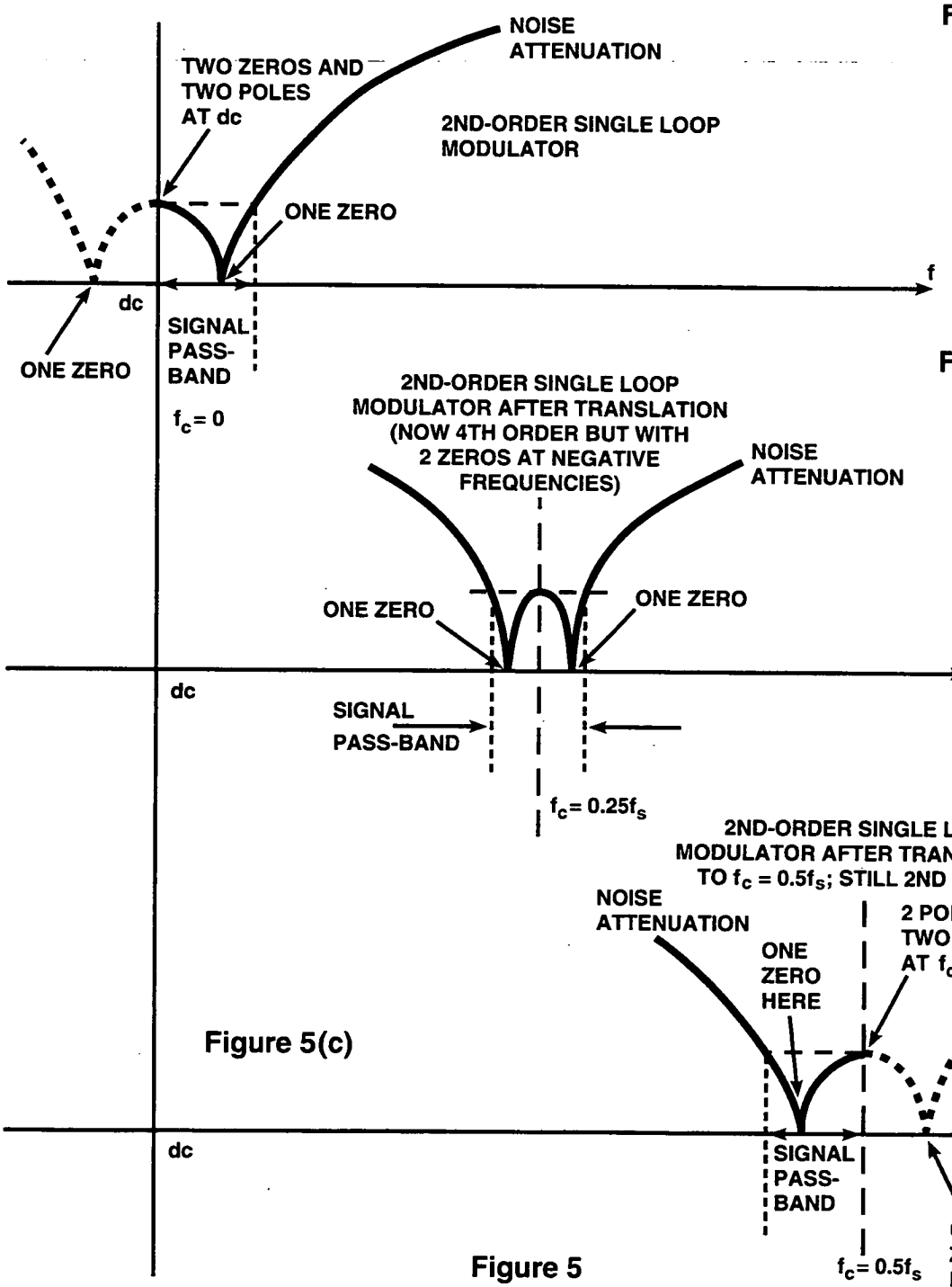


Figure 5(a)

Figure 5(b)

Figure 5(c)

Figure 5

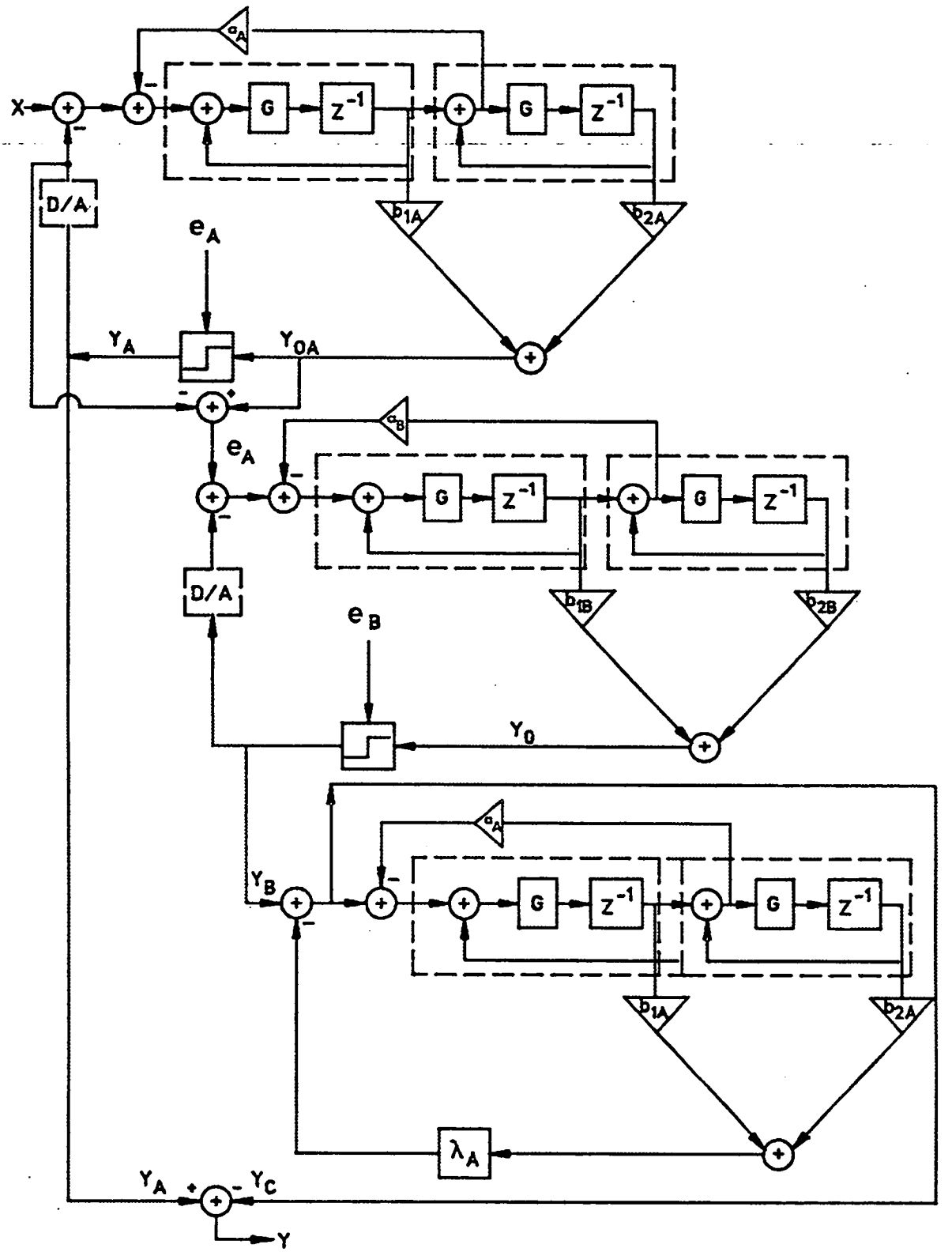
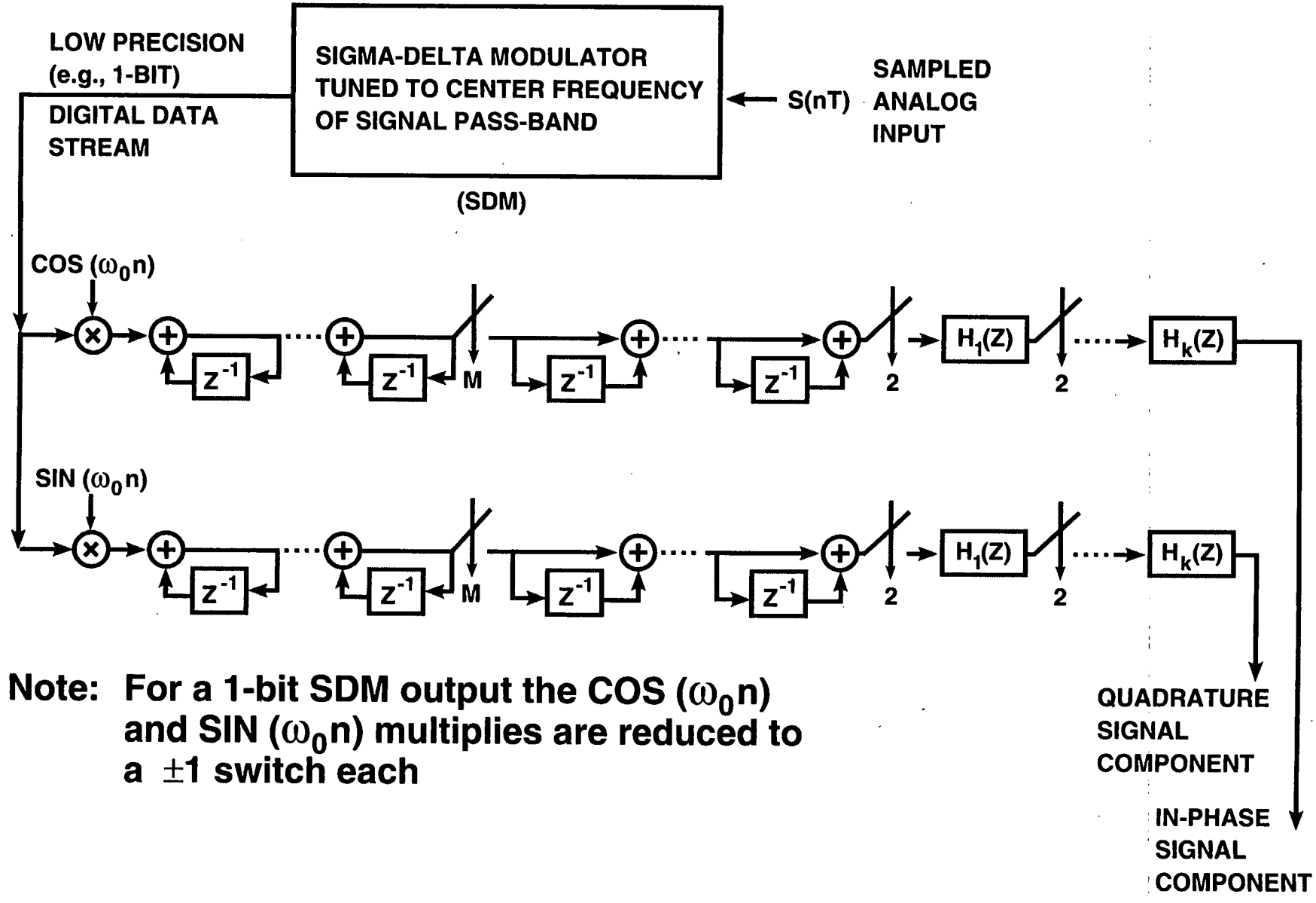


Figure 6



Note: For a 1-bit SDM output the COS (ω_0n) and SIN (ω_0n) multiplies are reduced to a $\pm$1 switch each

Figure 7

High Frequency Electronic Packaging Technology

Martin Herman, Lynn Lowry, Karen Lee, Elizabeth Kolawa, Ann Tulintseff
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

Kurt Shalkhauser
NASA Lewis Research Center
21000 Brookpark Road MS 54-8
Cleveland, OH 44135

John Whitaker
Ultrafast Science Laboratory
University of Michigan
Ann, Arbor, MI 48109

Melinda Piket-May
Electrical & Computer Engineering Department
University of Colorado
Campus Box 425
Boulder, CO 80309

ABSTRACT

Commercial and government communication, radar, and information systems face the challenge of cost and mass reduction via the application of advanced packaging technology. A majority of both government and industry support has been focused on low frequency digital electronics. However higher operating frequencies for both digital and analog circuits will be required for future systems. *By 1997, the projected domestic captives sales for Multichip Modules (MCMs) will be over a half billion dollars.* Therefore, it is critical to develop the technical infrastructure necessary to support this emerging industry. This paper discusses the kick-off of a JPL sponsored Director's Discretionary Fund project to specifically address the needs of high frequency packaging and our coordination with NASA's Lewis Research Center's ongoing efforts to provide commercial high frequency packaging technology. We are working with industry, universities, and DoD to characterize and analyze high frequency multichip module packages. Our emphasis is on Materials Science and Radio Frequency (RF) Engineering with a goal to develop CAD tools and characterization techniques which can be readily transferred to industry to accelerate the development of *manufacturable* high frequency packaging technology.

INTRODUCTION

It is projected that in the next 5-10 years, commercial applications in the high frequency regime (1GHz to 100 GHz) will expand in areas which include: cellular communications, telecommunications, automotive radar for intelligent vehicular highway systems, advanced computing, aircraft radar, direct broadcast satellites, communication networks for the information superhighway just to name a few. This is due to the desire to increase our information capacity and at the same time decrease system mass and cost. It was estimated [1] that by 1997 merchant market value and domestic captive sales for multichip module (MCM) packages would be a half billion dollars.

During a JPL-hosted High Density Packaging Workshop [2] the consensus among industrial participants was that NASA funding in the area of high frequency packaging technology would be timely and beneficial to U.S. industry as a stimulus. In addition, the majority of resources being invested by industry in high density packaging technology development are focused on low frequency (<200 MHz) applications (Figure 1). A challenge which industry, DoD, NASA and JPL must meet is the development of high-density, high-frequency packages which are manufacturable, provide excellent electrical/mechanical performance, and long term reliability. The JPL/NASA team offers a non-competitive forum to coordinate on-going and future efforts to help accelerate and in partnership with industry contribute to this technology.

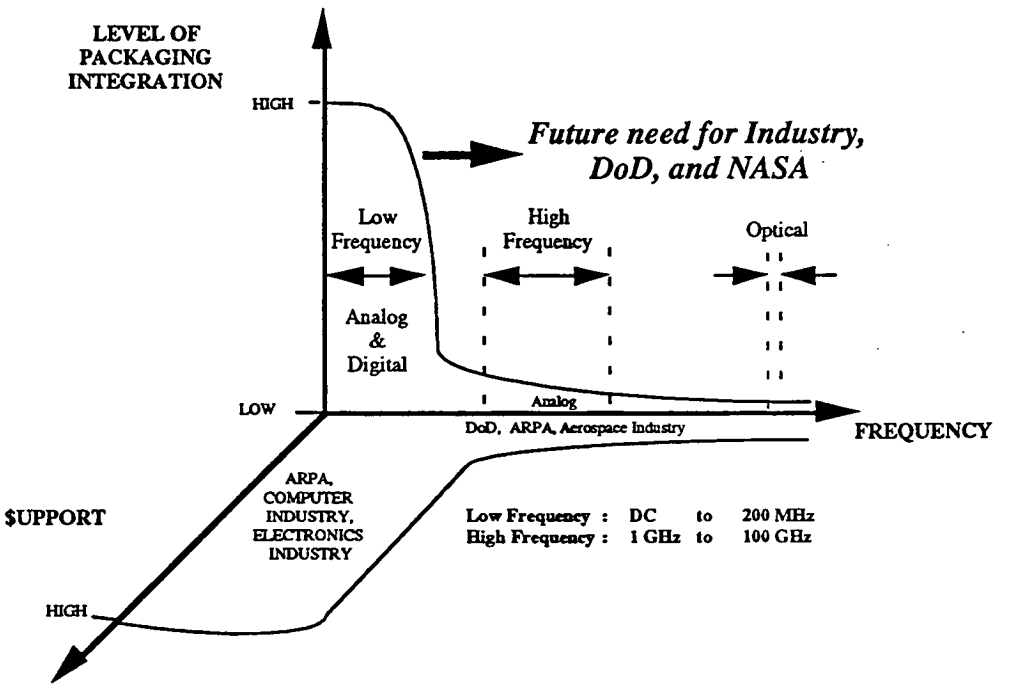


Figure 1. Current funding trend for advanced electronic packaging technology

We have initiated the development of high frequency packaging technology which has drawn a large response from the user community [3] and was documented for public review in [4]. In this paper we will describe new techniques to characterize MCMs using non-invasive optoelectronic probing techniques, development of 3-D electromagnetic CAD tool which will allow packaged devices and networks to be accurately analyzed thereby reducing development costs, the development of MMIC packages for phased array applications, and the use of novel multi-material systems to enable new package structures which are reliable and most of all manufacturable.

Background

Microelectronic packaging provides four basic functions:

- Power Distribution
- Signal Distribution
- Heat Dissipation
- Environmental Protection

High density electronic packaging refers to the incorporation of multiple integrated circuits, as bare die and passive components on a common interconnecting substrate. This format is commonly referred to as a multichip module (MCM). Significant improvements in circuit density are attainable (10-30X), as compared with printed circuit board technology.

Signal distribution becomes a major challenge as the frequency and packaging densities increase. Stray (RF) radiation in the form of crosstalk and package moding become limiting factors which must be addressed in order to preserve performance. The interconnecting networks into and out of the package must have low loss characteristics and be matched to the impedances of the adjacent networks. 3-D high density formats which use multilayer materials systems and embedded components further complicate signal propagation characteristics and thermal management issues.

Technical Challenges

In the development of a high-density high-frequency packaged module, the designer selects a variety of individual devices (transistors, diodes) and integrated circuits (Monolithic Microwave Integrated Circuits-MMICs) which have been characterized in an open air environment. The design is usually formulated using commercially available microwave circuit programs which are limited in their ability to correctly analyze the interaction of the active components within a closed environment of the package. Furthermore, almost every package configuration is unique to the particular problem at hand. Commercial CAD to analyze active RF networks in a packaged environment is not available. *This is a major gap which we have begun to address.* For high

frequency packaging networks we need to employ 3D electromagnetic analysis combined with non-linear effects of active circuits to determine operation and prevent stray radiation from degrading performance due to crosstalk or package moding. We must develop a technique which can be applied to a variety of MCM structures and active circuits. Finally, this technique should be amenable to being integrated to existing software packages in order to facilitate their use in industry.

The question of how does one characterize an enclosed (packaged) module when it is not performing as expected becomes more complicated the higher the system operating frequency. For low frequency, built-in-test (BIT) structures can be employed. However, for high frequencies where there are analog as well as digital signals, BIT structures may not give enough information on what the problem is and how to solve it. One could open the package and check around; however, the complete enclosed structure may be necessary in order to have the inhibiting phenomena to occur. The solution is a technique which can be non-invasive, preserve the enclosed package environment and provide both spatial and temporal resolution to probe the entire space and look at a full spectrum. *Our approach is to employ optoelectronic probing.* The technique is simple in principle and can be applied to a manufacturing line as well as in the development laboratories.

Although MMIC devices are maturing and becoming more available to potential users, their application in systems remains severely impeded by lack of high-frequency packaging. In order to successfully transition these small, fragile circuits from the laboratory environment to real-world systems, such as phased array antennas, engineers at NASA's Lewis Research Center and JPL have been developing advanced packaging through technically aggressive R&D programs. Novel solutions to provide high frequency packages for MMIC devices are presented.

Up to this point we have concentrated on the electrical aspects of high frequency packaging. Just as important are the material parameters which comprise the package and influence its reliability, manufacturability and cost. At high frequencies the electrical signal propagates closer to the dielectric/conductor interface, consequently the interface structure and chemistry significantly impacts the electrical performance. Thin film coatings for circuit protection and thermal management also come under this category. A variety of interface systems are being investigated and this constitutes a large portion of our effort.

RF PACKAGING TECHNOLOGY

As stated previously, the current work which is being funded concentrates on the development of optoelectronic techniques to characterize MCMs, 3-D electromagnetic simulator tool which can accurately analyze packaged networks and help decrease development costs and the status in the development of MMIC package technology.

Optoelectronic Probing

As industry develops advanced high frequency MCMs, there is an urgent need to characterize the entire module for functionality quickly and inexpensively. One approach is to add numerous detectors and I/O ports to the module in order to monitor each discrete device. This drives up development, fabrication and test costs, and increases mass. An alternative approach which has not been applied to electronic packaging involves the use of optoelectronic probes, based on either electro-optic or photoconductive principles. In electro-optic probing, a thin dielectric tip (fabricated from materials whose index of refraction changes with the intensity of an RF field passing through them) could be used to measure the amplitude and phase of electric fields within the volume of a package. Alternatively, photoconductive gates (which briefly pass current across a small gap charged by an electric field when illuminated by an ultrashort-pulse laser) could be integrated with extremely small metal lines on a probe to scan an entire package volume with very high sensitivity. The data obtained could be used to quickly identify faulty parts and anomalous RF performance within the package structure.

Even though optoelectronic probing has never been used in the test and characterization of high density packages, the concepts which we will rely on have been demonstrated in measurements of individual MMICs [5, 6].

The group involved in high-resolution, optically-based measurement techniques resides in the Ultrafast Science Laboratory at the University of Michigan. For more than ten years researchers in this group have worked to develop electro-optic - and more recently, photoconductive - sampling techniques that have been applied to test and characterization of short-relaxation-time materials (semiconductors and superconductors), passive and active devices, and analog and digital integrated circuits. Using this foundation, the Michigan laboratory is developing the system that will be utilized for measuring both the electric field and time-resolved signals within enclosed circuits. Optically-based techniques have been used extensively for measurements in open architectures, both to resolve transients as short as 150 fs and to map out fields within circuits. However, no external technique has before attempted to address problems such as the identification of resonances or failed devices within packaged MCMs.

Optoelectronic probes, which would be fed by optical fibers and should have very little influence on the operation of the circuit under test, would enter the packaged device through openings

much smaller than the wavelengths of the signals to be measured. Taking into account previous experience, it is anticipated that the probes should have sensitivity adequate to measure fields present in MCMs with analog sections driven with milliwatts of power. Furthermore, when used to obtain individual waveforms, a time resolution corresponding to a bandwidth of 100 GHz should be readily attainable. The probe will have dual-use capability from two standpoints: it will perform both field mapping and waveform acquisition; and it should be applicable to both analog and digital sections of MCMs.

The development process is taking place in two steps. Initially, in order to verify adequate sensitivity and to test the optoelectronic probe on a device with an analytically-known field pattern, existing probe designs are being mechanically modified in order to measure the field in a microwave cavity resonator. The second step will extend the applicability of the probe to a packaged MCM (most likely a transmit/receive module with analog and digital networks) which has its operation hindered by resonance/crosstalk effects.

Electromagnetic CAD Tool

Commercial CAD tools to analyze *active* RF networks in a packaged environment is not available. Therefore, we are currently investigating a new JPL concept called the Active Boundary Condition and employing the use of the JPL Cray capability to accelerate development. As part of the algorithm development, technology transfer to commercial microwave CAD packages will be important. The University of Colorado has experience with the integration of advanced electromagnetic software with commercial circuit analysis tools and will help accelerate this effort.

The basis of the analysis is a Finite Difference Time Domain (FDTD) algorithm which solves for Maxwell's equations. The FDTD techniques has received much attention in the open literature; however, the majority of this information is concerned only with passive structures. We will take advantage of the time domain aspects of FDTD and enable linear and nonlinear networks to be incorporated into the solution. Initial attempts to address this problem have idealized the active components modeled between two adjacent nodes. For low frequency applications this is probably an adequate solution; however, in the high frequency regime the physical device geometry and passive interconnect structure are distributed over many mesh points within the numerical grid thereby necessitating re-evaluation of the method of analysis.

MMIC Packaging for Phased Array Antenna Applications

NASA's Lewis Research Center's package development program focuses on MMICs operating in the 18.0 to 44.0 GHz frequency range, with interest in reconfigurable designs, flight qualification, and low-cost mass production. The packages are designed to withstand the challenging space environment of a communications satellite, but are directly applicable to a wide range of airborne and terrestrial systems.

Two contracts with industry examine various MMIC package configurations. The first contract, NAS3-25864, deals primarily with the development of a generic, or universal package and the attendant test fixturing that support a single MMIC chip. The package developed offers full hermeticity, mechanical device protection, heat removal from the device, and low insertion loss. This particular package is being carried through the environmental testing required for space qualification under NASA HQ Code Q funding.

The development of a single-chip package has moved rapidly from a research component to a marketable product. The package is constructed of high-purity ceramic with a metallic and base, with overall dimensions of only 0.28 by 0.28 by 0.05 inches. Radio frequency signals enter and exit the package through impedance-matched hermetic seals at opposite ends of the package. Ten bias and control lines are provided along the sides of the package. An additional feature of the package is its flexibility, whereby it may be refitted for different MMICs by modifications to only one of its four layers. Estimates place the cost of a typical package at approximately twenty dollars for small quantities, with substantial savings under mass production.

A second packaging effort, NAS3-25870, is examining the feasibility of multi-element, or multi-circuit packages as they would be applied to advanced phased array antenna systems (Figure 2). This approach permits several MMICs to be placed in close proximity to one another in a single package to achieve the tight spacing required between the radiating elements of an antenna. Each MMIC is installed in an individual compartment to eliminate crosstalk between chips and suppress electrical moding. An application-specific integrated circuit (ASIC) is included in the package to demultiplex a single serial data stream and thereby reduce the number of interconnect lines sent from a system controller. Interconnects between the ASIC and the RF MMICs are accomplished through multilayer interconnect located below the RF ground plane. This technique has permitted greatly-increased packaging density to be achieved.

The ongoing program will examine the performance of both single and multi-chip packages when operated in relevant space and aircraft environments. Additionally, emerging technologies such as fiber optics, multi-layer high-density interconnect, and self-correcting system-level integrated circuits (SLICs) are being developed for integration into antenna packaging.

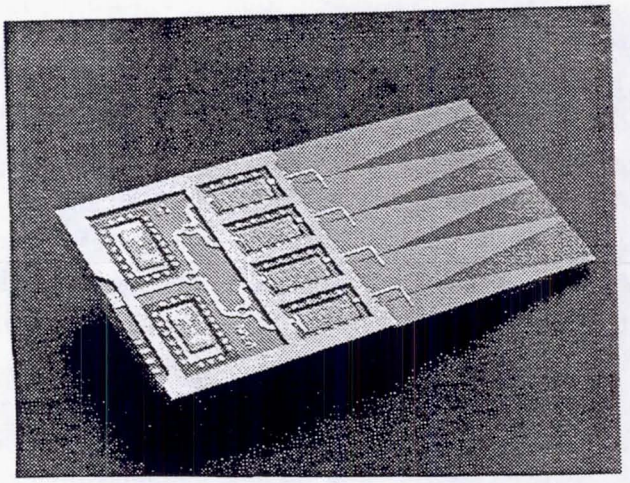


Figure 2. Multi-chip phased array module with integral radiating elements

MATERIALS SCIENCE

Our efforts in material science for high frequency electronic packaging focus on

- development of reliable conductor systems on a variety of substrates,
- integration of thin film capacitors into the packages,
- effective thermal management of high power-high frequency packages,
- environmental protection of packages using organic coatings.

The mechanical properties of thin film conductors as well as their interactions with dielectric substrates are a major source of reliability problems in high frequency packaging. For example, two types of stress intrinsic and extrinic are present in thin films. Intrinsic stress which, is a consequence of the deposition process, can vary from compressive to tensile, depending on the process parameters and is very difficult to predict. The major contribution to extrinsic stress is the thermal expansion mismatch between the conductor and the dielectric substrate. As a consequence of the stress relief process, delamination of the film/substrate interface, splitting of the film, or substrate cracking is observed for coatings in tension. The buckling of the film and the subsequent delamination of the films is associated with the compressive stress. These processes, which take place during thermal cycling or during MCM operation, are a major failure mode in electronic packaging. Chemical interactions and interdiffusion processes between metallic thin films and the substrate could also be a reliability problem. We have developed a novel Cr/CrN/Au conductor system which exhibits reduced intrinsic stress compared to the conventional Cr/Au bilayer. Due to the presence of a CrN diffusion barrier it also exhibits superior chemical and thermal stability. *We will continue to focus our efforts on the correlation between processing of thin film conductors and substrates and their performance in high frequency packages.*

High-frequency high-density MCM packaging technology imposes some critical restrictions on passive components, namely bypass and decoupling capacitors. The first restriction arises from the fact that the space available for capacitors (both in volume and area) is very limited. The second restriction is dictated by the power distribution and conditioning system. It is thus beneficial to directly integrate thin film bypass capacitors between the power and ground planes of the module. The dielectric constant of both decoupling and bypass thin film capacitors should be high and stable in the high frequency regime, with low leakage current and high breakdown strength. The highest dielectric constant is reported for perovskite ferroelectric capacitors but the deposition temperature or post annealing treatments required to fabricate perovskite structure are relatively high (above 400°C) and thus not acceptable for all MCMs. *We are in the process of developing of reliable thin film capacitors using deposition techniques which are compatible with manufacturing methods of high frequency multichip modules.*

The effective thermal management of the high-density high-frequency multichip modules is also a major problem. Conventional substrates used in high frequency applications, like alumina ceramics, must be replaced by high thermal conductivity substrates. Commercially available alumina has a good surface finish (about 1 μm) but its thermal conductivity is too low for packaging of high power high frequency devices (25 W/mK). On the other hand, thermal conductivity of commercial aluminum nitride ceramic substrates is high (220 W/mK), but the surface finish is poor (above 3 μm). *We have developed a new type of substrate for high frequency packaging [7].* The substrate consists of a highly polished bulk alumina ceramic substrate and a layer of aluminum nitride deposited on top of it using a thin film sputtering technique. The surface finish of the aluminum nitride film is the same as the surface finish of the alumina substrate (1 μm) and its thermal conductivity is one order of magnitude higher than that of alumina. The thickness of the aluminum nitride film can be adjusted to provide a required thermal heat spreading to a thermal sink. This new substrate takes advantage of both the surface finish of alumina and the thermal conductivity of aluminum nitride. Aluminum nitride can be replaced by a diamond substrate or diamond thin film but this technology is currently very expensive.

Current high frequency space electronics use semiconductor chips that are hermetically sealed in ceramic/metal packages. While providing good protection from severe environments, these packages significantly increase cost, mass, and volume of space electronics. Another serious problem associated with hermetically-sealed package materials (such as Kovar or Invar) involves the degradation of GaAs devices resulting from the interaction between hydrogen released from the package and donors in the channel region under the gate of the devices. One solution to this problem is to replace the hermetic package with an environmentally protective barrier coating. *We are working with industry to develop barrier coatings which will provide a weight/volume reduction by eliminating the welded metal cover from the package, while maintaining the radiation resistance and reliability currently associated with hermetically sealed packages.*

CONCLUSION

High-density high frequency electronic packaging is a key to achieving mass/volume reductions which will enable future NASA/JPL missions and complement the explosion in the commercial communication/computer/information industries. The work described in this paper is the beginning of a long term initiative to accelerate development of this emerging field. The development of accurate software tools, characterization techniques and material knowledge is aimed to help reduce design cycle time and development costs of high frequency packages. By establishing a close interaction with the outside community, we will accelerate the availability of high performance miniature high frequency packages.

ACKNOWLEDGMENTS

The work described here was carried out at the Jet Propulsion Laboratory, California Institute of technology under a contract with the National Aeronautics and Space Administration, and was funded through JPL's Director's Discretionary Fund.

We appreciate the support of Drs. M. Chahine , T. Cole, J. Layland and D. Rapp for their effort.

REFERENCES

[1] Frost and Sullivan, "The U.S. Market for Multichip Modules", Electronic Packaging and Production , October 1992.

[2] JPL-MCC Advanced Packaging Workshop, June 15-16 1993, Pasadena, CA

[3] M. Herman, K. Lee, L. Lowry, A. Carpenter, and P. Wamhof, "Hermetic Packages for Millimeter-Wave Circuits," NASA Tech Briefs, June 1994, p.24.

[4] M. Herman, K. Lee, L. Lowry, E. Kolawa, and A. Tulintseff, "Novel Techniques for Millimeter Wave Packages," submitted for review to the IEEE Transactions on Microwave Theory and Techniques

[5] J.F. Whitaker, J.A. Valdmanis, T.A. Jackson, K.B. Bhasin, R.Romanofsky, and G.A. Mourou, "External electro-optic probing of millimeter-wave integrated circuits," 1989 IEEE MTT-S International Microwave Symposium Digest, vol. 1, pp. 221-224.

[6] J. Kim, J. Son, S. Wakana, J. Nees, S. Williamson, J. Whitaker, Y. Kwon, and D. Pavlidis, "Time-domain network analysis of mm-wave circuits based on a photoconductive probe sampling technique," 1993 IEEE MTT-S International Microwave Symposium Digest, NewYork: IEEE, 1993, pp. 1359-1362.

[7] E. Kolawa, L. Lowry, M. Herman, and K. Lee, "Transmission Lines For High Frequency and High Density Packaging", International Society for Hybrid Microelectronics (ISHM) Conference and Exhibition on Multichip Modules, Boston MA, 1994.

AEROGELS FOR ELECTRONICS

L. W. Hrubesh

Chemistry and Material Sciences Department
Lawrence Livermore National Laboratory
Livermore, CA 94550

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

ABSTRACT

In addition to their other exceptional properties, aerogels also exhibit unusual dielectric and electronic properties due to their nano-sized structures and high porosities. For example, aerogels have the lowest dielectric constants measured for a solid material (having values approaching 1.0); they have exceptionally high dielectric resistivities and strengths (i.e., ability to insulate very high voltages); they exhibit low dielectric loss at microwave frequencies; and some aerogels are electrically conductive and photoconductive. These properties are being exploited to provide the next generation of materials for energy storage, low power consumption, and ultra-fast electronics. We are working toward adapting these unusual materials for microelectronic applications, particularly, making thin aerogel films for dielectric substrates and for energy storage devices such as supercapacitors. Measurements are presented in this paper for the dielectric and electronic properties of aerogels, including the dielectric constant, loss factor, dielectric and electrical conductivity, volume resistivity, and dielectric strength. We also describe methods to form and characterize thin aerogel films which are being developed for numerous electronic applications. Finally, some of the electronic applications proposed for aerogels are presented. Commercialization of aerogels for electronics must await further feasibility, prototype development, and cost studies, but they are one of the key materials and are sure to have a major impact on future electronics.

INTRODUCTION

Aerogels are very high porosity materials made by sol-gel chemistry and dried using special conditions to preserve the tenuous solid network [1,2]. The microstructure of the aerogels consists of particles and pores which are only fractions of the wavelength of visible light in size. Such structure is unique among common materials and many extraordinary properties result from it. For example, aerogels are known to exhibit the lowest thermal conductivity, sound velocity, and refractive index, of any bulk solid material [3]. The microstructure and the very high porosity of aerogels are also responsible for exceptional dielectric properties and electronic behavior. The dielectric properties of aerogels are affected by the large volume fraction of trapped gas in the pores and the high

concentration of adsorbed molecules on the abundant internal surfaces. This has been confirmed by measurements of the linear change of the dielectric properties with aerogel density, and the large effect on these properties attributed to adsorbed water [4,5]. The electrical conductivity of the dielectric aerogels is predictably low because the tenuous solid structure provides poor conduction paths and few charge carriers. The volume resistivity is expected to be high for the same reason. The dielectric strength of aerogels is also expected to be high due to the high volume resistivity and because the nano-sized pores confine the charge carriers to spaces that are about the same size as the mean-free-path for collisions. These properties show that aerogels are unusual dielectric materials and suggest that they can be used for many interesting applications which will be discussed in this paper. Some aerogels can be transformed into carbon aerogels which are electrically conductive. This property and the very large surface area available within them, make these particular aerogels especially useful for energy storage devices like capacitors and battaries.

The formation of thin aerogel films is a necessary step toward many electronic applications. We have developed methods to form aerogel films having thicknesses from 1 to 20 microns and we have successfully deposited metallized patterns on them. The characterization of the dielectric properties of such thin aerogel films has been challenging, although we have been able to apply optical techniques which allow the determination of some electronic properties. This paper presents measurement and development results for aerogels, and it also describes potential and anticipated applications of aerogel dielectrics and conductive aerogels.

MEASUREMENTS OF DIELECTRIC PROPERTIES

Dielectric Permittivity

Few measurements of the permittivity of aerogels have been made to date, but collectively, they cover a wide frequency range from 50 Hz to 40 Ghz and demonstrate the low values expected for such highly porous materials. Measurements of the real (dielectric constant) and imaginary (loss factor) parts of the complex permittivity were reported by da Silva, et.al. [4] for silica aerogels for frequencies between 50 and 10^5 Hz, and for temperatures of 1.6K to 300K. Hrubesh, et.al. [5] have measured permittivities for both the silica and organic aerogels, at microwave frequencies (i.e., 2 to 40 GHz) and at 298K.

At LLNL, the measurements of the dielectric properties were made at microwave frequencies using a cavity perturbation method. This method allows a sensitive measurement of changes in the resonant conditions of a dielectric filled microwave resonator. Such changes can be directly related to the real and imaginary parts of the relative permittivity of the dielectric. All measurements were made at 298K on aerogel samples which were either equilibrated at atmospheric conditions (as prepared), or were heated at 700K under a vacuum for 10 hours to remove adsorbed water (baked). A plot of the dielectric constants of silica aerogel measured at microwave frequencies, is shown in fig. 1 for the density range from 0.01 to 0.6 g/cc. It is seen that the dielectric constant (κ') varies linearly with density (ρ) over this density range. A least squares fit of the data gives $\kappa' - 1 = 1.60\ \rho$ for the as prepared aerogels, and $\kappa' - 1 = 1.48\ \rho$ for the baked aerogels. The difference in the slopes between the 'as prepared' and 'baked' samples is attributed to dispersion of the microwave radiation by interactions with polar molecules (mostly water and hydroxyls) on the internal surfaces of the aerogels. This effect contributes significantly

(~7% for silica) to the dielectric constant of aerogels. The surface water affects the loss tangent more than the dielectric constant as seen in fig. 2. The data for the loss tangent fits the relations, $\tan\delta = 0.172\,\rho$ for the 'as prepared' aerogels, and $\tan\delta = 0.004\,\rho$ for the 'baked' aerogels. The effect of water contributes ~70% of the loss tangent in silica aerogel. The dielectric properties have also been measured at LLNL for the purely organic aerogels, and the properties for those aerogels have also been determined to be linearly related to the densities.

It is notable that the dielectric constant for any silica aerogel having a density less than 0.6 g/cc, is less than the dielectric constant of teflon (≈2.0) which is the most common low dielectric material in current use. We measured a dielectric constant of 1.008 for a silica aerogel having a density of 0.008 g/cc. This is believed to be the lowest dielectric constant ever measured for a bulk solid material.

Three other properties of aerogels are important for applications of dielectrics in electronics; 1) sufficiently high thermal conductivity for heat dissipation, 2) a thermal expansion which closely matches the substrate to reduce chances for stress induced cracks, and 3) sufficient shear and compressive strength to support multiple layers. While aerogels are most notable for their exceptional thermal insulation property and are normally not considered as effective thermal conductors, the higher density silica aerogels (e.g., $\rho>0.5$ g/cc) are actually better thermal conductors than many polymer films used for low dielectrics. The measured coefficient for thermal expansion of silica aerogel is 3×10^{-6} cm/cm for the temperature range 275-323K. This value is similar to that for fused silica, suggesting that silica aerogel should be thermally compatible with the glassy substrates or coatings used in electronics packaging. Lastly, the shear strength has not been adequately measured for any aerogels to date. However, the compressive strength of aerogel has been measured and is strongly dependent on the density. The strength of aerogels for electronics applications should be an issue only for the lowest densities (i.e., $\rho<0.05$ g/cc).

Dielectric Conductivity and Volume Resistivity

The dielectric conductivity (σ) of aerogels is obtained from our microwave measurements by using the relation [7]; $\sigma = 5.5 \times 10^{-13}\,\kappa' \tan\delta\, f$ (ohm-cm), where κ' and $\tan\delta$ are the dielectric constant and the loss tangent, respectively, and f is the frequency. For the 'baked' silica aerogel with a density of 0.1 g/cc, the dielectric conductivity ranges from 1.1×10^{-4} to 8.1×10^{-6} $\Omega^{-1}cm^{-1}$ in the frequency range from 3 to 40 GHz. The typical volume resistivity for 'baked' silica aerogels (i.e., $2\pi f/s$) with the same density is 4.1×10^{15} Ω cm. Comparative values of dielectric conductivity for the 'as prepared' organic aerogel with a density of 0.1 g/cc are 2.6×10^{-4} to 1.9×10^{-5} $\Omega^{-1}cm^{-1}$, and 4.5×10^{14} Ω cm, for the typical volume resistivity. These values of volume resistivities for aerogels are comparable with the best of the polymer insulating materials (e.g., poly-tetrafluoroethylene and polyethylene) [8]. Aerogels should therefore be expected to exhibit very good dielectric strengths against high voltage breakdown.

Dielectric Strength

The dielectric strength of silica aerogels is higher than expected for a material that is so porous. This is likely attributable to the fact that pores in air-filled aerogels are of the same order of size as the mean distance for collisions of electrons. Thus, electrons in aerogel pores tend to collide with the solid before gaining sufficient kinetic energy to ionize

upon impact. We have measured the dielectric strength of air-filled silica aerogels at 300K. These measurements were made at 60 Hz on silica aerogels having different thicknesses. Breakdown voltages were registered for different thicknesses of silica aerogel. The average dielectric strength from these data is 128 kV/cm and it was determined to be essentially independent of the aerogel density. Measurements were only made on 'as prepared' aerogels; further measurements will be done to determine the effect of adsorbed water on the dielectric strength of silica aerogel. Higher values of dielectric strength are expected for 'baked' aerogels, but even the value reported here for silica aerogel is higher than for most ceramics (e.g., alumina is 110 kv/cm), though it is less than for pure polymers (e.g., 160 to 500 kV/cm) [9]. Aerogels should be effective, very lightweight insulators for high voltage applications, especially in vacuum applications.

THIN FILM AEROGEL PROCESSING

Sol-gel processes have been well developed, especially during the past decade or so, to produce advanced glasses, glass-ceramics, and ceramics, in thin films for electronics. There are several books and articles which describe the advantages and details of sol-gel films for electronic applications [10,11,12]. The formation of thin aerogel films might seem rather straightforward because of the considerable success in applying sol-gel coatings to surfaces, however, the rapid evaporation of solvent that accompanies the deposition processes causes the sols to form compact films as drying occurs, rather than forming a more porous gel. Therefore, special considerations and methods are needed to successfully form highly porous gels and dry them to make low density aerogels. Very little work has been reported for processing porous sol-gel films where the films are to remain porous for their application. Here we describe the special conditions and requirements which are needed to make highly porous aerogel films which exhibit low dielectric properties.

Most of the methods already developed for applying sol-gel thin films and coatings (e.g., dipping, spinning, spraying, etc.) generally apply to making true aerogel films. However, all of these methods have the common requirement that the gel must be formed under conditions in which the rate of evaporation is limited, both during and after the gel formation. To facilitate this we perform the coating processes within an enclosure that is maintained saturated with the vapor of the working solvent. The enclosure atmosphere also contains a partial pressure of ammonium hydroxide which helps to catalyze the gelation of the films. The most common aerogel films are silica, but we have also demonstrated our method with other metal oxides for which hydrolysis/condensation of the metal alkoxide is the predominant chemistry (e.g., zirconia, alumina, and tantala.) Generally, depending on the desired porosity of the final aerogel, we used two methods to prepare the precursor solution for the process. For gel porosities >95% we used the two-step, partial hydrolysis/condensation chemistry reported previously for silica aerogels [13]. For gels with other porosities, we prepared a single-step base-catalysed hydrolysis solution according to the typical method for silica as follows: mix tetramethoxysilane (TMOS), water, methyl alcohol, and ammonium hydroxide, in a molar ratio of 1:2:4:0.01. An additional amount of alcohol is added to this mixture to establish the ultimate porosity of the gel. The methods used to make thin aerogel films depend on the thickness desired and whether or not the film is to be bonded to a surface. In the following we discuss the various processing steps and procedures to form thin aerogel films on substrate materials for electronics.

Surface Preparation

From experience we found that most of the metal oxide gels adhere to glass or oxidized surfaces (probably through metal-oxide-metal bonds) but gels did not stick well to unprepared metal surfaces. Bonding to all oxidized surfaces was enhanced by etching the surface with a mild alkaline solution (e.g, KOH), then rinsing with alcohol immediately prior to film deposition. For the opposite case of non-bonding we treated the surfaces with a methylated silicate compound that serves as a release agent.

Spin, Dip, and Spray Coatings

Spin coating is essentially the same as that used to spin glass coatings for electronic applications, except that it is performed with the spin apparatus entirely within an enclosure that has a solvent saturated atmosphere. Film thicknesses are typically less than two micrometers. Special modifications were performed on the apparatus to prevent the possibility of an explosive hazard. Typical substrates are Pyrex glass slides and silicon wafers up to 3" diameter. The procedure for forming films is to meter a droplet of precursor solution onto the spinning substrate while its spin rate is increasing up to a maximum speed of 1850 RPM. The spinner is immediately turned off and stopped with a brake, so that the sample is subject to minimal loss of solvent during gelation. Typically, the gel will form within a few minutes, after which the substrate is manually removed from the spin apparatus and immersed in solvent. The substrates with films are stored submersed in solvent until ready for supercritical drying.

Dip coating is the simplest of the coating processes, but it is used only when all surfaces of a substrate material are to be coated. With this method, film thicknesses less than a few micrometers are obtainable depending on the viscosity of the precursor and the withdrawl rate. In our work, the precise thickness was not an important parameter, so the dipping and withdrawing procedure was performed manually without concern for controlling the rates. Typical substrates were Pyrex glass slides of various sizes. These were simply dipped into the prepared precursor solution, withdrawn, then placed edgewise and vertical in a holder which is located within the enclosure. The time necessary for a gel film to form was found by trial to be only a few minutes. After the film is gelled, the entire holder containing the slides is immersed in a beaker of solvent and it remains surrounded by liquid until ready for supercritical drying.

Spray coating has been used to put thicker single layer coatings on substrates like glass and silicon wafers. Films as thick as 80 micrometers have been acheived by this method. An aspirator is used to spray the precursor solution directly onto the substrate which is supported in a nearly vertical position within the enclosure. Excess solution drains by gravity, leaving a thick film which gels within a few minutes. These films have a varying thickness due to the draining, but the surface of the gel is smooth and continuous. After gelation, the substrate is manually immersed in solvent until ready for supercritical drying.

Supercritical Drying

All of the prepared gels were converted to aerogels using supercritical drying methods in order to prevent densification of the films. Either direct supercritical extraction (SCE) of the solvent was done at high temperature or an alternate low temperature extraction of carbon dioxide was performed after exchange of the original solvent. The method of drying chosen depended on the temperature stability of the substrate material. Gels on glass substrates were generally compatible with high temperature SCE whereas silicon wafers required a protective coating of silicon dioxide to survive the high temperature SCE of solvents. The procedure for direct SCE is to place the glass container holding the submersed (or encapsulated) samples directly into an autoclave. The autoclave is filled with additional solvent (usually, alcohol) and sealed. The temperature of the autoclave is raised at a rate of 0.3°C/min. while the pressure increases to about 120 bars and excess pressure above that is released. After the temperature reaches about 280°C, the pressure is released from the vessel at a controlled rate of 0.3 bar/min. until a pressure of about 1.5 bar is reached. The autoclave is then purged with air as the vessel cools. This cycle typically takes 24 hours. Alternatively, a similar cycle is performed at temperatures less than 45°C after the solvent in the gel-film is first exchanged with liquid carbon dioxide. The exchange time to remove solvent from the thin films less than 50 micrometers thick was typically 2 hours.

Metallization

Metallization of the aerogel was done by vapor or sputter deposition directly onto the as-prepared aerogel film surfaces. For some cases, low resolution patterning was achieved by depositing the metal through an appropriate mask. Photoresist patterning techniques were applied only after the aerogel surface was coated with approximately 1000Å of polymer to seal the pores of the aerogel.

Film Characterization

The morphology of the films and the film-substrate interface regions may be examined using a high resolution scanning electron microscope (SEM). The film thicknesses are measured using a stylus type profilometer. Film adherence on glass and silicon wafers are qualitatively determined using a tape stick test with Scotch® tape.

Refractive indexes of films were measured using an ellipseometer. This instrument will measure either the refractive index or the thickness of a thin dielectric layer if the other is known. The thickness measured by a profilometer is entered as input to the ellipsometer to calculate the refractive index of a film. Generally, it is difficult to apply this technique because the surface reflectivity of aerogels is typically less than 1% at a wavelength of 632.8nm. However, sufficient reflectivity is usually obtained at shallow angles of incidence. Other physical properties of the aerogel films were determined from this measurement as follows:

The film densities were determined from the measurement of the optical refractive index, n, using the relation, $\rho = (n-1)/.209$, reported by Henning and Svensson for silica aerogels [14]. The dielectric constant and the porosity of the films were determined indirectly from the density, after measurement of the optical refractive index. The porosity is determined from the relation: $\prod = 1- \rho/\rho_s$ where ρ_s is the density of the solid. So for silica aerogel, with $\rho_s = 2.19$ g/cc, $\prod = [(1.458 - n) / .458]$ is the percent porosity. Similarly, the dielectric constants for the silica aerogel films or sheets were determined after measurement of the optical refractive index using:$\kappa'=1+1.6\rho$ [15], so $\kappa'= 1+7.7\ (n-1)$.

Results

Using the methods described, we have successfully fabricated thin, flat, uniform silica aerogel films, having various porosities, with measured thicknesses from less than 1 to 20 micrometers on glass and silicon wafer substrates. The adherence of the films to the substrates has been qualitatively determined by a tape stick test. Generally, good film adherence was obtained for all aerogels with a porosity less than about 86% and mixed results were observed for porosities between 86% and 95%. For porosities >95%, the results were invalid because the aerogel is too weak to survive the application of the tape. These films have been examined using SEM to verify their aerogel microstructure.

We have measured the refractive indexes of deposited films and we have calculated their bulk densities to confirm that they are aerogels having densities in the range from about 0.04 - 0.46 g/cc, and porosities in the range from 77% to 98%. We have also calculated the dielectric constants of these films based on the measured refractive indexes. The dielectric constants span the range from about 1.06 to 1.7.

We developed processes to seal, pattern, and metallize thin aerogel films. An example of a high resolution pattern of gold metal on an aerogel film is shown in figure 3. This pattern of 4 micrometer wide strips was achieved using polymer coating and photoresist techniques. We have also sputtered thin (< 0.5 micrometer) metal layers directly onto the aerogel surfaces and electroplated thicker (> 1.0 micrometer) layers on the sputtered layers.

APPLICATIONS

In addition to their exceptional dielectric properties, aerogels exhibit other complementary properties for electronics such as low thermal expansion and adequate thermal conductivity and mechanical strength. Aerogels provide a unique set of properties and attributes to meet specifications for electronic applications. We list here some of the numerous applications for aerogels as dielectrics, some of which are under current development.

Aerogel Films

Thin aerogel films (≈2 micrometers) are formed on silicon wafers to provide a low dielectric constant substrate to reduce capacitance in integrated circuits. The dielectric constant for all aerogels is less than 2.0 and its value depends on the porosity, a controllable parameter for aerogel films. Dielectric constants of the aerogels well below 2.0 will enable significant improvements in the speed of integrated circuits. Thick organic aerogel films (≈0.5 millimeter) are formed by spray coating, dried by supercritical conditions, then pyrolyzed to carbon aerogel films for use in aerocapacitors. These devices can provide specific capacitances in excess of 45 Farads per gram of material [16].

Applications for the thin film dielectrics include: microwave striplines, microwave circuits such as used in radars and communications, low capacitance chip connectors, high-speed electronic conductors for both ultra large scale integrated circuits and for interconnections between computer chips, high-speed Gallium Arsenide test chips and associated electronic packages, extremely lightweight electronic packages, power transmission high voltage insulators, and spacers for electrodes in vacuum tubes.

Bulk Dielectrics

Aerogels, as bulk materials, are also useful for electronic and electrical applications, particularly where they provide effectively air-like properties of lightweight and high electrical insulation. Applications for the bulk dielectrics include; air-like suspension of microwave circuits, co-axial cable insulation, power transmission high voltage insulators, and microwave antenna.

Other Applications

Other applications in the microelectronics and electro-optics industries include low dielectric constant insulators for high-speed electronic conductors in large scale integrated circuits, for interconnections between computer chips, and as ultra-fast light modulators. Also, the conductive carbon aerogels, filled with electrolite fluids are useful for electrodes in supercapacitors and as batteries.

CONCLUSIONS

Aerogels exhibit very low dielectric permittivities as expected for such highly porous materials. However, the high porosity and high surface area also contribute to significant changes in the dielectric properties from adsorbed molecular species. This effect must be accounted for when considering electronic applications of the aerogels. The dielectric permittivities of aerogels are found to be linearly related to the density. This indicates that the properties are dominated by the trapped gas in the aerogels rather than by the solid matrix. The results presented here provide empirical relations for predicting the magnitude of the dielectric properties when the aerogel density or porosity is known.

It is possible and practical to form highly porous, true aerogel films on substrates using a variety of common deposition methods, if care is taken to slow evaporation of solvents during gelation, and if supercritically dried to preserve their tenuous structure. The aerogel films are good quality, bond well to the substrates, and are strong enough to survive other processing steps used to prepare them for specific applications. The exceptional dielectric properties of aerogels coupled with the ability to produce either bulk materials or thin films, suggest their use in many electronic applications. Already, aerogels are being developed for use in supercapacitors, microwave devices, and electronic packaging.

This paper provides information about the electronic properties of aerogels, discusses why they are exceptional, and offers a variety of applications for the use of aerogels in electronics. The author hopes this information will serve to stimulate commercial interest in these amazing and unusual materials called aerogels. While aerogels have many properties desireable for the electronic applications, further development is needed to determine their compatability with circuit processing and to determine manufacturing costs. Commercialization of aerogel dielectrics must await further feasibility, developmental, and cost studies, but they are expected to have a major impact on future electronics [17]. The commercialization of aerogels for supercapacitors is already underway through a spin-off company from LLNL.

References:

1) R.P. Iler, The Chemistry of Silica, (J. Wiley and Sons, N.Y., 1979), pp.537-539.
2) J. Fricke, "Aerogels," Sci. Am. 256, 92 (1988).
3) J. Fricke and A. Emmerling, "Aerogels - Preparation, Properties, Applications," in Chemistry, Spectroscopy, and Applications of Sol-Gel Glasses, eds. R. Reisfeld and C.K. Jorgensen, (Springer Series on Structure and Bonding, Vol. 77, Springer-Verlag, Heidelberg, Germany, 1991), p.37.
4) A. da Silva, P. Donoso, and M.A. Aegerter, "Properties of water adsorbed in porous silica aerogels," J. Non-Cryst. Solids 145, 168-174 (1992).
5) L.W. Hrubesh, L.E. Keene, and V.R. Latorre, "Dielectric properties of aerogels," J. Mater. Res. 8 (7), 101 (1993).
6) A. da Silva, D. I. dos Santos, and M.A. Aegerter, "Dielectric response of silica aerogels," J. Non-Cryst. Solids 96, 1159-1166 (1987).
7) A.R. von Hippel, Dielectric and Waves, (J. Wiley and Sons, N.Y., 1954), p. 28.
8) D.W. Van Krevelen, Properties of Polymers, (Elsevier, N.Y., 1990), pp. 330-333.
9) R.A. Flinn and P.K. Trojan, Engineering Materials and Their Applications, (Houghton Mifflin Co., Boston, 1981), pp. 614-615.
10) Sol-gel Technology for Thin Films, Fibers, Preforms, Electronics and Specialty Shapes, edited by L.C. Klein (Noyes, Park Ridge, N.J., 1988).
11) C.J. Brinker and G.W. Sherer, Sol-Gel Science (Academic Press, N.Y., 1990).
12) H. Dislich, in: Sol-gel Technology for Thin Films, Fibers, Preforms, Electronics and Specialty Shapes, ed. L.C. Klein (Noyes, Park Ridge, N.J., 1988) p. 50.
13) L.W. Hrubesh, T.M. Tillotson, and J.F. Poco, in: Chemical Processing of Advanced Materials, eds. L.L. Hench and J.K. West (Wiley, New York, 1992) p. 19.
14) S. Henning and L. Svensson, Phys. Scripta 23 (1981) 697.
15) L.W. Hrubesh, L.E. Keene, and V.R. Latorre, J. Mater. Res. 8 (1993) 1736.
16) S.T. Mayer, R.W. Pekala and J.L. Kaschmitter, J. Electrochem. Soc.,140(1993) 446.
17) Frost & Sullivan's Technology Impact Report, "Aerogels," #T047, p.115.

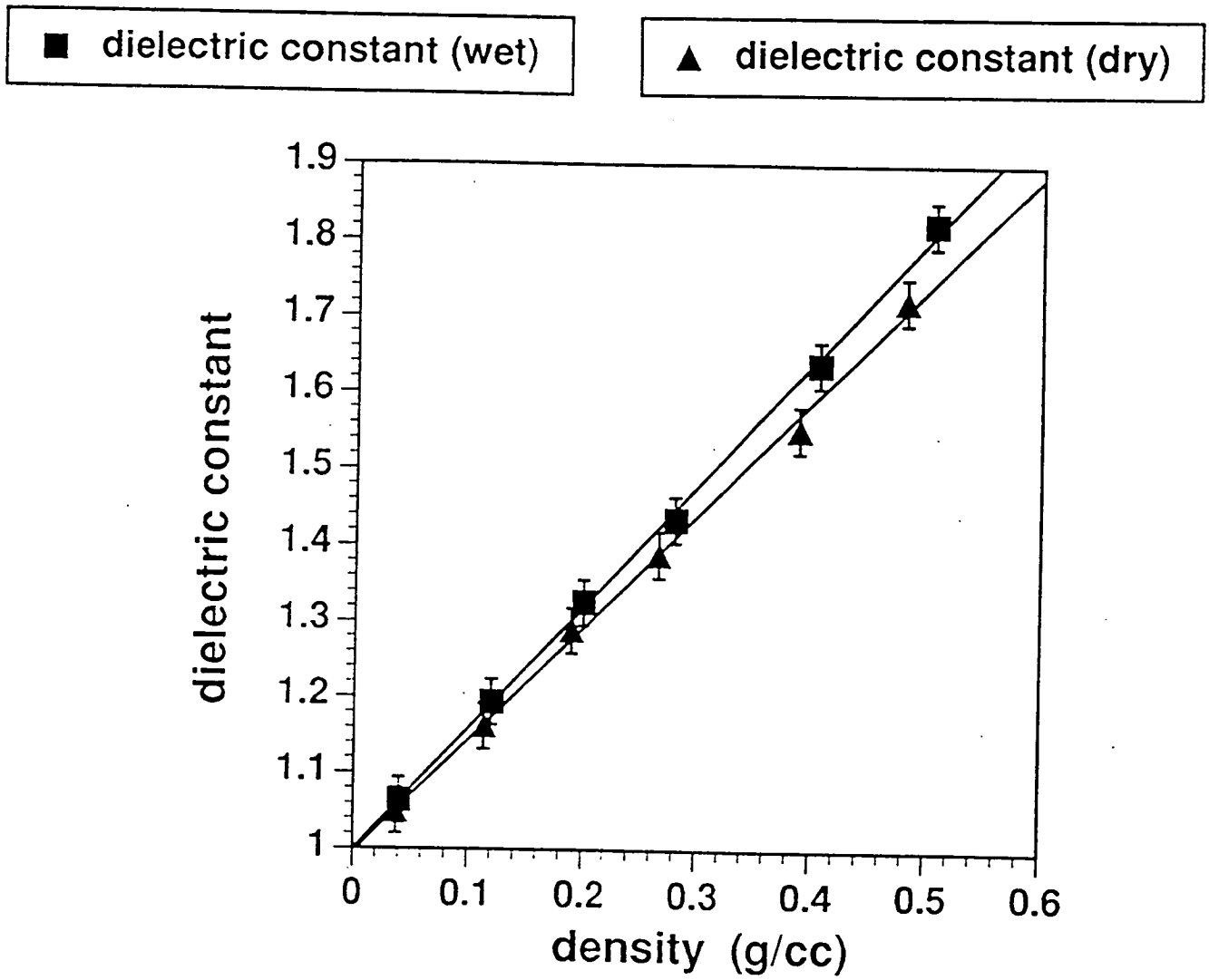


Figure 1. Plot of the dielectric constant of silica aerogel versus density, for moisture laden and dry samples, showing about 7% increase due to adsorbed water.

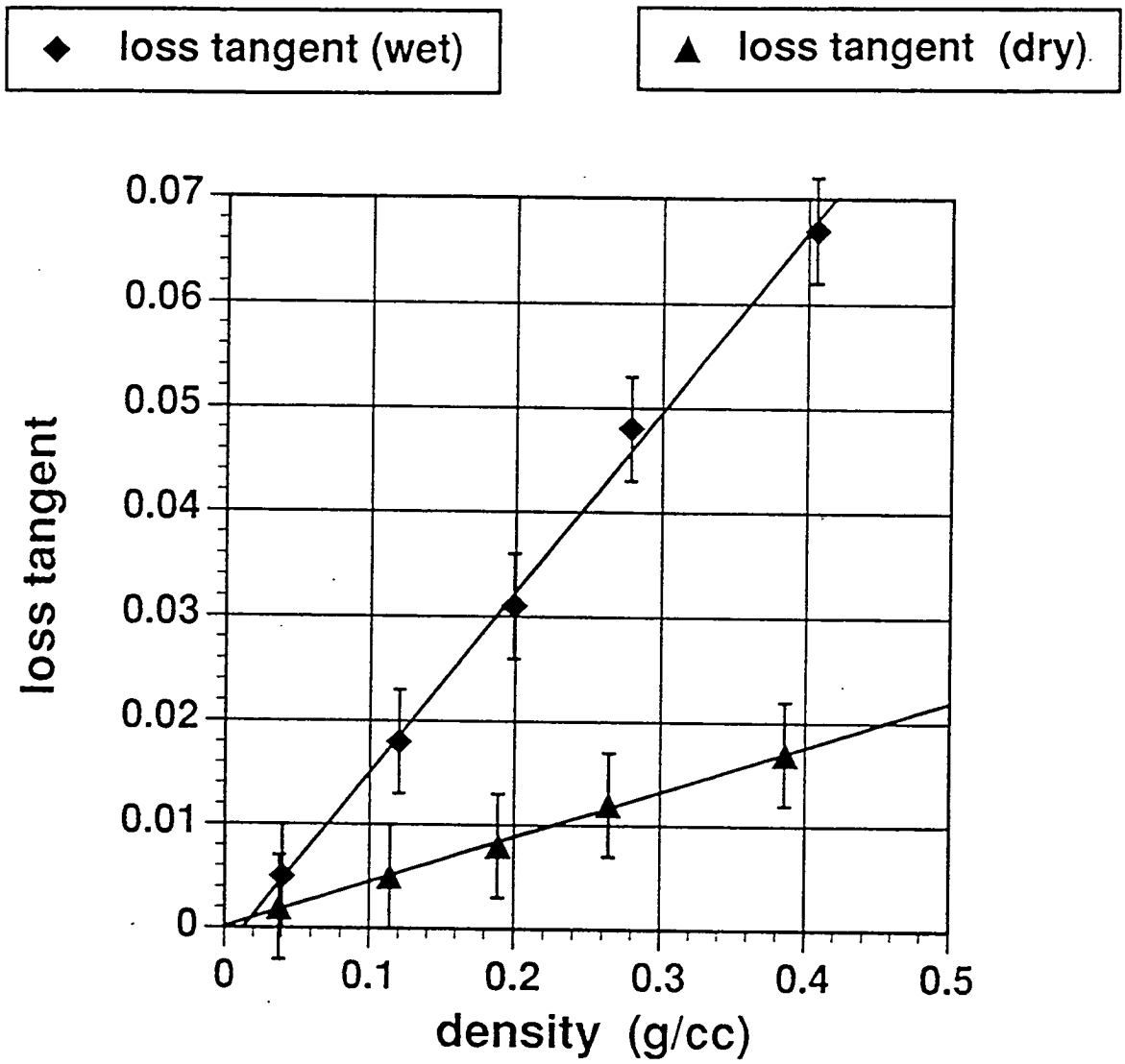


Figure 2. Plot of the loss tangent (at microwave frequencies) of silica aerogel versus density, for moisture laden and dry samples, showing about 70% increase due to adsorbed water.

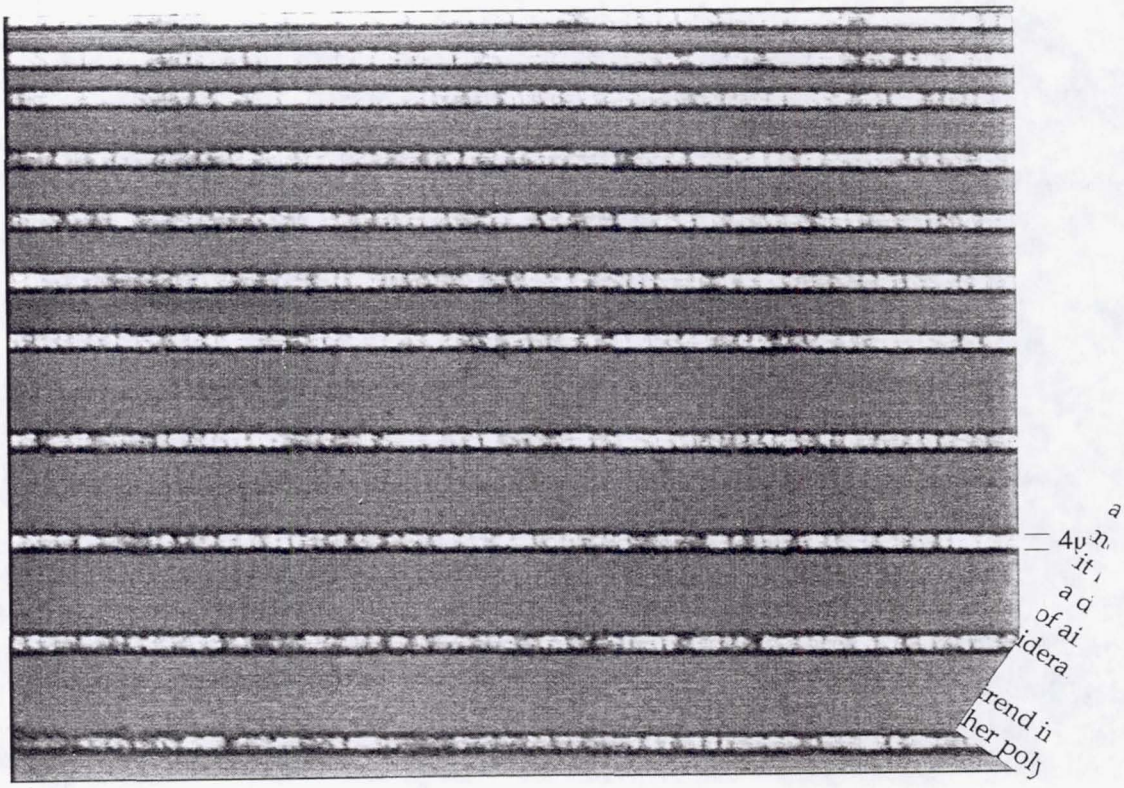


Figure 3. Photograph of 4 micrometer wide gold metal strips patterned on a thin silica aerogel film (≈ 1.3μm) which is on a silicon wafer. This pattern was produced by photoresist processing after the aerogel was first coated with about 100nm thick polymer which remains after the process.

A 3.5 W OUTPUT, DIODE-PUMPED, Q-SWITCHED 532 nm Nd:YAG LASER PUMPED BY FIBER-COUPLED DIODE LASERS

Hamid Hemmati and James R. Lesh
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Dr., M/S 161-135, Pasadena, CA 91109

ABSTRACT

A single Nd:YAG laser crystal was pumped by three 10-W fiber-coupled diode lasers. At 50 kHz pulse repetition frequency, average output powers exceeding 11 W of continuous-wave 1064 nm, and 3.5 W of 532-nm were achieved. A folded three-mirror cavity which compensated for thermal lensing in the laser crystal was utilized. One arm of the cavity contained the Nd:YAG rod and an acousto-optical Q-switcher while the other arm contained a frequency-doubling nonlinear crystal. The 532 nm output beam quality was 1.5.

INTRODUCTION

Compact lasers with high wall-plug efficiency and with average output power and pulse repetition frequency of greater than 2 W and 50 kHz, respectively, are needed for laser communication from outer planets to the earth. Other requirements for such a laser include: single spatial mode beam quality, visible to near infrared wavelength, short (ns level) pulse width, simple thermal management and low pulse jitter. Lasers with above characteristics also have applications in material processing & testing, resistor trimming, wafer marking, ranging, spectroscopy, film writing, printing, displays, holographics, and medical instrumentation.

With fiber-coupled lasers the diode laser can be remotely located relative to the laser resonator. This feature facilitates removal of heat generated by high power diode pump lasers and reduces complications due to thermal gradients in the laser's mechanical assembly caused by the diode pump lasers. Also, alignment of the pump laser(s) with the resonator is simpler with fiber-coupled diode lasers. Recently, a number of continuous-wave (cw) [1-4] and pulsed [5-8] solid-state lasers pumped with cw diode lasers have been reported. In this research we developed a very compact laser with greater than 11 W of cw output at 1064 nm and 3.5 W of near diffraction-limited 532 nm second harmonic at pulse repetition frequency (PRF) of 50 kHz, with the highest known wall-plug efficiency. An end-pumped scheme was selected since, generally diffraction-limited output can be obtained more efficiently with end-pumped rather than side-pumped lasers. A true wall-plug efficiency of 2.3% was obtained. Higher efficiencies are expected in the future.

DESIGN

To achieve over 2-W of 532 nm average output at pulse repetition frequency of 50 kHz, greater than 20 W of 809 nm pump laser power is required. At tens of Watt pump power levels, a significant thermally-induced lensing and birefringence is generated in most solid-state laser materials. Nd:YAG laser crystal was selected as this laser's active medium since it has lower thermal lensing coefficient than Nd:YVO_4 and significantly higher thermal fracture strength than Nd:YLF.

The approach proposed by Magni was followed to design the laser resonator [9]. The radii of curvature of the mirrors were selected such that the resonator supports a single-spatial-mode with large mode volume, has high alignment stability, compensates for the thermally-induced lens, and has low sensitivity to focal length fluctuations of that lens. A commercially available software (Paraxia™, Genesse Software) was used to model the resonator. With 27 W of continuous-wave 809 nm power focused to a spot size of 0.4 ± 0.05 mm in the crystal, two different resonators were identified to satisfy the requirements mentioned above: (1) a plano-concave resonator with the concave mirror having a radius of curvature of 100 cm, (2) a convex-concave resonator consisting of a 12 cm radius of curvature convex mirror and a concave mirror with radius of curvature of 100 cm. The plano-concave mirror was used in this set up due to availability of the mirrors.

EXPERIMENTAL SETUP

Figure 1A shows a schematic of the folded laser resonator. The frequency-doubled (green) intracavity beam is confined to a region of the cavity containing only the frequency-doubler. The folded resonator has been used earlier for both end-pumped and side-pumped configurations [10-12]. In this research, the output beam of each of the three 10-W fiber-coupled diode lasers (SDL-3450-P5) was partially collimated by an 8.6 mm focal length lens combination (Newport Research, F-L20) to a total beam diameter of 1.8 cm, measured 1 cm away from the lens. Following the approach by Fan et al [13], the three closely-spaced collimated beams were focused with an efficiency of 90% into one end of the Nd:YAG crystal using a single 2.35 cm focal length aspheric lens (Melles Griot, 01 LAG 115). The pump-spot radius of the focused beam at the laser crystal was 0.43 mm. Both end faces (7 mm in diameter) of the 7-mm long Nd:YAG rod were anti-reflection coated. The output wavelength of the lasers (at 25°C) ranged from 808 nm to 811 nm. The temperature of the crystal was maintained at 17 °C. The 100-cm radius of curvature concave input mirror had AR coating at 809 nm on the entrance face, and high reflectance (HR) at 1064 nm on the second surface. The flat fold mirror was HR-coated at 1064 nm and had high transmission at 532 nm second harmonic for 45° angle of incidence. The flat end mirror had dual HR coating at 1064 nm and 532 nm. The cavity length was 10.5 cm producing a fundamental spot size of approximately 0.52 mm at the input mirror.

To avoid the need to generate high voltages for electro-optical Q-switching, an acousto-optical Q-switcher was selected. The acousto-optical Q-switcher crystal had AR coating at 1064 nm on both surfaces. The Q-switcher was driven at 80 MHz center frequency with 1.4 W of RF power. It was always operated at above 10 kHz to avoid damage to the resonator optics and intra-cavity elements. A dual-wavelength AR-coated KTP ($KTiPO_4$) crystal was used for frequency-doubling.

RESULTS

Figure 2 illustrates the second harmonic and fundamental average output power, at 50 kHz pulse repetition frequency (PRF), as a function of the incident cw pump power. The cw 1064 nm and pulsed 532 nm lasing thresholds were 2.1 W and 2.5 W, respectively. With full pump power (30-W), over 11.7 W of cw 1064 nm power and 3.5 W of pulsed 532 nm power were obtained. Approximately 21.1 W of the total pump power was absorbed in the laser crystal. The optical-to-optical conversion efficiencies were 55% and 16.6% for 1064-nm and 532-nm, respectively.

Figure 3 shows a plot of the second harmonic average output power and laser pulse width as function of the PRF. The true wall-plug efficiency of the laser, considering all electrical power supplied to the pump diode lasers, the Q-switcher, and those for heat removal from the diode lasers and laser crystal, was 2.3%.

The ratio of the far-field beam diameter to the diffraction limited beam diameter, calculated for the same cavity waist, provides a measure of laser output beam quality The beam quality was a function of the pump power since the focal length of the thermally induced lens, and therefore the Fresnel number for the cavity, varied with pump power. The measured cw 1064 nm beam quality ($M^{2)}$ factor [14] was 2.2. The Q-switched 532 nm output beam quality was approximately 1.5 times the diffraction limit.

Acknowledgment: This work was conducted at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES:

1) S.C. Tidwell, J.F. Seamans, and M.S. Bowers, Opt. Lett., **18**, 116 (1993).

2) Y. Kenada, M. Oka, H. Massuda, and S. Kubota, Opt. Lett., **17**, 1003 (1992).

3) M.S. Keirstead and T.M. Baer, in *Digest of conference on Lasers and Electro-Optics* (Optical Society of America, Washington, DC, 1991, paper CFC3.

4) L. Marshall, A. Katz, and H. Verdun, in *Digest of conference on Lasers and Electro-Optics* (Optical Society of America, Washington, DC, 1993, paper CMF5.

5) H. Hemmati and J.R. Lesh, IEEE J. Quantum. Electron., **28**, 1018 (1992).

6) H. Plaessmann, S. A. Ré, J.J. Alonis, D. Vecht, and W.M. Grossman, Opt. Lett., **18**, 1420 (1993)

7) A.J.W. Brown, R. Mead, and W.R. Bosenberg, in *Digest of conference on Lasers and Electro-Optics* (Optical Society of America, Washington, DC, 1993, paper CMF7.

8) D.C. Shannon and R.W. Wallace, Opt. Lett., **16**, 318 (1991).

9) V. Magni, Appl. Optics, **25**, 107 (1986).

10) T.E. Dimmick, Opt. Lett., **14**, 677 (1989).

11) F. Hanson and D. Haddok, Appl. Optics, **27**, 80 (1988).

12) I.L. Bass and R.W. Presta, in Proc. SPIE, **1040**, 116 (1989).

13) T. Y. Fan, A. Sanchez, and W.E. DeFeo, Opt. Lett., **14**, 1057 (1991).

14) A.E. Siegman, in SPIE Proc., **1224**, 1 (1990).

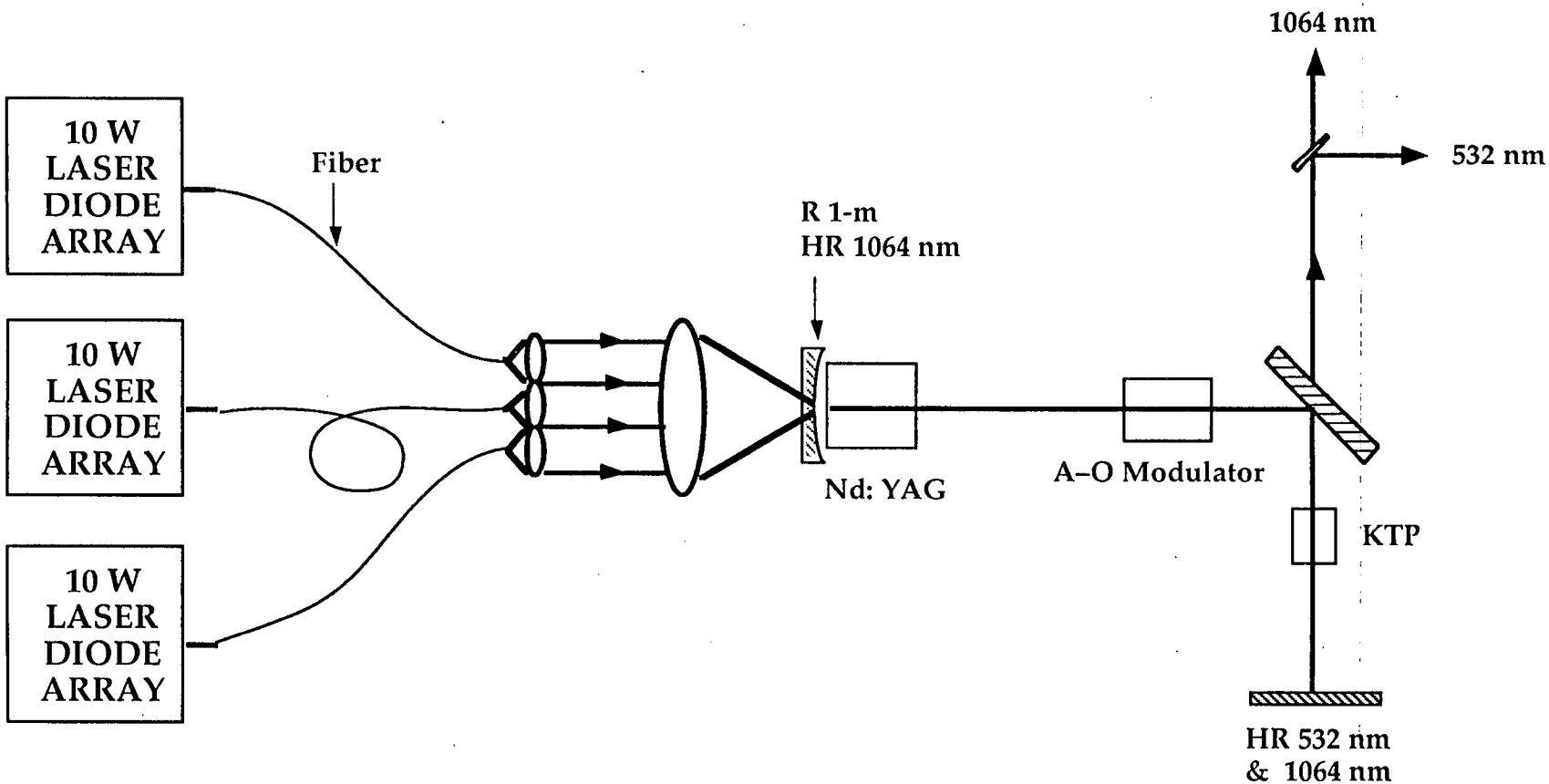


Figure 1. Schematic of the Experimental Setup

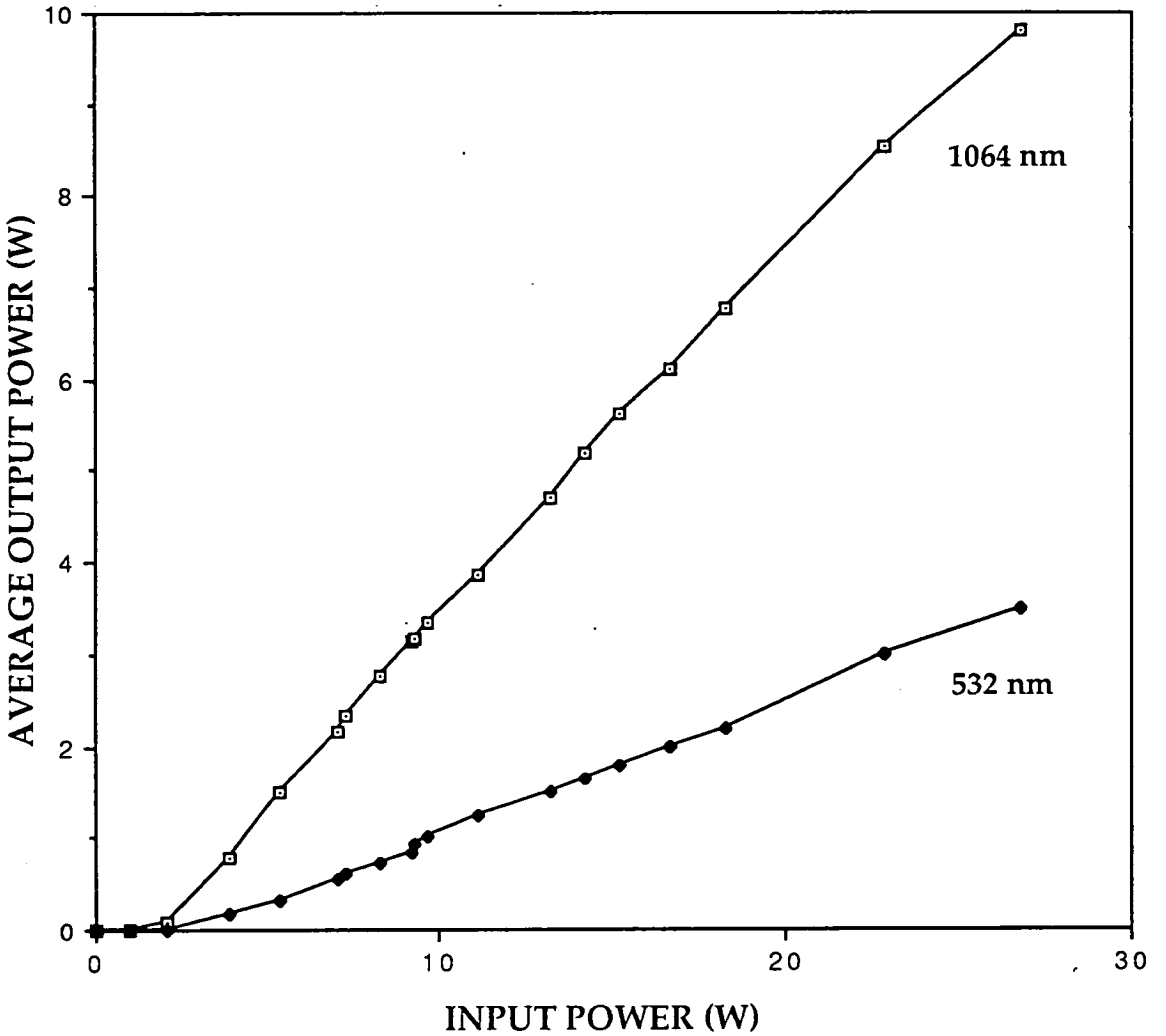


Figure 2. Continuous-Wave Output Power at 1064 nm and Pulsed 532 nm Output as a Function of the Incident Pump Power

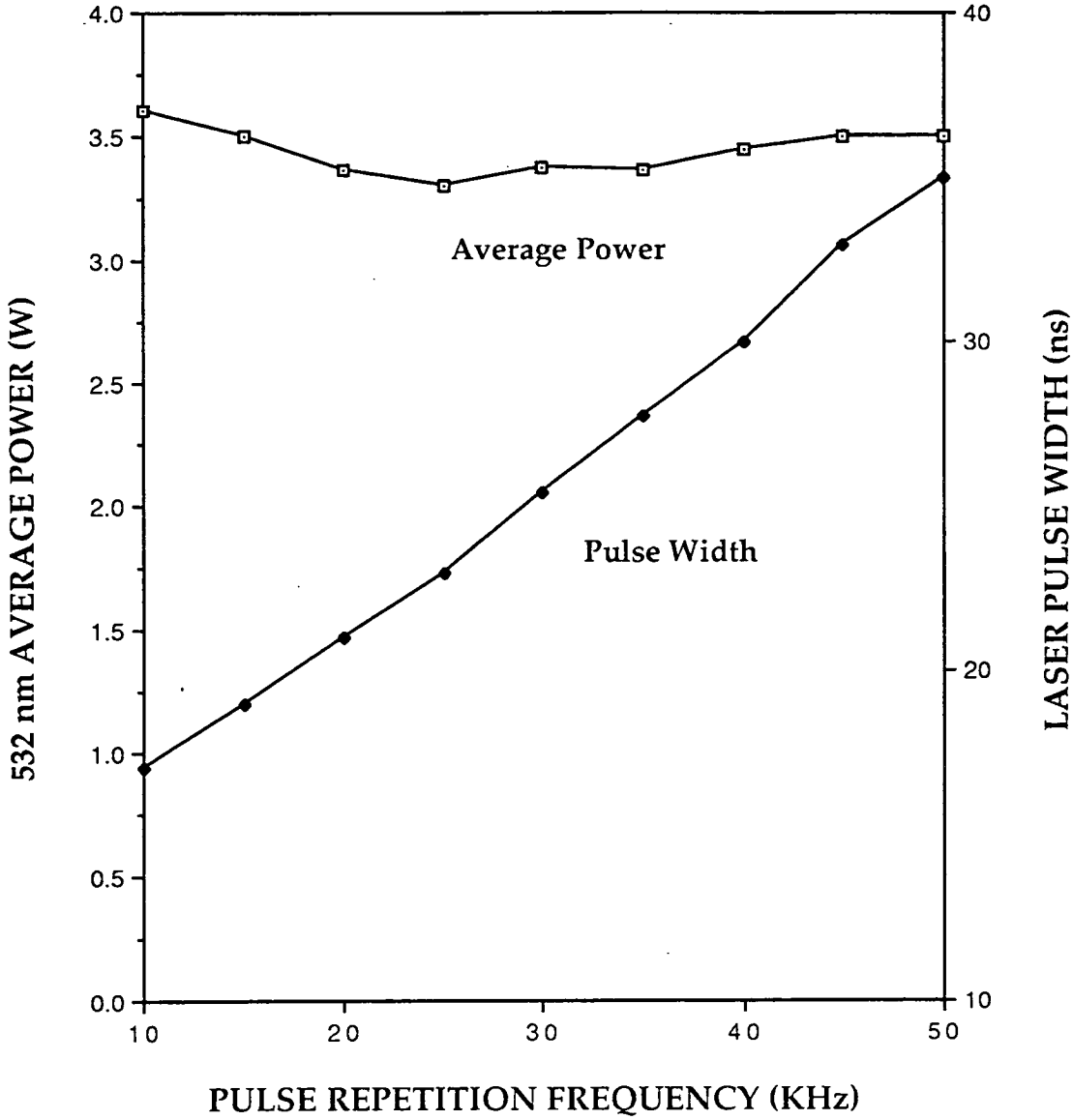


Figure 3. Average Output Power at 532 nm and Laser Pulse-Width as a Function of the Q-switch Pulse Repetition Frequency

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 2

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. 2

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 2 features papers on computers and software, virtual reality simulation, environmental technology, video and imaging, medical technology and life sciences, robotics and artificial intelligence, and electronics.

14. SUBJECT TERMS

Digital computers, computer programs, virtual reality, computerized simulation, earth environment, photometry, robotics, medical equipment, artificial intelligence, fiber optics

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